Investigation of the COVID-19 Research —A Big Data Approach

2020 5th International Conference on Universal Village (UV) | 978-1-7281-9523-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/UV50937.2020.9426209

Yong Xu
Institute of Machine Learning and Intelligent Science
Fujian University of Technology
Fuzhou, China
y.xu@fjut.edu.cn

Guojun Mao
Institute of Machine Learning and Intelligent Science
Fujian University of Technology
Fuzhou, China
19662092@fjut.edu.cn

Shan Huang
Institute of Machine Learning and Intelligent Science
Fujian University of Technology
Fuzhou, China
19872162@fjut.edu.cn

***Abstract*—The outbreak of novel coronavirus disease in 2019 (COVID-19) has drawn researchers' attention to find the causes and facts of it in hope of preventing its spread and saving patients' life. However, there are still lack of researches investigating this problem from a big data perspective. This paper tries to tackle this severe threat from a big data perspective to reveal the unknown facts and research trends concealed in the academic publications and to compare our findings with the traditional statistical methods. We downloaded 16, 560 publications from Web of Science and classified the most frequently mentioned keywords in the abstracts into seven different aspects. Then the cluster and strategic diagram methods were used to identify the core and mature research topics and trend. We found that although the vulnerable had been paid appropriate attention by researchers, undeveloped countries had not in this health catastrophe; lung was the most fragile organ to be infected and CT and RT-PCR were the most favorite diagnostic methods; and clinical and modelling methods were the most preferably used by researchers as medical and non-medical research tools etc. Strategic diagram revealed that instead of fever, respiratory distress and pulmonary symptoms/disorders were the most mature diagnosable symptoms. Our findings showed that this simple method proves itself as being applicable in bringing to light some unknown facts hidden behind the haphazard research data and revealing the future research trends.**

Keywords*—SARS-CoV-2, nCoV19, big data, medical science, clinical medicine, epidemiology, pandemic

I. INTRODUCTION

The unexpected outbreak of the current novel coronavirus disease in 2019 (COVID-19) caused by novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus has fully developed to a pandemic and become an urgent worldwide threat to the public health. Both of the spread speed and the case fatality rate stand high with the infection of such virus. Till now, more than 70 million people have been diagnosed with such infection and over 1.6 million people have died. The severity and the high infectivity of this disease have called for emergent actions by all countries from government, public health sectors and academia. Many researchers from all over the world have dedicated, from various aspects, to finding better ways of identifying and quarantining the suspected, treating the patients, reducing the death rate and suppressing the spread speed. Comprehensive reviews have also been provided to summarize the current research and point out some better directions of future research [1]-[3]. In this paper, we are going to investigate the research of COVID-19 from a big data perspective by analyzing the publications in this field in an attempt to uncover some useful clues, patterns and trends from the big pool of research data for the reference of future research.

Big data by definition refers a combination of large and diversely structured and unstructured data that may grow in increasing volumes on a day-to-day basis. One purpose of analyzing a big data is to search for some important information and relationship from different types of data, which generally can not be done by traditional methods, to aid some complex decision-making processes. Big data approaches have also been used among other methods to tackle the threat of COVID-19. In [4], the big data approach was used to identify the suspect from 627, 386 people who had been in contact with the passengers disembarked from the Diamond Princess cruise ship (the COVID-19 was spread on the ship at that time) in Taiwan and found that big data approach combined with some other smart tracing methods could curtail the required resources for suspected contact tracing during a severe epidemiological outbreak. The internet searched data was employed to investigate characteristics of symptoms in COVID-19 using a big data method by searching the symptom's keywords for the COVID-19 in [5]. In [6], the authors employed the artificial intelligence method based on big data to model the evolutionary process of COVID-19 by utilizing the patient data collected from other related diseases. Big data method had been used in these researches to find some useful information of a specific subject. However, big data approach can dig out much more useful information from all perspective of the COIVD-19 researches.

Therefore, the purpose of this paper is not an attempt to provide a technological review of the research in COVID-19 as there have been quite a few related technical or comprehensive review papers published so far. In fact, after having downloaded 16, 560 publications from Web of Science [7], we find that there were 1, 109 papers with the word review/overview/survey as their title which amounts to 6.9% of all the publications. Hence, we are not going to repeat the work already done by other colleagues, nor are we to apply this method to only one aspect

This work was supported by National Natural Science Foundation of China under Grant No. 61773415.

978-1-7281-9523-0/20/$31.00 2020 IEEE

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

of the research topic. Our purpose is to take a complete new measure from an omni-perspective point of view to cover all the research topics and reveal some unknown facts and trends inundated in the huge pool of research data using the big data approach. Our findings may sever as a useful reference for researchers and clinicians to tackle this severe health disaster and save more people in the future.

II. Materials, Methods and Data Preprocessing

16, 560 publications were downloaded from Web of Science [7] on 9 September 2020 by using 'SARS-CoV-2' as the keyword for subject search. For each paper, only the information of Author, Title, Journal, Abstract and Times of Cited was kept but not the full text of it. After all the publications have been downloaded, a basic work of data preprocessing was conducted to clean the data and remove some unusual ones. In this process, 17 papers published before January 2020 were removed; 120 responding letters, commentary papers, corrections or Editorials, etc. were also excluded. Some of the papers were removed either as they published in both a journal and the bioRxiv or medRxiv in which the bioRxiv or medRxiv ones were generally removed, or as they were included in the searched publications more than one time in which the ones with lower Times of Cited, without an abstract or a random one were removed. In this way, a total of 359 duplicates were removed in which 3 papers were repeated 15 times, 1 paper 5 times, 12 papers 4 times, 46 papers 3 times and 185 papers 2 times. Another one paper was removed as the searched title is different from the real title of it.

After such preprocessing of the data, a total of 16, 063 valid papers were used for analysis. Within these papers, 53 papers only had a non-English title which were translated to English by the authors; 43 papers were wrongly used the title in other languages as the English title and were adjusted by the authors to have a correct English title. By doing these, we can count the keywords in title more accurately and fairly in our research. Besides, 4 papers had an abstract in non-English languages and 3, 203 had no abstract. Therefore, only 12, 856 abstracts in all the publications were used for statistics.

We extracted the information from these downloaded publications by searching and counting the corresponding keywords from the title and abstract separately to mine some useful information for analysis. For each keyword, apart from the main form of the word, some other variants were also included in our statistic. For example, in searching the word epidemiology, epidemiologic, epidemiological and epidemiologically were also searched and in searching simulation, simulations, simulated, simulates and simulating were also included as well. Especially, in searching pains, apart from various pains, some other specific words such as muscle cramp, headache, sore throat, and myalgia etc. were also counted. In addition, apart from specifically mentioned, all the statistic results shown below were obtained from Abstracts of our searched publications.

III. Results

Based on the publications collected from the Web of Sciences, we identified some useful information, counted the frequency of all the words in the abstract and selected some of the most meaningful and the most frequently mentioned keywords from them. The results are given in the following subsections.

A. Basic Information

In order to find how many journals had published the papers in COVID-19 research, we counted them to a total of 3, 212. However, within them, 109 Journals had two or more different names, for example, the *Journal of Clinical Virology* was also shown as *Journal of Clinical Virology : the Official Publication of the Pan American Society for Clinical Virology* and the *Nature Reviews Immunology* was also shown as the *Nature Reviews. Immunology* etc., which amounts to 3.36% of the total. Those Journals all had their own Times of Cited and number of papers published. We then combined all the journals having different names into one single journal to make 3, 103 unique journals so that the statistics could be sound and reliable.

Fig. 1 shows the number of papers published by each of the most frequently published 100 Journals. The greatest number of papers published by a single journal is 329, which was the Journal of Medical Virology. Within the 3, 103 different Journals, the top 1% most published Journals (31 Journals) have published 3, 465 papers, which amounts to 21.57% of the total publications. The average number of papers published by the top 1% Journals is 111.77. In addition, the average number of papers published by the top 100 Journals is 57.8 and that by all the Journals is 5.17, which is only about 4.62% of the average number of papers published by the top 1% Journals. Besides, the Nature- branded primary research journals published 438 papers, which amounts to 2.73% of the total. Using the concept of Concentration Rate (CR), we find that the CR100 is only 36, which means that the publications were quite scatteredly distributed among different journals without great concentration although the top 1% journals seemed to have published quite a high proportion of the articles.

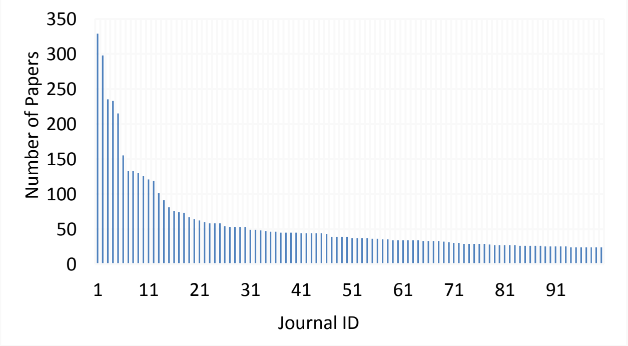

Fig. 1. The number of papers published by each of the most frequently published 100 Journals.

But if we look at the average number of times cited for each paper published in each month, we find that the highest average number of times cited for each paper was not seen in January but in March. Within all the publications, the most cited paper is Ref. [8], which was also published online on March 13, 2020 by Lancet and cited 2, 188 times by September 9, 2020. This shows that although some papers were published quite swiftly, the papers with higher quality or higher levels of interest were the ones that may need more time to conceive and analyze. Analyzing the percent of papers that had been cited at least one

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

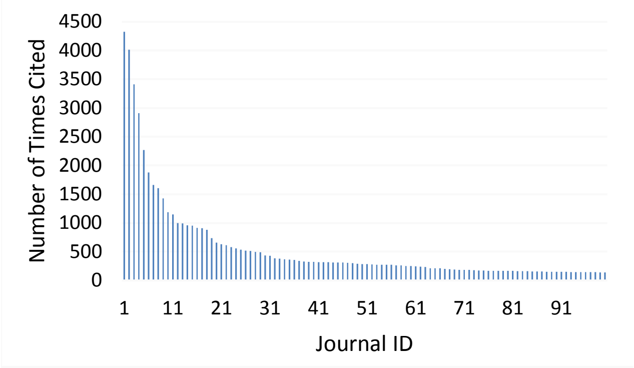

Fig. 2. The number of cited times in the top 100 cited Journals.

time in each month, we find that the highest percent of papers cited nonzero times also appeared in March.

In order to identify the actual impact of journals in this subject, Fig. 2 shows the total number of times cited in the most frequently cited 100 Journals. The total number of times cited for all journals is 73, 683 and the top 1% most cited Journals (31 Journals) was cited 39, 630 times, which amounts to 53.8% of the total number of times cited. The most cited journal was the *New England Journal of Medicine*, which had been cited 4, 326 times in our searched journals by September 9, 2020. However, there were 1, 962 journals being cited zero time, which amounts to 63.2% of the total number of journals. This means that researchers generally prefer referring to high quality journals and the gap of qualities among different journals are very significant. The Pearson's correlation coefficient between to top 50 cited journals per paper and the Impact Factor of the corresponding journals is -0.068 with a p-value > 0.1. This means that the impact of publications in this specific subject is totally different from other topics, which indicates that the research focus on some emergency topics could significantly affects a journal's impact.

B. *Time and Countries Mentioned in the Research*

The outbreak of COVID-19 was first noticed in Wuhan, China in December 2019 and then it gradually spread to other countries. We are interested in tracing the tracks of research along the time-line to identify how many number of times each month was mentioned in the researches. Fig. 3 shows the number of times each month being mentioned in the Abstract in all the publications. We find from this Fig. 3 that, with the sudden outbreak of this disease, December was the first month being mentioned many times. Then with the disease gradually spread to other countries, especially moved to the US in March, the next peak of month being mentioned was March. Therefore, the number of times each month being mentioned was consistent with the severity of this disease and the time point researchers would be most interested in, which can also be used as an indicator to trace the progress of this disease.

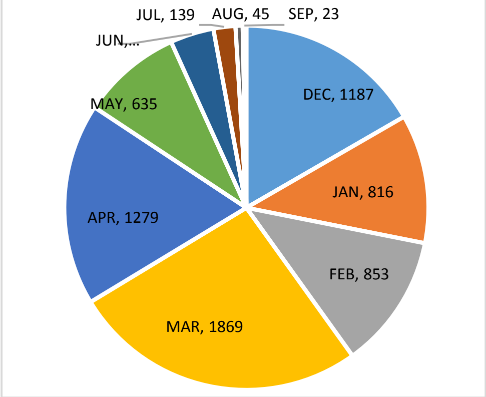

Fig. 3. The distribution of the times each month being mentioned in the Abstract.

In order to find whether there is a bias in mentioning different countries in the researches, plotted in Fig. 4 is the distribution of the top 50 most infected countries at the time of this research and the corresponding number of times the countries being mentioned in the Abstracts. We find from this Fig. 4 that there are some peaks and troughs in the number of times a country was mentioned. The highest peak is Italy (19th) and the 3 lower peaks are Spain (9th), UK (13th), and Germany (21st), respectively. On the other hand, the troughs are Russia (4th), Peru (5th), Columbia (6th), and Argentina (14th), respectively. There are even 3 countries that had not been mentioned in the literature, which are Qatar (28th), Dominica (31st), and Guatemala (39th). From these facts we find that although most of the heavily infected countries were generally received a high visibility in the research, the sad side is that developed countries were generally paid higher attention to than undeveloped or poor ones by academia. For example, Guatemala had more than 80, 000 infected cases and nearly 3,000 deaths but was not mentioned in the research at all while New Zealand had only less than 2, 000 infected cases and 24 deaths but was mentioned 37 times. Therefore, when there is a public health disaster, poor countries were generally not paid enough attention to and it is also the case in this pandemic event.

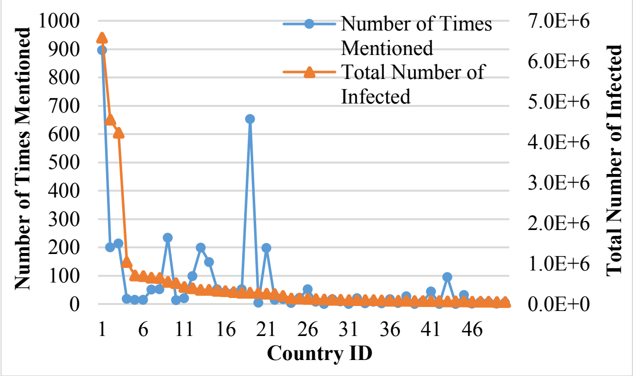

Fig. 4. The distribution of the top 50 most infected countries at the time of this research and the corresponding number of times the country being mentioned in the Abstracts.

C. *People Group Mentioned in the Research*

Researches showed that different groups of people may have different kinds of infection and/or death rates in this COVID-19 pandemic [9], [10]. Therefore, looking at how many times of different groups of people had been mentioned in the research may be interesting. Fig. 5 shows the relative percent of times each group of people being mentioned in the Abstracts. We find from this Fig. 5 that, the most frequently mentioned in the research was human (beings), followed by children and women

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

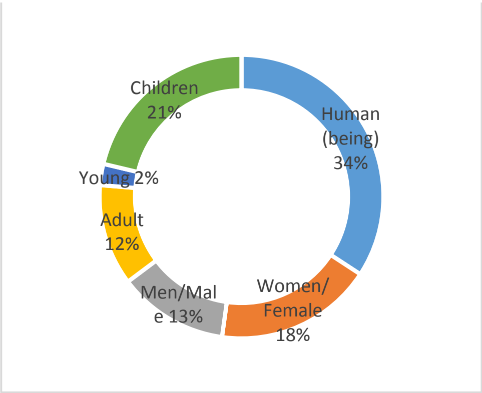

Fig. 5. The distribution of the relative percent of times each group of people being mentioned in the Abstracts.

group, and the least concerned was the youngest. This is consistent with the fact that the most vulnerable people should be paid more attention to in such a severe pandemic of coronavirus.

As for the vulnerable groups of people, we noticed that different groups of the vulnerable were also mentioned in the Abstracts which is shown in Fig. 6. We can see from this Fig. 6 that, the most frequently mentioned vulnerables were pregnant women and the elderly people. The baby/infant and newborn babies were not so frequently mentioned. This does not mean that babies were not paid enough attention to in this infection, but reflects the fact that babies were not so easily infected as the elderly or pregnant women as the range of their activities was relatively narrow. Therefore, this is also consistent with the concerns people often show to the vulnerable during public health emergency.

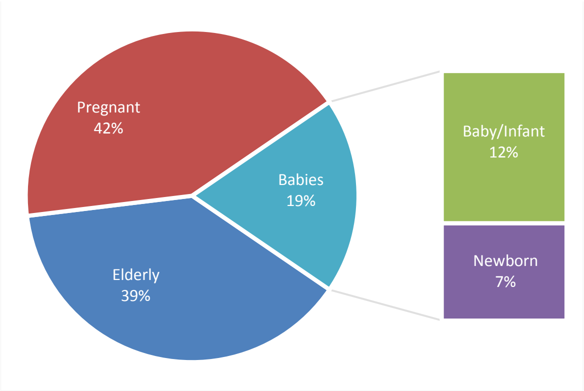

Fig. 6. The relative percent of times the vulnerable being mentioned in the Abstracts.

D. *Symptoms Mentioned in the Research*

Patients infected with COVID-19 may show a number of different symptoms during their infection [11]. In order to reveal how frequently each symptom had been mentioned in the research and whether the facts dug out by big data method is the same as analyzed by traditional statistic one, listed in Fig. 7 is the distribution of the number of times each symptom being mentioned in the Abstracts. We only listed the first 18 most

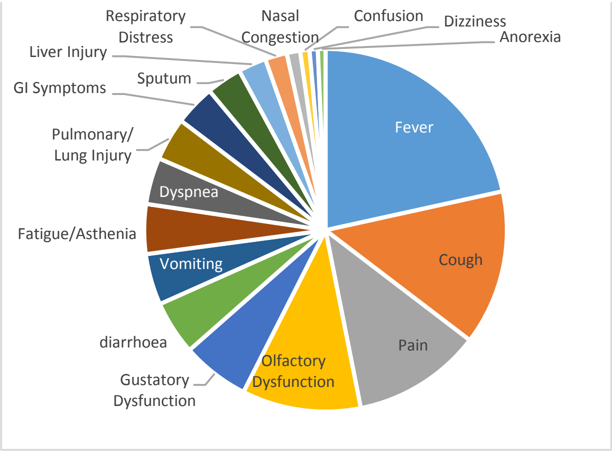

Fig. 7. The distribution of times each symptom being mentioned in the Abstracts.

frequently mentioned symptoms in descending order of the times mentioned in this Fig. 7. We find from this Fig. 7 that, the most prevalent symptom was fever, followed by cough and various pains, which is consistent with the findings in [12]. This prevalent symptom of fever has also been widely used as the simplest and cheapest way by many countries to preliminarily screen out the possible suspects. However, we also find using our big data method that olfactory and gustatory dysfunctions, diarrhoea, vomiting, fatigue/asthenia and dyspnea are among the most significant symptoms, some of which were also mentioned in [11]. This shows that our big data approach is indeed able to reveal some unknown facts embedded in the big pool of data.

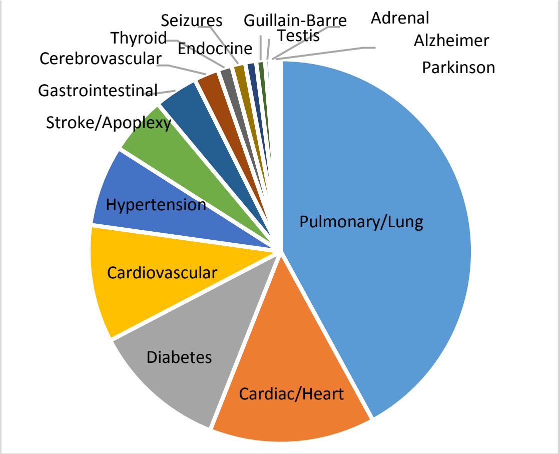

Fig. 8. The distribution of times each of the main complications being mentioned in the Abstracts.

E. *Complications Mentioned in the Research*

COVID-19 patients may be accompanied with a number of different complications during their infection [13]. Some of these complications are accompanied with underlying diseases which may be the major cause that had added death rate to this disease. Therefore, in order to reveal how frequently each complication had been mentioned in the research, we drew in Fig. 8 the distribution of the number of times each of the main complications being mentioned in the Abstracts. We find from this Fig. 8 that lung was the most fragile organ to be infected in

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

this coronavirus infection, followed by the heart. Patients with a history of diabetes, cardiovascular disease and hypertension are also among the most dangerous group. We also find from this Fig. that patients infected with this virus may also be complicated with testis infection.

F. *Diagnostic Methods Mentioned in the Research*

In order to tackle the virus and save patients' life, various diagnostic methods had been used in the researches. Plotted in Fig. 9 is the distribution of times each of the top 10 diagnostic methods being mentioned in the Abstracts. We find from this Fig. 9 that the most frequently used diagnostic method to identify the attack of this virus was the computed tomography (CT) method to check the possible pathological changes in patients' lung and other organs. The reverse transcription-polymerase chain reaction (RT-PCR), Immunoglobulin (IgG) and Immunoglobulin M (IgM) were also among the most frequently used diagnostic methods. The statistical importance of these methods identified by our big data methods were the same as analyzed by other traditional statistic methods in [14].

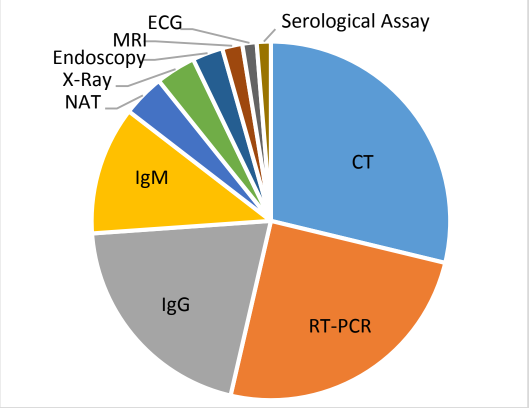

Fig. 9. The distribution of times each of the major diagnostic methods being mentioned in the Abstracts. CT: computed tomography, RT-PCR: reverse transcription-polymerase chain reaction, IgG: Immunoglobulin G, IgM: Immunoglobulin M, NAT: nucleic acid test, MRI: magnetic resonance imaging, and ECG: electrocardiogram.

G. *Research Methods Mentioned in the Literature*

In the course of tackling this disease, clinicians and researchers from different disciplines of the academia had used various methods to find the cause, screen the suspects, cure the patients and prevent the further spread of it, which can roughly be divided into medical and other computer-aided (non-medical) methods. Plotted in Fig. 10 is the distribution of times each of the major medical methods being mentioned in the Abstracts. We find from this Fig. 10 that most of the researchers utilized clinical methods to tackle this emergent public health disaster. This is quite reasonable as during the outbreak of a severe infectious disease, providing proper hospitalization for the patients is the most urgent and effective way of saving people's life. Then the next most frequently used methods were the epidemiological and pathogenesis researches. The former aimed to stop further spread of the disease and the latter to identify the origin and/or cause of this disease, which were also among the most urgent measures. We also find that there were quite a few researchers mentioned using the autopsy method to identify the distribution of virus in postmortem tissues for the definition of its clinical features in order to find a better intervention method to cure the patients.

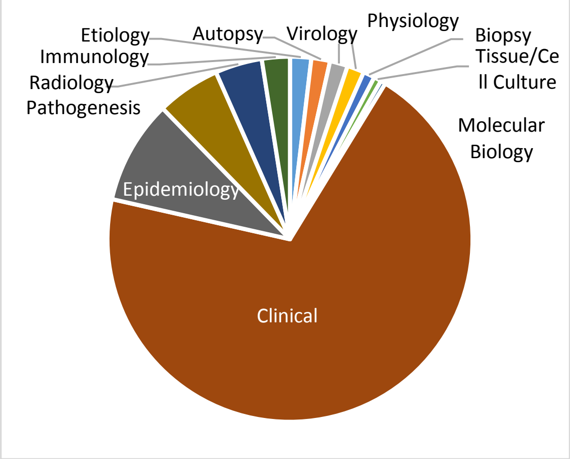

Fig. 10. The distribution of times each of the major medical research methods being mentioned in the Abstracts.

As for the computer-aided methods, drawn in Fig. 11 is the distribution of the most frequently mentioned methods in the literature. We can find from this Fig 11 that modelling was used by most of the researchers followed by correlation and simulation methods. The modelling techniques included the generalized linear and additive models, spatio-temporal models and differential equations, etc. [15]. The authors also found that hot weather may not have enough impact on the transmission of this disease to control the pandemic [15]. Therefore, computer-aided mathematical methods were able to reveal some unusual facts behind the complicated phenomena.

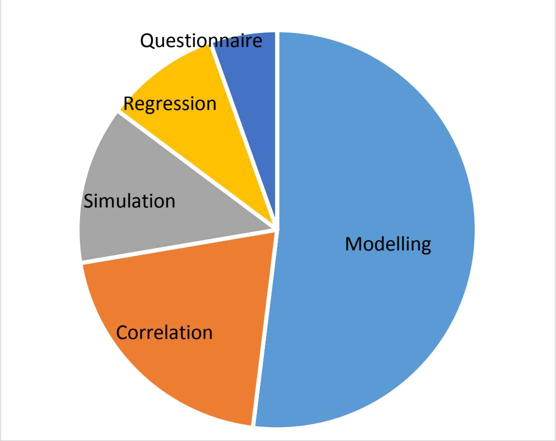

Fig. 11. The distribution of times each of the major non-medical research methods being mentioned in the Abstracts.

IV. Discussion

In order to find the possible relationships between any 2 keywords identified in Section III, we selected 65 notional keywords from the keywords mentioned above, i.e., the words excluding those related to month and people. Based on these words, we searched the abstracts again one by one using the 65*65 pairs of keywords to set up a Co-occurrence Matrix. That is, if any two keywords were simultaneously mentioned by the

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

TABLE I. LIST OF CLUSTERS ALONG WITH THE ADHESIVENESS FOR EACH KEYWORD

Cluster	Word ID	Keyword	Adhesiveness	Cluster	Word ID	Keyword	Adhesiveness
C0	29	Respiratory Distress	232.969	C5	16	Fever	73.719
	41	Clinical	207.547		17	Cough	56.625
	0	Pulmonary/Lung	91.188		18	Pain	32.75
	45	Immunology	69.484		6	Gastrointestinal	23.531
	63	Blood/Urine Test	41.969		23	Fatigue/Asthenia	21.969
	51	CT	36.203		21	Diarrhoea	21.734
	52	RT-PCR	34.484		24	Dyspnea	18.625
	44	Radiology	29.984		22	Vomiting	12.5
	55	NAT	17.547		26	GI Symptoms	11.469
	30	Nasal Congestion	16.141		33	Anorexia	3.266
	27	Sputum	9	C6	28	Liver Injury	20.391
	56	X-Ray	7.703		57	Endoscopy	4.328
C1	1	Cardiac/Heart	62.297		10	Endocrine	2.609
	4	Hypertension	33.172		8	Thyroid	1.484
	3	Cardiovascular	32.797		12	Testis	1.469
	2	Diabetes	32.469		59	ECG	1.297
C2	5	Stroke/Apoplexy	8.609		64	Ultrasonography	1.031
	7	Cerebrovascular	6.047		13	Adrenal	0.656
	9	Seizures	4.016	C7	42	Epidemiology	47.813
	32	Dizziness	3.672		46	Modelling	45.609
	11	Guillain-Barre Syndrome	3.234		47	Correlation	28.469
	31	Confusion	2.656		43	Pathogenesis	25.891
	58	MRI	2.406		39	Tissue/Cell Culture	19.984
	14	Alzheimer	0.938		49	Regression	17.5
	15	Parkinson	0.828		34	Etiology	11.703
	62	EEG	0.359		37	Physiology	9.609
C3	19	Olfactory Dysfunction	13.703		25	Pulmonary/Lung Injury	9.188
	20	Gustatory Dysfunction	9.766		36	Virology	8.984
C4	53	IgG	22.031		48	Simulation	8.922
	60	Serological Assay	15.672		35	Autopsy	7.125
	54	IgM	15.188		50	Questionnaire	5.609
	61	Immunochromatography	2.563		38	Biopsy	5.094
					40	Molecular Biology	1.125

same abstract, the corresponding element in the matrix will be added by 1. Based on this matrix, we can calculate the correlations between the keywords, the coefficient of which was worked out using the following Ochiai coefficient [16]:

$$O_{ij} = \frac{E_{ij}}{\sqrt{E_{ii}E_{jj}}} \tag{1}$$

where E_{ij} is the *ij*th element of the co-occurrence matrix.

Based on the correlation matrix, the keywords could be clustered which is listed in Table I. The adhesiveness of each word is also given in this Table, which is worked out from the following equation:

$$A_i = \frac{\sum_{j\neq i}^{n} E_{ij}}{n-1} \tag{2}$$

where E_{ij} is the *ij*th element of the co-occurrence matrix and n is the total number of keywords.

The keywords in each cluster were sorted by the value of adhesiveness in a descending order. Adhesiveness is the measure of the contribution of a keyword to the cluster and represents the degree of attraction of the keyword to others in the process of clustering. A higher adhesiveness means that the

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

corresponding keyword will have more prominent position in the cluster. In a cluster, the keyword with the largest adhesiveness is called the central word which plays an important role in determining the nature of the cluster.

By analyzing the semantic relationship among the keywords in each cluster and referring to the central word, each cluster in Table I can have a definite connotation. For example, cluster C0 represents clinical diagnosis of the COVID-19; C1 represents cardiovascular disease related complications; C2 represents cerebrovascular diseases related complications and diagnostic measures, etc. However, it is not able to determine the relative importance of each cluster using only the cluster information. In order to identify this, the strategic diagram should be used.

The strategic diagram was proposed by Law *et al.* [17]. It makes use of the information given in the co-occurrence matrix and cluster to evaluate the internal intensity of interaction inside a cluster (represented by density) and mutual influence among different clusters (represented by centrality). The strategic diagram is two dimensional with the mean centrality as the horizontal axis and the mean density as the vertical axis. In the strategic diagram, the centrality represents the strength of interaction between one research field (cluster) and other fields, and the density represents the strength of internal links within the cluster. The centrality (C_k) density (D_k) of any cluster k can be worked out by the following expressions:

$$C_k = \frac{\sum_{(i \in m_k) \cap (j \notin m_k)} E_{ij}}{|m_k|} \quad (3)$$

$$D_k = \frac{\sum_{(i \neq j) \cap (i,j \in m_k)} E_{ij}}{|m_k|} \quad (4)$$

where E_{ij} is the *ij*th element of the co-occurrence matrix, m_k is the set of elements in cluster k and $|m_k|$ is the number of elements in cluster k.

From (3) and (4) we can see that the greater the centrality is, the stronger the research field (cluster) is related to other fields and the more attractive the field is to other fields; and the higher the density is, the stronger ability the research field has to maintain and develop itself and the more stable and mature the research field is. By using the mean of these two parameters as the axes and the difference of each cluster's centrality and density to the corresponding means as the coordinates to draw the strategic diagram, the visualization map would divide the topic domain into four quadrants and any research field (cluster) would fall into one of them. For those in the first quadrant, they have high centrality and density, which indicates that they have extensive relations with other fields, that is, they are at the core of all fields, and high internal strength. If a cluster lies in the second quadrant, it has low centrality but high density, which means that it lies at the edge of the research although has been paid attention to and well studied. The fields in the third quadrant are at the edge of the research and immature. Those in the fourth quadrant mean that they are in the core attention but immature. Therefore, the strategic diagram can be used to describe the maturity and attractiveness of a research field and to show the evolutionary process of the research structure.

Fig. 12 shows the strategic diagram of the COVID-19 research. We find that cluster C0 stands out of all the other clusters lying far in the first quadrant, which is the most concerned and mature topic in COVID-19 research. Examining Table I, we find that the topic of this cluster is about the COVID-19 symptoms and related clinical diagnosis, including symptoms: Respiratory Distress, Pulmonary/Lung disorder, Nasal Congestion, and Sputum; clinical research methods: Radiology and Immunology; and diagnostic methods: Blood/Urine Test, CT, RT-PCR, Nucleic Acid Test, and X-Ray. Therefore, different from the most frequently mentioned symptom of fever, the respiratory distress and pulmonary symptoms/disorders are the most mature diagnosable symptoms. This is another piece of useful information that can only be mined out by using the strategic diagram approach.

Another two research fields (clusters) that are the core and relative mature are the C1 and C5. The former is about the cardiovascular-related complications and the latter is about observable symptoms. We find that fever is at last identified by the strategic diagram but its adhesiveness is much less than that of respiratory distress. Therefore, compared with frequency analysis, strategic diagram is far more important in identifying core and mature topics in a research big dataset.

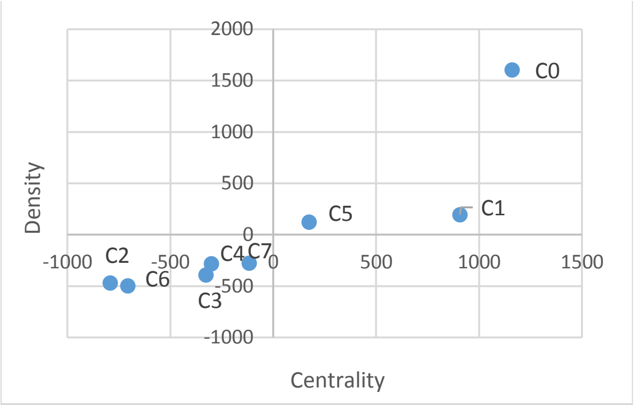

Fig. 12. Strategic diagram of the COVID-19 research.

Other clusters in strategic diagram all lie in the third quadrant, which means that those research fields were very diversified and relatively immature. This reflects the fact that due to the severity of this disease and the rapid spread of it, many researchers had tried every possible measure to stop the spread, identify the infected and rescue patient's life, which has led to no major focus of the research and most of the research fields still need more work to mature. However, with a better understanding of the nature of the disease and the progress and maturity of research, the elements in each cluster and the position of each cluster in the strategic diagram would both change. In this way, we can obtain some more stable and mature research fields to tackle this disease in the future.

V. Conclusions

The COVID-19 pandemic is still a big threat today to the health of the human beings. The number of the infected and the dead is still increasing day by day. Furthermore, many countries are still in a chaos due to this threat. Therefore, it is urgent to find some effective measures to suppress its rampancy. Among other methods, big data is the one used to process and analyze

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.

all the data collected instead of the traditional statistic methods limited to just a small set of samples to gain an overall perspective of the investigated events. This paper is the first attempt to use this method to reveal some unknown facts concealed in the disorganized huge pool of the COVID-19 research data.

Our results showed that we have dug out some of the important facts concealed in the research literature by using this new tool. We found that undeveloped countries had not been paid enough attention by researchers in this health catastrophe; vulnerable groups of people had been paid appropriate attention by researchers; lung was the most fragile organ to be infected but some other organs even the testis cannot be overlooked either; CT and RT-PCR were among the most frequently-used diagnostic methods; and clinical and modelling methods were the most favorite ones used by researchers as medical and non-medical research tools etc. By using strategic diagram, we also found that instead of fever, respiratory distress and pulmonary symptoms/disorders were the most mature diagnosable symptoms. These results showed that our new attempt of big data approach is useful for digging out some valuable information to guide the future research in this area.

Furthermore, with the coming of winter in the Northern Hemisphere, a new wave of the epidemic is raging in many countries. According to the latest statistics released by Johns Hopkins University on December 10, 2020 US Eastern time, 711, 562 new confirmed cases have been added in the past 24 hours, setting to a new record. Our results would be able to serve as a good reference for reducing the number of the infected and death rate. New research data collected in the course of the research would be able to allow us to mine out more useful data by using this big data approach to identify better measures to tackle this disease.

REFERENCES

[1] E. Ortiz-Prado, K. Simbana-Rivera, L. Gomez-Barreno, M. Rubio-Neira, L. P. Guaman, and N. C. Kyriakidis *et al*., “Clinical, molecular, and epidemiological characterization of the SARS-CoV-2 virus and the Coronavirus Disease 2019 (COVID-19), a comprehensive literature review,” Diagn Micr Infec Dis, vol. 98(1), 2020.

[2] D. S. Chauhan, R. Prasad, R. Srivastava, M. Jaggi, S. C. Chauhan, and M. M. Yallapu, “Comprehensive review on current interventions, diagnostics, and nanotechnology perspectives against SARS-CoV-2. Bioconjugate chem, 2020.

[3] A. S. Adly, A. S. Adly, and M. S. Adly, “Approaches based on artificial intelligence and the internet of intelligent things to prevent the spread of COVID-19: scoping review,” J Med Internet Res, 2020, vol. 22(8).

[4] C. M. Chen, H. W. Jyan, S. C. Chien, H. H. Jen, C. Y. Hsu, and P. C. Lee *et al*., “Containing COVID-19 among 627,386 persons in contact with the Diamond Princess Cruise Ship passengers who disembarked in Taiwan: big data analytics,” J Med Internet Res, 2020, vol. 22(5).

[5] H. J. Qiu, L. X. Yuan, Q. W. Wu, Y. Q. Zhou, R. Zheng, X. Huang, and Q. T. Yang, “Using the internet search data to investigate symptom characteristics of COVID-19: a big data study,” World J Otolaryng Head Neck. 2020.

[6] L. Lin and Z. Hou, “Combat COVID-19 with artificial intelligence and big data,” J Travel Med, 2020, vol. 27(5).

[7] Web of Science: www.webofknowledge.com/ (accessed on 8 September 2020).

[8] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, and Z. Liu *et al*., “Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study,” Lancet, 2020, vol. 395(10229), pp. 1054-1062.

[9] A. I. Moula, L. R. Micali, F. Matteucci, F. Luca, C. M. Rao, and O. Parise *et al*., “Quantification of death risk in relation to sex, pre-existing cardiovascular diseases and risk factors in COVID-19 patients: let's take stock and see where we are,” J Clin Med, 2020, vol. 9(9).

[10] S. J. Kang and S. I. Jung, “Age-related morbidity and mortality among patients with COVID-19,” J Infect Chemother, 2020, vol. 52(2), pp. 154-164.

[11] M. C. Grant, L. Geoghegan, M. Arbyn, Z. Mohammed, L. McGuinness, E. L. Clarke, and R. G. Wade, “The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries,” Plos One, 2020, vol. 15(6).

[12] G. Sacco, O. Briere, M. Asfar, O. Guerin, G. Berrut, and C. Annweiler, “Symptoms of COVID-19 among older adults: systematic review of biomedical literature,” Geriatr Psychol Neur, 2020, vol. 18(2), pp.135-140.

[13] I. Ahmad, and F. A. Rathore, “Neurological manifestations and complications of COVID-19: A literature review,” J Clin Neurosci, 2020, vol. 77, pp. 8-12.

[14] S. Manigandan, M. T. Wu, V. K. Ponnusamy, V. B. Raghavendra, A. Pugazhendhi, and K. Brindhadevi, “A systematic review on recent trends in transmission, diagnosis, prevention and imaging features of COVID-19,” Process Biochem, 2020, vol. 98, pp. 233-240.

[15] A. Briz-Redon and A. Serrano-Aroca, “The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques,” Prog Phys Geog. 2020.

[16] O. Akira, “Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions – II,” Bull Jap Soc Sci Fish, 1957, vol. 22 (9), pp. 526-530.

[17] J. Law, S. Bauin, J. P. Courtial, and J. Whittaker, “Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification,” Scientometrics, 1988, vol. 14(3-4), pp. 251-264.

Authorized licensed use limited to: Miami University Libraries. Downloaded on June 15,2021 at 17:06:01 UTC from IEEE Xplore. Restrictions apply.