Covid-19 knowledge deconstruction and retrieval: Solutions of intelligent bibliometrics

Mengjia Wu†
Australian Artificial Intelligence
University of Technology Sydney
Australia
mengjia.wu@uts.edu.au

Yi Zhang
Australian Artificial Intelligence
University of Technology Sydney
Australia
yi.zhang@uts.edu.au

Mark Markley
Search Technology, Inc.
United States
markmarkley@searchtech.com

Caitlin Cassidy
Search Technology, Inc.
United States
caitlin.cassidy@searchtech.com

Nils Newman
Search Technology, Inc.
United States
newman@searchtech.com

Alan Porter
Search Technology, Inc.
Georgia Institute of Technology
United States
aporter@searchtech.com

ABSTRACT

Covid-19 is an unprecedented challenge that disruptively reshapes societies and scientific research communities. Facing the knowledge flood brought by the overwhelming volume of research efforts, there still lacks a platform to link those to previous knowledge foundations and efficiently visualize and understand them. Aiming to fill this gap, we propose a research framework in this paper to assist scientists in identifying, retrieving, and visualizing the emerging Covid-19 knowledge. The proposed framework incorporates principal topic decomposition (PCD), text analytics-based knowledge model (KM), and the hierarchical topic tree (HTT) method to profile the research landscape, retrieve knowledge of specific interest, and visualize the knowledge structures. Initially, our topic analysis of 127,971 research papers published during 2020-2021 identified 35 research hotspots. Furthermore, we built up a knowledge model on the topic of vaccination and retrieved 92,286 research papers from the entire PubMed database as the knowledge foundation of this topic. Lastly, the HTT results of the retrieved papers highlighted multiple relevant disciplines, from whose branches we identified four future research directions: Monoclonal antibody treatments, vaccination in diabetic patients, vaccination effectiveness in SARS-CoV-2 antigenic drift, and vaccination-related allergic sensitization.

CCS CONCEPTS

• Document management and text processing • Information retrieval • Computing/technology policy

KEYWORDS

Covid-19, knowledge retrieval, intelligent bibliometrics

1 Introduction

Since the outbreak of the Covid-19 pandemic in 2020, global scientists have contributed more than 200,000 research papers to investigate the nature of this virus and help mitigate its negative impacts. However, the pandemic also comes with an information crisis [1, 2]. Apart from the problem of misinformation and rumors, the overwhelming and growing speed of research papers results in a severe information overload. This publication explosion challenges scientists, healthcare professionals, and the public to 1) keep up with the rapid accumulation of new knowledge, 2) accurately and comprehensively obtain knowledge on specific topics, and 3) understand the retrieved emerging new knowledge. Even though there are already some open literature datasets [3, 4] and search tools [5, 6] available online, a comprehensive framework is lacking to provide an effective solution for the challenges. Based on the current situation, we summarize three key unresolved research questions:

- Question 1 (Q1): What are the key topics of the Covid-19 knowledge system?
- Question 2 (Q2): How can we retrieve knowledge foundations for specific Covid-19 topics?
- Question 3 (Q3): How do we understand the retrieved knowledge on specific topics?

Existing efforts trying to address these research questions mainly consist of Covid-19 research topic analysis [7-10], literature-based discovery studies [11-13], and literature search tools [4, 5]. A common approach seen in current topic studies is using co-word clustering [7], or topic modeling [8, 9] on post-Covid literature (scholarly papers published after the Covid-19 outbreak) to identify and depict the research landscape. Such studies are useful in tracking new knowledge evidence but may overlook the internal relationships of new evidence with previously established coronavirus knowledge foundations. Knowledge foundations are significant to facilitate knowledge discovery [14-17]. Besides, most literature-based discovery studies on the global Covid-19 dataset [11, 13] do not dive into specific topics to discover targeted knowledge for people pursuing different interests, such as basic medical research, epidemiological models, and social impacts research, etc. Given that, combing topic analysis and literature-based discovery approaches could be a promising way to fill these two gaps. As for Covid-19 knowledge retrieval and

understanding, few of the available search tools provide visualizations or other efficient ways to assist users in comprehending the retrieved results [5, 6]. A concise and appropriate visualization could save their time in finding papers to follow or narrowing down their search scope. Aiming to fill the research gaps and provide a comprehensive solution to the three proposed research questions, we propose a research framework in this paper with its details as follows.

To answer Q1, we utilized principal component decomposition (PCD) to identify research topics from scientific literature, yielding a bird's eye view of Covid-19's knowledge system. For identified Covid-19 topics, we further employed text analytics and developed a text similarity-based knowledge model to retrieve relevant documents across the entire PubMed database, linking every identified topic with relevant pre-Covid literature, which can be regarded as the topic's knowledge foundation. By combing topic analysis and literature-based discovery, we composed our answer for Q2. Targeting Q3, we focused on hierarchy, a specific dimension of knowledge composition, to profile and visualize the hierarchical intellectual structure of the retrieved knowledge body of certain research topics. This can help researchers efficiently understand the retrieved knowledge foundation and further support knowledge discovery. In all, this data-driven study blends multiple, AI-empowered bibliometric approaches -- what we call in our pilot studies, "intelligent bibliometrics" [18] -- to reveal insights for Covid-19 knowledge deconstruction, effective retrieval, and understanding.

In the case study, we collected 127,971 research papers published in 2020 and 2021 from the PubMed database. Feeding those papers into our research framework, we first generated 35 PCD topics and revealed the different emphasis in different periods, changing from the epidemiological and clinical characteristics to the impacts of Covid on societies. Additionally, we constructed a knowledge model based on the most popular PCD topic of vaccination; then a global search was run against the entire PubMed records before 2020 to retrieve the knowledge foundations of this topic, ending up with 92,286 research papers as the knowledge foundations of this topic. Lastly, we utilized HTT to visualize the knowledge structures of the retrieved results; the HTT results highlighted multiple vaccination-related disciplines, including immunology, molecular biology, virology, etc. From the branches of those disciplines, we identified four future research directions: Monoclonal antibody treatments, vaccination in diabetic patients, vaccination effectiveness in SARS-CoV-2 antigenic drift, and vaccination-related allergic sensitization. We empirically evaluated the results by matching evidence identified from the literature and identified research evidence in the latest studies. This empirical case does not only demonstrate the reliability of our method but also derive insights to support potential COVID-related R&D and strategic management in funding agencies, research individuals, and institutions.

2 Data and methods

Our research framework is illustrated in Figure 1. In the following subsections, we will detail the data collection process and our three core methods, i.e., the method of principal component decomposition (PCD) for profiling the research landscape of Covid-19 research, a knowledge model (KM) for knowledge retrieval of specific Covid-19 topics, and the method of hierarchical topic tree (HTT) for identifying topic hierarchies in divergent research directions.

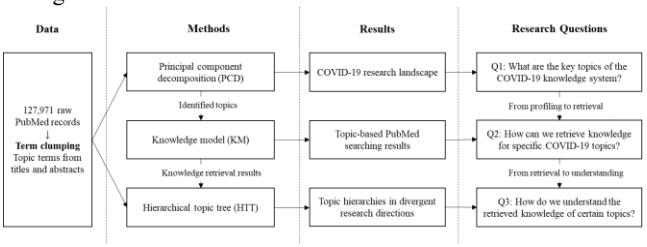

Figure 1: Research framework of the Covid-19 knowledge deconstruction and retrieval

To collect Covid-19 bibliographic data, we investigated multiple data sources in our pilot study [19] and ultimately decided on PubMed. Compared to the larger dataset with massive preprints like CORD-19 [3], most articles from PubMed are peer-reviewed and contain more curated metadata (e.g., MeSH Descriptors and Qualifiers) for our following systematic comparative analysis. By mining this globally largest and most comprehensive, open-source biomedical database, we retrieved 127,971 relevant research papers from 1/1/2020 – 1/1/2022.

We further applied the natural language processing function of VantagePoint[1] to extract topic terms from titles and abstracts. The list of extracted terms was cleaned to remove stop words, consolidate similar terms, and eliminate all terms appearing only once [20]. The term clumping process and stepwise results are given in Table 1.

[1] VantagePoint is a software platform for bibliometrics-based text analytics and knowledge management, owned by Search Technology Inc. More details can be found at the website: www.thevantagepoint.com.

Table 1: Stepwise results of term clumping

Step	Detail	# Terms
1	Extract terms from titles and abstracts using VantagePoint NLP function	1,603,542
2	Remove terms starting/ending with non-alphabetic characters	1,367,374
	Remove common terms in scientific articles, e.g., "research framework"	
	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions	
	Consolidate synonyms based on expert knowledge, e.g., "Covid-19" and "Covid"	
	Consolidate terms with the same stem, e.g., "severe patient" and "severe patients"	

2.1 Principal component decomposition (PCD)

PCD is essentially an iterative process of principal components analysis (PCA), which classifies and categorizes research papers to represent the research landscape [21]. By applying PCD to the extracted terms, we can retain the principal popular terms and conclude them as PCD research topics, with each paper assigned to the most semantically similar PCD topic.

2.2 Knowledge model (KM)-based document retrieval

Knowledge model (KM)-based document retrieval aims to find documents with high semantic similarities with given a collection of text [22]. Specifically, we construct a KM containing the subset of relevant papers and their corresponding topic terms; the top and bottom 50 terms with the highest/lowest average term frequency-inverse document frequency (TF-IDF) constitute the feature vector of this set. We then employ the KM to search related documents across the entire records in the PubMed database before 2020, containing over 30 million research papers at the time of retrieval. Based on document frequencies, we compute and rank the cosine similarity between the KM's feature vector and TF-IDFs extracted for those same terms for each document in the PubMed set. The ranking list is the outcome of this KM-based document retrieval model, indicating the semantic priority of external documents in PubMed to the given topic cluster.

2.3 Hierarchical topic tree (HTT)

Hierarchical topic tree analysis [23] is a network analytics-based method that identifies research topics and their hierarchies from scientific documents. This method identifies nodes with 1) notably high density and 2) relatively far distance to other high-density nodes as anchor leaves, and then assigns the rest of the nodes to their proximate anchor leaves to form node communities (research topics). This process is then iteratively applied to the sub-clusters of each layer to identify their sub-communities until no anchor leaves are found. Each iteration constitutes a topic layer of the final tree. Thus, a tree consists of anchor leaves and their node communities and sub-communities, representing the intellectual structure of a knowledge system. The stepwise processes of this method are given below:

Step 1: Construct the term co-occurrence network of documents and calculate the shortest distances of pairwise nodes.

$$G = (V, E)$$

where G is the term co-occurrence network, V is the set of term nodes and E is the set of co-occurrence edges.

$$w_{E_{ij\,(i\neq j)}} = \begin{cases} CF(V_i, V_j) & if\ V_i\ and\ V_j\ co-occur\ at\ lease\ once \\ 0 & otherwise \end{cases}$$

where $w_{E_{ij\,(i\neq j)}}$ is the edge weight of the edge connecting nodes V_i and V_j, $CF(V_i, V_j)$ represents the co-occurring weight (number of documents that V_i and V_j co-occur) of nodes V_i and V_j.

Step 2: Calculate the neighbor density for each node and generate the shortest distance of every node to its closet node with a higher neighbor density. Considering the scalability of our algorithm on this network, we used neighborhood density to proxy the density measures of each node.

$$\rho_{V_i} = \exp\left(-\frac{1}{\Gamma(V_i)} \sum_{V_j \in \Gamma(V_i)} \frac{1}{{w_{E_{ij}}}^2}\right)$$

in which ρ_{V_i} denotes the local density of node V_i, and $\Gamma(V_i)$ is the neighbor node set of V_i.

$$\delta_{V_i} = \begin{cases} \max\limits_{V_j \in \Gamma(V_i)} \left(\frac{1}{w_{V_i V_j}}\right) & if\ \rho_{V_i} = \max\limits_{V_j \in \Gamma(V_i)} (\rho_{V_j}) \\ \min\limits_{V_j\, \epsilon\, V_{\rho_{V_j} > \rho_{V_i}}} \left(\frac{1}{w_{V_i V_j}}\right) & otherwise \end{cases}$$

in which δ_{V_i} is the shortest distance from V_i to its closest neighbor node with larger local density.

Step 3: Locate community centers, V_c, with local density peak values according to the following formula; initialize them as community centers, and then allocate the rest of the nodes to the nearest V_c.

$$\rho_{V_c} > \varepsilon \max_{V_i \in \Gamma(V_c)} \rho_{V_i}$$

in which ε is the density threshold that decides the significance of a topic.

Step 4: Iterate Step 3 to the subcommunities until no centroid node can be found in any sub-communities. From the second iteration, an additional criterion will be added to guarantee the identified centroids for sub-communities are sparse to each other:

$$\delta_{V_c} > \frac{1}{w_{V_r V_c}}$$

in which V_r denotes the node centroid of its parent community.

Then every rest node in will be assigned to the closet centroid. By applying Steps 1-4, we will partition the co-occurrence network into a set of hierarchical communities. Every community and its subcommunities composes a branch of the final hierarchical topic tree.

3 Results

We first applied the descriptive bibliographic analysis to review the collected papers in terms of the monthly growth, institution ranking, and geographical distribution of Covid-19 research papers.

Figure 2 illustrates that the rate of literature growth was exponential in early 2020 and then stabilized in 2021, reflecting the disruptive influence of Covid-19 on scientific activities. We attribute the early stage burst of publications to the need to reveal novel knowledge and facts about this new virus and disease [10]. However, the reasons for the slowing rate can be much more complicated. Is it due to the research capacity limitation (funding, paper review process, journal), or is it indicating the start of knowledge convergence? How long will the effects of Covid-19 last on scientific communities? When will the decay period start? More explorations and efforts are needed to answer these questions.

Figures 3, 4 and 5 respectively profile the global distribution of the Covid-19 research papers among worldwide institutions and countries. In terms of the absolute number of papers published at the national level (Figures 3 and 5), the United States and China unsurprisingly hold leading positions, followed by Italy, India, Germany, Canada, etc. Diving into the institution level (Figure 4), we found that compared to the earlier China-led trends in Covid-19 research [24], the momentum for US institutions to lead in this domain is continuously growing [12]. This indicates that even though China has a substantial volume of papers published, among top institutions, individual Chinese universities and research institutions do not demonstrate equal strengths in competition with their global counterparts, particularly those from the States.

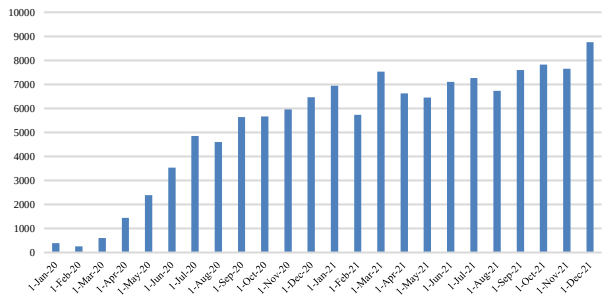

Figure 2: Monthly increasing trend of Covid-19 research papers

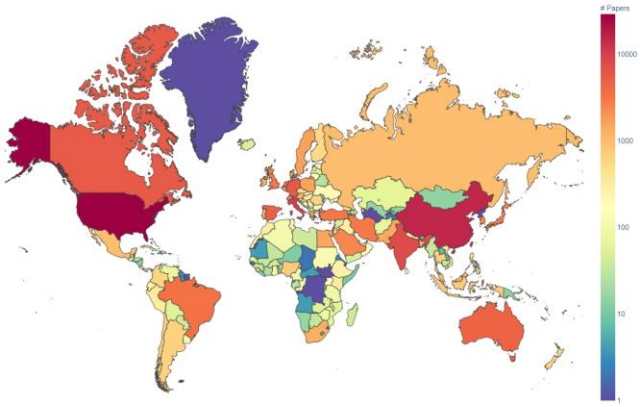

Figure 3: The geographical distribution of Covid-19 papers

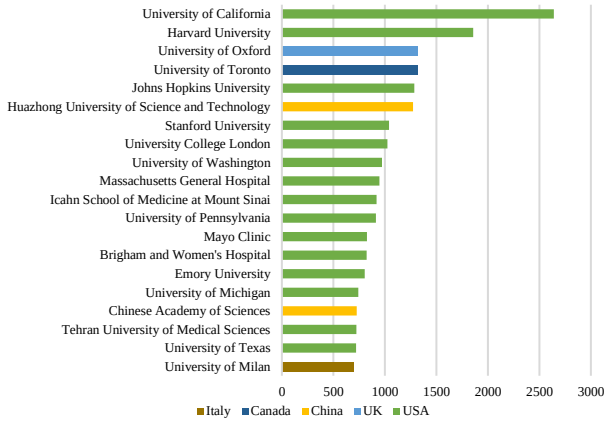

Figure 4: Top 20 prolific research institutions

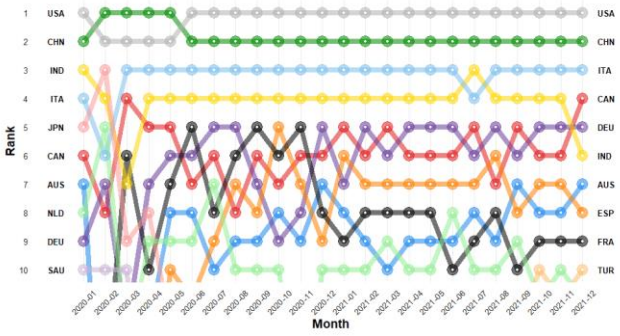

Figure 5: The ranking changes of countries

3.1 Research landscape of Covid-19

Feeding the extracted topic terms into the PCD algorithm, we generated 35 topics and present them in Figure 6, with each bubble representing a PCD research topic and the size denoting its associated paper count. We further applied cross-correlation

analysis to the generated topics, with the links denoting the substantial cosine correlation [25] degree above 0.5. The core topics in the middle of Figure 6 represent an internal strongly-connected topic cluster of clinical investigations, while the other topics appear to be relatively independent of each other.

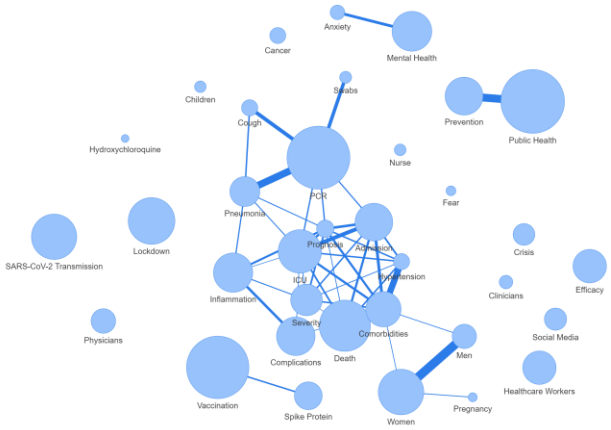

Figure 6: The distribution and cross-correlation of PCD topics

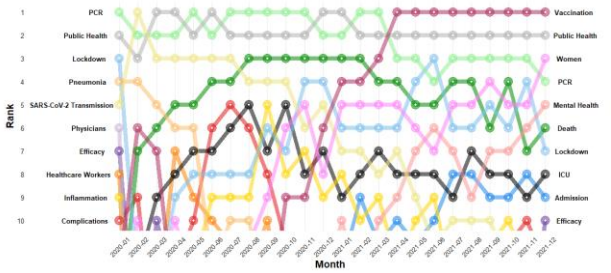

Figure 7: Monthly increasing trend of PCD topics

The monthly ranking changes of the top ten topics are given in Figure 7, indicating different stages of Covid-19 research. Among these topics, the rankings of *PCR* and *Public Health* maintained the top, while other topics show significant fluctuating trends.

At the beginning of the Covid-19 breakout in Wuhan, the PCD topics *Pneumonia*, and *SARS-CoV-2 Transmission* attracted massive attention, as first-hand clinical and epidemiological investigations were urgently needed to improve Covid treatments and control its transmission [26-29]. In such investigations and following clinical trials, the gender difference is an essential associated factor as indicated by the continuing ranking rise of PCD topics *Women* and *Men*; Additional attention was especially put on the female group to investigate the vulnerabilities of pregnant women or women at lactating ages [30]. With Covid-19 turned from regional transmissions into a global pandemic, scientists started to examine the social impacts of Covid-19 as illustrated by the rise of topics *Lockdown* [31, 32] and *Mental Health*. The former topic broadly covers the social impacts of lockdown measures on healthcare services [32], economy [33], education [34], environment [35], etc.; The latter topic discusses mental health issues among the general public [36, 37] and healthcare workers [38]. With the anti-epidemic activities gradually normalizing and becoming a part of daily life, rankings of topics *Death* and *ICU* relatively decrease steadily.

Notably, the change in *Vaccination*-related papers illustrates two waves of vaccine studies. The first wave appeared at the beginning of the Covid-19 breakout and peaked in February 2020. These early-stage papers majorly focus on reviewing past coronavirus vaccines, calling for rapid vaccine development procedures, and proposing possible developing approaches [39-42]. With the advent of multiple available vaccines, the next vaccination research wave emerged in the third quarter of 2020 and continued to rise. In addition to the massive basic medicine and clinical trial studies around the safety and efficacy of those vaccines [43-45], the rollout of vaccines also triggers researchers' concerns about the social implications, including vaccine hesitancy phenomena [46, 47], vaccine allocation strategy [48] and vaccination incentives [49, 50]. As vaccination is one of the most effective measures in preventing Covid-19, we will demonstrate how we used our knowledge model to retrieve historical knowledge of vaccination studies in the next section.

3.2 Knowledge model search results

This section demonstrates the utility of our knowledge model approach in retrieving historical knowledge from the entire PubMed database, using the most prominent PCD research topic, *Vaccination*, as an example. We selected 15,967 papers related to this topic and calculated the TF-IDF values of all the extracted terms of those papers, then a knowledge model was built up with its top 50 and bottom 50 term stems[2]. Further, we run the KM model against the entire PubMed database and retrieved 92,286 historical records out of the Covid dataset. In the next section, we will demonstrate how we could visualize the knowledge of the search results.

3.3 HTT results of knowledge modelling

We further mapped the 92,286 records to Open Academic Graph (OAG)[3] and retrieved 89,951 records with the field of study (FoS) values [51]. To efficiently understand and visualize the knowledge in the search results, we constructed a FoS co-occurrence network and ran our HTT algorithm on it. The density threshold was empirically set as 0.95. The topological characteristics of the FoS network are given in Table 2.

[2] The details of this KM are available at https://github.com/IntelligentBibliometrics/Covid_knowledge.
[3] https://www.aminer.cn/oag-2-1

Table 2: The characteristics of the FoS network

	Number	Weight			
		Max.	Min.	Avg.	Std.
Node	27,596	39,542	1	3.459	44.336
Edge	922,252	18,737	1	28.135	427.105
Average degree			66.840		

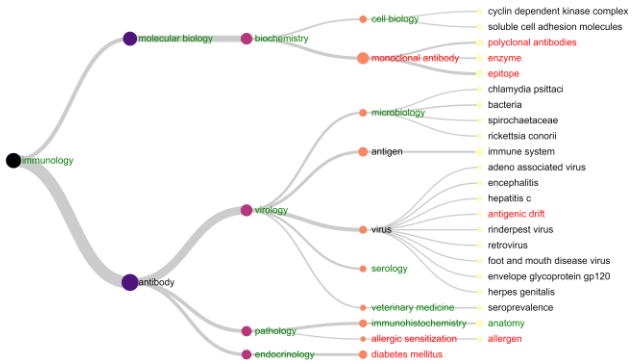

Figure 8: The HTT of retrieved results

We trimmed this HTT to retain the main knowledge body and presented it in Figure 8[4]. Immunology is the root topic of this HTT, indicating that vaccination studies are mostly constructed based on immunology knowledge. For the presented topics, we primarily highlighted some of them as discipline-level topics (green font) and entity-level topics (red font). We compared and contrasted the historical records (regarded as knowledge foundations) with the latest research and concluded the following insights.

- ***Monoclonal antibody***: This topic is positioned in the branch of *molecular biology – biochemistry*. Diving into this topic, we can trace a bunch of historical studies of developing monoclonal antibodies as treatments for existing human and animal coronaviruses, including severe acute respiratory syndrome (SARS) coronavirus [52, 53] and bovine coronaviruses [54, 55]. Such studies can provide instructive research clues for developing novel monoclonal antibody treatments for Covid-19. With the approvement of multiple monoclonal antibody treatments for Covid-19, more efforts will predictably be put into finding efficient methods of extracting and producing such monoclonal antibodies [56].
- ***Antigenic drift***: This topic exists in the *virus* branch, describing a natural phenomenon of antigen genetic mutations that also happens in the SARS-CoV-2 virus [57]. Medical experts can trace historical studies of influenza viruses [58, 59] and other possibly related viruses [60] in search results to infer and analyze the impacts of antigenic drift on vaccination implementations. The effectiveness and immune durability of current vaccines for various SARS-CoV-2 variants (including currently concerned Omicron) need deeper exploration [61, 62].
- ***Diabetes***: Locating on the endocrinology branch, this topic consists of historical papers clarifying the autoimmune-mediated beta-cell damage mechanisms [63], significant autoantigens [64], and different subtypes of type 1 diabetes [65, 66]. Recent studies reported that two types of diabetes were both associated with higher odds of Covid-19 hospital deaths [67, 68], and SARS-CoV-2 infection possibly induces negative effects on beta-cells [69-72]. Consequently, vaccination in diabetic patients has become a trending topic among vaccination studies. On the one hand, a lot of researchers have called for prioritizing vaccination in diabetic patients as they are more vulnerable to Covid-19 [73, 74]. On the other hand, associating the knowledge (especially for type 1 diabetes) from our search results with Covid vaccination is worth deeper exploration because current evidence is still limited [72, 75].
- ***Allergic sensitization***: Historical studies related to this topic majorly discussed the reactivity of immunoglobin E in allergic reactions [76-78], which can provide instructive insights for Covid-19 vaccination allergic studies [79, 80].

In conclusion, the HTT results yielded us an overview of the search results. The high-level HTT topics highlighted multiple disciplines relevant to vaccination studies, including *immunology*, *molecular biology*, *virology*, etc. The topics on those discipline branches reveal four future directions based on established knowledge foundations: Monoclonal antibody treatments, vaccine immunes effectiveness and durability, vaccination in diabetic patients, and vaccine allergy.

4 Discussion

Covid-19 brings a global public health pandemic and an overwhelming knowledge flood. Aiming to efficiently discover and utilize the knowledge laid in the massive Covid-19 scientific studies, we propose an incorporated research framework to 1) profile the Covid-19 knowledge landscape and research topics in both flat and hierarchical levels, 2) retrieve knowledge foundation related to specific topics, and 3) visualize the retrieved knowledge to support knowledge understanding and discovery. We anticipate our research methodology and key findings can support a) scientific researchers to quickly absorb emerging new knowledge and identify their future study directions and b) research

[4] The entire HTT can be found at https://github.com/IntelligentBibliometrics/Covid_knowledge/blob/main/Vacc_all.svg

policymakers to make informed decisions about research funding allocation.

4.1 Key findings

Q1: What are the key topics of the Covid-19 knowledge system?

We exploited PCD analysis to profile the Covid-19 research landscape. The results highlighted 35 research hotspots and research emphases during different periods. The changing trends in PCD topic rankings indicate that early Covid-19 studies investigated the clinical and epidemiology characteristics of Covid-19, while the subsequent studies threw more light on the societal impacts of Covid-19 on different people groups. Intriguingly, the change in PCD topic *vaccination* papers illustrates two waves of vaccination studies, respectively appeared at the start of the Covid outbreak and after the rollout of multiple available vaccines.

Q2: How can we retrieve knowledge foundations for specific Covid-19 topics?

We developed a text analytics-based knowledge model to discover the knowledge foundations and demonstrated its utility using the topic of *vaccination* in Section 3.3. Using this KM, we conducted a global search against the entire PubMed database and retrieved 92,286 papers that own high document semantic similarities with records in the topic, which were regarded as the knowledge foundations of this topic.

Q3: How do we understand the retrieved knowledge on specific topics?

We ran our HTT algorithm on the KM search results and uncovered the knowledge hierarchies of topics. At the top levels of the HTT, we identified multiple significant medical disciplines, including *immunology*, *molecular biology*, *virology*, etc. Apart from the disciplines, we also identified four directions worth more attention in future vaccination-related studies, which are respectively 1) monoclonal antibody treatments, 2) vaccination priority and immune responses in diabetic patients, 3) the effectiveness of vaccines on various SARS-CoV-2 mutations, and 4) vaccination allergy. Such insights can 1) inspire medical researchers to conduct future studies with knowledge reference and foundation and 2) assist scientific policymakers in making informed decisions about research funding allocation.

Nevertheless, there are still some limitations to our study. From the methodological perspective, there is a limitation stemming from our knowledge model and HTT approaches: The two methods both need parameter configurations; we empirically selected parameters in this case to achieve better results, but developing an automatic data-driven parameter fine-tuning process is a direction we are heading in the future; From the theoretical perspective, we profiled the knowledge landscape of Covid-19 and obtained knowledge foundations of the *vaccination* topic. Compared to obtaining literature-based evidence, it might be further interesting to validate the results with clinical trials and in-depth expert consultations.

ACKNOWLEDGMENTS

Yi Zhang and Mengjia Wu are supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

REFERENCES

[1] B. Xie *et al.*, "Global health crises are also information crises: A call to action," *Journal of the Association for Information Science and Technology*, vol. 71, no. 12, pp. 1419-1423, 2020.

[2] M. Chahrour *et al.*, "A bibliometric analysis of COVID-19 research activity: A call for increased output," *Cureus*, vol. 12, no. 3, 2020.

[3] L. L. Wang *et al.*, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020, doi: 10.48550/arXiv.2004.10706.

[4] Q. Chen, A. Allot, and Z. Lu, "LitCovid: An open database of COVID-19 literature," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1534-D1540, 2021.

[5] A. Trewartha *et al.*, "COVIDScholar: An automated COVID-19 research aggregation and analysis platform," *arXiv preprint arXiv:2012.03891*, 2020, doi: 10.48550/arXiv.2012.03891.

[6] E. Zhang, N. Gupta, R. Nogueira, K. Cho, and J. Lin, "Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned," *arXiv preprint arXiv:2004.05125*, 2020, doi: 10.48550/arXiv.2004.05125.

[7] A. Pourhatami, M. Kaviyani-Charati, B. Kargar, H. Baziyad, M. Kargar, and C. Olmeda-Gómez, "Mapping the intellectual structure of the coronavirus field (2000–2020): A co-word analysis," *Scientometrics*, vol. 126, no. 8, pp. 6625-6657, 2021.

[8] G. Colavizza, R. Costas, V. A. Traag, N. J. Van Eck, T. Van Leeuwen, and L. Waltman, "A scientometric overview of CORD-19," *PloS One*, vol. 16, no. 1, p. e0244839, 2021.

[9] B. X. Tran *et al.*, "Studies of novel coronavirus disease 19 (COVID-19) pandemic: A global analysis of literature," *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, p. 4095, 2020.

[10] Y. Zhang, X. Cai, C. V. Fry, M. Wu, and C. S. Wagner, "Topic evolution, disruption and resilience in early COVID-19 research," *Scientometrics*, vol. 126, no. 5, pp. 4225-4253, 2021.

[11] Q. Yu *et al.*, "Analyzing knowledge entities about COVID-19 using entitymetrics," *Scientometrics*, vol. 126, no. 5, pp. 4491-4509, 2021.

[12] M. Wu *et al.*, "Profiling COVID-19 genetic research: A data-driven study utilizing intelligent bibliometrics," *Frontiers in Research Metrics and Analytics*, vol. 6, p. 30, 2021.

[13] C. Wise *et al.*, "COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature," *arXiv preprint arXiv:2007.12731*, 2020, doi: 10.48550/arXiv.2007.12731.

[14] M. Haghani and M. C. Bliemer, "Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCoV literature," *Scientometrics*, vol. 125, no. 3, pp. 2695-2726, 2020.

[15] M. Haghani and P. Varamini, "Temporal evolution, most influential studies and sleeping beauties of the coronavirus literature," *Scientometrics*, vol. 126, no. 8, pp. 7005-7050, 2021.

[16] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, "Characteristics of SARS-CoV-2 and COVID-19," *Nature Reviews Microbiology*, vol. 19, no. 3, pp. 141-154, 2021.

[17] N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, and E. Petersen, "COVID-19, SARS and MERS: Are they closely related?," *Clinical Microbiology and Infection*, vol. 26, no. 6, pp. 729-734, 2020.

[18] Y. Zhang, A. L. Porter, S. Cunningham, D. Chiavetta, and N. Newman, "Parallel or Intersecting Lines? Intelligent Bibliometrics for Investigating the Involvement of Data Science in Policy Analysis," *IEEE Transactions on Engineering Management*, pp. 1-13, 2020, doi: 10.1109/TEM.2020.2974761.

[19] A. L. Porter, Y. Zhang, Y. Huang, and M. Wu, "Tracking and Mining the COVID-19 Research Literature," *Frontiers in Research Metrics and Analytics*, vol. 5, p. 12, 2020.

[20] Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, and N. C. Newman, "'"Term clumping" for technical intelligence: A case study on dye-sensitized solar cells," *Technological Forecasting and Social Change*, vol. 85, pp. 26-39, 2014.

[21] R. J. Watts, A. L. Porter, and C. Courseault, "Functional analysis: Deriving systems knowledge from bibliographic information resources," *Information Knowledge Systems Management*, vol. 1, no. 1, pp. 45-61, 1999.

[22] C. Cassidy, "Parameter tuning Naïve Bayes for automatic patent classification," *World Patent Information*, vol. 61, p. 101968, 2020.

[23] M. Wu and Y. Zhang, "Hierarchical topic tree: A hybrid model comprising network analysis and density peak search," presented at the 18th International Conference on Scientometrics and Informetrics, Belgium, 2021.

[24] C. V. Fry, X. Cai, Y. Zhang, and C. S. Wagner, "Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research," *PloS One*, vol. 15, no. 7, p. e0236307, 2020.

[25] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[26] Q. Li *et al.*, "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia," *New England Journal of Medicine*, 2020.

[27] R. Lu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding," *The Lancet*, vol. 395, no. 10224, pp. 565-574, 2020.

[28] D. Wang *et al.*, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China," *JAMA*, vol. 323, no. 11, pp. 1061-1069, 2020.

[29] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497-506, 2020.

[30] H. Chen *et al.*, "Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records," *The Lancet*, vol. 395, no. 10226, pp. 809-815, 2020.

[31] N. W. Ruktanonchai *et al.*, "Assessing the impact of coordinated COVID-19 exit strategies across Europe," *Science*, vol. 369, no. 6510, pp. 1465-1470, 2020.

[32] J. P. Shepherd, S. C. Moore, A. Long, L. M. M. Kollar, and S. A. Sumner, "Association between COVID-19 lockdown measures and emergency department visits for violence-related injuries in Cardiff, Wales," *JAMA*, vol. 325, no. 9, pp. 885-887, 2021.

[33] G. Bonaccorsi *et al.*, "Economic and social consequences of human mobility restrictions under COVID-19," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15530-15535, 2020.

[34] P. Engzell, A. Frey, and M. D. Verhagen, "Learning loss due to school closures during the COVID-19 pandemic," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[35] Z. S. Venter, K. Aunan, S. Chowdhury, and J. Lelieveld, "COVID-19 lockdowns cause global air pollution declines," *Proceedings of the National Academy of Sciences*, vol. 117, no. 32, pp. 18984-18990, 2020.

[36] M. Brülhart, V. Klotzbücher, R. Lalive, and S. K. Reich, "Mental health concerns during the COVID-19 pandemic as revealed by helpline calls," *Nature*, vol. 600, no. 7887, pp. 121-126, 2021.

[37] L. Shi *et al.*, "Prevalence of and risk factors associated with mental health symptoms among the general population in China during the coronavirus disease 2019 pandemic," *JAMA Network Open*, vol. 3, no. 7, pp. e2014053-e2014053, 2020.

[38] J. Lai *et al.*, "Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019," *JAMA Network Open*, vol. 3, no. 3, pp. e203976-e203976, 2020.

[39] E. Prompetchara, C. Ketloy, and T. Palaga, "Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic," *Asian Pacific Journal of Allergy and Immunology*, vol. 38, no. 1, pp. 1-9, 2020.

[40] J. Pang *et al.*, "Potential rapid diagnostics, vaccine and therapeutics for 2019 novel coronavirus (2019-nCoV): A systematic review," *Journal of Clinical Medicine*, vol. 9, no. 3, p. 623, 2020.

[41] S. F. Ahmed, A. A. Quadeer, and M. R. McKay, "Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies," *Viruses*, vol. 12, no. 3, p. 254, 2020.

[42] D.-G. Ahn *et al.*, "Current Status of epidemiology, diagnosis, therapeutics, and vaccines for novel coronavirus disease 2019 (COVID-19)," (in eng), *J. Microbiol. Biotechnol.*, vol. 30, no. 3, pp. 313-324, 2020, doi: 10.4014/jmb.2003.03011.

[43] F. P. Polack *et al.*, "Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine," *New England Journal of Medicine*, 2020, doi: 10.1056/NEJMoa2034577.

[44] S. J. Thomas *et al.*, "Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine through 6 months," *New England Journal of Medicine*, vol. 385, no. 19, pp. 1761-1773, 2021, doi: 10.1056/NEJMoa2110345.

[45] S. Xia *et al.*, "Effect of an inactivated vaccine against SARS-CoV-2 on safety and immunogenicity outcomes: Interim analysis of 2 randomized clinical trials," *JAMA*, vol. 324, no. 10, pp. 951-960, 2020.

[46] A. A. Dror *et al.*, "Vaccine hesitancy: The next challenge in the fight against COVID-19," *European Journal of Epidemiology*, vol. 35, no. 8, pp. 775-779, 2020.

[47] N. Biswas, T. Mustapha, J. Khubchandani, and J. H. Price, "The nature and extent of COVID-19 vaccination hesitancy in healthcare workers," *Journal of Community Health*, vol. 46, no. 6, pp. 1244-1251, 2021.

[48] R. Duch *et al.*, "Citizens from 13 countries share similar preferences for COVID-19 vaccine allocation priorities," *Proceedings of the National Academy of Sciences*, vol. 118, no. 38, 2021.

[49] H. Dai *et al.*, "Behavioural nudges increase COVID-19 vaccinations," *Nature*, vol. 597, no. 7876, pp. 404-409, 2021.

[50] P. Campos-Mercade, A. N. Meier, F. H. Schneider, S. Meier, D. Pope, and E. Wengström, "Monetary incentives increase COVID-19 vaccinations," *Science*, vol. 374, no. 6569, pp. 879-882, 2021.

[51] Z. Shen, H. Ma, and K. Wang, "A web-scale system for scientific knowledge exploration," *arXiv preprint arXiv:1805.12216*, 2018.

[52] E. Traggiai *et al.*, "An efficient method to make human monoclonal antibodies from memory B cells: Potent neutralization of SARS coronavirus," *Nature Medicine*, vol. 10, no. 8, pp. 871-875, 2004.

[53] Z. Zhu *et al.*, "Potent cross-reactive neutralization of SARS coronavirus isolates by human monoclonal antibodies," *Proceedings of the National Academy of Sciences*, vol. 104, no. 29, pp. 12123-12128, 2007.

[54] D. Deregt and L. A. Babiuk, "Monoclonal antibodies to bovine coronavirus: Characteristics and topographical mapping of neutralizing epitopes on the E2 and E3 glycoproteins," *Virology*, vol. 161, no. 2, pp. 410-420, 1987.

[55] A. A. Mockett, D. Cavanagh, and T. D. K. Brown, "Monoclonal antibodies to the S1 spike and membrane proteins of avian infectious bronchitis coronavirus strain Massachusetts M41," *Journal of General Virology*, vol. 65, no. 12, pp. 2281-2286, 1984.

[56] P. C. Taylor, A. C. Adams, M. M. Hufford, I. De La Torre, K. Winthrop, and R. L. Gottlieb, "Neutralizing monoclonal antibodies for treatment of COVID-19," *Nature Reviews Immunology*, vol. 21, no. 6, pp. 382-393, 2021.

[57] M. Yuan *et al.*, "Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants," *Science*, vol. 373, no. 6556, pp. 818-823, 2021.

[58] N. Pica *et al.*, "Hemagglutinin stalk antibodies elicited by the 2009 pandemic influenza virus as a mechanism for the extinction of seasonal H1N1 viruses," *Proceedings of the National Academy of Sciences*, vol. 109, no. 7, pp. 2573-2578, 2012.

[59] X. Yu *et al.*, "Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors," *Nature*, vol. 455, no. 7212, pp. 532-536, 2008.

[60] B. S. Coulson, K. Fowler, R. Bishop, and R. Cotton, "Neutralizing monoclonal antibodies to human rotavirus and indications of antigenic drift among strains from neonates," *Journal of Virology*, vol. 54, no. 1, pp. 14-20, 1985.

[61] E. Cameroni *et al.*, "Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift," *Nature*, vol. 602, no. 7898, pp. 664-670, 2022.

[62] T. Koyama, D. Weeraratne, J. L. Snowdon, and L. Parida, "Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment," *Pathogens*, vol. 9, no. 5, p. 324, 2020.

[63] T. L. Van Belle, K. T. Coppieters, and M. G. Von Herrath, "Type 1 diabetes: Etiology, immunology, and therapeutic strategies," *Physiological Reviews*, vol. 91, no. 1, pp. 79-118, 2011.

[64] J. M. Wenzlau *et al.*, "The cation efflux transporter ZnT8 (Slc30A8) is a major autoantigen in human type 1 diabetes," *Proceedings of the National Academy of Sciences*, vol. 104, no. 43, pp. 17040-17045, 2007.

[65] A. Imagawa, T. Hanafusa, J.-i. Miyagawa, and Y. Matsuzawa, "A novel subtype of type 1 diabetes mellitus characterized by a rapid onset and an absence of diabetes-related antibodies," *New England Journal of Medicine*, vol. 342, no. 5, pp. 301-307, 2000.

[66] G. Stenstrom, A. Gottsater, E. Bakhtadze, B. Berger, and G. Sundkvist, "Latent autoimmune diabetes in adults: Definition, prevalence, β-cell function, and treatment," *Diabetes*, vol. 54, no. suppl_2, pp. S68-S72, 2005.

[67] E. Barron *et al.*, "Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: A whole-population study," *The Lancet Diabetes & Endocrinology*, vol. 8, no. 10, pp. 813-822, 2020.

[68] N. Holman *et al.*, "Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in England: A population-based cohort study," *The Lancet Diabetes & Endocrinology*, vol. 8, no. 10, pp. 823-833, 2020.

[69] M. Apicella, M. C. Campopiano, M. Mantuano, L. Mazoni, A. Coppelli, and S. Del Prato, "COVID-19 in people with diabetes: Understanding

the reasons for worse outcomes," *The Lancet Diabetes & Endocrinology*, vol. 8, no. 9, pp. 782-792, 2020.

[70] S. Lim, J. H. Bae, H.-S. Kwon, and M. A. Nauck, "COVID-19 and diabetes mellitus: From pathophysiology to clinical management," *Nature Reviews Endocrinology*, vol. 17, no. 1, pp. 11-30, 2021.

[71] S. R. Bornstein *et al.*, "Practical recommendations for the management of diabetes in patients with COVID-19," *The Lancet Diabetes & Endocrinology*, vol. 8, no. 6, pp. 546-550, 2020.

[72] L. Marchand, M. Pecquet, and C. Luyton, "Type 1 diabetes onset triggered by COVID-19," *Acta Diabetologica*, vol. 57, no. 10, pp. 1265-1266, 2020.

[73] A. C. Powers, D. M. Aronoff, and R. H. Eckel, "COVID-19 vaccine prioritisation for type 1 and type 2 diabetes," *The Lancet Diabetes & Endocrinology*, vol. 9, no. 3, pp. 140-141, 2021.

[74] R. Pal, S. K. Bhadada, and A. Misra, "COVID-19 vaccination in patients with diabetes mellitus: Current concepts, uncertainties and challenges," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 2, pp. 505-508, 2021.

[75] S. K. Boddu, G. Aurangabadkar, and M. S. Kuchay, "New onset diabetes, type 1 diabetes and COVID-19," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 6, pp. 2211-2217, 2020.

[76] R. C. Aalberse, J. Akkerdaas, and R. Van Ree, "Cross-reactivity of IgE antibodies to allergens," *Allergy*, vol. 56, no. 6, pp. 478-490, 2001.

[77] P. Eibensteiner, S. Spitzauer, P. Steinberger, D. Kraft, and R. Valenta, "Immunoglobulin E antibodies of atopic individuals exhibit a broad usage of VH-gene families," *Immunology*, vol. 101, no. 1, pp. 112-119, 2000.

[78] M. Jenmalm, J. Van Snick, F. Cormont, and B. Salman, "Allergen-induced Th1 and Th2 cytokine secretion in relation to specific allergen sensitization and atopic symptoms in children," *Clinical & Experimental Allergy*, vol. 31, no. 10, pp. 1528-1535, 2001.

[79] N. G. Kounis *et al.*, "Allergic reactions to current available COVID-19 vaccinations: Pathophysiology, causality, and therapeutic considerations," *Vaccines*, vol. 9, no. 3, p. 221, 2021.

[80] B. Cabanillas, C. Akdis, and N. Novak, "Allergic reactions to the first COVID-19 vaccine: A potential role of Polyethylene glycol," *Allergy*, vol. 76, no. 6, pp. 1617-1618, 2020.