


Network graph representation of COVID-19 scientific publications to aid knowledge discovery

George Cernile,¹ Trevor Heritage,¹ Neil J Sebire ,² Ben Gordon,² Taralyn Schwering,¹ Shana Kazemlou,¹ Yulia Borecki¹

To cite: Cernile G, Heritage T, Sebire NJ, *et al.* Network graph representation of COVID-19 scientific publications to aid knowledge discovery. *BMJ Health Care Inform* 2021;**28**:e100254. doi:10.1136/bmjhci-2020-100254

Received 12 October 2020
Revised 01 December 2020
Accepted 11 December 2020

ABSTRACT

Introduction Numerous scientific journal articles related to COVID-19 have been rapidly published, making navigation and understanding of relationships difficult.

Methods A graph network was constructed from the publicly available COVID-19 Open Research Dataset (CORD-19) of COVID-19-related publications using an engine leveraging medical knowledge bases to identify discrete medical concepts and an open-source tool (Gephi) to visualise the network.

Results The network shows connections between diseases, medications and procedures identified from the title and abstract of 195 958 COVID-19-related publications (CORD-19 Dataset). Connections between terms with few publications, those unconnected to the main network and those irrelevant were not displayed. Nodes were coloured by knowledge base and the size of the node related to the number of publications containing the term. The data set and visualisations were made publicly accessible via a webtool.

Conclusion Knowledge management approaches (text mining and graph networks) can effectively allow rapid navigation and exploration of entity inter-relationships to improve understanding of diseases such as COVID-19.

INTRODUCTION

There is urgency to accelerate research that can help contain the spread of the COVID-19 epidemic, to ensure that those affected are promptly diagnosed and receive optimal care and to support research priorities in a way that leads to the development of global research platforms in preparation for the next disease epidemic, thus allowing for accelerated research, and research and development for diagnostics, therapeutics and vaccines and their timely access. In view of the urgency of this outbreak, the international community is mobilising to find ways to significantly accelerate the development of interventions.¹ Experts have identified key knowledge gaps and research priorities and shared scientific data on ongoing research, thereby accelerating the generation of critical scientific information to contribute to the control of the COVID-19 emergency.²

However, the pace and volume of research mean that it is hard to stay up to date with the growing body of new scientific papers about the disease and the novel coronavirus that causes it. To mitigate this, many organisations are hosting digital collections holding thousands of freely available papers that can help researchers quickly find the information they seek, and several studies have described or mapped the rapid evidence generation in this area.^{3–5} By one estimate, the COVID-19 literature published since January has reached more than 200 000 papers and is doubling every 30 days, one of the biggest episodes of disease-specific publications of scientific literature ever.⁶

One approach to navigating and searching such knowledge collections is through graph databases, which represent the connections between the semantic concepts with nodes, edges and other properties of the data.⁷ This allows semantic queries to search across the data set to find relationships between papers on any set of data points. Such a graph displayed in a visualisation tool gives an interactive overview of the nodes and connections between the concepts across the papers and allows one to move around and focus on what is interesting to the researcher.⁸

The aim of this short report is to demonstrate the feasibility of using a network graph approach for rapid navigation of the COVID-19 literature in a publicly available format and to present an openly available tool for exploring a COVID-19 knowledge data set.

METHODS

The COVID-19 Open Research Dataset (CORD-19) is a rapidly increasing open-source collection of scholarly articles related to the coronavirus which has been designed to facilitate the development of text mining



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Inspirata, Tampa, Florida, USA
²HDRUK, London, UK

Correspondence to

Professor Neil J Sebire;
neil.sebire@hdruk.ac.uk

and information retrieval systems.^{9,10} As of 8 August 2020, the data set has 207 311 papers from over 160 000 sources. The articles available include title, abstract, authors, source, publication date and in some cases full text.¹¹

We used proprietary natural language processing (NLP) and artificial intelligence (AI) engines, which leverage the heuristic segmentation approach (a fast heuristic search algorithm) and a knowledge-driven approach for concept identification, context determination, inferring and extraction of corresponding values and units. The engine works with domain-specific knowledge bases of clinical terms, concepts and rules that are tailored to the data to be extracted.¹²

In this study, we used a collection of 10 knowledge bases consisting of a core knowledge base and 9 domain-specific knowledge bases that were built using UMLS (Unified Medical Language System) terms and updated

with recently added terms specific to COVID-19: core oncology knowledge base, pharmacological substance (medications) (T121), virus (T005), therapeutic or preventive procedure (T061), sign or symptom (T184), disease or syndrome (T047), gene or genome (T028), immunological factor (T129), finding (T033), and body part, organ or organ component (T023).¹³

The title and abstract sections of all papers in the COVID-19 Dataset were processed against the various knowledge sources to extract discrete data from each paper and were stored in a database. Along with the discrete data, the following metadata were also stored: COVID-19 UID (unique identifier), title, abstract, body text, publication date, URL, authors, journal, knowledge base (which of the 10 available knowledge sources was used to extract the term, term category or question; ie, medication, virus, symptom), paper ID (identification

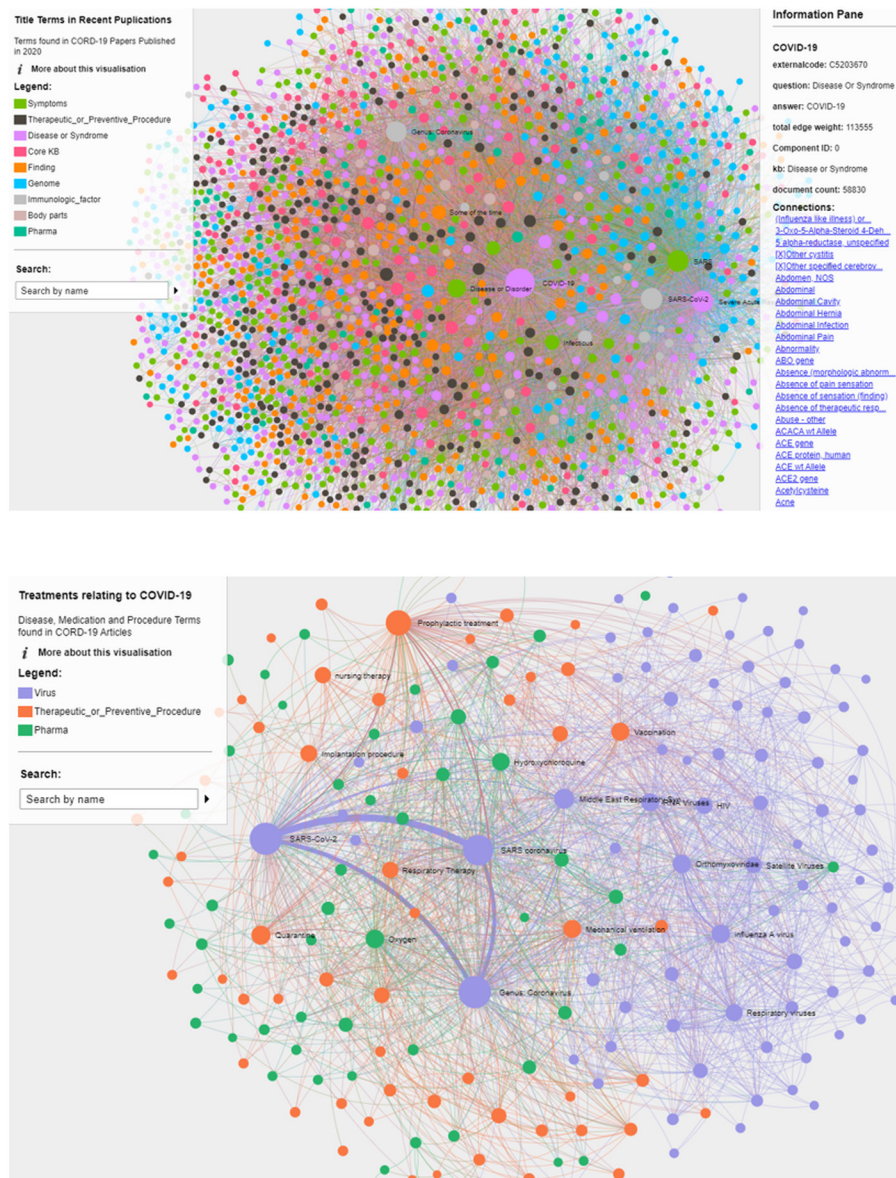


Figure 1 Example of network graphs including high-density network showing concepts associated with COVID-19 (top) and specific query treatment map for COVID-19 (bottom). COVID-19, COVID-19 Open Research Dataset; KB, knowledge base.

Table 1 Extracted concepts from COR-19 Dataset by knowledge base (semantic type) showing the number of unique terms found and the total number of extracted concepts from each knowledge base, as well as the number of papers containing terms from that knowledge base and the percentage coverage across the entire data set

Knowledge base	Unique terms	Extracted concepts	Papers	Coverage (%)
Body parts	1332	172 438	77 400	37
Core knowledge base	1434	338 552	102 037	49
Disease or syndrome	7195	507 819	152 402	73
Finding	5580	526 433	145 504	70
Genome	9395	419 413	86 073	41
Immunological factor	1845	130 996	45 912	22
Pharmacological substance	2599	58 308	30 494	15
Symptoms and side effects	8883	630 116	144 063	69
Therapeutic or preventive procedure	4923	332 260	111 277	54
Virus	1308	240 993	84 325	41
Total	44 494	3 357 328	195 958	94

Papers may contain multiple extracted concepts, and concepts may be found in multiple papers within the knowledge base; hence, we provide both all extracted concepts using the natural language processing tool in addition to the number of unique terms. COR-19, COVID-19 Open Research Dataset.

of the paper in the COR-19 Dataset from which the term was extracted) and source section (either title or abstract). Generic terms with little significance were determined, for example, ‘air’, ‘water’ and ‘virus’, and these were removed from the set of extracted concepts.

Networks created with the entire set of results and all the knowledge sources are very large with too many terms to visualise details in the data. For this reason, a subset of the data was selected to enable meaningful visual exploration by selecting a subset of the knowledge sources, paper sections and publication year for each network based on specific medical themes, for example, treatments, cardiology and so on. Duplicate terms (same terms found in multiple knowledge sources) were consolidated to remove redundant data. For example, ‘obesity’ is included in both the ‘symptoms and side effects’ and the ‘disease or syndrome’ knowledge sources.

For each term found in a paper, a link was created to every other term on the same paper. The culmination of these links for all papers resulted in a network structure where the weight of a connection between any two terms was determined by the number of papers linking the terms. Additional filtering was performed to refine the scope of the network and removal of noise to aid readability and navigation; for example, links with low weights were removed, as were links with terms that were disconnected from the rest of the network.

The open-source software tool Gephi was used to create a visualisation of the network using the collections of terms and connections that made up the network structure.¹⁴ Network nodes were coloured based on the knowledge source, with the size of the nodes proportional to the frequency of each term and the connection weight (edge thickness) based on the number of associated papers. The networks were exported and visualised in an

HTML (hypertext markup language) website using the Sigma JS JavaScript library.

RESULTS

A total of 207 311 publications from the COR-19 Dataset were processed using the NLP engine. In total 3 357 328 total entities were extracted from 195 958 of these papers, consisting of 44 494 unique terms. Four network graphs were generated using these extracted data: cardiological diseases, lung diseases, title network and treatment network (<https://nlp.inspirata.com/networkvisualisations/treatmentnetwork/#>) (figure 1). The filters applied to create each of the networks and the number of terms, edges and papers involved in each network are displayed in table 1 and online supplemental table 1.

DISCUSSION

Recently there have been several initiatives to explore knowledge graphs in medical data and with some applied to aspects of COVID-19-associated published literature.^{15 16} This study has demonstrated the feasibility of using a graph database approach to create a targeted concept association networks as an interactive way to allow users to easily navigate the rapidly growing COVID-19-related literature, and particularly as a way to understand and explore the relationships between key concepts within this corpus of literature articles, which is potentially widely applicable to other disease areas.

This approach is also applicable to any collection of scientific literature, such as PubMed or ClinicalTrials.gov, or proprietary document management systems. Specific lexical terms and knowledge sources can be used from

the UMLS collection or other publicly available sources and imported for use with NLP/AI engines.

One constraint of this knowledge mining approach is that the network size increases as more knowledge sources are added. As a consequence, methods to simplify the network to enable easier visual exploration are required, such as ‘pruning’.¹⁷ The concept is to remove a subset of the ‘least important’ edges while maintaining the overall graph connectivity, since it becomes more difficult to interactively explore without a priori knowledge of the specific knowledge sources as the network density increases. Another limitation is that the network only shows the first-level connections or the direct connection between papers and concepts. It does not find connections between concepts that span several papers, although this can be achieved by traversing the network visually.

We addressed these limitations of network size and the search for deep connections by implementing a breadth-first search on the network structure.¹⁸ Essentially this approach searches the graph data structure beginning at a root node by exploring all of the adjacent nodes at a given depth before moving to the nodes at the next subsequent level. This search type is efficient and can be applied across very large networks, even when all the knowledge sources are used simultaneously, and can find the shortest path connections (the trail of papers) between any concepts.

This study has demonstrated that an approach using graph databases and network analysis can be developed rapidly and is a useful approach to understanding large volumes of medical literature, quickly grasping the current state of our knowledge, and discovering previously unknown or unnoticed relationships between emerging medical concepts. The unusual circumstances of a global pandemic have given rise to the assembly of an unprecedented volume of medical literature, and this work demonstrates a powerful approach to condensing the literature into insights that help us fight this disease. Further development of this approach will enable ongoing analysis and deep searching of large collections of literature, such as PubMed, and application to other disease areas, as well as for target or biomarker discovery.^{19–21}

Contributors TH, GC, TS, SK and YB conceived the study and performed the data analysis and tool development. All authors contributed to the manuscript preparation and writing and critically improved the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests TH, GC, TS, SK and YB are employed by Inspirata, a company specialising in health data management, and carried out the work as part of their employment.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Neil J Sebire <http://orcid.org/0000-0001-5348-9063>

REFERENCES

- 1 Kambhampati SBS, Vaishya R, Vaish A. Unprecedented surge in publications related to COVID-19 in the first three months of pandemic: a bibliometric analytic report. *J Clin Orthop Trauma* 2020;11:S304–6.
- 2 Coronavirus disease, 2019. Available: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> [Accessed 3 May 2020].
- 3 Zyoud Sa'ed H, Al-Jabi SW. Mapping the situation of research on coronavirus disease-19 (COVID-19): a preliminary bibliometric analysis during the early stage of the outbreak. *BMC Infect Dis* 2020;20:561.
- 4 Liu N, Chee ML, Niu C, et al. Coronavirus disease 2019 (COVID-19): an evidence map of medical literature. *BMC Med Res Methodol* 2020;20:177.
- 5 Albahri AS, Hamid RA, Alwan JK, et al. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. *J Med Syst* 2020;44:122.
- 6 Brainard J. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science* 2020.
- 7 Lysenko A, Roznovã IA, Saqi M, et al. Representing and querying disease networks using graph databases. *BioData Min* 2016;9:23.
- 8 Fensel D, Şimşek U, Angele K. *Knowledge graphs*, 2020.
- 9 Kaggle. COVID-19 open research dataset challenge (CORD-19). Available: <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge> [Accessed 12 Oct 2020].
- 10 Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: the Covid-19 open research dataset. *ArXiv* 2020. [Epub ahead of print: 22 Apr 2020].
- 11 Semantic Scholar. [PDF] CORD-19: The Covid-19 Open Research Dataset. Available: <https://www.semanticscholar.org/paper/CORD-19%3A-The-Covid-19-Open-Research-Dataset-Wang-Lo/4a10dffca6dcce9c570cb75aa4d76522c34a2fd4> [Accessed 12 Oct 2020].
- 12 Inspirata Launches Cloud-Based Cancer and Clinical Data Extraction Software Service. Available: <https://www.inspirata.com/inspirata-launches-nlp-on-demand/> [Accessed 12 Oct 2020].
- 13 Unified medical language system (UMLS). Available: <https://www.nlm.nih.gov/research/umls/index.html> [Accessed 12 Oct 2020].
- 14 Gephi - The Open Graph Viz Platform. Available: <https://gephi.org/> [Accessed 12 Oct 2020].
- 15 Domingo-Fernández D, Baksi S, Schultz B, et al. COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* : 2020;12.
- 16 Das D, Katyal Y, Verma J. Information Retrieval and Extraction on COVID-19 Clinical Articles Using Graph Community Detection and Bio-BERT Embeddings. In: *ACL 2020 work NLP-COVID*, 2020.
- 17 Zhou F, Mahler S, Toivonen H. Simplification of networks by edge pruning. *Lect Notes Comput Sci* 2012;7250:179–98.
- 18 Wikipedia. Breadth-first search. Available: https://en.wikipedia.org/wiki/Breadth-first_search [Accessed 27 Nov 2020].
- 19 Shi L, Li S, Yang X. Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services. *Biomed Res Int* 2017.
- 20 Sharma S, Santosh T, Santra B. Incorporating domain knowledge into medical NLI using knowledge graphs. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2020*.
- 21 Xu J, Kim S, Song M, et al. Building a PubMed knowledge graph. *Sci Data* 2020;7.

Table 2: Filters applied to four networks with links to visualisations showing the filters and the number of terms, edges and papers involved in each network.

	Title Network	Treatment Network	Lung Diseases	Cardio Diseases
Knowledge Base Sections Publications years Edges weight cutoff Manual filtering Number of papers Number of nodes (terms) Number of edges (connections)	All	Therapeutic or prevention procedure; pharmacological substances; virus	Symptoms and side effects; disease or syndrome; virus	Symptoms and side effects; disease or syndrome; virus
	Title	Title; abstract	Title; abstract	Title; abstract
	2020	2015–2020	2015–2020	2015–2020
	5	5	10	25
	No	Yes	Yes	Yes
	34 032	36 801	52 258	35 282
	2,586	206	86	49
	16,735	1,739	994	277
Link	https://nlp.inspirata.com/NetworkVisualisations/TitleNetwork/	https://nlp.inspirata.com/NetworkVisualisations/TreatmentNetwork/	https://nlp.inspirata.com/NetworkVisualisations/LungNetwork/	https://nlp.inspirata.com/NetworkVisualisations/CardioNetwork/