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

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 generating the publication 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 200000 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
and information retrieval systems. 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.

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.

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).

The title and abstract sections of all papers in the CORD-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: CORD-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 number), and in some cases full text.

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). CORD-19, COVID-19 Open Research Dataset; KB, knowledge base.
of the paper in the CORD-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. 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 CORD-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. 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

---

Table 1 Extracted concepts from CORD-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.

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

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. CORD-19, COVID-19 Open Research Dataset.
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’. 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. 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.

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.

REFERENCES