Computational and Structural Biotechnology Journal 23 (2024) 2661–2668

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

CORACLE (COVID-19 liteRAture CompiLEr): A platform for efficient tracking and extraction of SARS-CoV-2 and COVID-19 literature, with examples from post-COVID with respiratory involvement

Kristina Piontkovskaya [a,b,1], Yulian Luo [a,1], Pia Lindberg [a,b,1], Jing Gao [a], Michael Runold [a,b], Iryna Kolosenko [a,2], Chuan-Xing Li [a,2], Åsa M. Wheelock [a,b,*,2]

[a] Respiratory Medicine Unit, Department of Medicine Solna and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden
[b] Department of Respiratory Medicine and Allergy, Karolinska University Hospital Solna, Stockholm, Sweden

ARTICLE INFO

Keywords:
Literature mining
COVID-19
Citation maps
MeSH maps

ABSTRACT

Background: During the COVID-19 pandemic a need to process large volumes of publications emerged. As the pandemic is winding down, the clinicians encountered a novel syndrome - Post-acute Sequelae of COVID-19 (PASC) - that affects over 10 % of those who contract SARS-CoV-2 and presents a significant challenge in the medical field. The continuous influx of publications underscores a need for efficient tools for navigating the literature.

Objectives: We aimed to develop an application which will allow monitoring and categorizing COVID-19-related literature through building publication networks and medical subject headings (MeSH) maps to identify key publications and networks.

Methods: We introduce CORACLE (COVID-19 liteRAture CompiLEr), an innovative web application designed to analyse COVID-19-related scientific articles and to identify research trends. CORACLE features three primary interfaces: The "Search" interface, which displays research trends and citation links; the "Citation Map" interface, allowing users to create tailored citation networks from PubMed Identifiers (PMIDs) to uncover common references among selected articles; and the "MeSH" interface, highlighting current MeSH trends and their associations.

Results: CORACLE leverages PubMed data to categorize literature on COVID-19 and PASC, aiding in the identification of relevant research publication hubs. Using lung function in PASC patients as a search example, we demonstrate how to identify and visualize the interactions between the relevant publications.

Conclusion: CORACLE is an effective tool for the extraction and analysis of literature. Its functionalities, including the MeSH trends and customizable citation mapping, facilitate the discovery of emerging trends in COVID-19 and PASC research.

1. Introduction

1.1. The COVID-19 crisis underscores the need for efficient literature mining tools

The emergence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, and the associated respiratory coronavirus disease (COVID-19), posed a significant threat to global public health, impacting populations worldwide. In response to this unprecedented crisis, a collective effort among clinicians, researchers, and scientific communities has fostered a remarkable effort in open-source dissemination of research findings. This aimed to expedite the resolution of the global health crisis by accelerating scientific progress, as evidenced by the large volume of literature dedicated to SARS-CoV-2 and COVID-19-related topics.

To date, over 400,000 papers are published on COVID-19. Even as

* Corresponding author at: Respiratory Medicine Unit, Department of Medicine Solna and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.
E-mail address: asa.wheelock@ki.se (Å.M. Wheelock).
[1] Contributed equally
[2] Senior authors contributed equally

https://doi.org/10.1016/j.csbj.2024.06.018
Received 10 April 2024; Received in revised form 19 June 2024; Accepted 19 June 2024
Available online 20 June 2024

the pandemic is winding down, an influx of 212 new publications was recorded in the past week alone (week 11 of 2024). Additionally, a substantial body of literature addresses previous coronavirus outbreaks, including those caused by SARS-CoV and MERS-CoV, which have afflicted populations with severe respiratory syndromes over the past decade. Furthermore, the emergence of post-acute COVID-19 syndrome in the aftermath of the COVID-19 pandemic also prompted a substantial number of publications that demand daily monitoring for the involved professionals. The exponential growth of COVID-19-related publications presents a challenge in staying updated on the status of the field. To facilitate access to relevant research, several open-access literature repositories have been established, such as LitCovid [1], the WHO COVID-19 Database [2], and the Coronavirus Knowledge Hub [3]. These hubs offer comprehensive collections of publications, featuring search and filtering functionalities akin to those found in databases like PubMed. However, navigating through the vast array of publications necessitates considerable human curation, requiring significant time investment and domain expertise. Given the dynamic nature of the field and the rapid influx of new research, there is a pressing need for refined and time-efficient literature mining tools. To address this demand, we present CORACLE (COvid liteRAture CompiLEr), a literature integration tool tailored to support researchers, clinicians, epidemiologists, policy-makers, and other involved personnel in navigating, tracking, filtering, summarizing, and prioritizing COVID-related literature. CORACLE aims to provide daily updates on the latest trends and findings in the COVID field, offering a valuable resource to aid in their pursuit of knowledge related to COVID-19 and post-COVID-19 syndrome (https://coracle.cmm.se).

2. Methodology

2.1. Structure and function of the CORACLE workflow

The objective of the CORACLE application is to deliver a fast, simple, and smooth user experience when navigating large numbers of COVID-19-related publications. It is a web-based application and does not require installation on the users' computers. To provide a smooth experience, this application follows the structure of a Single Page Application (SPA) with back-end and front-end separated. Fig. 1A shows a schematic overview of the project.

The JavaScript framework Vue.js was utilized for developing user interfaces, using Vite version 2.7.2 for building the Vue components. The structure of the components is shown in Fig. 1B. The five main components in the Main Page are controlled by Vue Router version 4.0.12. Components in the Tab Pane are only shown separately in different tabs and they are under the same path. In the Search Page component, Vuex version 4.0.2 is used for passing the change of data in the Filter component to Statistics, Citation Summary, Citation Map components, and their child components conveniently. Some reuse components that do not have specific functions are not shown in the Fig. 1B, such as the component for drawing loading status.

Flask version 2.0.2, a micro web framework written in Python, is used for building APIs. The extension flask_restful is used in this project to create RESTful APIs since it provides a cleaner way to parse arguments, format output, and organize the routing of calls to the API. After parsing the arguments, the Python script utilizes them to retrieve data from the database. Subsequently, the retrieved data is converted to the JSON format before being sent to Vue components via APIs. The extension flask_cors is employed to handle cross-origin requests, ensuring secure data exchange between the client and server components of the application.

A graph database management system, Neo4j version 4.4.2 is used to save and manage data. Unlike relational databases, graph databases use nodes and edges to store data. To store relationship information, relational databases use tables with one column that contains certain items and another column that contains the related items. For graph databases, relationship information is stored directly as edges. This difference in data structure makes graph databases faster than relational databases when the data is relationships. For the project, different kinds of relationships including the citation relationships and the relationships between articles and Medical Subject Headings (MeSH) terms are used. Therefore, a graph database is suitable for building this application. All information related to articles is fetched from NCBI's API E-utilities and parsed by the Python module ElementTree, then stored in the Neo4j graph database. Nodes labeled as *COVID* and *Article* are used to save basic article information including PMID, title, journal, publication date, and language. The PMID is saved as a node property called name, while other information is saved as different properties. MeSH and publication types are saved as nodes and connected to COVID nodes since they are in one-to-many relationships with articles. Each reference to an article is created as a node labeled only as *Article* since not all articles in references are COVID-19 related. Fig. 2 shows a scraped article as an example in Neo4j. Due to the data-intense nature and associated long calculation times required for the MeSH functions provided by CORACLE, the data

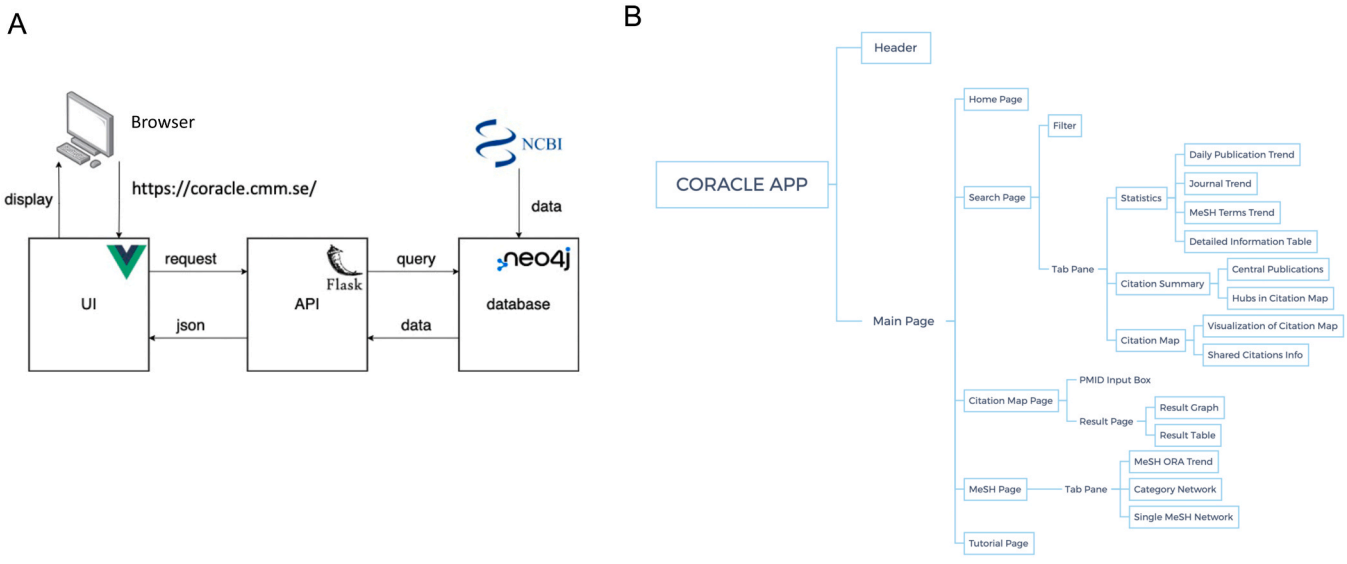

Fig. 1. Schematic overview of the CORACLE application. (A) Schematic overview of data processing in CORACLE. (B) An overview of modules currently available in CORACLE. Shapes with borders are components, and shapes without borders are for presenting the structure.

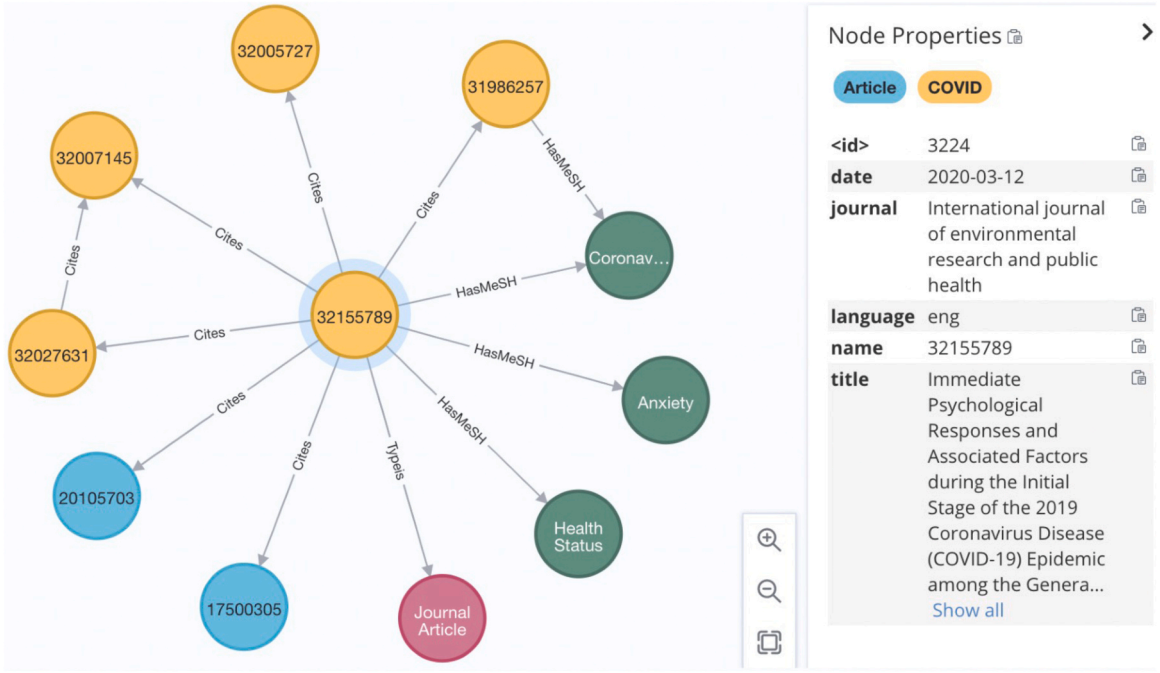

Fig. 2. An example of data in Neo4j (not real PMID).

used for the MeSH functions are pre-calculated and stored independently from the data scraped from NCBI E-utility. Given that these relationships do not change markably over time, monthly updates of these terms were deemed sufficient to provide accurate ORA and Fisher's exact test. To avoid interference with the data directly scraped, a new set of nodes and relationships are created for storing MeSH terms information.

2.2. Functions of CORACLE

CORACLE is Python [4] is program designed to provide four major functions:

1) It quickly finds new and important topics related to SARS-CoV-2/COVID-19, along with key papers, keywords and journals.
2) It customizes literature search using PMID (PubMed Identifier) lists, using citation relationships.
3) The program prioritizes identification of highly related publication using direct citation maps and subsequently, by indirect citation-similarity networks.
4) It uses keywords to understand connections between research areas.

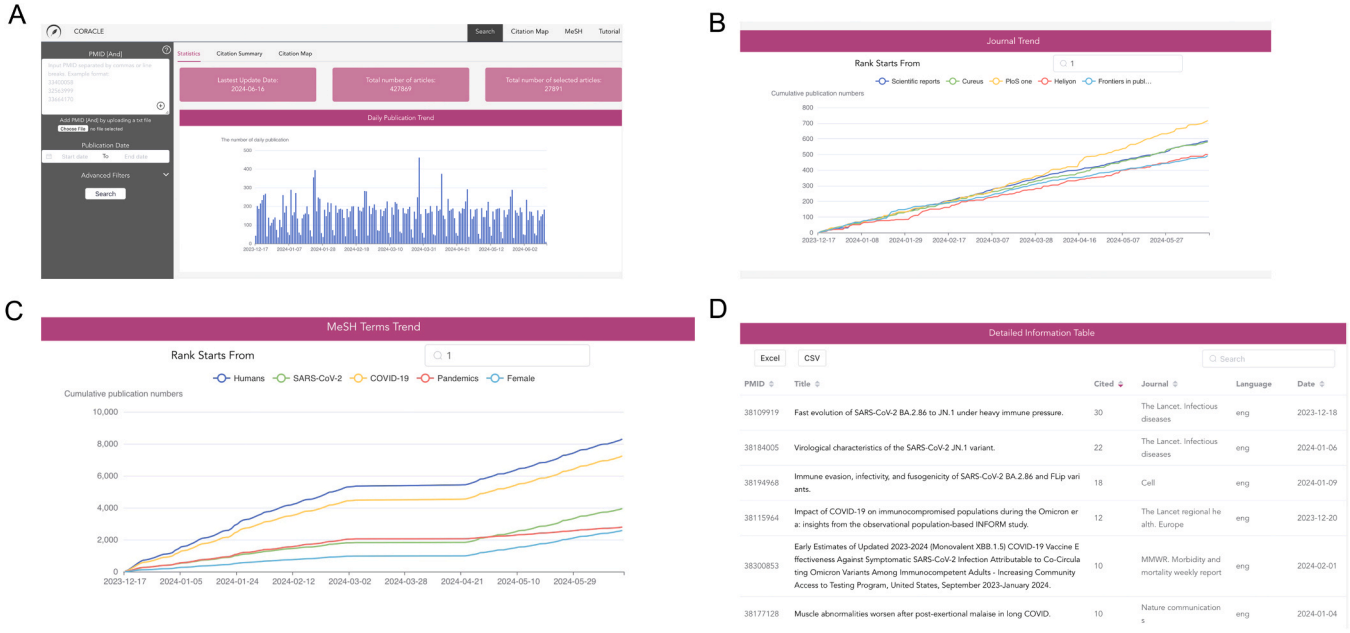

Fig. 3. A global view of CORACLE Statistics tab. (A) CORACLE input field (on the left) together with daily publication trends for the last 6 months for all COVID-19-related articles. (B,C) Printscreens from the "Statistics" window of CORACLE showing the journal trends (PloS one currently leading) and MeSH trends (COVID-19 and Human leading) over time. (D) A tabulated summary of the top 6 COVID-19-related articles globally.

5) It creates tables of relevant literature readable by both humans and machines.

CORACLE features a global filter block, visualization panels, and a data table. It also has an easy-to-use web-based interface. The Search-function includes five main filters: publication type, country, language, MeSH term, and PMIDs (see Fig. 3 and Tutorial for details).

Output is generated in four tabs. The STATISTICS tab summarizes publication trends, frequency by journal, and a table of the most cited articles. The CITATION MAP tab identifies key publications, using both upstream and downstream citations. The MeSH MAP tab shows network analyses of keyword annotations for publications.

Results are displayed as a network map, where each MeSH term is a node, and edges show the number of shared publications between these terms. All data tables are updated daily from PubMed and can be downloaded for further analysis.

3. Results

3.1. Literature mining example using CORACLE: post-acute sequelae of COVID-19

In the initial stages of the COVID-19 pandemic, healthcare efforts were primarily focused on the immediate recovery of survivors, particularly those suffering from respiratory distress syndrome and COVID-19-related pneumonia. This focus was a direct response to the urgent needs presented by the acute phase of infections. However, as the pandemic progressed, a significant observation was made worldwide. Individuals who had experienced mild COVID-19 symptoms, without requiring hospitalization, began to report persistent symptoms [5]. These included heart palpitations, vertigo, brain fog, cough, and breathing difficulties, persisting or recurring weeks to months after initial recovery [6]. Initially, these symptoms were often attributed to the psychological impact of the pandemic, but subsequent reports suggested possible dysregulations in the autonomic nervous system among this patient group [7].

The persistence of these symptoms, without following any previously known disease patterns, posed significant challenges for healthcare professionals. This was compounded by a lack of published research and clinical guidelines for diagnosing and treating what would eventually be termed Post-Acute Sequelae of SARS-CoV-2 infection (PASC), long COVID, or Post-Acute COVID-19 syndrome (PACS). The condition, characterized by symptoms such as fatigue, shortness of breath, and cognitive dysfunction, impacts daily life significantly. These symptoms can emerge post-recovery or persist from the initial illness, with severity fluctuating over time [8].

Our clinic has been actively monitoring non-hospitalized PASC patients, focusing particularly on respiratory symptoms and conducting comprehensive investigations. Despite normal spirometry results, radiological findings in some patients have shown air trapping, raising questions about the underlying mechanisms of PASC-associated respiratory issues. The absence of a clear pathophysiological understanding of PASC has made clinical consultations challenging, highlighting the need for systematic, evidence-based approaches.

To bridge the knowledge gap, especially regarding patients with initially mild symptoms, we undertook an extensive literature review. By using findings from our in-house developed system, CORACLE, we aim to provide a comprehensive overview of the publications related to PASC, with a particular focus on the respiratory system. This effort seeks to shed light on the long-term effects of PASC, including the poorly understood phenomena of dysfunctional breathing and its potential long-term impact on lung function.

By design, CORACLE undergoes daily updates to incorporate literature related to COVID-19. Consequently, the default configuration—in the absence of a manually specified list of PMIDs—provides an overview of recent developments within the domain over the preceding months. Between September 14, 2023, and March 14th, 2024 the database has registered 30,129 manuscripts on COVID-19. With winding down of the pandemic, an evident turn towards publications on long COVID, is noticeable. The most frequently cited amongst these is a seminal study on immune profiling in long COVID conducted by Klein et al., within the Mount Sinai- Yale Long COVID project [9]. This investigation undertakes a detailed immune profiling of peripheral blood mononuclear cells (PBMC) from a diverse cohort of 275 individuals, encompassing SARS-CoV-2-infected healthcare personnel, asymptomatic non-infected controls, convalescent controls, subjects with symptoms of long COVID, and individuals with persistent symptoms from a different study. The findings of the study demonstrate an elevated prevalence of non-conventional monocytes ($CD14^{low}CD16^{high}$) within the long COVID cohort, alongside a diminished frequency of conventional dendritic cells in comparison with other study groups. Moreover, the research delineated a relative increase in immune reactivity in cells from participants of the long COVID group, as shown by both autoantibody assays against endoproteome and in vitro T-cell stimulation assays. The group also attempted to utilize machine learning approaches to identify the features characteristic for long COVID. In summary, this work - identified as the most cited paper on COVID-19 in the last 6 months by CORACLE – demonstrates that there are immunological changes in the group of long COVID patients. The authors, however, themselves address the point that the groups were formed by the sampling convenience principle, hence the observed heterogenicity. The study was also performed on PBMCs although there are distinct organ-specific symptoms that require elucidation.

For a more targeted search and the demonstration of CORACLE utility, below we outlined example where we aimed to investigate the current state of research regarding to the lung function of individuals experiencing PASC, focusing on those who did not require hospitalization during the primary SARS-CoV-2 infection. As terminology surrounding this phenomenon varies among centers, with the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) predominantly employing the term "post COVID-19 condition." Within the PubMed database, the MeSH term "Post-Acute COVID-19 syndrome" is designated for this syndrome, however, other variations such as "long COVID-19," "PASC," and "long-haul COVID" are commonly used in the literature. Consequently, we attempted to include a comprehensive range of the terms in our search and inputted "Post-acute COVID-19 syndrome" OR "long COVID" OR "PACS" OR "Post-acute sequelae of COVID-19" AND "non-hospitalized" AND "lung function.".

The search for this inquiry yielded a total of 11 publications, with the identified publications compiled into a list of PMIDs and inputted into the CORACLE search interface (Fig. 4A**, left panel**) [10–20]. We retained the default settings allowing for broad inclusion without discrimination based on article type, journal, or publication date, however, these advanced filtering options are available for more nuanced analysis.

Subsequently, utilizing CORACLE, we conducted a detailed examination of the interactions among the identified publications concerning lung function in patients with long COVID-19. The platform's "Statistics" module provided an overview of daily publication trends (Fig. 4A**, right panel**), journal distribution (Fig. 4B), and temporal patterns of MeSH term utilization (Fig. 4C). Additionally, a comprehensive exportable table-like format facilitated detailed summary of publication statistics, including citation counts, publication details, and access links to PubMed (Fig. 4D).

Further, the "Citation Summary" feature presented a graphical depiction of citation networks, highlighting "hub" publications and their respective number of citations (Fig. 5A).

The "Citation Map" function provided an interactive visualization of citation networks, revealing central nodes within the literature network (Fig. 5**C**). Among these was the paper with PMID 34999762 (Fig. 5D), which gathered considerable citation attention regarding lung function in long-COVID-19 patients [10]. In-depth analysis of individual

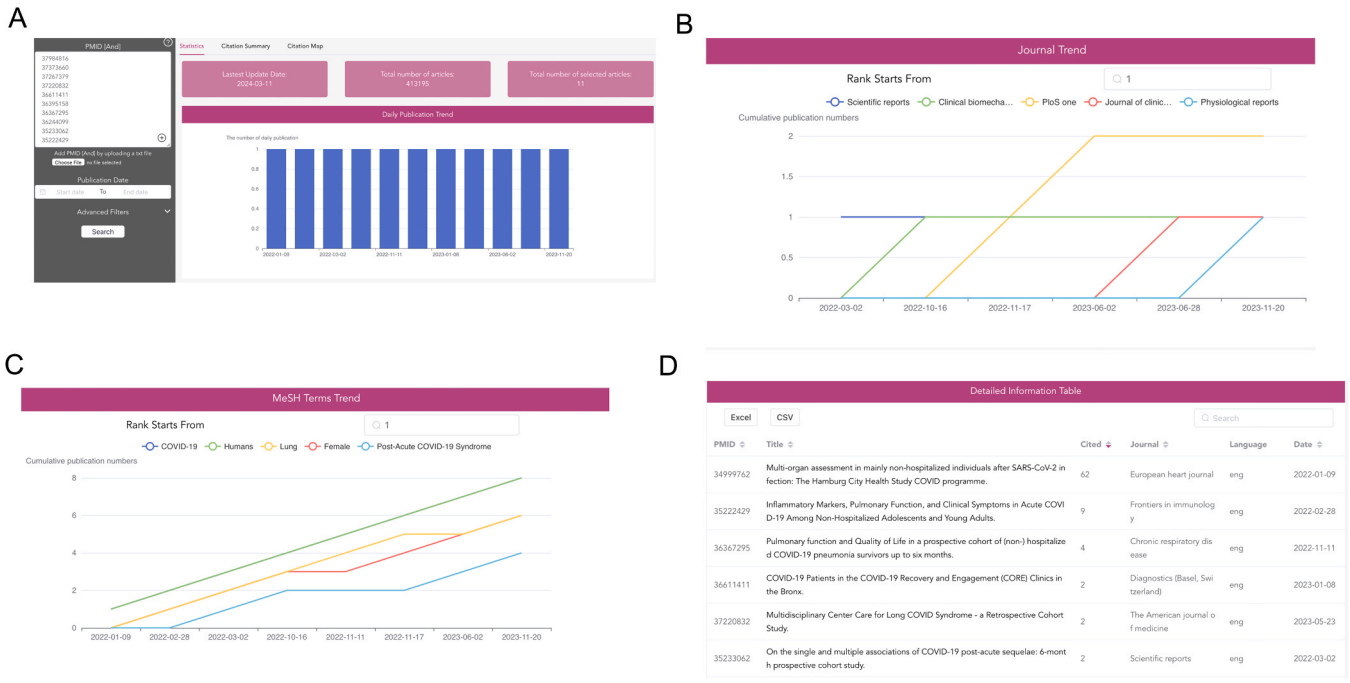

Fig. 4. An example of CORACLE utility for literature mining for lung function in patients with post-COVID-19 syndrome. STATISTICS module options. (A) A printscreen from CORACLE showing inputted PMIDs in the left and publication trends on the right. (B,C) Printscreens from the "Statistics" window of CORACLE showing the journal trends and MeSH trends over time. (D) A tabulated summary of the top 6 inputted articles.

Fig. 5. An example of CORACLE utility for literature mining for lung function in patients with post-COVID-19 syndrome. CITATION SUMMARY (A, B) and "CITATION MAP" module options. (A) A printscreen from CORACLE showing the relation between the number of references and number of citations for the inputted publications (B) An exportable summary of all references in the current network (includes both the inputted articles and the articles referenced in them). (C) The citation map for the given search. Dots are hubs, triangles are cited papers. (D) A zoomed-in view on the most cited paper.

publications revealed notable insights, such as the study by PMID 33428867, which exhibited considerable citation impact (highest degree) on the papers dealing with lung function in long-COVID-19 patients (Fig. 5C) [21]. Degree indicates the number of input articles which either refer to the article in the total section (indegree), or refer to another input article in the search (outdegree). It should, however, be noted that this paper was later retracted and the journal awaits submission of a corrected version. If desired by the user, CORACLE provides an option of excluding specific PMIDs from building citation maps.

In this example, the highest-cited publication (Fig. 5A, 5B) was the paper by Petersen et al. from *Eur Heart J* (with 62 citations) entitled "Multi-organ assessment in mainly non-hospitalized individuals after SARS-CoV-2 infection: The Hamburg City Health Study COVID programme" [10]. This paper examines 443 patients and 1328 matched controls from general population on average 9,6 months after the acute infection. Using body plethysmography (which includes spirometry measurements), the authors identified mild changes in total lung volume and in total airway resistance, as well as a number of mild cardiological

changes.

The second most cited paper (a hub with the PMID: 35222429) is by Berven et al [11]. This longitudinal study examines adolescents and young adults aged 12–25 who experienced a mild course of acute COVID-19 infection, compared to age- and sex-matched controls without a recorded history of COVID-19. The study is part of the Long-Term Effects of COVID-19 in Adolescents (LoTECA) project, conducted in two major hospitals in Southeast Norway. The median period between the positive PCR test and measurements was 18 days, establishing baseline data for future observations in this specific age group. Notably, inclusion criteria for the project required confirmation of SARS-CoV-2 infection through a PCR test. In total, the study included 405 cases and 109 matched controls (PCR-negative for SARS-CoV-2 infection, with additional testing for SARS-CoV-2 antibodies).

All included individuals underwent a multiplex cytokine array analysis to measure a panel of inflammatory cytokines, as well as spirometry to measure forced vital capacity (FVC) and forced expiratory volume in one second (FEV_1). Participants also completed a questionnaire consisting of 24 symptoms rated on a 5-point severity scale.

While significant differences in plasma inflammatory cytokine levels were observed between the COVID-19 and control groups, no significant differences were detected in spirometry measurements. Symptom severity ratings were higher among individuals with detected COVID-19 infection; however, a positive correlation was only observed with female sex and age. Importantly, no correlations were found between pulmonary function parameters and symptom severity in this study.

Another prospective cohort study is presented by de Roos et al. [14] within the next largest hub in the identified network of publications. In this study, the authors examine the patients 3 and 6 months after the infection and investigate diffusion capacity (DLCO) and quality of life (QoL). Although the total group of included individuals is heterogeneous, the patients are sub-divided according to the severity of the acute infection during the data analysed. All non-hospitalized patients ($n = 59$) – the group we were interested in by limiting our initial literature search - were categorized as moderate cases. 44 % of patients in this group had decreased DLCO at 3-months follow up. At 6-month follow up, there was a small yet significant improvement in DLCO for the moderate COVID-19 group, however, it did not normalize neither in severe nor in moderate patient groups at the latest observation time point.

From the ”Citation map” block, it is also apparent that the paper with PMID 36611411 [20] entitled “COVID-19 Patients in the COVID-19 Recovery and Engagement (CORE) Clinics in the Bronx” from Diagnostics (Basel) has the highest number of references (Fig. 5B) and thus can be used for collection of the background articles. This study investigates various clinical parameters, including pulmonary function tests, lung imaging, and symptom assessment, in patients approximately 5 months after experiencing COVID-19 infection. The findings reveal that patients commonly report the onset of multiple new symptoms, such as dyspnea, fatigue, low exercise tolerance, and cognitive difficulties, irrespective of whether they were hospitalized during the acute phase of COVID-19 infection. However, abnormalities in lung function values (including FVC, FEV_1, DLCO%, and oxygen saturation (SpO_2) following a 6-minute walking test) were observed exclusively in the group of patients who had been hospitalized during the acute phase. Likewise, the presence of lung opacity persisting after COVID-19 infection was observed solely among patients who had been hospitalized during the acute phase, whereas non-hospitalized patients displayed normal lung imaging findings. Consequently, despite both hospitalized and non-hospitalized patients reporting comparable frequencies and severities of symptoms, lung function abnormalities were exclusively identified in the hospitalized cohort.

Collectively, a brief analysis of the three most cited papers examining lung function in non-hospitalized patients following COVID-19 infection reveals a lack of consistent evidence supporting a uniform decline in lung function across multiple medical centers. Notably, the most cited papers predominantly originate from early studies, published between 2021 and 2022, during the emergence of post-COVID-19 as a distinct medical concern, when its public health implications were not yet fully understood. [21].

This demonstration of literature tracking for lung function in non-hospitalized patients with long COVID-19 exemplifies the utility of CORACLE, which offers versatile applications across various research domains. To the best of our knowledge, our approach to citation map construction, extraction, and analysis represents a novel approach to study and visualize relationship between research articles.

3.2. Building MeSH term networks in CORACLE

An independent computational module designated "MeSH" is available for rapid analysis of interactions among Medical Subject Headings (MeSH) terms in the context of COVID-19. This module by default displays MeSH terms identified through overrepresentation analysis at the time of update. Users can modify term categories, p-value thresholds, time frames, and the volume of analyzed literature (Fig. 6A).

In the Category Network Tab, interactions between MeSH terms are depicted via a chord diagram, with correlations determined through precalculated Fisher’s exact tests as exemplified on Fig. 6B with the default settings. The diagram’s line thickness signifies the co-occurrence frequency of term pairs within the literature corpus. Interaction with the diagram, such as hovering over connections, reveals the count of articles for each term pair. By default, connections among the top 300 most prevalent term pairs are shown, though adjustments are possible for both category and minimum article count for displayed terms. Elevating the maximum article count threshold to 4000 allows for the exclusion of commonly occurring terms (e.g., "Humans," "COVID-19," "SARS-CoV-2"), enabling a more nuanced examination.

The Single MeSH Network Tab employs a force-directed graph to visualize associations with a selected MeSH term. Initially, this graph displays the 50 most prevalent associated term pairs, with a complete list downloadable in.csv format. The selected input term is marked as a square, with associated terms as circles, and line width reflects co-occurrence frequency. User interaction with the graph, such as hovering to reveal article counts for term pairs, is supported. Adjustments to the minimum article count criteria allow for focused analysis, with a 4000 maximum count limit to exclude prevalent terms for targeted inquiry. Fig. 6C illustrates the MeSH network for "Post-acute COVID-19 syndrome" across all categories and specifically within "Chemicals and Drugs" (Fig. 6D) and "Diseases" (Fig. 6E) categories, offering insight into prevalent diseases and drugs related to "Post-acute COVID-19 syndrome” term usage.

4. Discussion

4.1. Limitations of the CORACLE platform

The CORACLE platform, while robust has some limitations. PubMed sometimes inconsistently reports, which can lead to misleading citation statistics, and incorrect assumption that some articles lack references.

The search function within relies on MeSH terms for article retrieval, despite the standardized indexing of biomedical literature, this can be too broad. Emerging topics like the Omicron variant and other viral strains (Alpha, Beta, Epsilon) and certain sublineages may not be effectively captured since they are grouped under a general SARS-CoV-2 MeSH term.

The CORACLE platform depends on NCBI’s API E-utility for data, which can be incomplete. This can lead to issues regarding missing references, inconsistent publication dates, and complexities from XML-based publication date formats. Book-type articles, which lack a ’PubMedPubDate’ tag, need alternative date extraction methods.

MeSH terms added manually after publication can lead to data gaps. While updates are periodically made through data retrieval scripts, their

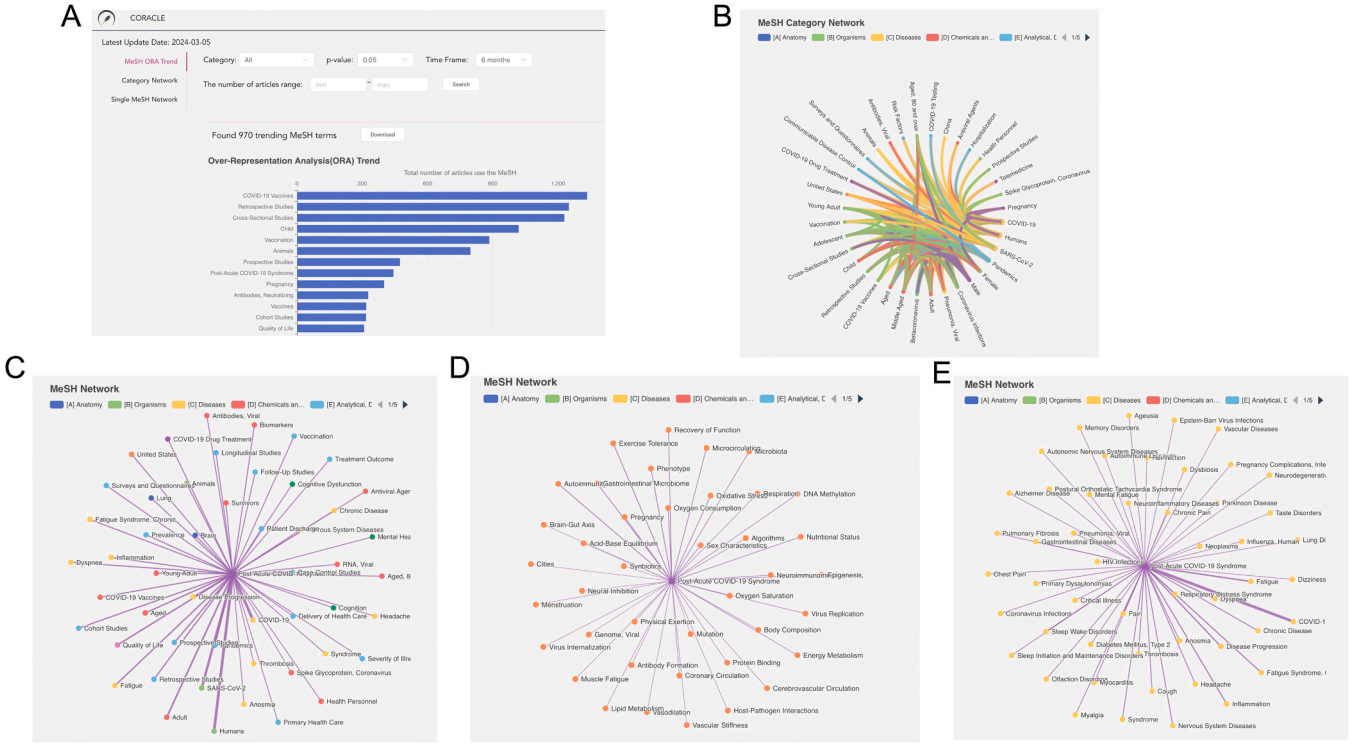

Fig. 6. MESH usage in CORACLE (A) MeSH over-representation analysis trends through a default setting. (B) MeSH category network analysis (default, all MeSH terms). (C-E). Single MeSH network for the term "Post-Acute COVID-19 Syndrome" for all terms (C), categorized for "chemicals and drugs" (D) and "Diseases" (E).

effectiveness depends on the availability of updated data from NCBI databases.

In summary, CORACLE provides valuable insights into COVID-19 research, but its limitations highlight the need for ongoing improvements to stay effective and relevant in the fast-changing field of COVID-19 and post-COVID-19 syndrome research.

4.2. Future CORACLE developments and applications

In a fast evolving research field, keeping up with the scientific literature can be challenging for both clinicians and researchers. This is especially true in a field such as SARS-CoV-2/COVID-19, where the number of new publications grows quickly.

To help, we developed CORACLE, a tool that offers interactive, multi-level filtering and integration of COVID-19 literature. It uses network-based analysis and MeSH/keyword searches to provide additional insights. CORACLE allows for detailed searches using keywords and custom PMID list, offering a time-efficient and comprehensive overview of current and emerging research. This tool helps experts in the field prioritize and stay updated on publications.

In the future, CORACLE aims to include papers from related fields like SARS-CoV-1 and MERS. Extract and predict COVID-19-related genes and pathways to aid vaccine development and identify new therapeutic targets, especially for vulnerable patients with COPD and severe asthma. Coracle also aims to explore the potential to identify drug target genes and pathways from new literature using established pathway databases, enhancing data mining efforts [22–24].

Daily updates to the databases and personalized prioritization, along with future plans to include other conditions and integrate various pathway tools to identify novel pharmaceutical targets, we aim to be of aid for clinicians and researchers dealing with large literature influx.

Author contributions

Conception and design: CXL, ÅMW; analysis and interpretation: CXL, ÅMW; drafting of manuscript: KP, PL, IK, C-X. Li, ÅMW; Python processing of PubMed and LitCovid data: YL; Tutorials and example development: KP, PL, IK, CXJ, JG, ÅMW.

CRediT authorship contribution statement

Pia Lindberg: Writing – original draft, Investigation. **Jing Gao:** Investigation. **Michael Runold:** Writing – review & editing, Conceptualization. **Iryna Kolosenko:** Writing – review & editing, Writing – original draft, Supervision, Investigation. **Kristina Piontkovskaya:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation. **Yulian Luo:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Chuan-Xing Li:** Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Conceptualization. **Åsa Maria Wheelock:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

We would like to extend our appreciation to Dr. Zhiyong Lu and Dr. Qingyu Chen from National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM) and National Institutes of Health (NIH) for their supports of the batch download of LitCovid data, Marika Ström and Benedikt Zöhrer from Karolinska Institutet for their assistance with beta-testing; Daniel Uvehag from Center for Molecular Medicine for his help with the platform and Nicole Wagner for the help with proofreading the manuscript. The project was supported by the Swedish Heart-Lung Foundation (grant IDs: 20190017 and 20210053) and the Swedish Research Council (grant IDs: 2018-00520 and 2021-06546).

References

[1] Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. 193–193 Nature 2020;579. 193–193.
[2] WHO COVID-19 database. ⟨https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/⟩; (2024).
[3] Coronavirus Knowledge Hub. ⟨https://coronavirus.frontiersin.org/⟩; (2024).
[4] Van Rossum, G. & Drake Fred L. Python 3 Reference Manual; (2009).
[5] Global Burden of Disease Long COVID Collaborators et al. Estimated Global Proportions of Individuals With Persistent Fatigue, Cognitive, and Respiratory Symptom Clusters Following Symptomatic COVID-19 in 2020 and 2021. JAMA 328, 1604–1615; (2022).
[6] Townsend L, et al. Persistent poor health after covid-19 is not associated with respiratory complications or initial disease severity. Ann Am Thorac Soc 2021;18:997–1003.
[7] Nalbandian A, et al. Post-acute COVID-19 syndrome. Nat Med 2021;27:601–15.
[8] Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV. A clinical case definition of post-COVID-19 condition by a Delphi consensus (Preprint at) Lancet Infect Dis 2022;vol. 22:e102–7. https://doi.org/10.1016/S1473-3099(21)00703-9.
[9] Klein J, et al. Distinguishing features of long COVID identified through immune profiling. Nature 2023;623:139–48.
[10] Petersen EL, et al. Multi-organ assessment in mainly non-hospitalized individuals after SARS-CoV-2 infection: the Hamburg City Health Study COVID programme. Eur Heart J 2022;43:1124–37.
[11] Lund Berven L, et al. Inflammatory markers, pulmonary function, and clinical symptoms in acute COVID-19 among non-hospitalized adolescents and young adults. Front Immunol 2022;13:837288.
[12] Jiménez-Rodríguez BM, et al. On the single and multiple associations of COVID-19 post-acute sequelae: 6-month prospective cohort study. Sci Rep 2022;12:3402.
[13] de Oliveira TCP, Gardel DG, Ghetti ATA, Lopes AJ. The Glittre-ADL test in non-hospitalized patients with post-COVID-19 syndrome and its relationship with muscle strength and lung function. Clin Biomech 2022;100:105797.
[14] de Roos MP, et al. Pulmonary function and Quality of Life in a prospective cohort of (non-) hospitalized COVID-19 pneumonia survivors up to six months. 14799731221114272 Chron Respir Dis 2022;19. 14799731221114272.
[15] Yazji B, Voduc N, Mulpuru S, Cowan J. Pulmonary sequelae of SARS-CoV-2 infection and factors associated with persistent abnormal lung function at six months after infection: prospective cohort study. PLoS One 2022;17:e0277624.
[16] Eligulashvili A, et al. COVID-19 patients in the COVID-19 Recovery and Engagement (CORE) clinics in the Bronx. Diagnostics 2022;13.
[17] Bailey J, et al. Multidisciplinary center care for long coVID Syndrome-A retrospective cohort study. Am J Med 2023. https://doi.org/10.1016/j.amjmed.2023.05.002.
[18] Wong AW, et al. Use of latent class analysis and patient reported outcome measures to identify distinct long COVID phenotypes: a longitudinal cohort study. PLoS One 2023;18:e0286588.
[19] Steinmetz A, et al. Longitudinal clinical features of Post-COVID-19 patients-symptoms, fatigue and physical function at 3- and 6-month follow-up. J Clin Med 2023;12.
[20] Njøten KL, et al. Relationship between exercise capacity and fatigue, dyspnea, and lung function in non-hospitalized patients with long COVID. Physiol Rep 2023;11:e15850.
[21] Huang C, et al. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. Lancet 2021;397:220–32.
[22] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
[23] Xu D, et al. DTMiner: identification of potential disease targets through biomedical literature mining. Bioinformatics 2016;32:3619–26.
[24] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinf 2015;16.