

Article

Use of Artificial Intelligence in Processing of COVID-19-Related Scientific Literature

Michal Ivantysyn ¹, Dagmar Vaclavikova ¹ and Peter Musil ^{2,*}

¹ APEL, Dunajska 52, 811 08 Bratislava, Slovakia; info@apelnet.sk

² Faculty of Medicine, Comenius University in Bratislava, Spitalska 24, 811 08 Bratislava, Slovakia; peter.musil@fmed.uniba.sk

* Correspondence: peter.musil@fmed.uniba.sk; Tel.: +421-2-901-19397

Abstract: The COVID-19 pandemic has resulted in an unprecedented acceleration in scientific production across multiple disciplines. The vast number of publications available makes it challenging for healthcare professionals and researchers to keep up with the current state of knowledge regarding COVID-19. This article presents covid19-help.org, a free expert-curated database designed to increase the availability of relevant original data related to COVID-19 treatment and prevention via immunization. To accelerate the process of identifying relevant original scientific publications and to simplify annotation of their content, the database uses our artificial intelligence in medical literature (AIM.lit) tool. The article provides an overview of the covid19-help.org database design, the criteria used to select publications, and the use of the AIM.lit tool. The database allows users to easily search and filter records, provides concise information on individual substances and their mechanisms of action, lists relevant original scientific publications with annotations, and offers links to external resources. The AIM.lit tool increases the speed of publication selection and extraction of basic relevant information, without compromising the validity of the data. The technology and experience gained from creating the covid19-help.org database and its tools could also be useful in other areas where scientific information organization is a challenge.

Keywords: COVID-19; literature processing; AI

1. Introduction

The COVID-19 pandemic has imposed an unprecedented burden on healthcare workers and healthcare provision [1]. Generally, scientific production in biomedicine is rapidly increasing. More than 1 million papers every year are indexed in the PubMed database, with the number increasing by 8-9% annually [2]. In response to the pandemic, there has been an extreme acceleration in scientific production across various disciplines. Currently, there are more than 350,000 publications associated with the keyword “COVID-19” alone indexed in PubMed (<https://pubmed.ncbi.nlm.nih.gov/?term=COVID-19>). The complexity and the volume of the data impede overwhelmed healthcare professionals, as well as researchers, from gaining a comprehensive understanding of the current state of knowledge regarding the management of this disease.

Several international initiatives are accumulating research data related to COVID-19 and allow their further analysis by users [3]. However, simple access to practically usable information is often complicated by the absence of annotations, insufficient or too general definitions of criteria for the articles' inclusion in the database, data redundancy, or the need for advanced data mining tools to extract information.

Our free database covid19-help.org was designed to increase the availability of relevant data for healthcare professionals, scientists, and the general public. The aim of covid19-help.org is to shorten the time needed to identify relevant original scientific publications (presenting data regarding the effectiveness of substances in relation to the treatment of COVID-19 or prevention of SARS-CoV-2 infection via immunization) and their conclusions. To make this approach feasible, we have developed and applied models of artificial intelligence (AI) to assist research paper selection and text-mining.

2. The covid19-help.org Database

Prior to the COVID-19 outbreak, we had built a database collecting scientific data on the effectiveness of therapies for cystic fibrosis with respect to the disease-causing mutations (<https://cf-help.org/>). The explosion in the number of scientific papers on COVID-19 [4] inspired us to use our experience and create covid19-help.org.

The number of published peer-reviewed papers immediately after the onset of the pandemic was limited. Therefore, our database also listed relevant preprints, which were gradually replaced by final versions after the peer-review process completion. By mid-2020, enough peer-reviewed papers were available. For quality assurance, only publications indexed in the LitCovid database were considered for covid19-help.org. LitCovid indexes publications from PubMed related to COVID-19 [5].

Covid19-help.org presents only publications fulfilling defined criteria: i) open access to the full text in English; ii) the publication provides experimental or observational data on the efficacy of a defined substance (or group of substances) in the treatment of COVID-19 or prevention of (severe) SARS-CoV-2 infection by immunization; iii) the data in question are original (i.e., not previously published, apart from preprint servers). This filter ensures that the publications which are irrelevant, inaccessible, or do not provide original experimental or observational results, are eliminated and the browsing dataset for the user is greatly simplified.

The workflow for publishing an entry on covid19-help.org is following: i) identification of the publication based on selection criteria; ii) evaluation and annotation of the content; iii) linking the entry to the records of defined substances. If a record is not found in covid19-help.org, it is created.

The covid19-help.org database allows users to easily search and filter records for substances, references, or other components of the database that provide general or interesting information linked to COVID-19. The database provides concise general information on individual substances (including synonyms, commercial products, or chemical structures) and describes their mechanism of action and the level of evidence regarding their putative anti-COVID-19 activity. Links to external sources of information (e.g., PubChem, DrugBank, or Wikipedia) are listed. The substance entry indicates when there is an approval for the use of the substance by selected drug regulatory authorities (European Medicines Agency, The Food and Drug Administration, or Medicines and Healthcare Products Regulatory Agency) and offers links to registered clinical trials (ClinicalTrials.gov) if available.

3. Use of AI in Text-mining

Initially, the identification of publications was carried out exclusively manually by experts. However, due to the increasing volume of scientific production, as well as the expansion of annotations, we have developed AI models that accelerate the process. AI enables identification of relevant publications based on the defined criteria using their abstracts or even titles, where abstracts are not available. The sensitivity of these models in the process of validation was more than 94% and the specificity was more than 90%. The annotation of publications is further accelerated by an AI model that helps to identify key sentences in the text that succinctly summarize the results. Based on validation studies, the current model reaches the sensitivity of 85 % with a false positivity rate below 43 %. This model is still being improved and we hope to decrease the false positivity rate below 20 % without impeding the sensitivity. We estimate the current models to decrease the time needed to extract the desired basic information from LitCovid by 80% without compromising validity of the data.

The testing set of all references available for processing (bulk export from LitCovid available until June 2022) contained 245,822 references. Based on our definition of relevance, AI identified 30,086 papers as relevant via abstract and further 1,128 via title where abstract was not available. This means that 12.7% of the papers were estimated to contain original data fulfilling the criteria for inclusion in covid19-help.org database.

The key content assigned to individual substances is a list of relevant original publications, which are sorted based on whether their conclusions support or contradict the assumption that the

respective substance has potential for the treatment or prevention of (severe) COVID-19. Most publications are briefly annotated (e.g., with information on the results, associated SARS-CoV-2 variants, dosing, or sample size). Available are tags which identify the type of study, the type of tested substances, the severity of COVID-19 in the assessed sample before treatment, and many other parameters.

As of March 2023, we provide records for 1647 substances, of which 501 are already used in the management of other diseases. Covid19-help.org lists various categories of substances, from small molecules, through biopolymers and defined mixtures, to cell products. The publications listed provide different levels of evidence, ranging from in silico modelling of interacting structures, through in vitro and animal models, to clinical studies. The included studies focus on the activity of substances in various phases of infection, including prevention of infection by humoral immunity, inhibition of internalization of viral particles by cellular machinery, inhibition of viral replication by virus-encoded proteins, or suppression of the excessive inflammatory reaction triggered by SARS-CoV-2 infection.

The data within the covid19-help.org database are freely available, exportable in .xls and .csv formats, and accessible for use by other software tools through API. In just one year in 2022, covid19-help.org attracted more than 17,000 users from all inhabited continents.

4. Conclusions

Despite the availability of vaccines and selected antivirals [6, 7], given the persisting risk of emergence of resistant SARS-CoV-2 variants [8], it is necessary to continue intensive research and development of safe and effective pharmacological tools for the prevention and management of COVID-19. Treatment and immunization options are especially needed for vulnerable populations [9]. Continuous mapping of the scientific output in this area therefore remains essential for correct epidemiological and individual decision-making. The experience gained in creating the covid19-help.org database and its tools, including AI models, could be also applied in other areas where organization of scientific information limits the development of knowledge.

Author Contributions: Conceptualization, M.I. and D.V.; writing—original draft preparation, M.I.; writing—review and editing, D.V. and P.M.; supervision, P.M. All authors have read and agreed to the published version of the manuscript.

Funding: The article was created thanks to support from the project “Development and implementation of immunomodulatory cell therapy in the fight against the COVID-19 pandemic” within the Operational Program Integrated Infrastructure, project ITMS2014 code: 313011ATT8.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: <https://covid19-help.org/>

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gnanapragasam, S. N., Hodson, A., Smith, L. E., Greenberg, N., Rubin, G. J., & Wessely, S. (2022). COVID-19 survey burden for health care workers: Literature review and audit. *Public Health*, 206, 94–101. <https://doi.org/10.1016/j.puhe.2021.05.006>
2. Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535(7612), Article 7612. <https://doi.org/10.1038/nj7612-457a>
3. Centers for Disease Control and Prevention. (2020, October 9). COVID-19 Databases and Journals. Stephen B. Thacker CDC Library. <https://www.cdc.gov/library/researchguides/2019novelcoronavirus/databasesjournals.html> (accessed 13/04/2023)
4. Riccaboni, M., & Verginer, L. (2022). The impact of the COVID-19 pandemic on scientific research in the life sciences. *PLOS ONE*, 17(2), e0263001. <https://doi.org/10.1371/journal.pone.0263001>
5. Chen, Q., Allot, A., & Lu, Z. (2021). LitCovid: An open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1), D1534–D1540. <https://doi.org/10.1093/nar/gkaa952>

6. Li, M., Wang, H., Tian, L., Pang, Z., Yang, Q., Huang, T., Fan, J., Song, L., Tong, Y., & Fan, H. (2022). COVID-19 vaccine development: Milestones, lessons and prospects. *Signal Transduction and Targeted Therapy*, 7(1), Article 1. <https://doi.org/10.1038/s41392-022-00996-y>
7. Singh, M., & de Wit, E. (2022). Antiviral agents for the treatment of COVID-19: Progress and challenges. *Cell Reports Medicine*, 3(3), 100549. <https://doi.org/10.1016/j.xcrm.2022.100549>
8. Chakraborty, C., Bhattacharya, M., & Sharma, A. R. (2022). Emerging mutations in the SARS-CoV-2 variants and their role in antibody escape to small molecule-based therapeutic resistance. *Current Opinion in Pharmacology*, 62, 64–73. <https://doi.org/10.1016/j.coph.2021.11.006>
9. Salasc, F., Lahlali, T., Laurent, E., Rosa-Calatrava, M., & Pizzorno, A. (2022). Treatments for COVID-19: Lessons from 2020 and new therapeutic options. *Current Opinion in Pharmacology*, 62, 43–59. <https://doi.org/10.1016/j.coph.2021.11.002>