DIGITAL HEALTH

Digital Health
Volume 8: 1–12
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076221133696
journals.sagepub.com/home/dhj
SAGE

A topic trend analysis on COVID-19 literature

Sara Urru[1], Veronica Sciannameo[2], Corrado Lanera[1], Silvano Salaris[1], Dario Gregori[1] and Paola Berchialla[2]

Abstract

Objective: In the past 2 years, the number of scientific publications has grown exponentially. The COVID-19 outbreak hugely contributed to this dramatic increase in the volume of published research. Currently, text mining of the volume of SARS-CoV-2 and COVID-19 publications is limited to the first months of the outbreak. We aim to identify the major topics in COVID-19 literature collected from several citational sources and analyze the temporal trend from November 2019 to December 2021.

Methods: We performed an extensive literature search on SARS-Cov-2 and COVID-19 publications on PubMed, Scopus, and Web of Science (WoS) and a structural topic modelling on the retrieved abstracts. The temporal trend of the recognized topics was analyzed. Furthermore, a comparison between our corpus and the COVID-19 Open Research Dataset (CORD-19) repository was performed.

Results: We collected 269,186 publications and identified 10 topics. The most popular topic was related to the clinical pictures of the COVID-19 outbreak, which has a constant trend, and the least popular includes studies on COVID-19 literature and databases. "Telemedicine", "Vaccine development", and "Epidemiology" were popular topics in the early phase of the pandemic; increasing topics in the last period are "COVID-19 impact on mental health", "Forecasting", and "Molecular Biology". "Education" was the second most popular topic, which emerged in September 2020.

Conclusions: We identified 10 topics for classifying COVID-19 research publications and estimated a nonlinear temporal trend that gives an overview of their unfolding over time. Several citational databases must be searched to retrieve a complete set of studies despite the efforts to build repositories for COVID-19 literature. Our collected data can help build a more focused literature search between November 2019 and December 2021 when carrying out systematic and rapid reviews and our findings can give a complete picture on the topic.

Keywords

COVID-19, information retrieval, text mining, science communications, open science, research practice

Submission date: 20 June 2022; Acceptance date: 30 September 2022

Introduction

The volume of published research has exponentially grown in the past 2 years. To give some figures, in 2020, the total number of publications indexed by PubMed was more than 1.4 million, corresponding to a 15% increase over 2019.[1]

The COVID-19 outbreak hugely contributed to this unprecedented number of publications. On one side, during the pandemic, many countries adopted a lockdown, which forced many researchers to home working, thus focusing more on writing papers than carrying out experiments in laboratories.[2] On the other side, the fast spread of COVID-19 became the new priority, so publication time was sped up,[3] producing a flood of new research focused on coronavirus. Accordingly, the need to mine such a volume of SARS-CoV-2 and COVID-19 publications emerged.[4]

[1]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padua, Italy
[2]Center of Biostatistics, Epidemiology and Public Health, Department of Clinical and Biological Sciences, University of Torino, Turin, Italy

Corresponding author:
Paola Berchialla, Center of Biostatistics, Epidemiology and Public Health, Department of Clinical and Biological Sciences, University of Torino, Regione Gonzole 10, Turin, 10043 Orbassano, Italy.
Email: paola.berchialla@unito.it

Special repositories were created to collect the scientific literature on Coronavirus research. An example is COVID-19 Open Research Dataset (CORD-19) collected at the Allen Institute for AI,[5] which includes the literature corresponding to the following query run in PubMed, PMC, medRxiv, bioRxiv, and World Health Organization (WHO) repository:

> “COVID-19"[All Fields] OR (“coronavirus"[MeSH Terms] OR “coronavirus"[All Fields]) OR “Corona virus"[All Fields] OR “2019-nCoV"[All Fields] OR “SARS-CoV"[All Fields] OR “MERS-CoV"[All Fields] OR “Severe Acute Respiratory Syndrome"[All Fields] OR “Middle East Respiratory Syndrome"[All Fields]”.

As seen emerging from the query, CORD-19 contains not only publications about COVID-19 but also on coronaviruses and viruses in general.[6] Another COVID-19-specific database is LitCovid, which comprises all the relevant publications indexed in PubMed.[7] LitCovid organizes the relevant literature into curated categories or research topics. By the end of December 2021, nine categories were identified: (1) “Mechanism”, comprising publications pointing up the causes of infections and the possible drug mechanisms of action; (2) “Transmission”, for publications related to the ways of transmission in human interactions; (3) “Diagnosis”, comprising publications related to symptoms and tests; (4) “Treatment”, related to treatment strategies, therapeutic procedures, and vaccine development; (5) “Prevention”, considering strategies and management for prevention and control; (6) “Case reports”; (7) “Forecasting”, to predict the temporal trend of SARS-CoV-2 spread; (8) “Long Covid”, related to long-term COVID-19 syndrome, and finally, (9) “General”, a category which covers all the other eight and includes the articles that cannot fit the previous filters. While LitCovid focuses on COVID-19 research, CORD-19 is more generic.[7] However, neither database retrieves data from larger citational repositories such as Web of Science (WoS) or Scopus.

The massive amount of literature available on a specific research theme also stimulated several analyses based on text analysis and network analysis to compare the scientific production among countries, detect the collaborations among institutes and carry out bibliometric analyses.[6,8] These investigations within the Coronavirus research allowed more detailed analysis of the content and the publications’ characteristics. Earlier works[6,8–15] are limited to a narrow range of time and do not consider the evolution over time of literature. To get a clear picture, it is necessary to reconstruct the timeline of COVID-19.[16] The first cases emerged at the end of 2019 when a cluster of pneumonia cases was reported in Wuhan on December.[16] Transmission most likely started beforehand; in fact, dating back to 17 November, 2019, patient zero was a 55-year-old man from the province of Hubei.[17] In January, the first COVID-19 death was registered in China, and then Wuhan was placed in quarantine.[16] As cases of COVID-19 spread worldwide, the WHO declared that the outbreak constituted a Public Health Emergency of International Concern (PHEIC).[18] In March, the Italian Prime Minister announced the first lockdown in Europe, and other countries followed soon afterwards.[16,19] The lockdowns limited the spread of COVID-19 but brought several economic and health impacts, especially a rising in mental illness associated with the isolation.[20–22] After the lockdown, not all economic and social activities resumed as before the emergency, among which were the schools.[23] The advent of vaccines helped reduce these burdens.[24]

The problem we address in this work is the rapid growth of literature since (1) quality of publications is not guaranteed,[3,25,26] (2) the search about a specific topic is time expensive and the results include noisy data,[4] and (3) repositories on COVID-19 literature built so far still lack available data.[5,7] Previous studies[6,8–15] tried to give an overview about the COVID-19 literature, but they were mainly published in 2020 and cannot therefore include later data. Furthermore, the approaches used aimed to describe the data instead of modeling them. The purpose of our work is to extend the analysis of the literature for a wide range of time and databases using Structural Topic Model (STM),[27] which allows to account for time, improving the latest findings and providing a complete repository focused on COVID-19 to facilitate future research on the theme.

Methods

Data source

We retrieved publications related to SARS-Cov-2 and COVID-19 research from PubMed, Scopus, and WoS, starting from 1 November 2019, when the first cases were discovered to 7 December 2021, when the analyses were performed. For the data extraction, we used the following R packages: easyPubmed,[28] rscopus,[29] scopusAPI,[30] and wosr.[31]

Search strategy

To retrieve the publications related to SARS-Cov-2 and COVID-19 research, we ran the following query on the paper titles, abstracts, and keywords, which was used to track the literature about the 2019 Coronavirus in the PubMed LitCovid web-based system:

> (“COVID-19” OR “COVID-19"[MeSH Terms] OR “COVID-19 Vaccines” OR “COVID-19 Vaccines"[MeSH Terms] OR “COVID-19 serotherapy” OR “COVID-19

serotherapy"[Supplementary Concept] OR “COVID-19 Nucleic Acid Testing” OR “COVID-19 nucleic acid testing"[MeSH Terms] OR “COVID-19 Serological Testing” OR “COVID-19 serological testing"[MeSH Terms] OR “COVID-19 Testing” OR “COVID-19 testing"[MeSH Terms] OR “SARS-CoV-2” OR “sars-cov-2"[MeSH Terms] OR “Severe Acute Respiratory Syndrome Coronavirus 2” OR “NCOV” OR “2019 NCOV” OR ((“coronavirus"[MeSH Terms] OR “coronavirus” OR “COV”) AND 2019/11/01[PDAT] : 3000/12/31[PDAT]))

We adapted the query for WoS and Scopus using the Polyglot Search Translator.[32]

Data preprocessing

We considered only articles with DOI, to guarantee uniqueness and publication, and abstract, to perform the text analysis. Duplicates, that is, papers with the same DOI, were removed, as well as non-English abstracts. For the latter purpose, Google’s Compact Language Detector (CLD) algorithm for language identification was applied. Following basic recommendations,[33] we adopted a combination of CLD2, based on a neural network model,[34] and CLD3, which uses a Bayesian approach.[33] We retained only the abstracts of articles, reviews, and proceeding papers classified as English by both algorithms. Then we lowercased text and removed punctuation, numbers, stop words, and words with less than three characters. Finally, we stemmed the remaining words, that is, suffixes were removed, so that words that differed only by their ending were treated as one.

Data analysis

We performed STM[27] on the preprocessed text of the retained abstracts. STM is a generative process of word counts. The word distribution, that is, the frequency and the co-occurrence of words, generates the topic distribution, that is, the probability that a set of words occurs in the text. In turn, the topic distribution generates documents; in other words, each document is seen as a distribution of topics. The innovation of STM regarding the standard topic model methods such as Latent Dirichlet Allocation (LDA) is the possibility of allowing for correlation among topics and including metadata. By assumption, topics are not necessarily considered independent from each other, and covariates can be used to estimate the topic prevalence, that is, the distribution of the topics, and the topic content, that is, the distribution of words in each topic. Since we were interested in topic trends over time, a trimester indicator was included to estimate the topic prevalence assuming a nonlinear relationship: we applied a B-spline basis with four degrees of freedom, that is, a smooth curve joining polynomial functions.

We ran STM over a range of 8 to 20 topics, and we chose the better performing one according to two specific metrics: semantic coherence (SC) and exclusivity of words. SC measures the co-occurrence of the most frequent words in a topic. Exclusivity, instead, measures how much a word is topic-specific and is computed using the FRequency and EXclusivity (FREX) metrics. FREX is a univariate measure that combines the importance of a word considering exclusivity and frequency together.[27,35] SC and FREX are calculated for every topic. Then, the average over all the topics is used to choose the best model fit.

Data analysis was performed in R version 4.1.0[36] using the *stm* package.[27]

Comparison with CORD-19

As CORD-19[5] is the largest repository of COVID-19 publications, a comparison with our corpus is performed to highlight the weaknesses and the strengths of our search strategy, counting the number of articles included in CORD-19 but not in our corpus and vice versa.

Results

Up to 7 December 2021, we retrieved a total of 295,313 publications from Scopus, 199,317 from PubMed, and 216,039 from WoS using the adapted query of LitCovid. We included only English language abstracts of articles, reviews, and proceeding papers. After removing abstracts without DOI and duplicates, we obtained a corpus of 269,186 documents (Figure 1).

The number of publications has continually grown: a peak in January 2021 can be seen in Figure 2 because in Scopus and WoS, many articles are indexed the first day of the year.

The corpus was preprocessed to remove useless characters, symbols, and terms. We started from a vocabulary of 357,781 terms extracted from our corpus. Then we excluded words that appeared in less than 100 documents corresponding to 0.04% of the corpus. Finally, a vocabulary of 8813 words was used to calculate topics and document distributions.

To determine the adequate number of topics, we ran STM for a number of topics ranging between 8 and 20. Overall, 10, 11, and 12 topics maximized both SC and exclusivity. We chose the parsimonious model with 10 topics.

Topic description

We assigned a label to each topic, according to the most frequent and exclusive words (Figure 3).

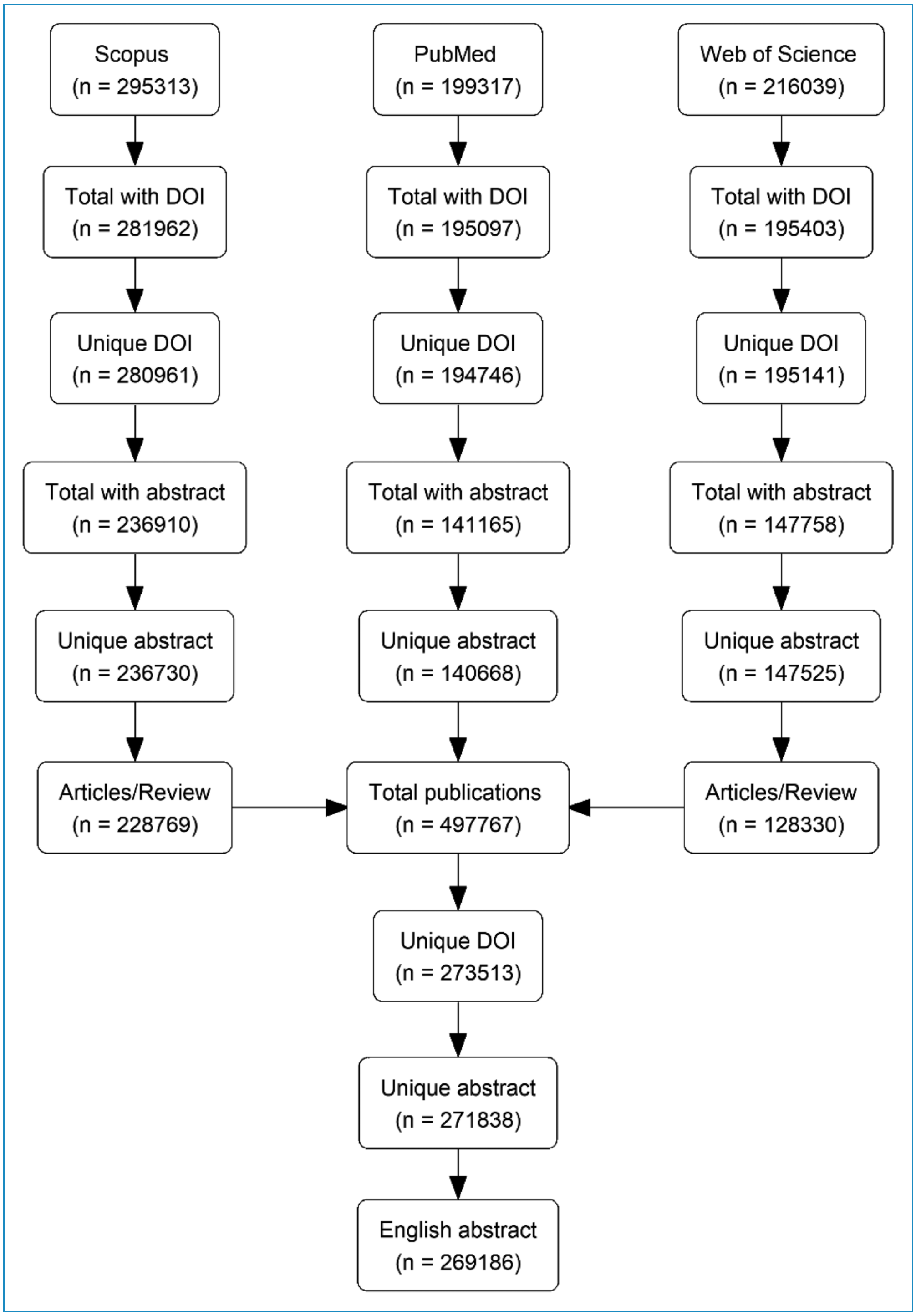


Figure 1. Data were extracted from Scopus, PubMed, and Web of Science (WoS). Duplicates based on DOI and abstract were removed together with non-English texts.

Topic 1 (clinical presentation) has the highest expected proportion (13.9%) and corresponds to the published papers that focused on the clinical picture of the COVID-19 outbreak. It is the most general topic on COVID-19, which started to be relevant at the beginning of the outbreak and has remained constant until now (Figure 4).

Topic 2 (education) covers a proportion of 12.1% and focuses on schools and students; it is characterized by words related to online learning. Topic 3 (vaccine

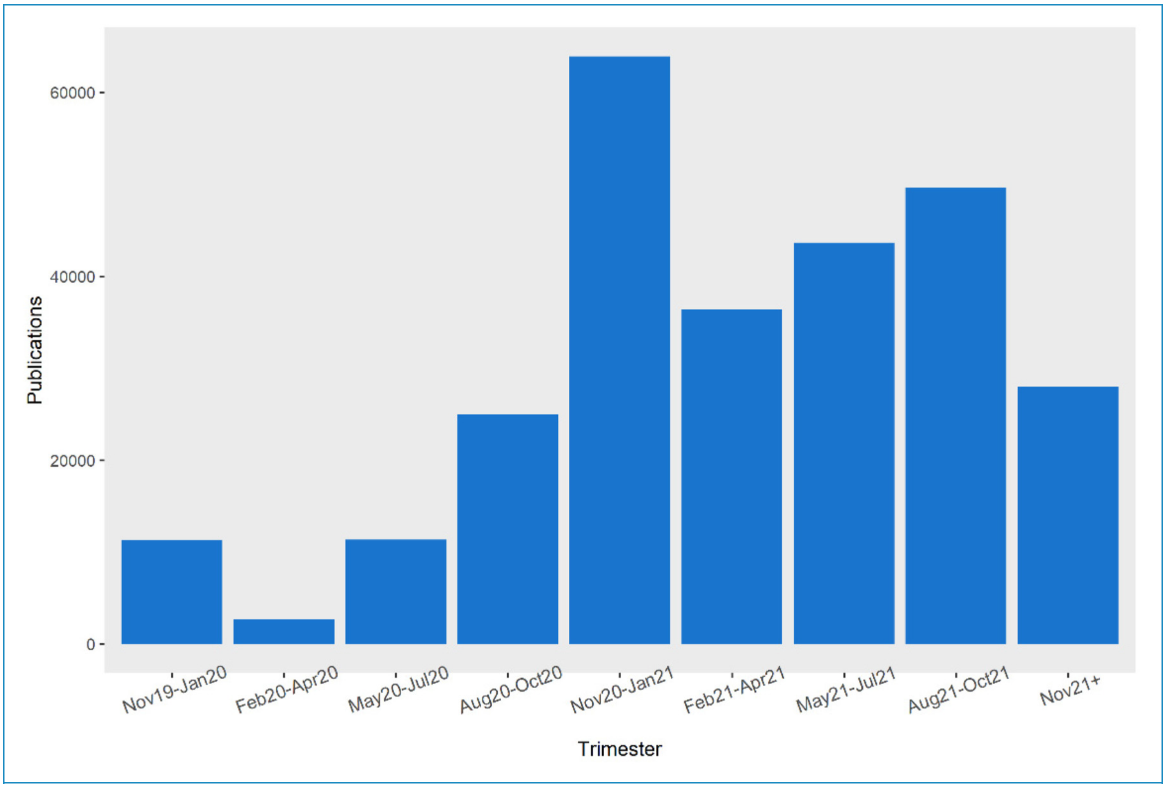


Figure 2. Time distribution of publications.

development), with 10.5% of documents, comprises the research related to vaccine development, which dramatically grew in the first months of the pandemic (Figure 4), following the increasing knowledge on the virus (Topic 3).

Topics 4, 5, and 6 share approximately the same proportion of articles, around 10%. In more detail, Topic 4 (epidemiology) is composed of epidemiological articles tracking and describing the COVID-19 spread over time and across borders. Topic 5 (telemedicine) is about telemedicine and telehealth as novel approaches to take care of patients in a safe manner. Topic 6 (COVID-19 impact on mental health) describes the COVID-19 impact on mental health as an increasing level of anxiety and depression in patients and, more generally, in people.

At the beginning of the pandemic, several researchers contributed in developing predictive models to forecast the unfolding of the SARS-CoV-2 spread, thus topic 7 (forecasting) is still an increasing topic, and it includes 9.6% of the literature.

Topic 8 (molecular biology), with 8.9% of documents, explains the characteristics of the SARS-CoV-2 virus such as its structure and genomic and how to identify its presence in the human body.

Topic 9 (COVID-19 impact on the economy) comprises papers related to the impact of COVID-19 on the economy from the markets to tourism (8.8%).

Finally, topic 10 (literature analysis) shares the smallest set of documents (3.3%), and it includes the analyses of COVID-19 literature: systematic reviews, meta-analyses, and databases. It predictably grew, albeit slightly, alongside COVID-19 literature as a whole.

Figure 3 shows the expected proportions of topics and the most frequent and exclusive words for each topic, while Figure 4 shows the topic trends from November 2019 to November 2021.

Comparison with CORD-19

CORD-19 is the largest citational database related to COVID-19 publications; for this reason, a comparison with the literature included in other databases can be helpful to underline the weaknesses and strengths of both CORD-19 and our corpus. CORD-19 collected 845,575 publications up to 29 November 2021. Applying the same strategy that we followed to get the corpus for our analysis (selecting articles, reviews, and proceeding papers with abstracts and excluding duplicates), the number of publications has been reduced to 299,075. This number is slightly greater than our corpus (271,838). When comparing the two datasets (Figure 5) we observed that almost half of the publications included in CORD-19 are not present in our corpus. This is due to the more general query used for retrieving publications in CORD-19 and the inclusion of

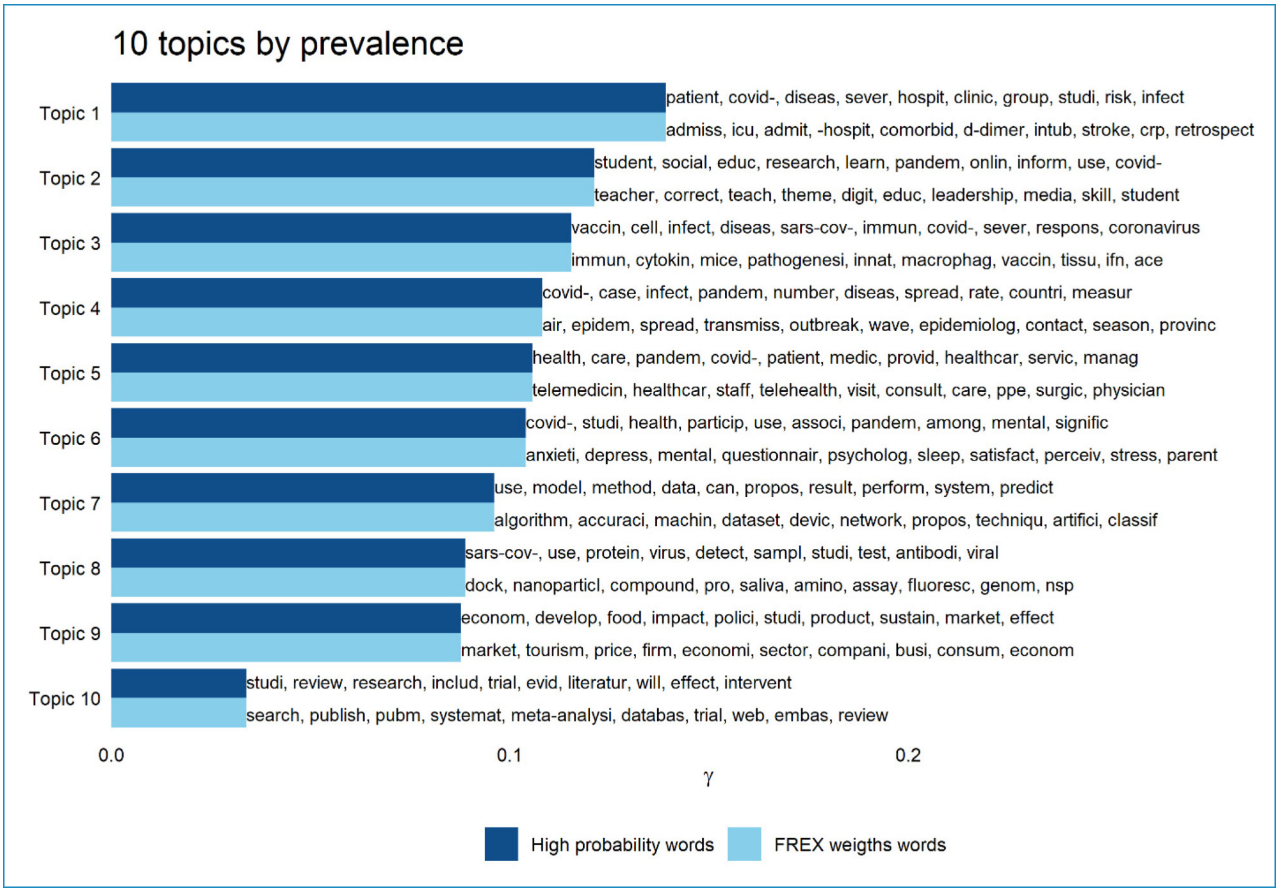


Figure 3. Topic prevalence for the model with 10 topics. High probability word and FRequency and Exclusivity (FREX) terms are shown. Topic 1 (Clinical presentation), Topic 2 (Education), Topic 3 (Vaccine development), Topic 4 (Epidemiology), Topic 5 (Telemedicine), Topic 6 (COVID-19 impact on mental health), Topic 7 (Forecasting), Topic 8 (Molecular biology), Topic 9 (COVID-19 impact on economy), and Topic 10 (Literature analysis).

7334 preprint articles and 28,281 WHO reports. Nevertheless, our corpus contains many articles not included in CORD-19. Only 2202 out of 128,330 publications indexed in WoS and 33,727 out of 228,769 in Scopus are included in CORD-19. Table 1 shows the distribution of publications, both from our corpus and CORD-19, according to their occurrence in the specific databases considered (Scopus, WoS, and PubMed). Each row of Table 1 reports the number of publications appearing in all the databases indicated in the "Source" column and none of the other databases. For example, if one investigates Scopus, they can notice that 75,533 publications are included in all the other repositories as well, 67,503 are in Scopus only, 33,727 are both in Scopus and CORD-19, and so on.

Discussion

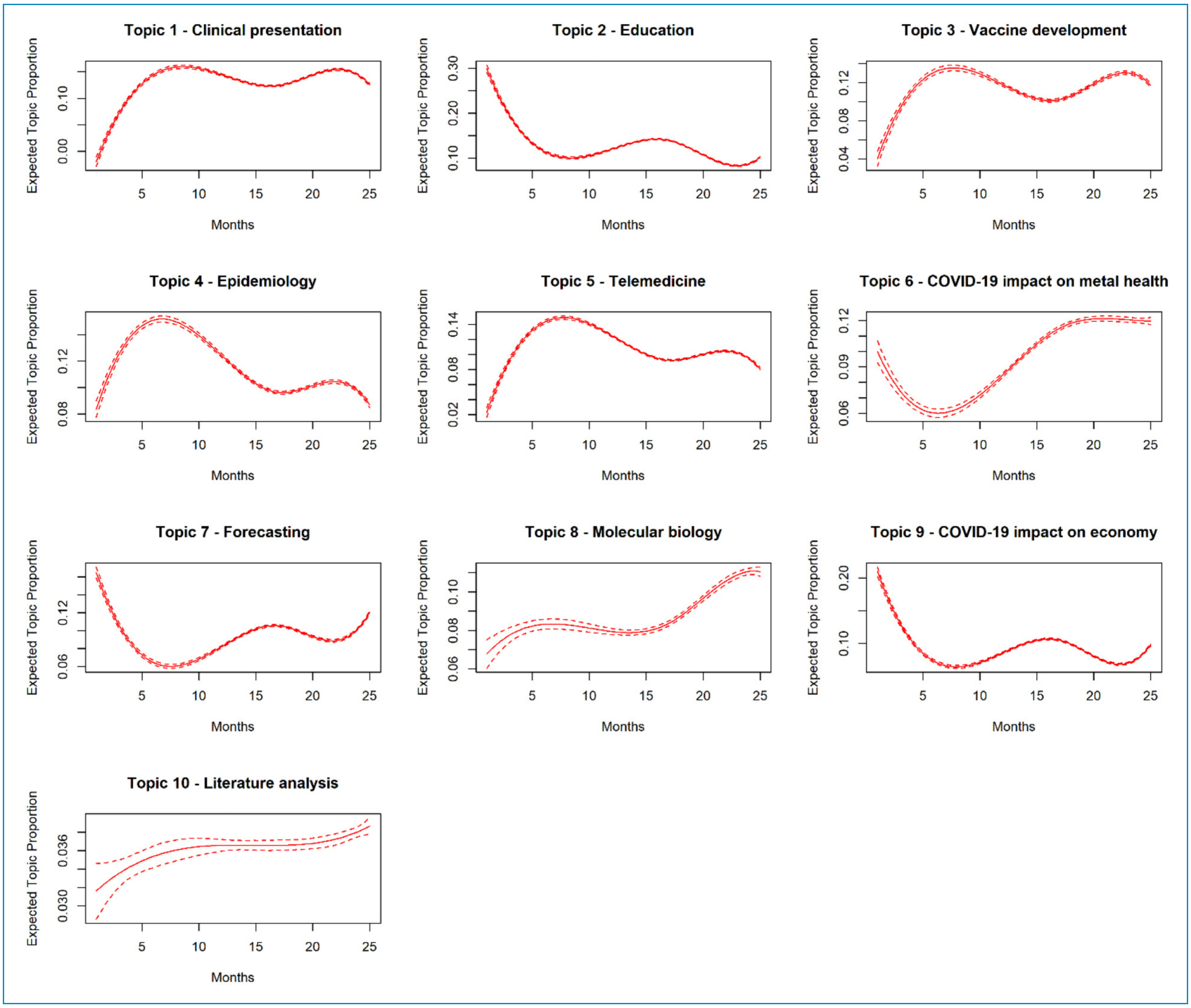
Topic evolution over time

In this work, we performed an analysis based on topic modeling on COVID-19 publications' abstracts retrieved from PubMed, WoS, and Scopus. We identified 10 topics and studied their distribution over time. The most popular topic was "Clinical presentation", while the less recurrent was "Literature analysis". The temporal trend of the former remained constant after rapid growth, probably due to the need to provide an updated clinical picture of the disease.[37] The latter includes studies on COVID-19 literature and databases that increased over time as expected to grow as the pandemic progresses. Rising topics over time were also the "COVID-19 impact on mental health" and "Forecasting".

"COVID-19 impact on mental health" started to grow in May 2020, after several countries loosened isolation measures .[21,38,39] Depressive symptoms showed up in Europe, especially in women and elderly populations.[20,22]

"Forecasting" started its growth in October 2020; at this point, a large amount of data on COVID-19 was available and almost a 1000 models were performed and published in Scopus and WoS.[40]

"Epidemiology" was increasing in the early phase when there was the need to understand how the virus spread to


Topic 1 - Clinical presentation

Topic 2 - Education

Topic 3 - Vaccine development

Topic 4 - Epidemiology

Topic 5 - Telemedicine

Topic 6 - COVID-19 impact on metal health

Topic 7 - Forecasting

Topic 8 - Molecular biology

Topic 9 - COVID-19 impact on economy

Topic 10 - Literature analysis

Figure 4. Topic trends by month since the beginning of the pandemic.

make decisions about distancing, mask wearing, and other prevention measures. It peaked in June 2020 when restrictions started to be less severe in several countries, like for instance New Zealand which was declared “virus-free”, and France which reopened its borders with most European countries.[16]

“Telemedicine” arose in April 2020 when lockdown measures took place in many countries (e.g. China, Italy, the UK, and Spain). It comprises publications related to technology services applied as a remote monitoring tool for managing of the COVID-19 patients.[41]

“COVID-19 impact on economy” grew to start from October 2021 when businesses restarted. Lockdown halted most economic activities by reducing production, consumption, and employment in most sectors. On the contrary, the pharmaceutical industry benefited considerably from the vaccine demand.[42]

“Vaccine development” became immediately urgent, reflected by its rapid growth. The topic remained popular except for a slight drop between October 2020 and March 2021, coincident with the starting of vaccine administration to the population in the United States[43] and Europe.[44]

“Education” suffered greatly because of restrictions. During the first lockdowns, schools were not prepared for alternative ways of learning, but later online platforms became very popular.[45]

“Molecular biology” had an increasing trend justified by the appearance of new variants.[46]

Related works

As we highlighted in the introduction, the massive quantity of COVID-19 publications stimulated several literature analyses. In Table 2, the main related works are reported.

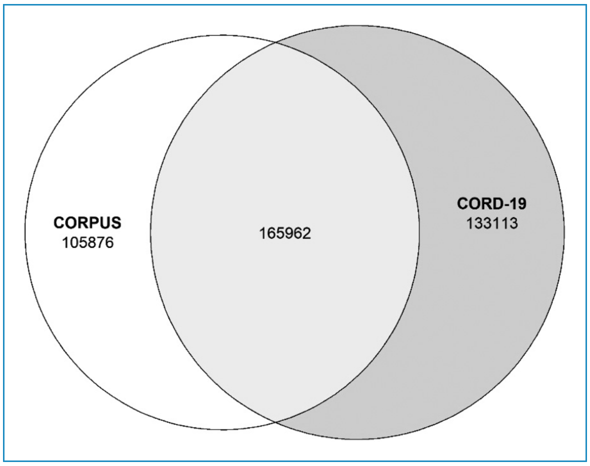


Figure 5. Comparison between coverage of COVID-19 Open Research Dataset (CORD-19) and coverage of our corpus of articles, reviews, and proceeding papers indexed in PubMed, Scopus, and Web of Science (WoS).

Table 1. Distribution of publications among COVID-19 Open Research Dataset (CORD-19), PubMed, Scopus, and Web of Science (WoS).

Sources	Number of publications
Scopus □ PubMed □ WoS □ CORD-19	75,533
Scopus	67,503
Scopus □ CORD-19	33,727
Scopus □ PubMed □ CORD-19	20,479
PubMed □ CORD-19	16,544
Scopus □ WoS	13,062
PubMed □ WoS □ CORD-19	11,571
Scopus □ PubMed □ WoS	8824
WoS	8770
Scopus □ WoS □ CORD-19	5906
PubMed	3246
Scopus □ PubMed	2798
WoS □ CORD-19	2202
PubMed □ WoS	1673

Table 2. Characteristics of studies that analyze COVID-19 research literature.

Study	Database	Time range	Method
Alga et al. 2020[10]	PubMed	February 2020–June 2020	LDA
Bai et al. 2020[11]	Kaggle	January 2020–March 2020	DTM
Cernile et al. 2020[15]	CORD-19	until August 2020	Network analysis
Colavizza et al. 2020[6]	CORD-19	until July 2020	Network analysis
Dastani et al. 2020[14]	LitCovid	until February 2021	LDA
Dehghanbanaki et al. 2020[8]	Scopus	December 2019–April 2020	Bibliometric analysis
Ebadi et al. 2020[12]	PubMed, ArXiv	January 2020–May 2020	STM
Resnik et al. 2020[13]	CORD-19	until May 2020	SpaCy
Wang et al. 2020[9]	PubMed	January 2020–July 2020	Network analysis
Our work	PubMed, Scopus, WoS	November 2019–December 2021	STM

CORd-19, COVID-19 Open Research Dataset; DTM, Dynamic Topic Model. LDH, Latent Dirichlet Allocation; WoS, Web of Science.

Table 2 shows that LDA[47] was the most used method[6,10,13,14,48] to analyze text and classify publication. STM, the method for topic modeling we used, was applied only by Ebadi et al.[12] who considered PubMed and ArXiv literature and included the month of publication to estimate the prevalence of topics using a linear relationship. Dynamic Topic Model (DTM)[49] is another approach used to analyze the change of topics in a text corpus, which has been used by Bai et al.[11] to track the evolution of topics over time in COVID-19 research.

Previous studies[6,10,12–14] performed topic models using a concatenated string of titles, keywords, and abstracts of

COVID-19 publications, which can be redundant and produce a bias in calculating word frequency. LDA does not allow the inclusion of information other than text, so the temporal evolution of the topics presented by Alga et al.[10] and Colavizza et al.[6] is just qualitative. On the contrary, STM has less restrictive assumptions; topic prevalence and content can be modeled, considering authors' country, journal, and publication time.

Together with LDA, network analysis has been used often with two aims: mapping collaborations between countries and institutions[8,9] which showed that the top four productive countries in the early pandemic were the United States, China, Italy, and the UK,[8,9,50] and getting a term map using co-occurrence words from a set of Mesh terms,[9] UMLS terms,[15] or title, abstract, and keywords.[8,50]

Bibliometric analysis was performed by Colavizza et al.,[6] using citation clustering and altimetric analysis to describe how the research was perceived on social media, and by Dehghanbanadaki et al., considering journals and their impact factors, citation scores, and H-index.[8]

Despite the development of open search databases such as CORD-19 and LitCovid, which facilitate the extraction of COVID-19 publications with a string query already prepared, many researchers conducted their search on other sources.

The number of topics

Wang and Hong[9] identified only four topics using a co-occurrence network of MeSH Terms: epidemiology and public health interventions, virus infection and immunity, clinical symptoms and diagnosis, drug treatments, and clinical studies. Each of their topics includes several topics which are distinguished in our analysis, for example, their first two topics also included vaccine and mental health issues that we identified as different topics. Clinical symptoms and diagnosis were in the spotlight during the early stage of the pandemic; thus, the different times we collected published research could explain the dissimilarity. The drug treatment topic was barely identified in Wang and Hong [9] and did not emerge as a single topic in our analysis. We agree with Wang and Hong's conclusion that there is still little research in this area.

Colavizza et al.,[6] in their analysis of CORD-19, identified 15 topics that were merged into broader categories: clinical medicine, coronavirus outbreak, epidemics, immunology, molecular biology, and public health. This overlap of the topics hints that a smaller number would probably lead to better classification. Furthermore, the topic on coronavirus outbreak refers to the broad literature on coronavirus in CORD-19 and not only to COVID-19.

Quality issues of research

Overall, 10 topics seem to be a good solution. To further support this, the results of the models we performed with 11 and 12 topics were quite similar, but the topics were not clearly defined. Interestingly we noticed that the former identified a small group of retracted articles, pointing out the quality issue of some rushed COVID-19 research.[1] In fact, one of the FREX words characterizing one topic of the model with 11 topics was "withdraw" (Supplemental Figure S1). We delved into it by isolating the publications whose probability of containing this topic was higher than all the other topics; we found 37 articles whose abstract or title were characterized by at least one of the following: "withdrawal notice", "withdrawn", "retraction", and "correction" (Supplemental Table S2). Publication process was sped up during the pandemic[3] because of the need to inform the community as fast as the virus spread, but little attention was given to details[4] causing a "pandemic of publications" as well. In particular, the retraction of a study concluding that COVID-19 patients taking chloroquine or hydroxychloroquine were more likely to die and a study made a major scandal in the scientific community,[51] but also had great consequences on several randomized clinical trials which stopped administering the drug to patients and recruiting them. Quality of research is questioned also from a statistical point of view: Wynants et al.[52] reviewed the prediction models for COVID-19 and concluded that they are affected by poor reporting, high risk of bias. and optimistic performances.

Strengths and limitations

To our knowledge, our work is the first considering both publications from a 2-year range of the pandemic and with an application of STM. STM allowed the inclusion of time to compute the topic prevalence. In this way, we also modelled the topic prevalence change over time, allowing for a potential non-linear trend.

STM has been proposed to overcome LDA limitations. However, its weaknesses still need to be investigated. Interpretability is a crucial point: topics are generated from a bag of words, but the context is not considered; hence the assigned labels could be misinterpreted. It is fundamental to find the most suitable set of words used to calculate topic and document distributions, so particular attention in the preprocessing phase is needed.[53] Moreover, since citational databases include non-English publications, a more accurate algorithm for language filtering than CLD2 and CLD3 are suggested to improve the quality of the results such as the convolutional neural network model proposed in 2020 by Vo and Khoury.[54]

When analyzing a vast corpus, machine learning algorithms can better understand the content than map term or graph network.

Deep learning approaches such as BERT,[55] neural networks, and their variations can be applied to the analysis of COVID-19 literature, as shown by teams who participated in the BioCreative IV challenge for the multilabel topic classification of LitCovid.[56,57] Our corpus focuses on COVID-19 and includes data from several citational databases such as Scopus, PubMed, and WoS. LitCovid is daily updated and COVID-19 specific, but it is limited to PubMed data. CORD-19 is an extensive database updated once a week. However, it focuses on coronaviruses and viruses in general, making it challenging to filter COVID-19 literature, and lacks much of the research indexed in Scopus and WoS. Future development could add data from other sources and analyze the preprint articles.

Conclusions

Despite the efforts to build repositories for COVID-19 literature, several citational databases must still be searched to get complete knowledge.

This work identified the main topics in COVID-19 literature collected from several citational sources. We analyzed the temporal trend from November 2019 to December 2021, obtaining an up-to-date overview of COVID-19 research literature. In doing it, we gathered a corpus of articles, reviews, and proceeding papers that focuses on COVID-19 and are indexed on Scopus, PubMed, and Web of Science. Our corpus provides a more extensive coverage than the other repositories, such as LitCovid, which is limited to PubMed data, and CORD-19, which focuses on coronaviruses and viruses in general, making it challenging to filter COVID-19 literature, and lacks much of the research indexed in Scopus and Web of Science. Thus, our findings can help build a more focused literature search when carrying out systematic and quick reviews.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Contributorship: Conceptualization: PB, VS, and SU; data analysis: SU; writing—original draft preparation: SU; writing—review and editing: VS, SS, CL, and PB; methodology: PB and DG; data interpretation: SU, VS, PB, and CL; supervision: PB and DG.

Ethical approval: Not applicable, because this study does not contain data on human subjects.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the University of Turin (BERP_RILO_21_01).

Guarantor: PB.

Data availability: Data are available upon request from the corresponding author.

ORCID iDs: Dario Gregori https://orcid.org/0000-0001-7906-0580
Paola Berchialla https://orcid.org/0000-0001-5835-5638

Supplemental material: Supplemental material for this article is available online.

References

1. Berchialla P, Urru S and Sciannameo V. The effect of COVID-19 on scientific publishing in Italy. *Epidemiol Prev* 2021; 45: 449–451.
2. Squazzoni F, Bravo G, Grimaldo F, et al. Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals. *PLoS One* 2021; 16: e0257919.
3. Horbach SPJM. Pandemic publishing: medical journals strongly speed up their publication process for COVID-19. *Quant Sci Stud* 2020; 1: 1056–1067.
4. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science* 2020. https://www.science.org/content/article/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat (accessed 3 August 2022).
5. Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: the Covid-19 open research dataset. *ArXiv*. 2020. arXiv:2004.10706v2.
6. Colavizza G, Costas R, Traag VA, et al. A scientometric overview of CORD-19. *PLoS One* 2021; 16: e0244839.
7. Chen Q, Allot A and Lu Z. Litcovid: an open database of COVID-19 literature. *Nucleic Acids Res* 2021; 49: D1534–D1540.
8. Dehghanbanadaki H, Seif F, Vahidi Y, et al. Bibliometric analysis of global scientific research on Coronavirus (COVID-19). *Med J Islam Repub Iran* 2020; 34: 51.
9. Wang J and Hong N. The COVID-19 research landscape: measuring topics and collaborations using scientific literature. *Medicine (Baltimore)* 2020; 99: e22849.
10. Älgå A, Eriksson O and Nordberg M. Analysis of scientific publications during the early phase of the COVID-19 pandemic: topic modeling study. *J Med Internet Res* 2020; 22: e21559.
11. Bai Y, Jia S and Chen L. Topic evolution analysis of COVID-19 news articles. *J Phys Conf Ser* 2020; 1601: 052009.
12. Ebadi A, Xi P, Tremblay S, et al. Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. *Scientometrics* 2021; 126: 725–739.
13. Resnik P, Goodman KE and Moran M. Developing a curated topic model for COVID-19 medical research literature, https://openreview.net/forum?id=6Huoz_DkT2 (2020, accessed 11 January 2022).
14. Dastani M and Danesh F. Iranian COVID-19 publications in LitCovid: text mining and topic modeling. *Sci Program* 2021; 2021: e3315695.
15. 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.

16. Ryan JM. *COVID-19: two volume set*. London: Routledge, 2021.
17. Davidson H. First COVID-19 case happened in November, China government records show - report. *The Guardian*, 13 March 2020, https://www.theguardian.com/world/2020/mar/13/first-covid-19-case-happened-in-november-china-government-records-show-report (2020, accessed 5 August 2022).
18. Listings of WHO's response to COVID-19, https://www.who.int/news/item/29-06-2020-covidtimeline (2020, accessed 5 August 2022).
19. CNN Editorial Research. COVID-19 pandemic timeline fast facts. *CNN*, https://www.cnn.com/2021/08/09/health/covid-19-pandemic-timeline-fast-facts/index.html (2021, accessed 5 August 2022).
20. Adams-Prassl A, Boneva T, Golin M, et al. The impact of the coronavirus lockdown on mental health: evidence from the United States. *Econ Policy* 2022; 37: 139–155.
21. Pancani L, Marinucci M, Aureli N, et al. Forced social isolation and mental health: a study on 1,006 Italians under COVID-19 lockdown. *Front Psychol* 2021; 12.
22. García-Prado A, González P and Rebollo-Sanz YF. Lockdown strictness and mental health effects among older populations in Europe. *Econ Hum Biol* 2022; 45: 101116.
23. atzatzev. COVID-19 and school closures: one year of education disruption. *UNICEF DATA*, https://data.unicef.org/resources/one-year-of-covid-19-and-school-closures/ (2021, accessed 5 August 2022).
24. Bauer S, Contreras S, Dehning J, et al. Relaxing restrictions at the pace of vaccination increases freedom and guards against further COVID-19 waves. *PLoS Comput Biol* 2021; 17: e1009288.
25. Peyrin-Biroulet L. Will the quality of research remain the same during the COVID-19 pandemic? *Clin Gastroenterol Hepatol* 2020; 18: 2142.
26. Alexander PE, Debono VB, Mammen MJ, et al. COVID-19 coronavirus research has overall low methodological quality thus far: case in point for chloroquine/hydroxychloroquine. *J Clin Epidemiol* 2020; 123: 120–126.
27. Roberts ME, Stewart BM and Tingley D. Stm: an R package for structural topic models. *J Stat Softw* 2019; 91: 1–40.
28. Fantini D. easyPubMed: search and retrieve scientific publication records from PubMed, https://CRAN.R-project.org/package=easyPubMed (2019, accessed 11 January 2022).
29. Muschelli J. rscopus: Scopus Database 'API' Interface, https://CRAN.R-project.org/package=rscopus (2019, accessed 11 January 2022).
30. Belter C. scopusAPI, https://github.com/christopherBelter/scopusAPI (2022, accessed 11 January 2022).
31. Baker C. wosr: clients to the 'Web of Science' and 'InCites' APIs, https://CRAN.R-project.org/package=wosr (2018, accessed 11 January 2022).
32. Clark JM, Sanders S, Carter M, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. *J Med Libr Assoc* 2020; 108: 195–207.
33. Ooms J. Google Inc. cld3: Google's Compact Language Detector 3, https://CRAN.R-project.org/package=cld3 (2021, accessed 11 January 2022).
34. Ooms J. Dirk Sites. cld2: Google's Compact Language Detector 2, https://CRAN.R-project.org/package=cld2 (2020, accessed 11 January 2022).
35. Bischof JM and Airoldi EM. Summarizing topical content with word frequency and exclusivity. Edinburgh, Scotland, UK, 2012.
36. R Core Team. R: a language and environment for statistical computing. 2021.
37. Carvalho T, Krammer F and Iwasaki A. The first 12 months of COVID-19: a timeline of immunological insights. *Nat Rev Immunol* 2021; 21: 245–256.
38. Pieh C, Budimir S, Delgadillo J, et al. Mental health during COVID-19 lockdown in the United Kingdom. *Psychosom Med* 2021; 83: 328–337.
39. Benke C, Autenrieth LK, Asselmann E, et al. Lockdown, quarantine measures, and social distancing: associations with depression, anxiety and distress at the beginning of the COVID-19 pandemic among adults from Germany. *Psychiatry Res* 2020; 293: 113462.
40. Rahimi I, Chen F and Gandomi AH. A review on COVID-19 forecasting models. *Neural Comput Applic* 2021: 111. DOI: 10.1007/s00521-020-05626-8.
41. Haleem A, Javaid M, Singh RP, et al. Telemedicine for healthcare: capabilities, features, barriers, and applications. *Sens Int* 2021; 2: 100117.
42. Coronavirus: how the pandemic has changed the world economy. *BBC News*, 24 January 2021, https://www.bbc.com/news/business-51706225 (2021, accessed 4 August 2022).
43. First COVID vaccine is administered in the US. *BBC News*, https://www.bbc.com/news/av/world-us-canada-55307642 (2020, accessed 5 August 2022).
44. COVID-19 vaccination campaigns launched across Europe, https://www.aa.com.tr/en/europe/covid-19-vaccination-campaigns-launched-across-europe/2090420 (2020, accessed 5 August 2022).
45. Metcalfe LS and Perez I. Blinded by the unknown: a school's leader's actions to support teachers and students during COVID-19 school closures. *J Sch Adm Res Dev* 2020; 5: 49–54.
46. Telenti A, Hodcroft EB and Robertson DL. The evolution and biology of SARS-CoV-2 variants. *Cold Spring Harb Perspect Med* 2022; 12: a041390.
47. Blei DM, Ng A and Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
48. Sonmez E and Codal KS. Determination of research trends in COVID-19 literature using topic model approach. In: 14th IADIS international conference information systems 2021, 2021, pp. 161–169.
49. Blei DM and Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning, pp.113–120. New York, NY, USA: Association for Computing Machinery.
50. Sepúlveda-Vildósola AC, MejÍa-Arangué JM, Barrera-Cruz C, et al. Scientific publications during the COVID-19 pandemic. *Arch Med Res* 2020; 51: 349–354.
51. Servick K and Enserink M. The pandemic's first major research scandal erupts. *Science* 2020; 368: 1041–1042.
52. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *Br Med J*; 369: m132. Epub ahead of print 7 April 2020. DOI: 10.1136/bmj.m1328.
53. Symeonidis S, Effrosynidis D and Arampatzis A. A comparative evaluation of pre-processing techniques and their

interactions for twitter sentiment analysis. *Expert Syst Appl* 2018; 110: 298–310.

54. Vo D-T and Khoury R. Language identification on massive datasets of short messages using an attention mechanism CNN. In: 2020 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). Epub ahead of print 2020. DOI: 10.1109/ASONAM49781.2020.9381393.
55. Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Vol. 1, pp. 4171–4186, Minneapolis, MN.
56. Dong H, Wang M, Zhang H, et al. KnowLab at BioCreative VII Track 5 LitCovid: ensemble of deep learning models from diverse sources for COVID-19 literature classification. *KnowLab at BioCreative VII Track 5 LitCovid: Ensemble of deep learning models from diverse sources for COVID-19 literature classification* 2021, pp.310–313.
57. Labrak Y and Dufour R. Team LIA/LS2N at BioCreative VII LitCovid track: multi-label document classification for COVID-19 literature using keyword based enhancement and few-shot learning. In: *BioCreative VII challenge evaluation workshop*. Virtual Conference, United States, https://hal.archives-ouvertes.fr/hal-03426326 (2021, accessed 11 January 2022).