Scientometrics (2022) 127:1609–1642
https://doi.org/10.1007/s11192-021-04260-y

Discovering temporal scientometric knowledge in COVID-19 scholarly production

Breno Santana Santos[1,2] · **Ivanovitch Silva**[1] · **Luciana Lima**[3] · **Patricia Takako Endo**[4] · **Gisliany Alves**[1] · **Marcel da Câmara Ribeiro-Dantas**[5]

Received: 7 October 2021 / Accepted: 22 December 2021 / Published online: 16 January 2022

Abstract
The mapping and analysis of scientific knowledge makes it possible to identify the dynamics and/or growth of a particular field of research or to support strategic decisions related to different research entities, based on bibliometric and/or scientometric indicators. However, with the exponential growth of scientific production, a systematic and data-oriented approach to the analysis of this large set of productions becomes increasingly essential. Thus, in this work, a data-oriented methodology was proposed, combining Data Analysis, Machine Learning and Complex Network Analysis techniques, and Data Version Control (DVC) tool, for the extraction of implicit knowledge in scientific production bases. In addition, the approach was validated through a case study in a COVID-19 manuscripts dataset, which had 199,895 articles published on arXiv, bioRxiv, medRxiv, PubMed and Scopus databases. The results suggest the feasibility of the proposed methodology, indicating the most active countries and the most explored themes in each period of the pandemic. Therefore, this study has the potential to instrument and expand strategic decisions by the scientific community, aiming at extracting knowledge that supports the fight against the COVID-19 pandemic.

Keywords Scientometrics · Bibliometrics · COVID-19 · Pandemic · Data Science

Mathematics Subject Classification 58-00 · 58-06 · 68T09 · 68T10 · 68T99 · 94-11 · 94A16

Introduction

COVID-19 is an acute respiratory infection caused by SARS-CoV-2 coronavirus, potentially severe, highly transmissible and of global distribution (Mohamadian et al. 2021; Taleghani and Taghipour 2020). The coronaviruses belong to *Coronaviridae* family, *Nidovirales* order, and they have been known since the decade of 1930 and infect a high variety of species. However, they were only identified causing disease in humans in the

✉ Breno Santana Santos
breno.santos.038@ufrn.edu.br; breno1005@hotmail.com

Extended author information available on the last page of the article

decade of 1960 (Mohamadian et al. 2021; Taleghani and Taghipour 2020). They are enveloped viruses composed of a simple chain of ribonucleic acid (RNA) and commonly divided into 4 genres, according to their common features: α, β, γ and δ (Alsharif and Qurashi 2020; Mohamadian et al. 2021). The viruses belonging to the gender α (alphacoronovirus) generally cause common colds, while those named SARS-Cov (Severe Acute Respiratory Syndrome—epidemic in 2002, in China), MERS-Cov (Middle East Respiratory Syndrome—epidemic in Saudi Arabia, in 2012) and SARS-Cov-2 (Coronavirus disease 2019—COVID-19) belong to the gender β, so called betacoronaviruses (Alsharif and Qurashi 2020; Mohamadian et al. 2021; Taleghani and Taghipour 2020; Tiwari et al. 2021).

This virus has harmed the global economy and exhausted the health care system to a degree that has not been seen since the 1918 influenza outbreak (Tornberg et al. 2021). Since its origin in Wuhan, China in December 2019, it has caused more fatalities than SARS-Cov and MERS-Cov together (Mahase 2020). According to the World Health Organization (WHO) COVID-19 Dashboard (World Health Organization 2021), as of October 1, 2021, there had been more than 233 million confirmed cases and around 4,7 million reported deaths across the world. In any case, with the advancement of the process of immunization of the world population at the end of 2020, there is a trend of decline in cases and deaths by COVID-19 in practically all regions of the globe. Although challenges such as inequality in access to vaccines and the resurgence of variants are present in the current stage of the pandemic (Stein 2021).

In addition to being an unprecedented high mortality event since the Spanish Flu, the COVID-19 pandemic has the particularity of being the first in which information technologies have taken a leading role in the dissemination of both useful health information and fake news calls. Shortly after the announcement of the COVID-19 outbreak as a pandemic, the UN Secretary-General launched the United Nations Communications Response Initiative to combat the spread of fake news related to the SARS-CoV-2, in April 2020. The called infodemic is characterized by an overabundance of information, both online and offline (World Health Organization 2020), including misinformation that can cost lives as they contribute to poor adherence to health authorities' guidelines in central aspects of the pandemic such as social distancing, disease management and more recently, adherence to vaccination (Moscadelli et al. 2020; Gu 2021).

In this context, the science has played an important role in the dissemination of reliable and useful information about the various dimensions of the pandemic, with a rapid dissemination of articles and preprints throughout the pandemic (Nowakowska et al. 2020). Since the beginning of COVID-19 pandemic, many research communities and agencies dedicated their efforts to fight against this infectious disease with their own resources and means. One of these efforts was the sharing of coronavirus-related scientific research as openly as possible (Colavizza et al. 2021; Grammes et al. 2020; Tornberg et al. 2021).

To date, a huge number of studies related to COVID-19 have been generated, aiming to contribute to the fight against the pandemic (Colavizza et al. 2021; Haghani and Bliemer 2020; Maalouf et al. 2021; Şenel and Topal 2020). However, this fact negatively affects several researchers in monitoring recent contributions in their respective areas of expertise. Thus, to get around this problem, the researchers can use the tools available in Bibliometrics and Scientometrics, which have quantitative methods applicable to scientific literature (Haghani and Bliemer 2020; Şenel and Topal 2020).

Bibliometrics can be understood as the study of quantitative aspects of the production, dissemination and use of registered information, as well as this area allows the development of patterns and mathematical models to measure such processes, which can support

decision-making and prepare forecasts based on their results (Sugimoto and Larivière 2018; Vinkler 2010). On the other hand, Scientometrics, considered the metric of science, is defined as the study of the quantitative aspects of science, either as a discipline or as an economic activity. Furthermore, it uses quantitative indicators to measure a specific discipline of science (Şenel and Topal 2020; Sugimoto and Larivière 2018; Vinkler 2010).

So, aiming to extract relevant insights from the large set of publications and to support decision making to fight against the current pandemic, several scientometric and bibliometric studies have been published for this purpose (Casado-Aranda et al. 2021; Colavizza et al. 2021; Farooq et al. 2021; Gul et al. 2020; Haghani et al. 2020; Malik et al. 2021; Okhovati and Arshadi 2021; Tornberg et al. 2021; Zhang et al. 2020). In other words, the recent scientific contributions confirm that the academic response to COVID-19 is both massive and multifaceted, i.e., these studies explore several scenarios and fields of research.

However, there are few bibliometric analyses exploring the temporal aspect of the pandemic, based on keywords or other indicators in COVID-19 scientific publications (Cai et al. 2021; Ebadi et al. 2021; Haghani and Varamini 2021; Lauper et al. 2021), using a combination of indexed and preprints databases, using a process that is totally auditable and reproducible. This is relevant due to the differences observed in the pandemic during the past two years, such as waves in number of new cases, new COVID-19 variants, new public policies and interventions, new proposed treatments and so on.

Therefore, aiming to fill this research gap, this work proposes a fully auditable and reproducible methodology of temporal scientometric analysis of COVID-19 scientific publications. Also, we perform an experimental evaluation of the proposed approach from a case study, using a dataset of manuscripts related to COVID-19 (Santos et al. 2020), available on PubMed and Scopus databases, in addition to the arXiv, bioRxiv and medRxiv preprint servers. The results demonstrate the feasibility of the proposal indicating the main countries that produce COVID-19 related works, the main research focus and other interesting insights. The present methodology has the potential to instrument and expand strategic decisions of the scientific community, aiming at knowledge extraction that supports the fight against the COVID-19 pandemic. Therefore, the main contributions of this work are:

- Methodology of scientific production analysis related to COVID-19 publications;
- A temporal window overview of COVID-19 scientific contribution;
- Experimental process for the evaluation of proposed approach.

The rest of this paper is organized as follows. "Related works" section describes related works. "Scientometric approach" section explains the proposed methodology to extract scientometric knowledge from COVID-19 related manuscripts. In "Materials and methods" section, we detail the experimental evaluation of our approach. "Results and discussion" section discusses the obtained results. "Threats to validity" section details the threats to the validity of our study. Finally, in "Conclusion and future works" section, we present the conclusions and discussions of future works.

Related works

There are many studies that have performed a scientometric/bibliometric evaluation of literature related to the 2019 novel coronavirus (COVID-19), with each one focusing in specific aspects, subject areas or contexts, but all of them have contributed significantly to

combat this terrible pandemic and to expand the scientific knowledge. Among these works, the following studies stand out: Cai et al. (2021), Casado-Aranda et al. (2021), Colavizza et al. (2021), Ebadi et al. (2021), Fassin (2021), Grammes et al. (2020), Lauper et al. (2021), Malik et al. (2021), Rodríguez-Rodríguez et al. (2021) and Tornberg et al. (2021).

In Cai et al. (2021), the authors performed a scientometric analysis focused on exploring the trends in coronavirus research during the pandemic. As data source, they used the Scopus, PubMed, Web of Science (WoS) and some preprint servers databases. This study was a continuation of their previous study, about the behavior of COVID-19 production in terms of countries' collaborations. According to the authors, they identified that the most affected nations tended to produce the greatest number of coronavirus articles, with output closely coupled to the rate of infection. Moreover, the USA remained the single largest contributor to the global publication output. Contrary to China's dominance in the initial months of the pandemic, China's contribution fell as the national COVID-19 caseload dropped. Despite having used a slightly distinct dataset and other types of analysis, some results are in line with those obtained by this study, for example, countries' contributions, specially the essential role of China, USA, Italy and United Kingdom in the academic production related to SARS-CoV-2.

Next, in Casado-Aranda et al. (2021), a descriptive and visual quantification of scientific research on the virus and its effect on the environment was performed, in order to offer a first straightforward report on the evolution of publications combining the effect of COVID-19 on the environment since the outset of the pandemic, as well as to identify the main lines of research that were surging as a result of the crisis. All articles related to the COVID-19 pandemic were retrieved from the Scopus and Web of Science databases and analyzed using the SciMAT software. According to the authors, they identified that several environmental studies correlated to a sharp decrease of air pollutants (e.g., NO_2 and CO_2) and an increase of O_3 during the COVID-19 lockdown, as well as others had explored how the monitoring of environmental settings could serve to prevent and predict such outbreaks and, consequently, improve public health. The authors also highlighted that the potential use of artificial intelligence and smart devices could serve in monitoring the mobilization of citizens in urban and tourism destinations, and thus playing a vital role in preventing an advance of the pandemic. Finally, some results are in line with those obtained by this study, for example, the main topics focused by researchers, and countries' contributions.

In Colavizza et al. (2021), it was performed a scientometric analysis of the COVID-19 Open Research Dataset (CORD-19), which have publications from the Medline, PubMed, and WHO databases, in addition the arXiv, medRxiv and bioRxiv preprint servers. This dataset covers other research topics beyond COVID-19 and coronaviruses, i.e., beyond the core of research directly on COVID-19 and coronaviruses, it contains many articles on related yet distinct streams of virus research, such as on influenza, molecular biology and public health. So, the authors explored the potential of CORD-19 using the main scientometric indicators. Even though they had used a slightly distinct dataset and other types of analysis, some results are in line with those obtained by this study, for example, the main topics focused by researchers.

Next, in Ebadi et al. (2021), the authors used natural language processing and machine learning models and techniques to understand and characterize the overview of COVID-19 research by identifying main themes and their temporal behaviors, within the time-frame of January-May 2020, on PubMed and arXiv databases. The authors found out that, in the beginning of pandemic, the scientific community seemed to focus more on the pandemic aspect of the disease and its acute, imminent danger to public health. Over time, however, the attention had been gradually drawn to longer-term and chronic impacts on the public,

such as mental health. Other interesting finding was that the research community continuously focused on the vulnerable and high-risk populations who were in danger of severe illness from COVID-19, for example, high-risk groups, and pregnant women. Even though they had used other slightly distinct types of analysis, some results are in line with those obtained by this study, for example, the main topics focused by researchers, and countries' contributions.

In Fassin (2021), the author performed a bibliometric analysis, aiming to analyze the disruptive impact of the explosion of COVID-19 production with respect to several bibliometric indexes, for example, the impact factor and h-index. As data source, WoS data was used, and the author compared the COVID-19 production with the data of publications related to the general topic "cancer", available until the year 2020, as well as restricted for the year 2020 alone. According to the author's results, the theme of the research was important for impact and citations for articles in applied sciences. The salience of the topic, the magnitude of the problem at study and the urgency to find solutions were drivers for citations. An exceptional phenomenon (i.e., the "explosion" of research publications related to COVID-19) had a disruptive impact on bibliometric indicators, as well as the higher the specialization, the higher the possible impact of a disruptive phenomenon.

Next, in Grammes et al. (2020), the authors performed a scientometric study that aimed at providing profound insights into the current scientific SARS-CoV-2 research landscape, including correlating the severity of the COVID-19 outbreak with its related scientific output per region/country during the pandemic, as well as assessing international collaboration. All articles related to the COVID-19 pandemic were retrieved from Web of Science and analyzed using the web application SciPE (science performance evaluation). According to their findings, the United States, the United Kingdom, China, and Italy were the leading nations in terms of the number of publications, as well as the countries severely affected by the COVID-19 pandemic (e.g., Italy) and those with a generally high research output (e.g., the United States) contributed significantly to the literature base. Other interesting result was that the main types of publications were articles, editorial materials and letters. Finally, in terms of international cooperation, the United States were most active while China was underrepresented. Note that some results are in line with those obtained by this study, for example, countries' contributions.

In Lauper et al. (2021), it was performed a monthly-temporal scientometric analysis, as this study, of manuscripts published in the Annals of Rheumatic Diseases journal—belongs to the British Medical Journal (BMJ)—, which were related to COVID-19, in order to analyze and map the area of Rheumatology during the first months of the pandemic. According to the authors' findings, most of the publications on COVID-19 were letters, correspondences and correspondence responses. In line with the beginning of the pandemic, initially, there was a growing trend of publications with respect to the use of antimalarial drugs as a potential preventive therapy or treatment, and, after many studies performed, this type of treatment fell into disuse. Other point analyzed by the authors was the use of disease-modifying antirheumatic drugs and glucocorticoids as a mean of therapy. Finally, the authors discussed about the potential use of telemedicine as an invaluable tool for most patients and health professionals, except for older patients or those with higher disease activity, that demonstrated dissatisfaction about this type of approach, beyond the difficulties in its implementation in developing countries due to limited internet access.

Next, in Malik et al. (2021), the authors performed a scientometric and temporal evaluation of scholarly production related to all coronaviruses variants, extracted from the WoS, using the R-Bibliometrix package, to explore a wide range of indicators. Generally, an increasing trend in terms of numbers of publications was observed over the years, led by

the USA, China, United Kingdom, European countries, and a few other developed countries and the majority published in the last 2 decades. In addition, articles (53.4%) were the most common publication type. Journal of Virology, BMJ, and Virology were leading sources while BMJ and Lancet showed increased contributions recently. Finally, the top 20 countries contributed to >89% of the total number of documents, that, from the point of view of authors, suggested a lack of global efforts. It is important to highlight that some results of Malik et al. (2021) match the results of this work.

Similarly to this study, Rodríguez-Rodríguez et al. (2021) performed a monthly-temporal scientometric analysis of manuscripts related to SARS-CoV-2 and the use of emerging technologies (Artificial Intelligence, Machine Learning, Internet of Things, Blockchain, and others), available on WoS and Scopus databases. They analyzed how these emerging technologies are helping the task force to fight against to the pandemic. According to their findings, the United States, China and the United Kingdom were the leading nations in terms of the number of publications. The authors also extracted topics of research interest and observed that the most important current lines of research focused on patient-based solutions (e.g., diagnosis, drug discovery, patient, and others). They also identified the most relevant journals (e.g., Lancet, Nature, Plos One, and Cell), and the most influential authors by an analysis of citations and co-citations. Even though they had used a slightly distinct dataset and other types of analysis, some results are in line with those obtained by this study, for example, the main topics focused by researchers, and countries' contribution.

Finally, in Tornberg et al. (2021), the traditional bibliometrics (citation count and impact factors) and new bibliometrics (Altmetric and PlumX scores) of the top 100 COVID-19 articles with the highest Altmetric scores were characterized and compared. Its results demonstrated that citation count weakly correlated with Altmetric score and strongly correlated with PlumX score, with regard to articles analyzed. Other finding was that the current literature had noted that journals with a high Twitter presence also had high Altmetric and PlumX scores, as well as it was observed the strong correlations between citation count and Mendeley citations, and between citation count and Dimensions citations. Again, some of their results are in line with what we present in the current study, for example, that the most relevant type of production was journal article.

Finally, in Table 1, a comparative summary among the characteristics of the works aforementioned in this section are presented. Thus, it is possible to visualize their main similarities and differences, where each row represents one particular study and each of the five columns indicates topics (or features) and how they were covered in a given research. The features analyzed are: Databases, Country, Keyword, Temporal and DVC.[1]

All studies presented in Table 1 have distinct characteristics, and neither of them contemplate all predetermined topics, except this proposed work. In fact, as stated previously, it can be seen in the literature (through the Databases feature) that there is a predominance of indexed databases, and preprint servers are still little explored, unlike the proposed study. Likewise, the Keyword feature also shows that most works focus on extraction of thematic from the keywords and/or words contained in the title and abstract's publications. However, for the Country feature, few recent studies have performed this type of analysis, possibly, for being a widely used analysis. With relation to the Temporal feature, it is possible to note that few studies explore the temporal aspect of their analyses, and when they use this type, they basically use the annual granularity, unlike this work that uses a monthly

[1] Data Version Control.

Table 1 Summary of related studies

Study	Features				
	Databases	Country	Keyword	Temporal	DVC
Cai et al. (2021)	Scopus, PubMed, WoS and preprints	Yes	No	Quarterly	No
Casado-Aranda et al. (2021)	WoS and Scopus	Yes	Yes	No	No
Colavizza et al. (2021)	CORD-19	No	Yes	Yearly	No
Ebadi et al. (2021)	PubMed and arXiv	Yes	Yes	Monthly	No
Fassin (2021)	WoS	No	No	Yearly	No
Grammes et al. (2020)	WoS	Yes	Yes	No	No
Lauper et al. (2021)	Annals of Rheumatic Diseases journal	No	Yes	Monthly	No
Malik et al. (2021)	WoS	Yes	Yes	Yearly	No
Rodríguez-Rodríguez et al. (2021)	WoS and Scopus	Yes	Yes	Monthly	No
Tornberg et al. (2021)	Scopus and Altmetric Explorer	No	No	No	No
Proposed work	Scopus, PubMed and preprints	Yes	Yes	Monthly	Yes

granularity. Finally, for the last feature used for comparison, the use of Data Version Control (DVC), aiming to automate the analysis process and making it auditable and reproducible, has not been used by any other work for this purpose.

In this way, the relevance and the relationship of each one of the presented works and the work proposed here are highlighted. Thus, from the discussion presented previously, it is clear that there are still gaps to be explored in this area, where knowledge is still needed in order to direct efforts to fight the pandemic.

In the next section, the proposed methodology developed in this study will be shown.

Scientometric approach

The proposed method to extract scientometric knowledge from COVID-19 manuscripts is presented in Fig. 1. Firstly, it is necessary to collect the scientific production data related to the new coronavirus. In a previous work (Santos et al. 2020), we generated a dataset with manuscripts related to COVID-19 available in arXiv, bioRxiv, medRxiv, PubMed, and Scopus databases. These data can provide an overview of scientific research activities, making it possible to identify countries, scientists and research groups that are most active in this task force to combat the pandemic. Therefore, for this current work, this dataset was updated and now it has works published from December 2019 to 15 July 2021 (date of retrieval).

In order to analyze the relevance of these studies and their sources, from the impact factor to other metrics, the data related to the 2020 InCites Journal Citation Reports (JCR) were collected and used by convenience, which are available in Web of Science (2021). Both datasets were preprocessed and merged, using the Pandas Data Wrangling tool, generating the final dataset used in this work. The pipeline to preprocess and generate the dataset is versioned with the DVC tool, making the process easy to reproduce and audit (Santos et al. 2020; Iterative 2021). The Data Version Control (DVC) tool provides data science

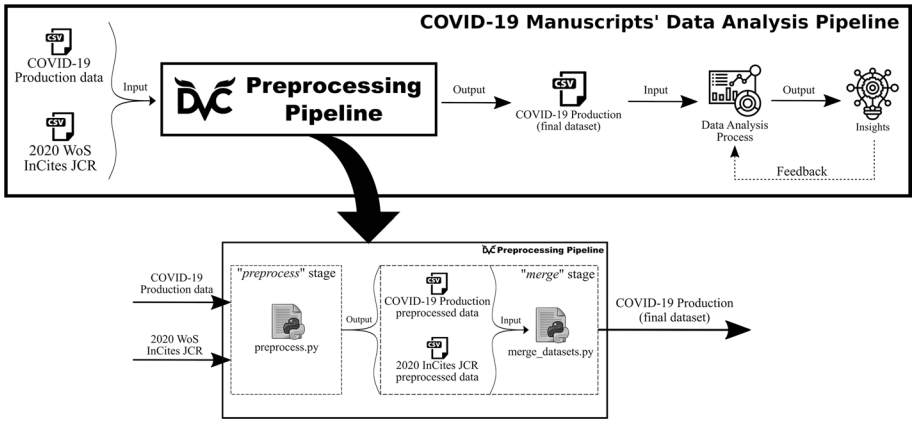

Fig. 1 Scientometric data analysis pipeline for COVID-19 manuscripts

workflow reproducibility and consistency, and it is Git-compatible, offering lock-free, local branching, and versioning. Furthermore, DVC is used to version data and data pipelines, following the same rationale used to version source code (Iterative 2021).

Finally, from the final dataset, it is possible to analyze the scholarly production related to COVID-19, making it possible to investigate researchers' efforts made during the pandemic until the retrieval date. Furthermore, based on techniques of data visualization (within the data analysis process), one can analyze and evaluate this extracted knowledge and, whether such patterns and information were inadequate, one can reapply or refine the data analysis process until the insights obtained are satisfactory, since the methodology is interactive and iterative. It is interactive because the researcher participates during the whole process of analysis, and it is iterative because of the steps repeated in the data analysis process until the relevant results are obtained.

With the proposed approach properly presented, its empirical evaluation will be detailed in the next section.

Materials and methods

This section describes a case study of our approach, which was based on an experimental process such as that presented by Santos et al. (2015) and Wohlin et al. (2012). The next subsections focus on the definition and planning of this empirical evaluation. The last subsection presents its operation process.

Goal definition

The main goal of this study is to analyze the scholarly production related to COVID-19 during the pandemic. This goal is formalized using the GQM (Goal-Question-Metric) template proposed by Basili and Weiss (1984) and presented by van Solingen and Berghout (1999): **Analyze** a scientific production dataset related to COVID-19 **with the purpose of** characterization and understanding it **with respect to** scientometric indexes and behavior of scholarly production **from the point of view of** researchers, affiliations and research

groups that work with studies associated to the pandemic **in the context of** preprints and published peer-reviewed manuscripts.

Planning

This subsection details the design of the case study.

Research questions

The research questions we want to explore in this work are:

- Based on the 2020 impact factor of WoS InCites, do the most peer-reviewed published manuscripts have the highest level of quality?
- For the most peer-reviewed published manuscripts, what is the minimum impact factor considered by their authors?
- Are the journals with the highest impact factor those that are the most cited ones?
- Are there journals without impact factor that have high quality, based on their number of citation?
- Which countries are most relevant in terms of scientometric indicators?
- Were the countries with the highest number of articles the ones with the most citations?
- Were the countries most affected by the pandemic the most productive, based on the number of manuscripts?
- Can the citation mean be considered as an indicator of the quality of a country's productions?
- What is the influence of impact factors of the journals in the amount of publications grouped by countries?
- What were the most explored research focuses during the pandemic? And which are the most explored by period?
- What is the minimum number of occurrences a keyword must have to also be among the most cited ones?
- Is there a difference among the keywords regarding the number of manuscripts and total of citations during the analyzed period?

The metrics to evaluate these questions are: (1) number of manuscripts by period, i.e., per month and year; (2) total number of manuscripts; (3) total citations by period; (4) cumulative total citations; (5) citation mean; and (6) the 2020 impact factor.

Participant and artifact selection

For convenience, the manuscript dataset produced by Santos et al. (2020) was chosen because the PubMed and Scopus databases have several studies in Health, specially the former, and other fields. Moreover, the arXiv, bioRxiv and medRxiv preprint servers were chosen for the same reason, i.e., they host works in progress in Biological and Health Sciences (bioRxiv and medRxiv), as well as Exact Sciences and Engineering (arXiv). In addition to this dataset, the 2020 InCites JCR data were selected, for convenience too, to analyze the relevance of the manuscripts.

Instrumentation

The materials and/or resources used in this work are:

- Python Data Science ecosystem (pandas,[2] NumPy,[3] Matplotlib,[4] seaborn,[5] scikit-learn[6] and others), provided by Anaconda platform[7] or Google Colab[8];
- Google Colab's Jupyter Lab;
- NetworkX[9] e Gephi,[10] library and tool, respectively, for modeling, visualization, analysis e manipulation of complex networks;
- Data Version Control;
- The manuscripts dataset related to COVID-19 and the proposed methodology, both discussed in the "Scientometric approach" section;
- The 2019 Revision of World Population Prospects dataset, which is made available by the United Nations (UN) from its open data service[11];
- The Jupyter Notebooks that contain all source code to perform the data analysis, which are available at GitHub repository.[12]

Operation

This subsection describes the preparation and execution of this empirical evaluation. The operation process was done initially with the configuration of the environment for the case study and planning of data collection.

Firstly, the scholarly dataset produced by Santos et al. (2020) was updated using the same pipeline and artifacts (Jupyter Notebooks and Python scripts) used by the authors to collect and preprocess the data. Next, the analysis pipeline was defined, as detailed in "Scientometric approach" section, and the updated data were merged with the 2020 InCites JCR dataset.

With the dataset properly preprocessed, the analysis process earlier discussed was performed (see "Scientometric approach" section) with the required artifacts (see "Planning" section).

Finally, after the execution, the analyses results were obtained, which were based on the metrics previously discussed (see "Planning" section). It is worth mentioning that these results are used to answer the research questions of this work.

[2] https://pandas.pydata.org.
[3] https://numpy.org.
[4] https://matplotlib.org.
[5] https://seaborn.pydata.org.
[6] https://scikit-learn.org.
[7] https://www.anaconda.com.
[8] https://colab.research.google.com.
[9] https://networkx.org.
[10] https://gephi.org.
[11] https://data.un.org.
[12] https://github.com/breno-madruga/analysis-covid-manuscripts.

Table 2 Number of manuscripts by database

Database	No. of manuscripts	Percentage (%)
arXiv	3499	1.75
bioRxiv	3222	1.61
medRxiv	12,402	6.20
PubMed	53,008	26.52
Scopus	127,764	63.92

The duplicated manuscripts among the databases were removed during the preprocessing step

Results and discussion

This section presents the results of the empirical evaluation that answers the research questions of this study. It is divided into three parts: (i) general production analysis; (ii) the main contribution of countries; and (iii) the thematic analysis based on the keywords.

General production analysis

We performed a general analysis of COVID-19 manuscripts with respect to the scientometric indexes aiming to characterize and understand this dataset, as well as investigate the relevance of journal publications according to the JCR impact factor for 2020. So, firstly, a distribution of manuscripts per database used was generated.

Table 2 presents the importance of indexed databases, PubMed and Scopus, to spread and disclose the scientific information related to COVID-19, corresponding to almost 91% of the documents contained in this dataset. The use of this kind of database is preferable since the publications were peer-reviewed and indexed by highly rigorous and responsible vehicles regarding the dissemination of quality information (Casado-Aranda et al. 2021; Grammes et al. 2020; Haghani and Varamini 2021).

Other important fact is the growing use of preprints to deliver in advance the scientific knowledge to support the force task to combat the COVID-19 pandemic, the same pattern identified by Callaway (2020) and Colavizza et al. (2021). Thus, for the COVID-19 pandemic, the medRxiv preprint together with arXiv have their importance in disseminating the information related to COVID-19, as well as the technological solutions that support the decision making process of the health agencies and professionals.

Aiming to analyze the contribution of each database during the period (month and year) of the pandemic, the distribution of manuscripts by database and period was also analyzed, as seen in the Fig. 2. Again, it is noticeable the importance of indexed databases, specially PubMed, since it is fully related to Health Sciences, to disclose the studies that associate public health, medicine and correlated fields with the COVID-19 outbreak. The Scopus database also performs an essential role in publishing scientific multidisciplinary studies.

A constant participation of preprints during the pandemic period is noticeable, in addition to a similar representation of each preprint server for each period analyzed, especially between March 2020 and June 2021.

The dataset was characterized by year and type of production (see Table 3). The type *Trade Journal* includes technical articles related to the products and solutions of medical and pharmaceutical industries; *Conference Proceedings* refers to the papers

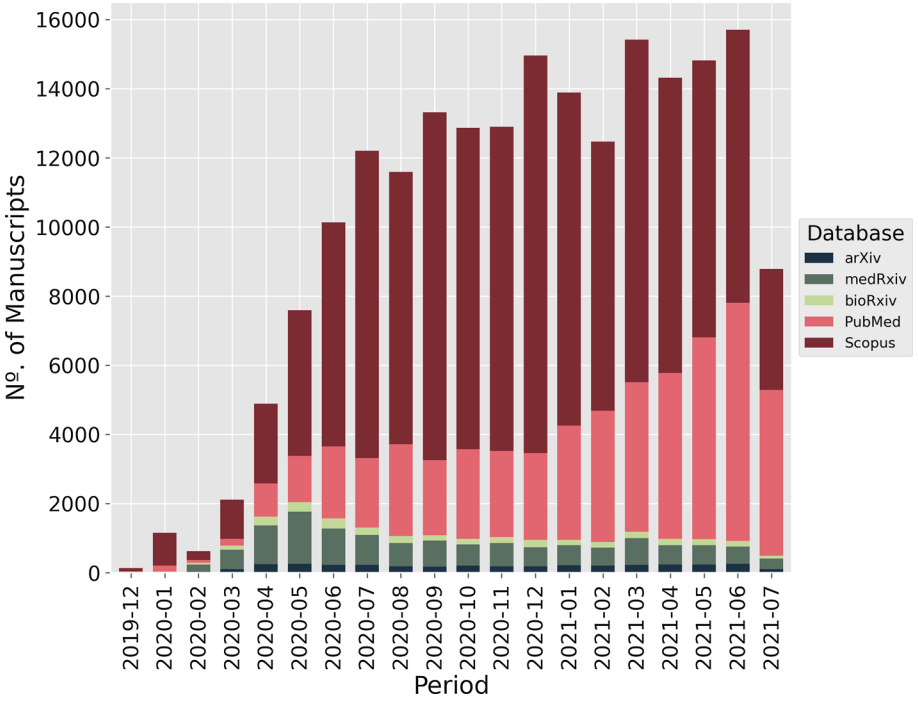

Fig. 2 Distribution of manuscripts by period and database

Table 3 Number of manuscripts by type and year

Year	NM	Source of production				
		Trade journal	Conference proceedings	Preprint	Unknown	Journal
2019	130	0	3	0	42	85
2020	104,346	331	2,636	12,752	19,229	69,235
2021	95,419	71	1,848	6,371	33,737	53,099

NM No. of manuscripts

that were presented and published in conference proceedings, while *Journal* represents the journal articles. This classification of source was based on the same categorization defined by Scopus, except the types *Preprint* and *Unknown*, which correspond to: *Preprint* represents the manuscripts uploaded to arXiv, bioRxiv and medRxiv preprint servers; while *Unknown* corresponds to the articles of the PubMed database, because it has no feature that classifies the means/vehicles of a publication.

In line with some findings in several studies, such as those performed by Grammes et al. (2020) and Malik et al. (2021), as shown in Table 3, the most common source is journals, due to its high rigor and reliability in the process of disseminating scientific information. In 2020, the importance of preprints in the dissemination of knowledge

Table 4 Journal classes

Impact factor (IF)	Journal class
$IF \geq 5$	A
$3 < IF < 5$	B
$1 \leq IF \leq 3$	C
$0 < IF < 1$	D
$IF = 0$	E

Table 5 Number of manuscripts, total of citation and citation mean by journal class and year

Journal class	Year											
	2019				2020				2021			
	NM	TC	μ_{TC}	σ_{TC}	NM	TC	μ_{TC}	σ_{TC}	NM	TC	μ_{TC}	σ_{TC}
A	30	1653	55.10	83.17	24,244	968,537	39.95	234.18	16,816	62,785	3.73	18.16
B	36	477	13.25	16.98	15,545	237,520	15.28	50.53	14,525	26,537	1.83	4.39
C	12	188	15.67	25.53	15,129	139,216	9.20	34.08	12,428	16,887	1.36	4.03
D	0	0	0.00	0.00	1396	2873	2.06	6.86	813	237	0.29	0.72
E	7	36	5.14	7.36	13,252	59,790	4.51	18.60	8588	5398	0.63	2.70

NM no. of manuscripts; *TC* total of citation; μ_{TC} citation mean; σ_{TC} citation standard deviation

related to COVID-19 becomes perceptible, when compared to the quantity of manuscripts indexed in the PubMed database (i.e., the type *Unknown*). Moreover, based on the number of manuscripts published/uploaded between 2020 and the retrieval date (July 15, 2021), the year 2021 will most likely surpass 2020.

In order to analyze the relevance of articles published in journals based on the 2020 impact factor of Web of Science, we performed a categorization of impact factor values, according to the intervals defined by convenience, which are presented in the Table 4.

In order to define the value of *journal class* for a manuscript, the journal must have an ISSN/e-ISSN.[13] Therefore, only the manuscripts that met this criterion were analyzed, where this sample corresponded to the set of articles whose type of production was journals and trade journals, all belonging to the Scopus database, totaling 122,821 peer-reviewed published articles.

As can be seen in the Table 5, from the number of manuscripts by year and journal class, it is noticeable that the researchers have published their studies in journals whose impact factor is greater than or equal to one, highlighting, specially, those ones that are considered quite relevant (classes A and B) to the scientific community. These findings are aligned with those found by Casado-Aranda et al. (2021), Lauper et al. (2021) and Malik et al. (2021). Moreover, these values are extremely relevant in their respective scientific scope, due to their high citation history, as well as by their very rigorous and serious peer-review process. Therefore, these factors are fundamental to allow greater credibility of the published studies.

[13] ISSN stands for "International Standard Serial Number", while e-ISSN corresponds to the electronic version of a ISSN.

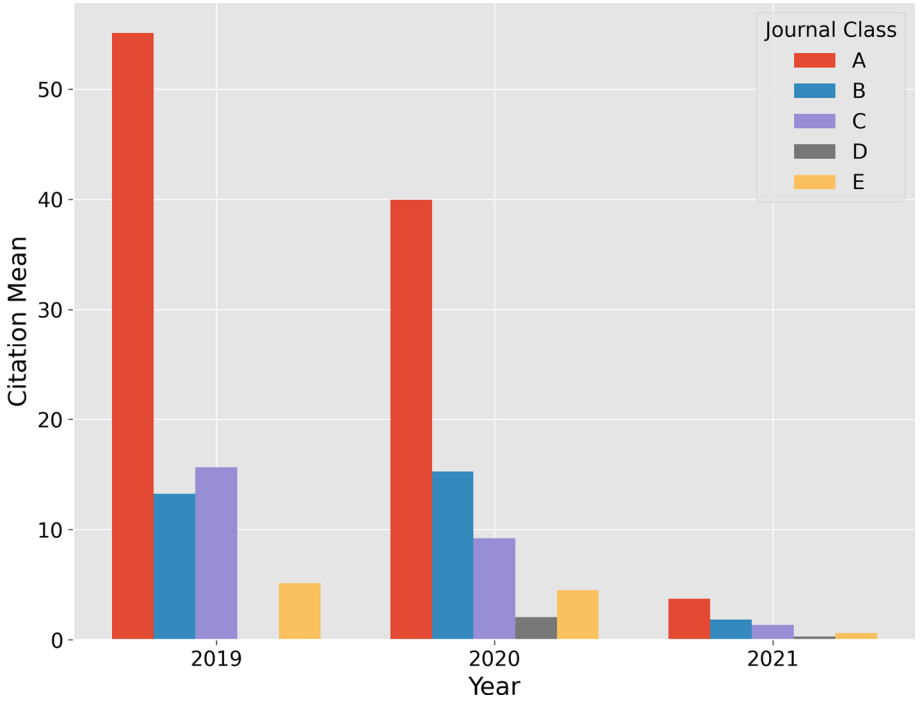

Fig. 3 Citation mean by journal class and year

As a result of the high relevance and impact of these journals, their articles tend to be highly cited, as can be seen, for each year, in the Fig. 3 and in the columns *Total of Citation (TC)* and *Citation Mean* (μ_{TC}) of the Table 5. This behavior is expected since most researchers seek and reference quality and high-value works for a particular field of research.

According to the Table 5 and best illustrated in Fig. 3, other interesting finding is a prevalence of class E journals over class D ones. Due to the unprecedented nature of the COVID-19 phenomenon, this fact leads to two possible situations: (i) the researchers publish their works in journals without impact factor, aiming to guarantee the pioneering in a particular field of study, since these means are possibly less competitive than those of classes A, B or C; or (ii) these journals were created recently to attend a specific scope or emerging area, and are gaining importance and highlight in their scientific community. Therefore, for this fact, further investigation is necessary.

Countries' contribution

In this subsection, we analyzed the countries that contributed the most in terms of scientometric indexes associated with their population rate—an approach similar to that used by Di Girolamo and Reynders (2020)—, in addition to the quality of their publications based on the impact factor in 2020. We extracted the set of countries, based on the affiliations/authors of each article that had this information, so the numbers presented do not represent

Table 6 Top 35 countries

Country	NM	TC	μ_{TC}	σ_{TC}	Rate CP	Rate MP
North Macedonia	47	1379	29.34	88.24	66,025.06	2250.31
China	**20,263**	**566,194**	27.94	253.87	39,135.37	1400.58
Honduras	66	1810	27.42	115.94	18,003.66	656.49
Viet Nam	471	9852	20.92	126.20	10,023.09	479.18
Bolivia	85	1556	18.31	97.17	13,173.27	719.62
Congo (Democ. Rep.)	115	2061	17.92	70.27	2263.25	126.29
Paraguay	56	997	17.80	91.34	13,806.50	775.49
Singapore	2317	38,747	16.72	91.71	656,794.49	39,275.11
Estonia	156	2467	15.81	54.74	185,799.69	11,748.99
Netherlands	3568	56,092	15.72	91.87	326,486.29	20,767.72
Mozambique	86	1311	15.24	58.73	4125.51	270.63
Venezuela	120	1760	14.67	82.28	6186.69	421.82
United Kingdom	**15,886**	**220,319**	13.87	81.93	323,160.67	23,301.35
Germany	7197	**99,472**	13.82	104.08	118,514.76	8574.78
Denmark	1511	20,208	13.37	74.40	347,921.60	26,014.92
Switzerland	3735	49,674	13.30	72.22	570,648.51	42,907.20
United States	**38,418**	**504,055**	13.12	69.68	151,507.43	11,547.57
Russia	1330	17,318	13.02	169.53	11,864.11	911.15
Austria	1773	22,929	12.93	141.82	253,473.67	19,600.02
Libya	66	833	12.62	60.96	11,971.37	948.51
South Korea	1835	22,956	12.51	59.88	44,582.20	3563.70
France	7051	84,879	12.04	70.19	129,745.95	10,778.15
Sweden	2055	22,968	11.18	56.45	226,374.28	20,254.23
Belgium	2567	27,859	10.85	64.78	239,585.34	22,076.01
Australia	6942	74,501	10.73	81.88	289,626.62	26,987.40
Italy	**15,689**	**164,554**	10.49	47.99	272,302.18	25,961.99
Japan	3611	30,559	8.46	56.80	24,202.53	2859.89
Spain	7556	61,578	8.15	48.00	131,498.85	16,135.72
Canada	9471	75,606	7.98	43.16	198,972.06	24,924.80
Brazil	5265	31,507	5.98	35.80	14,695.88	2455.77
Israel	2619	14,543	5.55	30.14	166,245.84	29,938.65
Iran (Islamic Rep.)	4484	23,457	5.23	20.47	27,584.62	5273.03
Turkey	3,638	16,589	4.56	28.68	19,450.23	4,265.47
India	**13,567**	54,279	4.00	22.36	3893.78	973.25
Saudi Arabia	9639	19,582	2.03	21.74	55,509.73	27,323.99

In columns NM and TC, the numbers highlighted in bold represent the top 5 values for a respective metric/column

NM no. of manuscripts; *TC* total of citation; μ_{TC} citation mean; σ_{TC} citation standard deviation; *Rate CP* total of citation per 100,000 inhabitants; *Rate MP* no. of manuscripts per 100,000 inhabitants

the whole dataset, but rather a sample considered quite significant, as around 86.6% of the records had this information.

Table 6 presents the top 35 countries by number of manuscripts (NM), total of citation (TC), mean (μ_{TC}) and standard deviation (σ_{TC}), as well as the population rates

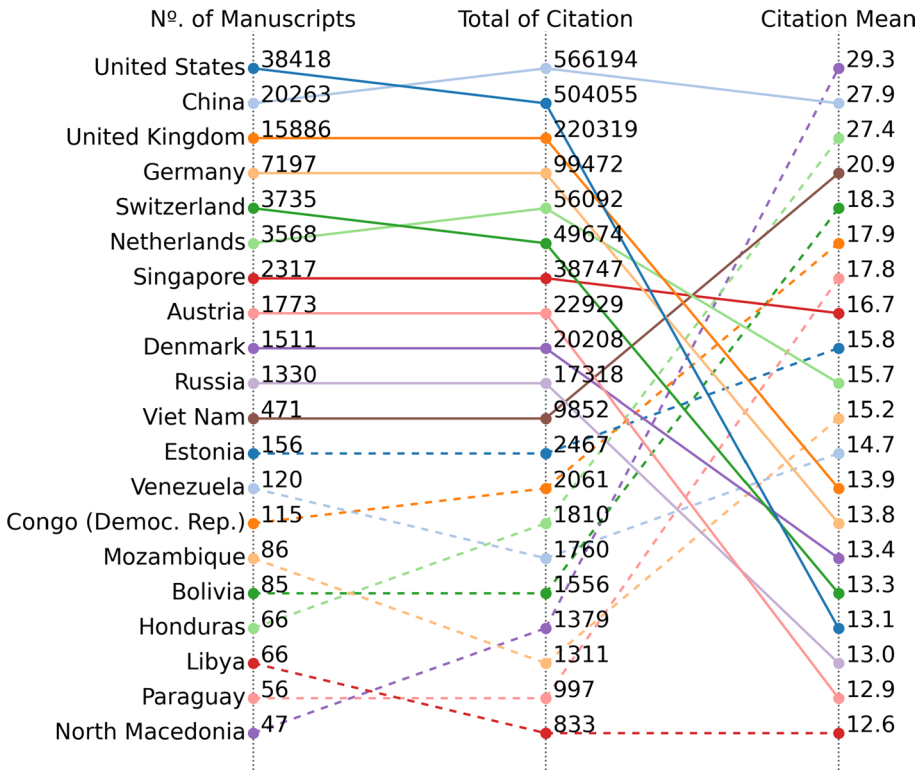

Fig. 4 Top 20 Countries by No. of manuscripts, total of citation and citation mean

related to total of citation (Rate CP) and number of manuscripts (Rate MP). The data is ordered decreasingly by the column *citation mean* (μ_{TC}), and the countries present in this table are the union of the top 20 ones for the *NM*, *TC*, and μ_{TC} metrics.

The countries that contributed the most, in terms of number of manuscript and total number of citations, were those that were heavily affected by the pandemic, such as China, United States, United Kingdom, and Italy. This result is in line with those of the studies carried out by Casado-Aranda et al. (2021), Grammes et al. (2020) and Malik et al. (2021). Another interesting fact is that, despite Brazil being one of the most affected countries by the pandemic, it does not have sufficient resources for conducting research, when compared to the already mentioned countries (Monteiro et al. 2022).

Another interesting fact is that the metric μ_{TC} does not represent the relevance of studies performed by a country, therefore, for emphasizing this, it is showed the standard deviation of the metric *total of citation* (σ_{TC}). For example, North Macedonia has a mean of 29.34, which is higher than China's (27.94). However, China has a total of citations approximately 410 times larger than North Macedonia. So, the studies performed by China are more relevant and more cited in comparison to North Macedonia.

It is important to highlight that the calculation of CP and MP rates was based on the 2019 Revision of World Population Prospects dataset (see "Planning" section). In addition, also for this calculation, it was used the population estimates for the year 2021.

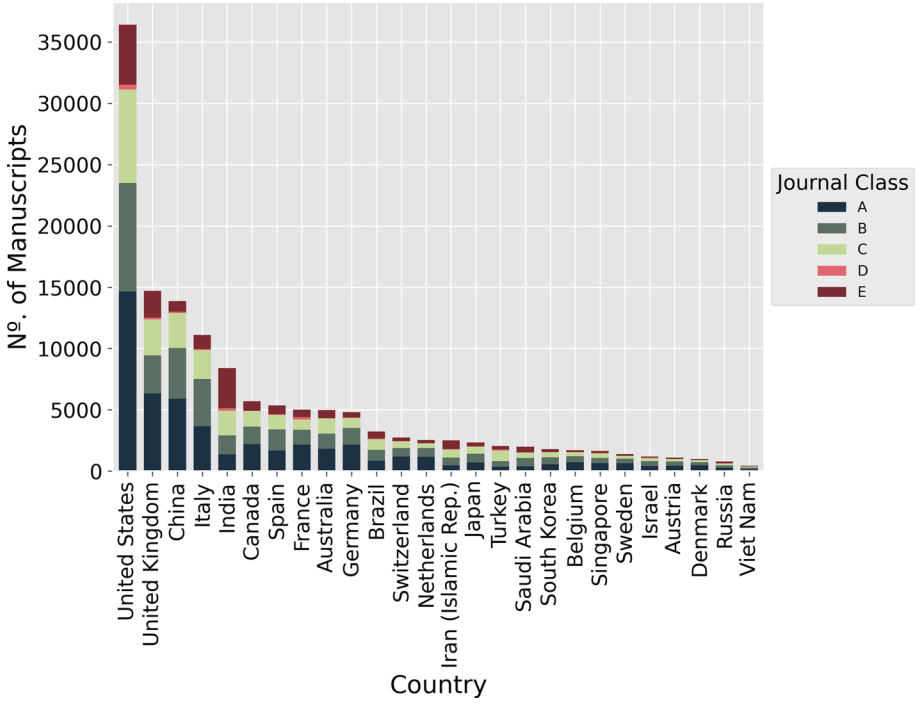

Fig. 5 No. of manuscripts by country and journal class

In order to facilitate the visual comprehension of the data of Table 6, it was generated a slope chart with the top 20 countries by number of manuscripts, total of citation and their means (see Fig. 4).

China is the country that contributed the most in terms of number of studies and citations. This was possibly because it was the pioneer in studies related to the COVID-19. Still, the United States has the highest number of manuscripts, but apparently they have an irregular distribution of citations, based on their citation mean, because it is lower than ones of China, the United Kingdom and Germany. Although Italy does not appear because of the low citation mean, it has studies that are widely cited and considered relevant for works carried out later on for the dissemination of knowledge linked to COVID-19.

We also analyzed the relevance of journal articles belonging to the same countries of Table 6, as seen in the Fig. 5. It is noticeable that the United States is the country with the highest number of relevant productions based on the journal classes previously defined. They have published a lot in journals of classes A, B, and C. The same pattern is found in the United Kingdom, Italy and China.

Brazil has more manuscripts with classes B and C, especially of the latter type. A possible cause for this case could be the excessive cuts in research funding and the lack of incentives for education and research, through the policy of the current government of this country (Monteiro et al. 2022).

It is important to highlight that the countries *Estonia*, *Venezuela*, *Mozambique*, *Libya*, *Bolivia*, *North Macedonia*, *Congo (Democ. Rep.)*, *Honduras*, and *Paraguay* did

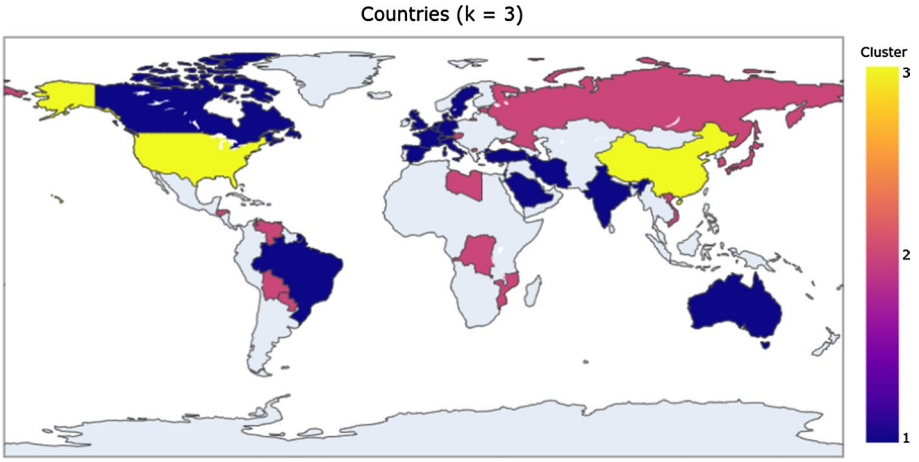

Fig. 6 Top 35 countries grouped by K-means (k = 3)

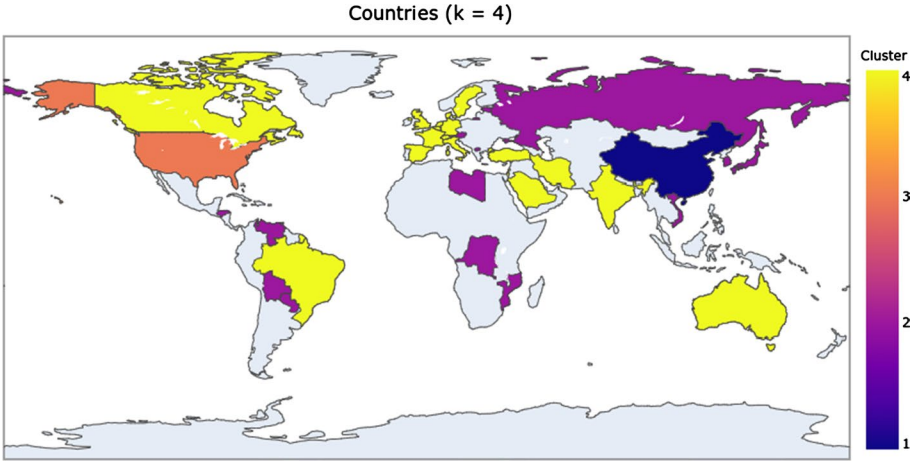

Fig. 7 Top 35 countries grouped by K-means (k = 4)

not appear on the Fig. 5, because their numbers of manuscripts by journal class were not relevant for this chart.

Finally, it was performed a clustering analysis for the same set of countries of Table 6. For the clustering process, apart from the features that appeared in the Table 6, for each country, the number of journal articles for each class previously defined was considered. Firstly, we used the K-means algorithm, varying the parameter k for the values 3, 4, and 5. The results of K-means clustering can be seen in Figs. 6, 7, and 8.

According to the Fig. 6, for K-means with the parameter $k = 3$, the United States and China remained in the same group (Cluster 3). This was probably because of their higher number of manuscripts and total of citation, as well as having high number of journal articles with classes A, B, and C. These countries can be considered as outliers in terms of contribution regarding the COVID-19 pandemic. For the Clusters 1 and 2,

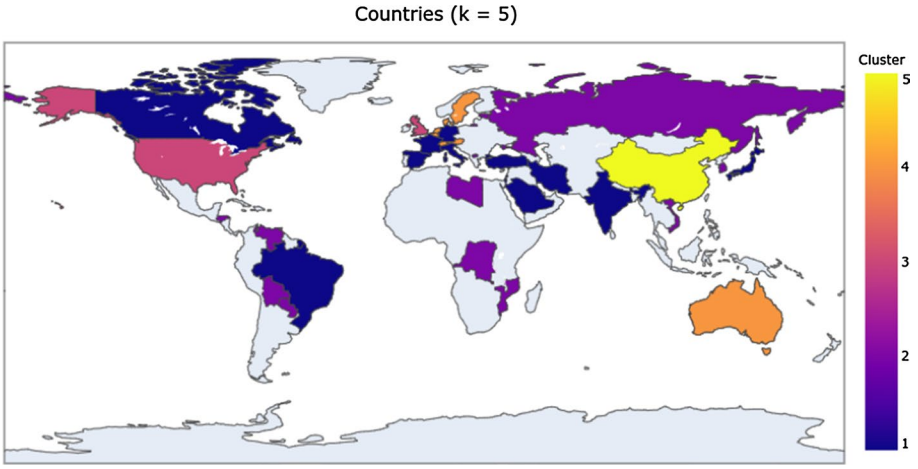

Fig. 8 Top 35 countries grouped by K-means (k = 5)

possibly the distribution of articles for each journal class has influenced the formation of these groups.

After comparing the results of the clustering in the Figs. 6 and 7, it was noticed that the group that had the United States and China (Cluster 1, in Fig. 6) was divided into two new clusters, generating Clusters 1 and 3, as can be seen in the Fig. 7. Maybe, the difference between the number of articles for each journal class influenced the clustering process, especially the number of articles with class A, where this difference is noticeable. Moreover, the other countries remained in their respective groups.

Finally, for K-means with the parameter $k = 5$ (Fig. 8), the result of clustering is totally different to those with $k = 3$ or $k = 4$. Probably, the algorithm must have captured the subtle differences among the top 35 countries, therefore, its results are a little complex to understand just by analyzing the generated clusters.

Main keywords related to COVID-19

In this subsection, we analyze the research focus based on the manuscripts' keywords, also in terms of scientometric indexes. From this analysis, it is also possible to identify some broad thematic areas, because some keywords are thematic areas or very characteristic of their subject areas. Similar to the countries analysis, we extracted the keywords from the feature *auth_keywords* that had this information, so the numbers presented do not represent the whole dataset, but rather the sample that was considered, which had 107,830 manuscripts.

A preprocessing process for the keywords was performed, in order to extract the focus of each research related to a manuscript. The keywords that corresponded to synonyms of COVID-19 and other types/variations of coronavirus were removed, in addition to terms that represented the places where the studies were carried out, such as China, Wuhan, Brazil and Africa. So, it enabled to extract the essence of each work.

This subsection is divided into two parts: (i) general keywords analysis; and (ii) advanced keywords analysis.

Table 7 Top 50 keywords

Keyword	NM	TC	μ_{TC}	σ_{TC}
Priming	8	16,763	2095.38	2891.34
Epidemic potential	3	3429	1143.00	0.00
Viral receptor	5	5540	1108.00	1269.26
Oral-fecal transmission	1	1032	1032.00	0.00
Pilot project	1	885	885.00	0.00
pcr cycle threshold	2	1744	872.00	0.00
ace2 tissue distribution	1	785	785.00	0.00
ctd	2	1552	776.00	0.00
Fibrin degradation product	3	2088	696.00	1205.51
Cardiovascular metabolic diseases	1	695	695.00	0.00
Entry	26	17,370	668.08	1813.48
Fingerstick blood	1	661	661.00	0.00
Rapid igm-igg combined test	1	661	661.00	0.00
Biocidal agents	2	1308	654.00	919.24
Proprotein convertase furin	1	644	644.00	0.00
t cell reduction	1	622	622.00	0.00
Maturation protease	2	1224	612.00	0.00
Diagnosis and interventions	1	561	561.00	0.00
Open-label nonrandomized control study	1	518	518.00	0.00
Potential interventions	1	493	493.00	0.00
Laboratory	117	**27,030**	231.03	563.94
Neutralization	144	21,687	150.60	806.33
Spike	275	22,534	81.94	585.38
Diagnostics	309	17,706	57.30	325.22
tmprss2	400	21,853	54.63	482.95
Outbreak	865	**37,624**	43.50	210.49
rt-pcr	620	19,238	31.03	228.39
ace2	1541	**45,440**	29.49	256.38
Thrombosis	877	**25,544**	29.13	139.63
Transmission	997	18,351	18.41	84.69
Stress	1488	24,720	16.61	107.86
Hydroxychloroquine	1198	19,672	16.42	113.66
Infection	1508	23,643	15.68	70.05
Depression	**2340**	**34,224**	14.63	92.28
Anxiety	**2714**	**37,458**	13.80	87.90
Treatment	1224	16,066	13.13	55.60
Epidemiology	**3379**	**34,408**	10.18	48.12
Pregnancy	1242	12,552	10.11	34.46
Children	1511	14,572	9.64	39.47
Mortality	**2961**	**27,990**	9.45	45.41
Inflammation	**1891**	17,362	9.18	30.97
Mental health	**3592**	**29,161**	8.12	39.22
Public health	**4131**	**32,372**	7.84	30.17
Lockdown	**2139**	14,255	6.66	23.05

Table 7 (continued)

Keyword	NM	TC	μ_{TC}	σ_{TC}
Social distancing	1233	7130	5.78	16.88
Telehealth	1486	8110	5.46	27.36
Cancer	1216	6340	5.21	22.13
Telemedicine	**2775**	13,983	5.04	22.42
Vaccine	**1801**	7354	4.08	14.89
Machine learning	1240	4987	4.02	13.14

In columns NM and TC, the numbers highlighted in bold represent the top 10 values for a respective metric/column

NM no. of manuscripts; *TC* total of citation; μ_{TC} citation mean; σ_{TC} citation standard deviation

General keywords analysis

Firstly, Table 7 presents the top 50 keywords by number of manuscripts (NM), total of citation (TC), together with its mean (μ_{TC}) and standard deviation (σ_{TC}), considering the whole period of the pandemic. The data of this table is ordered decreasingly by the column *citation mean* (μ_{TC}), as well as the keywords in this table are the union of the top 20 ones for the *NM*, *TC*, and μ_{TC} metrics.

The main focus based on keywords were the several aspects of public health and epidemiology, virus containment means (telemedice, telehealth and lockdown), the lockdown and virus' effects (inflammation, mortality, thrombosis, mental health, depression, and anxiety) and the results related to the vaccines, in terms of number of documents and total of citation.

Aiming to analyze the main focus based on keywords by each period, we extracted, by period, the three most frequent keywords and the three most cited ones, as shown in Table 8. In the first four months, the research focuses were the exploratory studies related to COVID-19 (the understanding of the virus transmission curve, the comparison between COVID-19 and its predecessors, and the laboratory analysis of the virus and its preliminary results), as well as the several works that highlighted the COVID-19 crisis in their respective countries/cities.

Between April 2020 and January 2021, the research focuses were the general aspects of medicine, epidemiology and public health linked to COVID-19, the alternative means to reduce the spread of the virus (the use of telemedicine, telehealth, lockdown, and social distancing), the analysis of substances and other drugs (chloroquine and hydroxychloroquine) to fight the pandemic, beyond the impact of the lockdown effects (mental health, depression, and anxiety).

Between February 2021 and July 2021, again, the focus remained on the general aspects of medicine, epidemiology and public health linked to the characteristics of COVID-19. However, some specific studies were performed during the pandemic, for example, those that involved pregnant women and people with asthma or anaphylaxis. Again, the alternative means to support the force task against the virus were explored, for example, telemedicine and deep learning. Finally, in the last two months, the works involving vaccines have gained their deserved recognition, probably because some countries are considering the application of a third dose as a booster for their population's immunity.

Table 8 Top 3 keywords by no. of manuscripts and total of citation

Period	Keywords and metrics			
	Keyword	NM	Keyword	TC
2019–12	Dromedary camels	10	Disease control	672
	Evolution	9	Infectious disease epidemiology	672
	Feline infectious peritonitis	8	Mathematical modelling	672
2020–01	Epidemiology	27	Diagnostics	14,135
	Public health	24	Outbreak	14,066
	Outbreak	21	rt-pcr	14,046
2020–02	Epidemiology	23	Respiratory disease	1802
	Outbreak	16	Clinical practice guideline	1800
	Virology	10	Evidence-based medicine	1800
2020–03	Epidemiology	36	Clinical features	12,278
	Outbreak	31	Laboratory	12,278
	Infection	30	Outcomes	12,278
2020–04	Public health	119	ace2	18,226
	Epidemiology	83	Spike	16,937
	Outbreak	83	tmprss2	16,814
2020–05	Public health	127	ace2	9276
	Epidemiology	110	Critical illness	7547
	Treatment	94	Clinical practice guidelines	7535
2020–06	Public health	172	Thrombosis	17,925
	Epidemiology	158	Anticoagulant	14,699
	Mental health	112	Antiplatelet	14,040
2020–07	Public health	207	Hydroxychloroquine	9514
	Telemedicine	161	Anxiety	8086
	Epidemiology	156	Mortality	8020
2020–08	Public health	190	Infection	6475
	Mortality	150	Anosmia	6459
	Epidemiology	149	Olfaction	6251
2020–09	Public health	264	Olfaction	3302
	Telemedicine	255	Head and neck surgery	3300
	Mental health	218	Somatosensation	3300
2020–10	Public health	278	Mental health	2419
	Mental health	226	Depression	2188
	Telemedicine	199	Anxiety	2010
2020–11	Public health	242	Mortality	2111
	Mental health	235	Public health	1556
	Mortality	215	Healthcare workers	1517
2020–12	Mental health	287	Anxiety	2410
	Public health	260	Mental health	2405
	Epidemiology	252	Depression	2357
2021–01	Epidemiology	271	Mortality	1242
	Anxiety	246	Depression	1139
	Public health	244	Anxiety	996

Table 8 (continued)

Period	Keywords and metrics			
	Keyword	NM	Keyword	TC
2021–02	Public health	299	Mortality	958
	Epidemiology	276	Infection	950
	Mental health	263	Pregnancy	892
2021–03	Public health	383	Asthma	865
	Mental health	338	Anaphylaxis	842
	Epidemiology	302	Psychological impact	826
2021–04	Public health	450	Mental health	453
	Mental health	383	Transmission	421
	Epidemiology	249	Epidemiology	377
2021–05	Mental health	358	Inflammation	372
	Epidemiology	303	Telemedicine	273
	Public health	302	Deep learning	233
2021–06	Mental health	437	bnt162b2	320
	Mortality	361	Infectivity	316
	Public health	346	Vaccine	270
2021–07	Vaccine	213	Epidemiology	800
	Epidemiology	200	Autoimmune diseases	658
	Public health	197	Glucocorticoids	651

NM no. of manuscripts; *TC* total of citation

Although the data referring to the three most cited keywords per period are shown in Table 8, only the chart of the three most frequent keywords per period will be presented (see Fig. 9). Thus, the main reason for this is the complexity of visualizing the results of the keywords for the metric of citations total per period.

As seen in Fig. 9, most studies actually focused on aspects of public health and epidemiology. Again, another focus was the exploration of studies involving mental health, possibly investigating the impact of the pandemic on the mental health of patients or relatives of victims, the effects of lockdown, among others. However, the works that reported evidences related to the vaccines only gained highlight in the last month in the period analyzed, possibly because of studies analyzing the possibility of applying a third dose as an immunity booster.

Advanced keywords analysis

Aiming to provide a deep perspective of the research focus, it was performed a complex network analysis applied over the whole manuscripts' keywords data. According to Zweig (2016), a network is a collection of objects, in which some pairs of these objects are connected by some kind of link. These objects are often referred to as nodes, as well as the relationships among these nodes are called links or edges. Particularly, a complex network is a set of related (connected) entities, where its structure is considered non-trivial (Menczer et al. 2020; Zinoviev 2018).

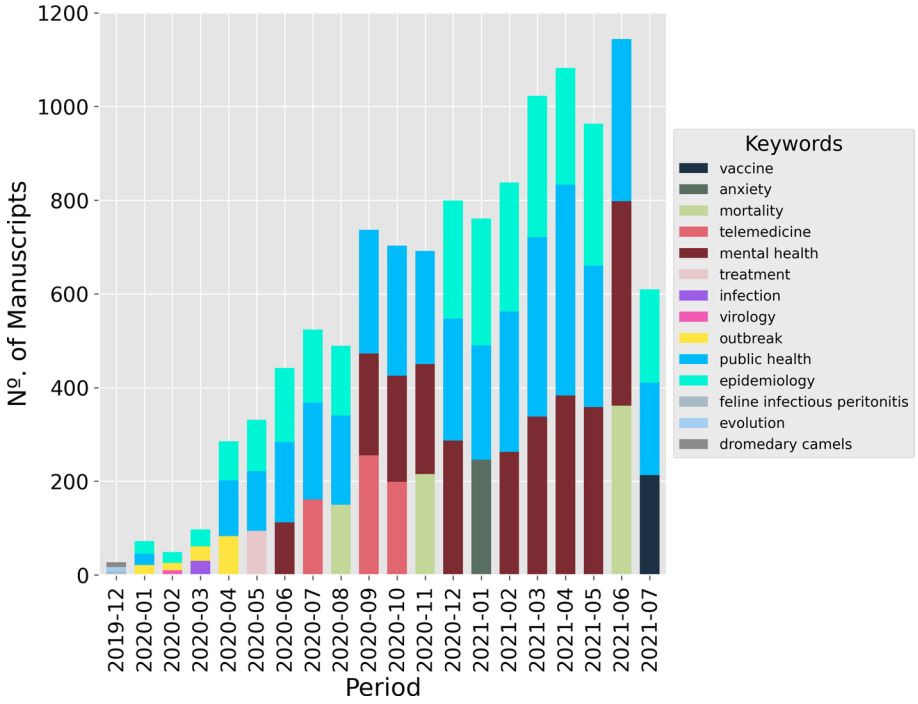

Fig. 9 No. of manuscripts by keyword and period

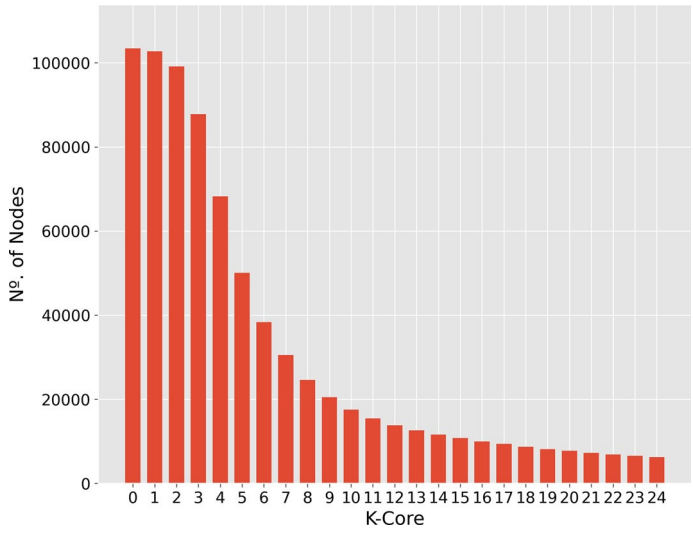

Fig. 10 Analysis of the top 25 k-cores of keywords network

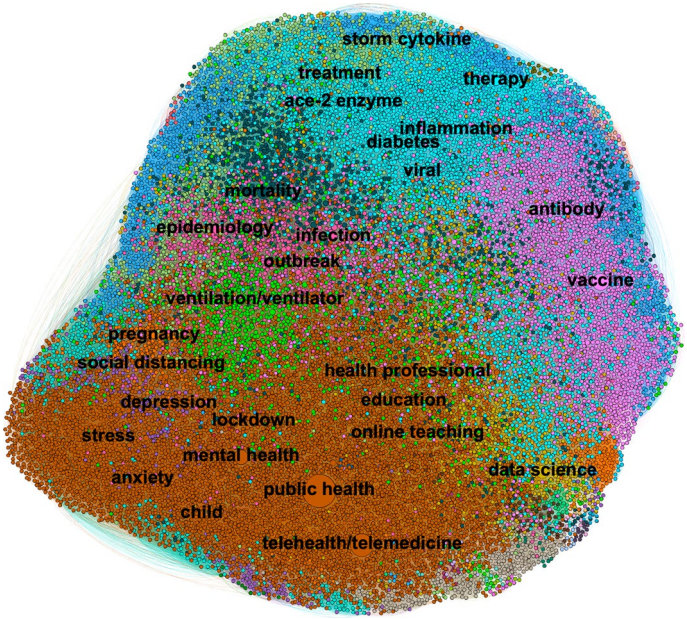

Fig. 11 Keywords' subnet with 8-core

The generated keyword network has 103,435 nodes (keywords) and 709,026 links (i.e., the co-occurrence between two keywords in the same article), and the number of manuscripts and total of citation were used as nodes' and edges' attributes. In addition to the preprocessing explained above, just for the generation of the network, the keywords were also submitted to textual preprocessing, in order to reduce textual variations of similar terms.

So, aiming to extract the most relevant keywords in this complex network, the relationship between the network cores and their number of nodes was analyzed using the k-core measure (see Fig. 10). *k-core* or *core of order k* is considered the cohesive subsets of nodes among whom there are relatively strong, direct, intense, frequent, or positive ties, where their number of neighboring nodes is equal to or greater than *k* (Batagelj and Zaversnik 2003; Menczer et al. 2020).

As seen in Fig. 10, an interesting behavior occurs when the k-core is greater than or equal to 8, once the number of keywords (nodes) seems to decrease more quickly. Because of that and aiming to extract a more representative subnet, the one whose k-core is equal to 8 was selected and illustrated in Fig. 11. Thus, this subnet has 24,627 nodes and 445,372 edges. Moreover, several related nodes of this subnet were manually merged in Gephi tool, aiming to simplify the structure of this subnet, resulting in a new network with 22,296 nodes and 384,572 edges, which was explored in this study.

In Fig. 11, the nodes of this subnet were grouped using the Modularity algorithm (Blondel et al. 2008)—implemented on Gephi—and they were colored based on their resulting clusters. Aiming to facilitate the visualization of this network, it was only showed the label of nodes whose number of manuscripts be equal to or higher than 830, as well as the nodes' size and edges' weight were determined by this attribute (number of manuscripts).

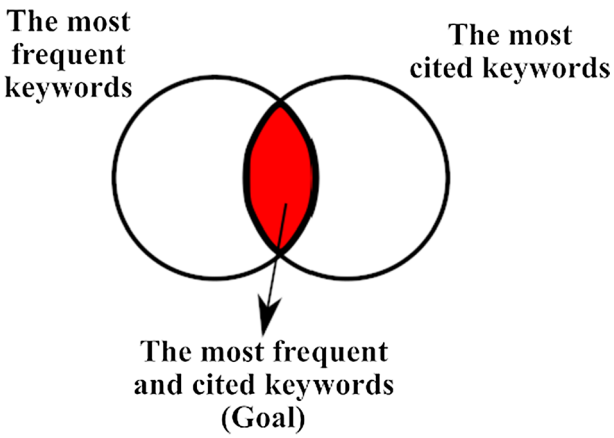

Fig. 12 Overview of outlier keywords analysis

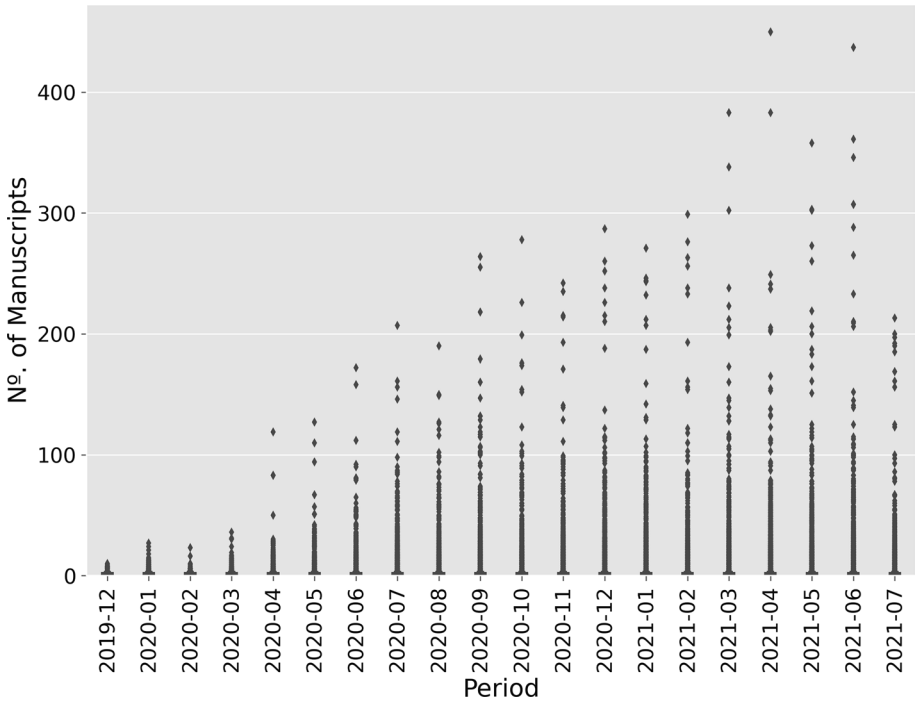

Fig. 13 Boxplot of keywords' occurrences on the manuscripts contained in the dataset

According to Fig. 11, again, several studies specially focused their research on several aspects related to the public health, while other works explored the mental and psychological effects of pandemic (mental health, anxiety, stress, depression, mortality).

A lot of studies that showed insights related to vaccines, as well as those that investigated antibodies against COVID-19 and others that explored the role of health professionals during the pandemic. It was also investigated several means of preventing the spread of

COVID-19, for example, lockdown, social distancing, online teaching, telemedicine and telehealth.

Next, in order to find out whether the most frequent keywords would be the most cited ones too (see Fig. 12), firstly, it was checked the existence of keywords considered as outlier from the number of manuscripts that they appeared. According to the Fig. 13, for each period, there are many keywords with a great number of manuscripts that they appeared. However, this raises the following question: what is the minimum number of occurrences for a keyword to also be one of the most cited?

It was found out the most frequent keywords and the most cited ones from the Algorithm 1, from Tukey's rule (Bruce et al. 2020), based on the interquartile distance. These sets of keywords are considered as outliers from their respective metric (number of manuscripts or total of citation) in each period analyzed. Outliers are data that differ drastically from all others, they are points outside the normal curve, i.e., they are values that are outside of normality (Igual and Seguí 2017).

Algorithm 1: Finding out the outliers in a dataset

input : Data D, feature f
output: Data D updated
procedure `FindOutliers`(D, f)
  get the unique list of periods $P \subset D$
  foreach $p_k \in P$ **do**
    forall *the feature* $f_k \in D$ *appeared in* p_k **do** calculate the 1^{o} quantile $Q1$;
    forall *the feature* $f_k \in D$ *appeared in* p_k **do** calculate the 3^{o} quantile $Q3$;
    forall *the feature* $f_k \in D$ *appeared in* p_k **do** calculate the interquartile range IQR $(Q3 - Q1)$;
    forall *the feature* $f_k \in D$ *appeared in* p_k **do**
      if $(f_k < Q1 - 1.5 \times IQR)$ *or* $(f_k > Q3 + 1.5 \times IQR)$ **then**
        D[$is\_outlier$] $\leftarrow True$
      else
        D[$is\_outlier$] $\leftarrow False$
      end
    end
  end
  return D

With the outliers determined, in the Algorithm 2, each unique value of outliers' number of manuscripts was tested, checking whether there was a difference between the sets of keywords. After a few iterations, it was found out that the minimum number of occurrences of a keyword for the same keyword to be among the most cited ones was 69.

Algorithm 2: Minimum occurrence for the most cited keywords

input : Keyword Occurrence Data X, Keyword Citation Data Y
output: minimum occurrence m

procedure `GetMinimumOccurrence`(X, Y)
  $X \leftarrow$ FindOutliers(X, number of occurrences)
  $occ\_out \leftarrow \mathsf{X}[keyword,\ number\ of\ occurrences] \mid \mathsf{X}[is\_outlier] = True$
  sort ascendingly occ_out by keyword and number of occurrences
  remove the duplicates in occ_out, keeping the last record of the duplicates
  $Y \leftarrow$ FindOutliers(Y, number of citation)
  $cit\_out \leftarrow \mathsf{Y}[keyword,\ number\ of\ citation] \mid \mathsf{Y}[is\_outlier] = True$
  get the unique list of number of occurrences $M \subset X$
  $m \leftarrow -1$
  foreach $m_k \in M$ **do**
    if $\{\mathsf{cit\_out}[keyword]\} - \{\mathsf{occ\_out}[keyword] \mid \mathsf{occ\_out}[number\ of\ occurrences] \geq m_k\} \equiv \varnothing$ **then**
      $m \leftarrow m_k$
  end
  return m

In addition to the previous analysis, it was analyzed how discrepant the number of manuscripts and total of citation among the keywords would be for each period analyzed. So, based on the Equation 1, the multivariable Gini index of the keywords for each analyzed period was calculated, using the metrics *number of manuscripts* and *total of citation*.

The Gini index is used to measure the inequality of a set of observations, defined as the mean of absolute differences between all pairs of observations for some measure. Its values are measured between 0 and 1, i.e., the minimum value is 0 when all measurements are equal and the theoretical maximum is 1 for an infinitely large set of observations where all measurements but one has a value of 0, which is the ultimate inequality (Brown 1994).

$$G_m = 1 - \sum_{i=0}^{k-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i) \tag{1}$$

Therefore, as seen in Fig. 14, for the metrics analyzed, the Gini index was higher than 0.5 in all periods. This shows that there is a considerable difference among keywords from the metrics *number of manuscripts* and *total of citation*. In other words, the keywords do not have the same number of manuscripts and total of citation, i.e., there is a great inequality between keywords, based on these two metrics.

However, between May 2020 and February 2021, this inequality declined modestly, possibly indicating that researchers were focusing on similar or correlated themes. Also, it is noticeable to highlight that, from the official start of the COVID-19 pandemic—declared by WHO—, this inequality was decreasing, possibly indicating a convergence of efforts and studies more focused on the pandemic.

On the other hand, from the official start of vaccination, this inequality went back up, apparently retracting the effects of vaccines in different aspects, where some topics stood out more among the others. Another interesting fact is the two peaks at the ends, possibly, retracting of the unprecedented nature of the disease and the results linked to vaccines, respectively.

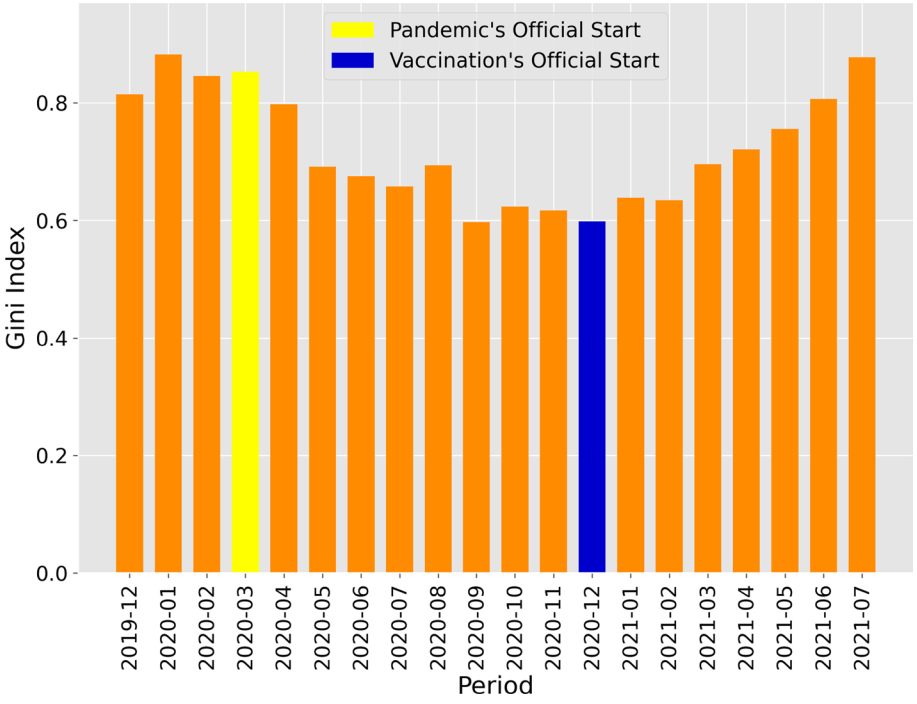

Fig. 14 Gini Index of number of manuscripts and total of citation by period

Conclusive discussion

In summary, despite the importance of indexed databases for the dissemination of scientific research, during the COVID-19 pandemic, preprints have also played an essential role in this scientific dissemination, in addition to being used to ensure a certain scientific precedence for the authors' studies. However, it is important to follow the work progress in order to verify if it will be approved after peer-review.

As expected, most researchers have published their studies in journals whose impact factor is greater than or equal to one, highlighting specially those ones that are considered quite relevant (impact factor is greater than three) to the scientific community. On the other hand, there is a considerable amount of manuscripts that were published in journals without impact factor.

The countries that contributed the most, in terms of number of manuscript and total of citation, were those that were heavily affected by the pandemic, such as China, United States, United Kingdom, and Italy. On the other hand, despite Brazil being one of the countries most affected by the pandemic, the current resources for research are scarce, which have harmed publications from that country, according to the impact factors of the journals of their manuscripts.

Regarding keyword analysis, several studies have focused their research on several aspects related to the public health, while other works explored the mental and psychological effects of the COVID-19 pandemic (mental health, anxiety, stress, depression, mortality). Also, many studies have showed insights related to vaccines, have explored the role of

health professionals during the pandemic, and have investigated several means of preventing the spread of COVID-19, for example, lockdown, social distancing, online teaching, telemedicine and telehealth.

With the results properly presented, the threats to validity of this study will be explained in the next section.

Threats to validity

The threats to the validity of the present study are:

- *Collection Bias* A threat linked to the collection process is the possible existence of relevant productions that were published after the date of retrieval, in addition to the failure to contemplate other sources of complementary data, which may also contain relevant studies to support fight COVID-19. Thus, to mitigate this threat, the main production bases widely used by the scientific community were used, in addition to the use of preprint repositories.
- *Indexing Bias* Since the data contained in the datasets used are cataloged and maintained by third parties, this threat could not be mitigated as it is beyond the scope of the approach proposed by this work. Another fact related to this threat is the existence of duplicate publications with slightly different titles, hosted in indexed bases and preprints. Once a study, previously hosted in a preprint database, has been officially published under a slightly different title, it is understood that the author is solely responsible for updating/removing the preprint version. Therefore, this is beyond the scope of this work.
- *Ethical approval* Since this was a metadata analysis of published work, ethics committee approval was not required.
- *External validity* a scientometric analysis of published and ongoing works related to SARS-CoV-2 was carried out for the period from December 2019 to 15 July 2021. However, it is possible that some relevant work was not yet indexed or is indexed on bases other than those used. Therefore, it is not possible to generalize the conclusions obtained to understand the behavior of scientific production related to the pandemic. However, the results are quite relevant to outline future investigations regarding COVID-19.

Conclusion and future works

With the exponential growth of scientific production related to the new coronavirus, bibliometric and scientometric studies started to need a systematic process that allows analyzing this large set of data. Thus, the aforementioned work proposed a Data Science-oriented methodology, combining techniques of Data Analysis, Machine Learning, Complex Network Analysis and DVC, in order to extract relevant and implicit knowledge and patterns in these scientific productions, as well as the experimental validation of this approach through a case study.

From the proposed approach, it was possible to analyze, characterize and understand, descriptively and temporally, the behavior of scientific productions linked to SARS-CoV-2, based on scientometric indexes and other related metrics. Therefore, the results

demonstrated the feasibility of the proposal, indicating the main countries, research focus and other interesting insights. The presented methodology has the potential to instrument and expand strategic and proactive decisions of the scientific community aiming at knowledge extraction that supports the fight against the pandemic.

As future works, it is intended to expand the context of analysis, increasing the quantity and diversity of analyzes and databases, enabling a general and complete view of the research linked to COVID-19.

Author contributions The authors contributed equally to all parts of the study analysis, development, writing and interpretation.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials The COVID-19 manuscripts data generated and used during this research are openly available in Santos et al. (2020). The 2019 Revision of World Population Prospects dataset is openly available in https://data.un.org. Finally, the 2020 InCites JCR data are openly available in Web of Science (2021).

Code availability The code and additional datasets are hosted in a private repository on GitHub (see "Planning" section). On the other hand, after the article is published and available to the public, we intend to make it public. During the review process, we may send these files via email, by contacting the corresponding author.

Declaration

Conflict of interest All authors declare that they have no conflicts of interest.

References

Alsharif, W., & Qurashi, A. (2020) .Effectiveness of covid-19 diagnosis and management tools: A review. *Radiography*

Basili, V. R., & Weiss, D. M. (1984). A methodology for collecting valid software engineering data. *IEEE Transactions on Software Engineering SE, 10*(6), 728–738. https://doi.org/10.1109/TSE.1984.5010301

Batagelj, V., & Zaversnik, M. (2003). An o(m) algorithm for cores decomposition of networks. arXiv preprint cs/0310049 http://arxiv.org/abs/cs/0310049

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10,* P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008.

Brown, M. C. (1994). Using gini-style indices to evaluate the spatial patterns of health practitioners: Theoretical considerations and an application based on alberta data. *Social Science& Medicine, 38*(9), 1243–1256.

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media.

Cai, X., Fry, C. V., & Wagner, C. S. (2021). International collaboration during the covid-19 crisis: Autumn 2020 developments. *Scientometrics, 126*(4), 3683–3692.

Callaway, E. (2020). The covid-19 crisis could permanently change scientific publishing. *Nature, 582*(7811), 167–168.

Casado-Aranda, L. A., Sánchez-Fernández, J., & Viedma-del Jesús, M. I. (2021). Analysis of the scientific production of the effect of covid-19 on the environment: A bibliometric study. *Environmental Research, 193*, 110416.

Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T., & Waltman, L. (2021). A scientometric overview of cord-19. *PLoS ONE, 16*(1), e0244839.

Di Girolamo, N., & Reynders, R. M. (2020). Characteristics of scientific articles on covid-19 published during the initial 3 months of the pandemic. *Scientometrics, 125*(1), 795–812.

Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R., & Wong, A. (2021). Understanding the temporal evolution of covid-19 research through machine learning and natural language processing. *Scientometrics, 126*(1), 725–739.

Farooq, R. K., Rehman, S. U., Ashiq, M., Siddique, N., & Ahmad, S. (2021). Bibliometric analysis of coronavirus disease (covid-19) literature published in web of science 2019–2020. *Journal of Family& Community Medicine, 28*(1), 1.

Fassin, Y. (2021). Research on covid-19: A disruptive phenomenon for bibliometrics. *Scientometrics, 126*(6), 5305–5319.

Grammes, N., Millenaar, D., Fehlmann, T., Kern, F., Böhm, M., Mahfoud, F., & Keller, A. (2020). Research output and international cooperation among countries during the covid-19 pandemic: Scientometric analysis. *Journal of Medical Internet Research, 22*(12), e24514.

Gu, C. F. Y. (2021). Influence of public engagement with science on scientific information literacy during the covid-19 pandemic: Empirical evidence from college students in china. *Science Education (Dordrecht), 30,* 1–15.

Gul, S., Ur Rehman, S., Ashiq, M., & Khattak, A. (2020). Mapping the scientific literature on covid-19 and mental health. *Psychiatria Danubina, 32*(3–4), 463–471.

Haghani, M., & Bliemer, M. C. (2020). Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across sars, mers and 2019-ncov literature. *Scientometrics, 125*(3), 2695–2726.

Haghani, M., & Varamini, P. (2021). Temporal evolution, most influential studies and sleeping beauties of the coronavirus literature. *Scientometrics* pp 1–46.

Haghani, M., Bliemer, M. C., Goerlandt, F., & Li, J. (2020). The scientific literature on coronaviruses, covid-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review. *Safety Science, 129,* 104806.

Igual, L., & Seguí, S. (2017). *Introduction to data science*. Berlin: Springer.

Iterative. (2021). DVC: Data Version Control - Git for Data & Models. https://doi.org/10.5281/zenodo.3677553.

Lauper, K., Bijlsma, J. W., & Burmester, G. R. (2021). Trajectories of covid-19 information in the annals of the rheumatic diseases: The first months of the pandemic. *Annals of the Rheumatic Diseases, 80*(1), 26–30.

Maalouf, F. T., Mdawar, B., Meho, L. I., & Akl, E. A. (2021). Mental health research in response to the covid-19, ebola, and h1n1 outbreaks: A comparative bibliometric analysis. *Journal of Psychiatric Research, 132,* 198–206.

Mahase, E. (2020). Coronavirus: Covid-19 has killed more people than sars and mers combined, despite lower case fatality rate. *BMJ*. https://doi.org/10.1136/bmj.m641.

Malik, A. A., Butt, N. S., Bashir, M. A., & Gilani, S. A. (2021). A scientometric analysis on coronaviruses research (1900–2020): Time for a continuous, cooperative and global approach. *Journal of Infection and Public Health, 14*(3), 311–319. https://doi.org/10.1016/j.jiph.2020.12.008

Menczer, F., Fortunato, S., & Davis, C. A. (2020). *A first course in network science*. Cambridge: Cambridge University Press.

Mohamadian, M., Chiti, H., Shoghli, A., Biglari, S., Parsamanesh, N., & Esmaeilzadeh, A. (2021). Covid-19: Virology, biology and novel laboratory diagnosis. *The Journal of Gene Medicine, 23*(2), e3303.

Monteiro, R. P., de Holanda Coelho, G. L., Hanel, P. H., Vilar, R., Gouveia, V. V., & de Medeiros, E. D. (2022). The dark side of brazil: Effects of dark traits on general covid-19 worry and responses against the pandemic. *Personality and Individual Differences, 185*, 111247. https://doi.org/10.1016/j.paid.2021.111247

Moscadelli, A., Albora, G., Biamonte, M. A., Giorgetti, D., Innocenzio, M., Paoli, S., Lorini, C., Bonanni, P., & Bonaccorsi, G. (2020). Fake news and covid-19 in italy: Results of a quantitative observational study. *International Journal of Environmental Research and Public Health, 17*, 2–13. https://doi.org/10.3390/ijerph17165850

Nowakowska, J., Sobocińska, J., Lewicki, M., Lemańska, Żaneta, & Rzymski, P. (2020). When science goes viral: The research response during three months of the covid-19 outbreak. *Biomedicine& Pharmacotherapy, 129*, 110451. https://doi.org/10.1016/j.biopha.2020.110451

Okhovati, M., & Arshadi, H. (2021). Covid-19 research progress: Bibliometrics and visualization analysis. *Medical Journal of the Islamic Republic of Iran, 35*, 20.

Rodríguez-Rodríguez, I., Rodríguez, J. V., Shirvanizadeh, N., Ortiz, A., & Pardo-Quiles, D. J. (2021). Applications of artificial intelligence, machine learning, big data and the internet of things to the covid-19 pandemic: A scientometric review using text mining. *International Journal of Environmental Research and Public Health, 18*(16), 8578.

Santos, B. S., Júnior, M. C., da Paixão, B. C., Santos, R. M., Nascimento, A. V. R., dos Santos, H.C., Wallace Filho, H., & de Medeiros, A. S. (2015). Comparing text mining algorithms for predicting irregularities in public accounts. In: SBSI (pp. 667–674).
Santos, B. S., Silva, I., da Câmara, Ribeiro-Dantas M., Alves, G., Endo, P. T., & Lima, L. (2020). Covid-19: A scholarly production dataset report for research analysis. *Data in Brief, 32*, 106178. https://doi.org/10.1016/j.dib.2020.106178
Şenel, E., & Topal, F. E. (2020). Holistic analysis of coronavirus literature: A scientometric study of the global publications relevant to sars-cov-2 (covid-19), mers-cov (mers) and sars-cov (sars). Disaster Medicine and Public Health Preparedness (pp. 1–8).
Stein, F. (2021). Risky business: Covax and the financialization of global vaccine equity. *Global Health, 17*, 2–11.
Sugimoto, C. R., & Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford: Oxford University Press.
Taleghani, N., & Taghipour, F. (2020). Diagnosis of covid-19 for controlling the pandemic: A review of the state-of-the-art. *Biosensors and Bioelectronics* (p. 112830).
Tiwari, A., So, M. K., Chong, A. C., Chan, J. N., & Chu, A. M. (2021). Pandemic risk of covid-19 outbreak in the united states: An analysis of network connectedness with air travel data. *International Journal of Infectious Diseases, 103*, 97–101.
Tornberg, H. N., Moezinia, C., Wei, C., Bernstein, S. A., Wei, C., Al-Beyati, R., et al. (2021). Assessing the dissemination of covid-19 articles across social media with altmetric and plumx metrics: Correlational study. *Journal of Medical Internet Research, 23*(1), e21408.
van Solingen, D. R., & Berghout, E. W. (1999). *The Goal/Question/Metric Method: A practical guide for quality improvement of software development*. New York: McGraw-Hill.
Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Amsterdam: Elsevier.
Web of Science. (2021). Incites journal citation reports. Available in: https://jcr.clarivate.com/. Accessed 10 Jul 2021.
Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Berlin: Springer.
World Health Organization. (2020). Managing the covid-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. https://bit.ly/39No6bA. Accessed 29 September 2021.
World Health Organization. (2021). Who coronavirus (covid-19) dashboard. https://covid19.who.int/. Accessed 9 July 2021.
Zhang, L., Zhao, W., Sun, B., Huang, Y., & Glänzel, W. (2020). How scientific research reacts to international public health emergencies: A global analysis of response patterns. *Scientometrics, 124*, 747–773.
Zinoviev, D. (2018). Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret. Pragmatic Bookshelf.
Zweig, K. A., et al. (2016). *Network analysis literacy*. Berlin: Springer.

Authors and Affiliations

Breno Santana Santos[1,2] · Ivanovitch Silva[1] · Luciana Lima[3] · Patricia Takako Endo[4] · Gisliany Alves[1] · Marcel da Câmara Ribeiro-Dantas[5]

Ivanovitch Silva
ivanovitch.silva@ufrn.br

Luciana Lima
luciana.lima@ufrn.br

Patricia Takako Endo
patricia.endo@upe.br

Gisliany Alves
gisliany.alves.094@ufrn.edu.br

Marcel da Câmara Ribeiro-Dantas
marcel.ribeiro-dantas@curie.fr

[1] Postgraduate Program in Electrical and Computer Engineering, Federal University of Rio Grande

do Norte, Natal, RN, Brazil

[2] Federal University of Sergipe, Itabaiana, SE, Brazil

[3] Federal University of Rio Grande do Norte, Natal, RN, Brazil

[4] University of Pernambuco, Recife, PE, Brazil

[5] Institut Curie (UMR168), Sorbonne Université (EDITE), Paris, France