PlatCOVID: A Novel Web Tool to Analyze, Curate and Share COVID-19 Literature

Lucas André Cavalcanti Brandão (✉ **lucas.cbrandao@ufpe.br**)
Universidade Federal de Pernambuco https://orcid.org/0000-0002-8588-4559

Almerinda Agrelli
Universidade Federal de Pernambuco

Lucas Bernardo
Universidade Federal de Pernambuco

Francesco Paparella
Universita degli Studi di Roma La Sapienza

Ronald Moura
IRCCS materno infantile Burlo Garofolo

Sergio Crovella
Universita degli Studi di Trieste Dipartimento di Scienze Mediche e Chirurgiche e della Salute

Software

Keywords: PlatCOVID, COVID-19, SARS-CoV-2, Text Mining, Syntax Analysis, Data Analysis

Posted Date: July 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42169/v1

Abstract

Background

In the attempt to face the COVID-19 pandemic, the global scientific community has been expending great efforts to produce useful and reliable data aiming to help patients, physicians and guiding public health policies. A huge amount of information is being released every week, making impossible for a single person (or even for a research group) to read everything and get constantly updated on the scientific literature concerning COVID-19 and its etiological agent, SARS-CoV-2. Therefore, we developed PlatCOVID (www.platcovid.com), a Web platform designed to analyze, cluster, classify and discuss COVID-19 literature available on LitCovid (NCBI).

Results

PlatCOVID has been created as a novel COVID-19 hub able to add features of text mining and syntax analyses methods, such as word and sentence atomization and tokenization, clusterization and classification. The main division of the literature comprehends five categories: 1) Diagnosis; 2) Epidemiology; 3) Clinical, Signs & Symptoms; 4) Transmission; and 5) Treatment & Prevention. Consequently, it is possible to reduce the amount of text to be read with minimal loss of information, identifying target subjects by mining as new insights arise, enhancing data analysis efficiency. PlatCOVID has been designed with central panels (Gene, Drug and Tissue panels) to easily gather and share with the scientific community important COVID-19 information.

Conclusions

Although most of the text mining and syntax analysis is made by using automated computing processes, the final results must to be humanly curated. With this in mind, PlatCOVID allows researchers to be part of our effort to curate, analyze, discuss and rate each matter of interest (helpus.platcovid.com). We welcome user feedback for further enhancement.

Background

During the last months the world has been distraught by the COVID-19 pandemic, a disease responsible for an aggressive and acute respiratory syndrome. In fact, COVID-19 patients present other not specific symptoms such as fever, myalgia, headache, lymphopenia, hyposmia, and hypogeusia [1]. Although in a minor fraction of the cases, other tissues may be affected by SARS-CoV-2, causing diarrhea, nausea and vomiting, suggesting the susceptibility of the gastro-enteric system to the infection [2]. Moreover, proteinuria and acute renal tubular damage in COVID-19 patients indicate a kidney impairment [3], and elevated troponin T and N-terminal pro B-type natriuretic peptide levels imply a possible cardiovascular injury [4].

COVID-19 has been causing thousands of deaths worldwide. Several specialists had declared the COVID-19 as the most important health issue of the century with millions of human beings directly or indirectly affected. The course of this pandemic situation showed us that science communication and sharing have never been so important as it is now. To produce objective management guidelines for patients with COVID-19 and deal with the high demand for hospital beds, effective and reliable scientific data are required.

Thousands of scientific reports about COVID-19 have been published and the number of articles is still increasing, as reported by LitCovid [5]. Considering the current scenario, the speed of the publication process could be a pitfall, since the methodology accuracy and the relationship between results and conclusion could be sometimes mistaken.

In the face of this great amount of information, what has science been doing to efficiently translate this information into health policies? What has been the path, if one, followed so far? How to pass along this massive quantity of information to our leaders/policy makers in a comprehensive way? How to make them reach general population? In order to have an overview of science's response to this complex situation, we designed PlatCOVID [6], a tool to analyze and categorize, automatically, the whole literature about COVID-19, allowing scientists to discuss and classify published data. We believe that professional engagement will accelerate the curation of the literature. Finally, our platform will gather dynamic information aiming to build a scientific consensus to assist our policy managers in decision-making processes.

Implementation

PlatCOVID is free Web platform that allows to analyze and curate scientific data, enabling the identification of useful information concerning the 2019 new coronavirus (SARS-CoV-2) pandemic. The platform aims to provide scientific consensus about COVID-19 issues by analyzing, discussing and classifying published scientific data, making possible to assist and guide health care policies. Therefore, it is addressed mainly to scientists, academical staff, specialists in the field and health professionals.

The compiler design process is divided into two phases: lexical analysis and syntax analysis. The lexical analysis or “tokenization” is the process of breaking up a sequence of characters into pieces called “tokens”. The syntax analysis or “Parsing” comes after the lexical analysis and analyzes the syntactical structure of the given input (source code or a program). It does so by building a data structure that may be called a “Parse tree” or “Syntax tree” [7].

Using the combination of rentrez [8], easyPubMed [9], pubmed.mineR [10], tokenizers [11] and blogdown [12] R packages, and Hugo [13] it was possible to download and analyze scientific literature and develop the platform. Firstly, we download LitCOVID curated database [5]. After that, a secondary search was done according to five categories: *Diagnosis, Treatment, Epidemiology, Transmission and Clinical* & *Signs* & *Symptoms*. For this categorization process we used Mesh [14] and DeCS [15] terms list. Then, we selected articles that had available abstracts. The analysis of the abstracts was performed by the linguistic structured by the level of sentence and word tokenization using the pubmed.mineR and tokenizer. The online map was built up by tmap [16] and sp [17] R packages. Common words and numerals were extract from the results (**Supplementary Table 1**). All analyzes were developed in R environment and all script and data (.Rdata) are accessible at our github repository [18].

To facilitate the screening of publications, we assembled panels for the genes, tissues and drugs involved in COVID-19. A FAQ section is available with tutoring and information about how to curate data.

Results And Discussion

On 6 of July, 2020, the search found 1405,7 abstracts from 26,980 published articles. As expected, we observed an exponential increase in publication never seen in the recent scientific literature history (Fig. 1). These articles were published manly as Journal Articles (60.8%), Letters (17.09%), Editorials (6.84%), Reviews (6.51%) and Comments (2.03%). We excluded articles without available abstract (12,923) and applied the word and sentence tokenization methodology. Then, using the countrycode R package [19], we calculated how many times a country was cited in the abstract and the article filiation. United States (43.59%), United Kingdom (16.63%), China (11.25%), Italy (5.71%) and Spain (5.35%) were the main source of scientific literature. About 82.53% of articles analyzed came from these five countries.

Using the atomization process, 75,368 words/terms were found. Of these, 7,899 common words were excluded, remaining 67,469 words. The ten most cited terms are demonstrated in Table 1. After that, we selected the 50 most recurrent words in the abstracts to continue the investigation (**Supplementary Table 2**). Our analysis suggests that the scientific focus, until now, has been to summarize the main clinical symptoms of COVID-19. It is also possible to infer that many articles were driven to describe the virus spreading. The other scientific efforts discussed were about the transmission, prevention, treatment, health care management and diagnosis of SARS-CoV-2 and COVID-19.

Table 1
The ten most cited words in COVID literature.

Global	n	Diagnose	n	Treatment	n	Epidemiology	n	Transmission	n	Signs	n
disease	11738	disease	776	treatment	1935	disease	377	transmission	991	disease	3192
pandemic	10117	diagnosis	700	disease	1660	clinical	308	disease	616	clinical	2864
health	9418	clinical	649	clinical	1366	health	264	infection	520	pandemic	2599
infection	7760	infection	566	pandemic	1256	infection	241	health	519	health	2288
clinical	7598	pandemic	548	severe	1116	epidemiological	236	pandemic	455	infection	2118
respiratory	6821	respiratory	460	infection	1069	pandemic	235	respiratory	400	severe	1954
severe	6767	treatment	458	care	958	respiratory	203	during	390	respiratory	1838
care	6453	severe	444	respiratory	920	severe	197	virus	382	during	1713
during	6242	study	415	health	879	study	188	risk	380	care	1668
risk	5311	during	406	during	812	risk	145	study	282	study	1663

Since our platform is based on published data, we are not reporting available pre-prints articles. We chose to analyze and share literature that had undergone a strict reviewing process, thus reporting validated findings.

Based on global words tokenization/atomization from the analyzed abstracts, we categorized the studies in five categories. Respectively, 1,999, 4,260, 1,038, 1,834 and 8,584 abstracts were classified in the categories as *Diagnosis, Treatment, Epidemiology, Transmission and Clinical* & *Signs* & *Symptoms* (**Supplementary Fig. 1**). Twenty-eight articles hit all five criteria simultaneously (Fig. 2) and 3,374 abstracts were not categorized. *Diagnosis* studies have been focusing on clinical diagnosis of the acute symptoms, mainly respiratory. The terms "PCR" or "qPCR" were rarely found in the abstracts. Curiously, a small quantity of molecular diagnosis was cited and consequently discussed. We are sensitive to this matter, since molecular or antibody detection tests (qPCR and ELISA/CLIA, respectively) are considered golden standard for diagnosis. *Treatment* focused in the clinical treatment of the severe acute respiratory syndrome and pneumonia. Health care management was highly mentioned. The use of antivirals was suggested, but no specific drugs were found to be relevant. The words “therapy”, “drugs”, “trials” and “effective” indicate that investigations into forms of treatment are currently being conducted. Despite that, we implement a Panel Drug at PlatCOVID to list all cited drugs. *Epidemiology* studies have been focusing on clinical and infection features of the disease as well as on the transmission risks. Epidemiological data from pneumonia status seems to be relevant to medical prevention and treatment during COVID-19 pandemic. *Transmission* studies have reported how the disease is transmitted by respiratory routes. The terms "transmission", "disease", and "infection" were highly cited in the abstracts, suggesting that forms of infection play an important role in epidemic transmission. Articles categorized as *Clinical & Signs & Symptoms* were the most abundant in the general analysis. In detail, these studies discussed the severe acute respiratory conditions and pneumonia symptoms in the infected group, being "acute", "pneumonia" and "lung" common terms used to describe patient's clinical condition.

Moreover, the most frequent terms (Table 1) indicate the importance of determining the clinical aspects of the infection. Taking all these findings into account, the primary scientific response during the pandemic seems to be focused into the report of main clinical signs and symptoms in order to extend this information to appropriate treatment and patient management. Nevertheless, a new perspective in molecular treatment and diagnosis shall be critical to face COVID-19.

The translation scientific language is a continuous challenge. The scientific perception and fake news circulating with dramatic frequency in the media and social networks could misunderstand the real meaning of scientific evidence. Thus, we implemented a Web platform dedicated to COVID-19 scientific literature that is able to automatically analyze, classify and evidence the important information of published articles.

Conclusions

Aware of the computational limitations to study scientific article linguistic and semantic, we invite scientists and all specialists in the field to join us and help mining and curating COVID-19 literature. The categorization, classification and discussion of scientific issues led by professionals in the field should be translated to help guiding public health measures and policy managers’ decisions in controlling and managing this pandemic.

Availability And Requirements

All data is available at www.platcovid.com. The source code created, data analyzed and results are available in the platcovidsource repository from our github, (https://github.com/bio-hub/platcovidsource).

Project name

PlatCOVID

Project home page: https://www.platcovid.com

Operating system(s)

Web

Programming language: R language

Other requirements: None

License

GNU GPL.

Any restrictions to use by non-academics

None

Abbreviations

COVID-19
Coronavirus Disease 2019
SARS-CoV-2
Severe Acute Respiratory Syndrome Coronavirus 2
LitCovid
Literature hub about the 2019 new coronavirus
NCBI
National Center for Biotechnology Information
FAQ
Frequently asked questions
PCR
Polymerase Chain Reaction
qPCR
Quantitative Polymerase Chain Reaction
ELISA
Enzyme-Linked Immunosorbent Assay
CLIA
Chemiluminescence Immunoassay

Declarations

Ethics approval and consent to participate

We wish to confirm that there are no known conflicts of interest associated with this publication.

Consent for publication

Not applicable

Competing interests

None.

Funding

This study has been supported by the following grants: ISE-EMH (Italian-Slovenian Ecosystem for Electronic and Mobile Health) from European Community and RC03/20 BioHub - A High-throughput Platform For OMICs Data Analysis And Integration from the Italian Ministry of Health. Both, ISE-EMH and RC03/20 supported the computational stations and requirement to perform the current analysis. RC03/20 project also support Ronald Moura as a recipient of a senior fellowship. Lucas Brandão is recipient of a senior fellowship from the Brazilian National Council for Scientific and Technological Development (CNPq). None of these funding, cited above, were involved in the design of the study and collection, analysis, and interpretation of data.

Authors’ contribution

1. Study Design: Lucas André Cavalcanti Brandão
2. Data Collection: Lucas André Cavalcanti Brandão, Almerinda Agrelli, Lucas C. Bernardo, Ronald Rodrigues de Moura.
3. Software development: Lucas André Cavalcanti Brandão, Francesco Paparella, Ronald Rodrigues de Moura.
4. Curation: Lucas André Cavalcanti Brandão, Almerinda Agrelli, Lucas C. Bernardo, Ronald Rodrigues de Moura.
5. Manuscript Preparation: Lucas André Cavalcanti Brandão, Almerinda Agrelli, Sergio Crovella.
6. Literature Search: Lucas André Cavalcanti Brandão, Almerinda Agrelli, Lucas C. Bernardo, Ronald Rodrigues de Moura.

All authors have read and approved the entire manuscript. There are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

Acknowledgements

This study has been supported by the following grants: ISE-EMH (Italian-Slovenian Ecosystem for Electronic and Mobile Health) from European Community and RC03/20 BioHub - A High-throughput Platform For OMICs Data Analysis And Integration from the Italian Ministry of Health. Lucas Brandão is recipient of a senior fellowship from the Brazilian National Council for Scientific and Technological Development (CNPq). Ronald Moura is recipient of a senior fellowship from RC03/20 project.

References

1. Tu YF, Chien CS, Yarmishyn AA, Lin YY, Luo YH, Lin YT, et al. A review of SARS-CoV-2 and the ongoing clinical trials. Int J Mol Sci. 2020;21. doi:10.3390/ijms21072657.
2. Puelles VG, Lütgehetmann M, Lindenmeyer MT, Sperhake JP, Wong MN, Allweiss L, et al. Multiorgan and Renal Tropism of SARS-CoV-2. N Engl J Med. 0:null. doi:10.1056/NEJMc2011400.
3. Wang S, Zhou X, Zhang T, Wang Z. The need for urogenital tract monitoring in COVID-19. Nat Rev Urol. 2020;17:314–5. doi:10.1038/s41585-020-0319-7.
4. Nicin L, Abplanalp WT, Mellentin H, Kattih B, Tombor L, John D, et al. Cell type-specific expression of the putative SARS-CoV-2 receptor ACE2 in human hearts. Eur Heart J. 2020;41:1804–6. doi:10.1093/eurheartj/ehaa311.
5. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature. 2020;579:193. doi:10.1038/d41586-020-00694-1.
6. PlatCOVID. https://www.platcovid.com. Accessed 03 July 2020.
7. Details CI. Modern compiler design. 2013. doi:10.5860/choice.50-3309.
8. Winter DJ. rentrez: An R package for the NCBI eUtils API. R J. 2017;9:520–6. doi:10.32614/rj-2017-058.
9. Fantini D. Package ‘ easyPubMed.’ 2019. https://cran.csiro.au/web/packages/easyPubMed/easyPubMed.pdf.
10. Rani J, Shah ABR, Ramachandran S. pubmed.mineR: an R package with text-mining algorithms to analyse PubMed abstracts. J Biosci. 2015;40:671–82.
11. Mullen A, Benoit L, Keyes K, Selivanov O, Arnold D, Fast J. Consistent Tokenization of Natural Language Text. J Open Source Softw. 2018;3:655. doi:10.21105/joss.00655.
12. Xie Y. blogdown. Create Blogs and Websites with R Markdown. 2020. R package version 0.20. https://github.com/rstudio/blogdown.
13. Hugo. https://gohugo.io. Accessed 03 July 2020.
14. MeSH (Medical Subject Headings). https://www.ncbi.nlm.nih.gov/mesh/. Accessed 03 July 2020.
15. Health Sciences Descriptors: DeCS. http://decs.bvsalud.org/I/homepagei.htm. Accessed 03 July 2020.
16. Tennekes M. Tmap. Thematic maps in R. J Stat Softw 2018;84. doi:10.18637/jss.v084.i06.
17. Roger SB, Edzer P, Virgilio G-R. Applied spatial data analysis with R, Second edition. 2013. https://asdar-book.org.
18. R Core Team. R: A language and environment for statistical computing. 2014. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
19. Arel-Bundock V, Enevoldsen N, Yetman C. Countrycode. An R package to convert country names and country codes. J Open Source Softw. 2018;3:848. doi:10.21105/joss.00848.
20. Additional, Files.
21. Additional file 1. List of common words excluded from syntax analysis.
22. Additional file 2. The most frequent words cited in COVID literature.
23. Additional file 3. Workflow of PlatCOVID categorization.

Figures

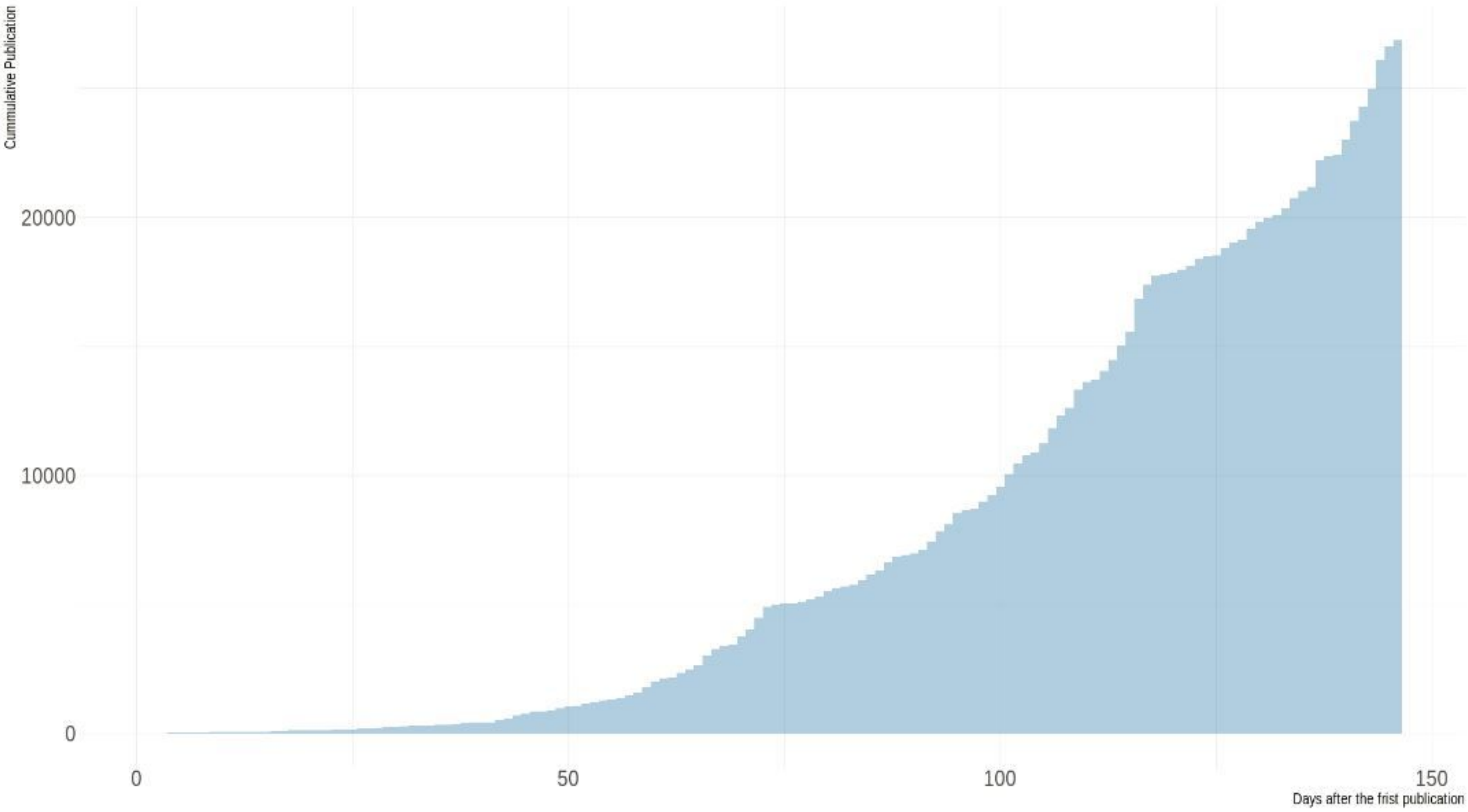

Figure 1

Accumulative publication about COVID and SARS-CoV-2.

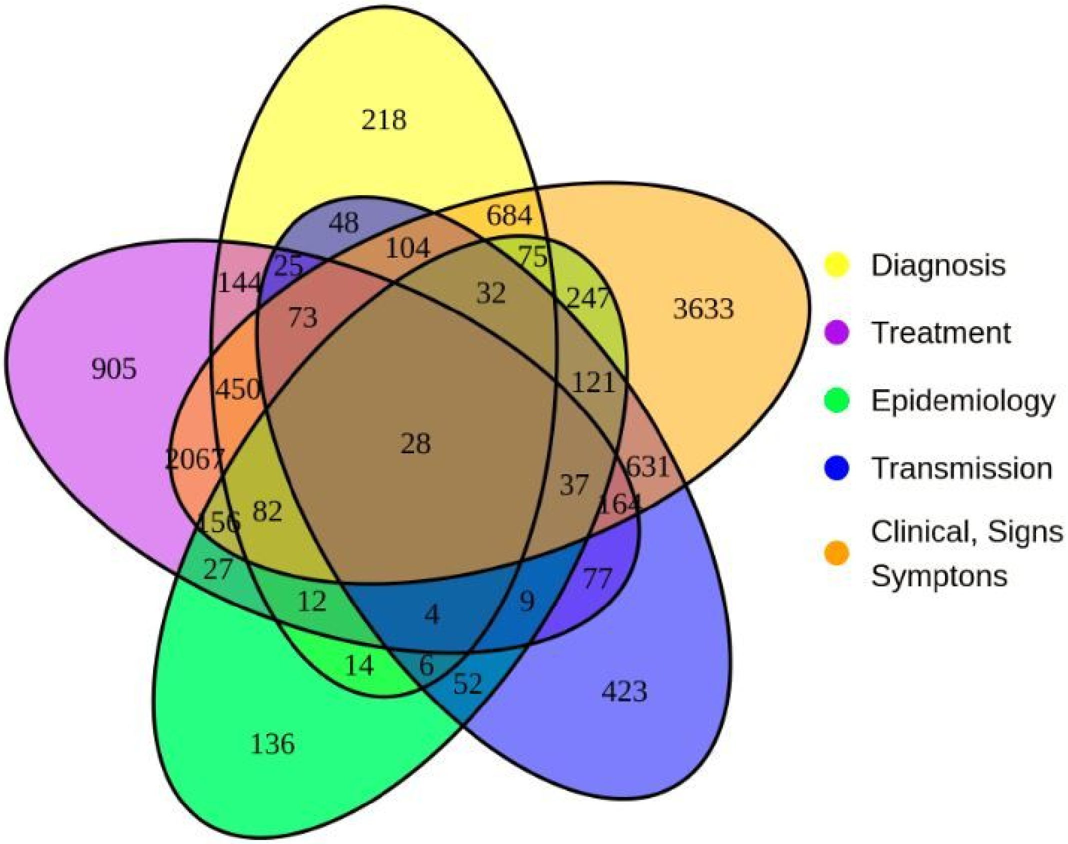

Figure 2

Venn diagram for the five categories in PlatCOVID.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supp3.png
- supp2.xlsx
- supp1.xlsx