The effect of COVID-19 on scientific publishing in Italy

Gli effetti di COVID-19 sulla pubblicazione scientifica in Italia

Paola Berchialla,[1] Sara Urru,[2] Veronica Sciannameo[1]

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

Corresponding author: Paola Berchialla; paola.berchialla@unito.it

Since the beginning of the COVID-19 outbreak, a flood of Coronavirus research has spurred, and it has produced an unprecedented number of publications, favoured by rapid publishing, reduced fees, and specific COVID-19 paper calls that many biomedical journals have provided.[1] Let us look over some numbers. LitCOVID,[2] a daily updated literature hub tracking relevant COVID-19 related articles indexed in PubMed, at the end of October 2021 reported having collected a database of 187,206 citations. The COVID-19 Open Research Dataset (CORD-19),[3] another corpus of tens of scholarly articles about COVID-19 and SARS-CoV-2 regularly updated with new research published in peer-reviewed journals and preprint repository, counted more than 280,000 scientific papers at the same date. We can fairly say that, during all the COVID-19 pandemic, as researchers we have been daily exposed to scientific information we were used to getting in biomedical conferences that happened yearly. To give a sense of context to the volume of such scientific production, in 2020, the total number of publications indexed by PubMed was more than 1.4 million, corresponding to a 15% increase over 2019. This increase in publications was not only due to COVID-19 related articles submission. The same year, a growing number of non-COVID-19 related articles has been submitted to scientific journals, probably explained by the fact that many researchers, because of lockdown, were forced to work from home and could focus on writing down papers than conducting experiments in laboratories.[4]

However, disease-specific comparator searches in citation database like PubMed suggest that the pandemic almost drove the sharp increase in publication volume. Searching how many articles focused, for example, on cardiovascular diseases that are not COVID-19 neither SARS-CoV-2 related supports this observation. In 2017, cardiovascular diseases were the leading cause of years of life lost.[5] Searching on PubMed, the number of articles retrieved rose from 111,769 in 2019 to 119.646 in 2020, accounting for only a 7% increase. Similarly, such pattern occurs for almost any other disease.

An interesting scientometric analysis performed on novel Coronavirus publications in 2020 shows that Italy has played a significant part in COVID-19 research, ranking as the 5th most productive country in terms of publications.[6] This result is striking since Italy has never ranked in the top five publishing countries before the pandemic. A similar pattern, even if at a smaller size, is displayed in other countries like Brazil and Hong Kong: the widespread and the severe impact of the pandemic can explain it.

Let's look a bit more in detail at the figures of Italian COVID-19 research published so far. To have a reasonable idea of the volume of the Italian COVID-19 research, we get the data from two main sources: Scopus and PubMed. Applying the COVID-19 article filters used to collect relevant articles in the LitCOVID database from November 2019 to 20th October 2021, from PubMed we identified 16,088 relevant articles with at least one author with an Italian affiliation. Performing the same research in Scopus, using the automatic term mapping feature to create the search string, we identified 18,146 articles. After deduplication, we finally obtained 19,331 relevant articles. As a matter of interest, the majority of COVID-19 papers were published in Open Access high-ranked journals, like *International Journal of Environmental Research and Public Health*, *Journal of Clinical Medicine*, *Critical Care*, only to mention a few.

Looking at the temporal trend, a growing number of publications started from March 2020, coinciding with the first wave, and peaked in August 2020. Then, a second peak oc-

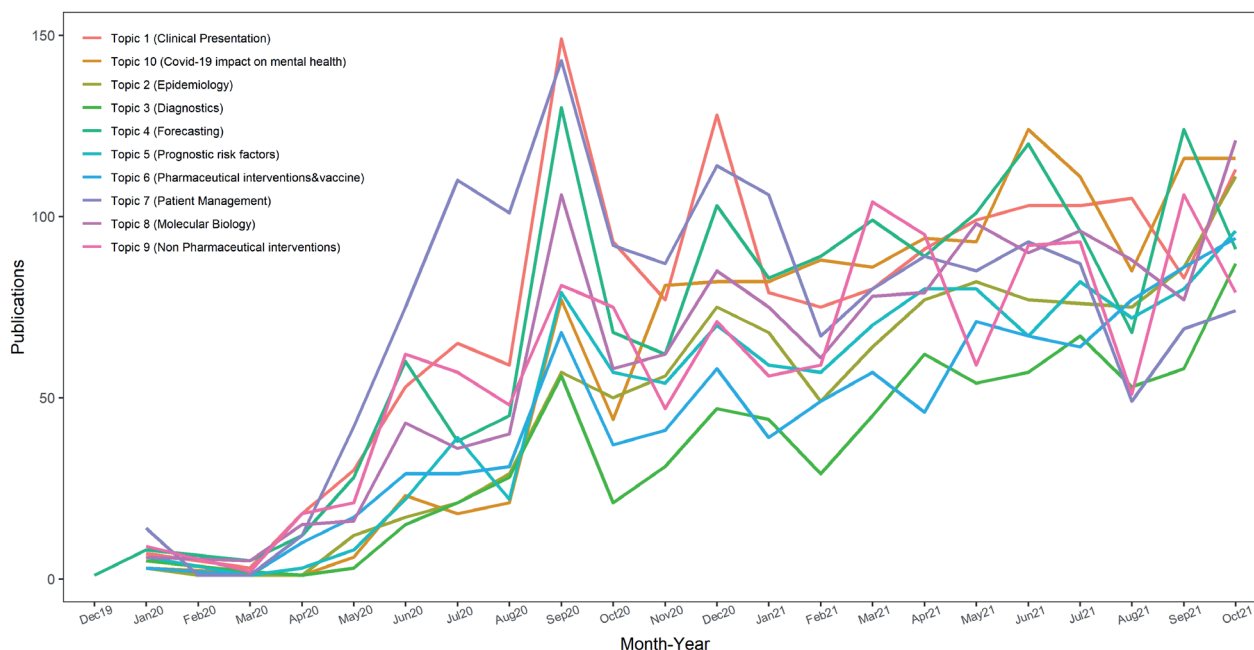**Figure 1.** Topic trends.
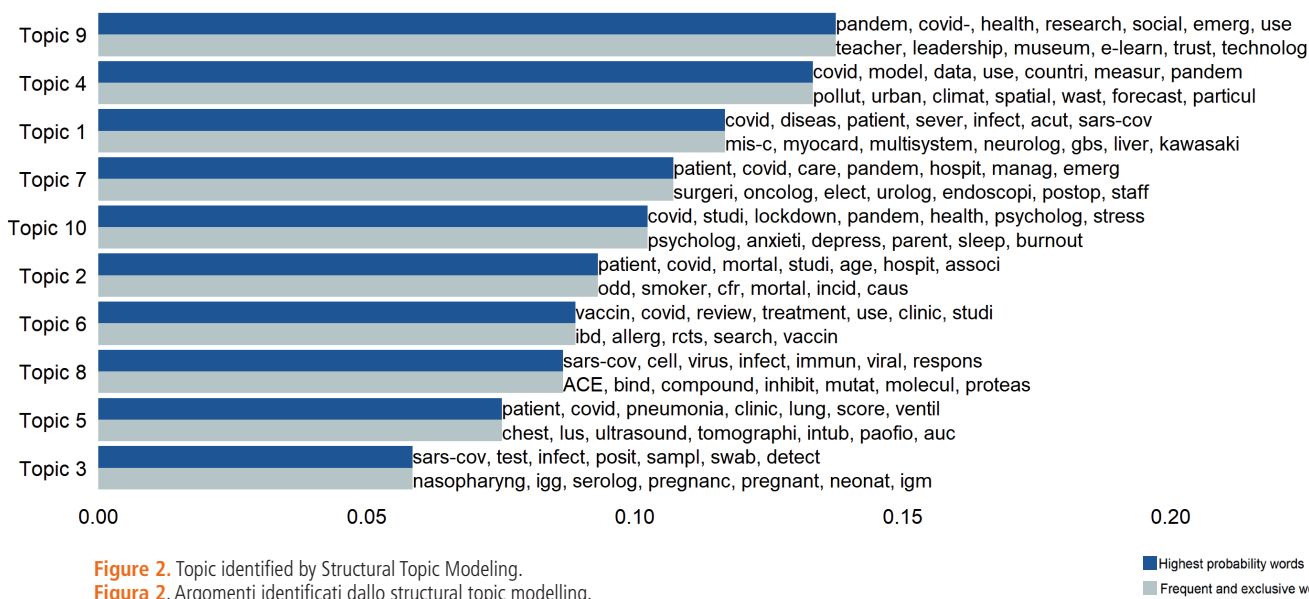**Figura 1.** Andamento degli argomenti trattati

Figure 2. Topic identified by Structural Topic Modeling.
Figura 2. Argomenti identificati dallo structural topic modelling.

curred in December 2020, at the end of the second wave. Finally, an almost constant number of publications (about 1,000) was maintained from February to August 2021. Figure 1 depicts the temporal trend of publications.

LitCOVID identifies not only COVID-19 articles, but also organizes the relevant literature into curated categories or research topics. The categories foreseen are mechanism to which are assigned papers associated to MeSH terms like: • metabolic networks and pathway, pathogenesis, pathologic process; • transmission (based on MeSH terms disease transmission, infectious, replication); • treatment (randomized controlled trials, therapeutics, therapy); • diagnosis (diagnosis, diagnostic imaging, diagnostic equipment, diagnostic services etc.); • prevention (prognosis, treatment outcome, prevention and control); • case report (ambulatory care facilities, case reports, clinic, patient management); • forecasting (trends, forecast).

These broad categories are intertwined, meaning that the same article can belong to more than one and, thus, they are not helpful to get a sense of how publications pattern

EDITORIALE

changed over time. With this objective, we further analysed 13,825 citations, out of the 19,331 found, whose abstract was available and applied a Structural Topic Modeling to identify more granular topics.

Topic modelling is a widely-used method to uncover latent topics within text. The most popular approach to the topic modelling is the Latent Dirichlet Allocation, a probabilistic method that assumes each topic is a distribution over words and each document is a mixture of corpus-wide topics.[7] Unfortunately, Latent Dirichlet Allocation works on some restrictive assumptions, like topics within a document which are independent of one another and that can be modelled entirely based on the text of the document.[8] Structural Topic Modelling is a different approach that extends Latent Dirichlet Allocation allowing to incorporate the document metadata, like the date of publication, into the data generating process of the corpus.[9]

Based on a dictionary of 18,140 words, we identified ten topics using the Structural Topic Modelling, which we characterized using the words with the highest probability and the words that are both frequent and not shared by other topics (frequent and exclusive words). (Figure 2). At the beginning of the epidemic in Italy, several publications focused on the novel Coronavirus SARS-CoV-2 biology and its potential origin and evolution (Topic 8). Meanwhile, other studies appeared on diagnostics about the serological and nucleic acid method to testing individuals (Topic 3). A large corpus of research articles investigated public-health measures (behavioural and policy-based) to contain the spread of pandemic (Topic 9), patients' management (Topic 7); forecast the unfolding of the epidemic, including the description of the risk factors for infection and transmission, estimation of parameters like the reproductive number or incubation period (Topic 4); and the study of association with COVID-19 and mortality of resident population (Topic 2). In this regard,

the Italian National Statistics Institue released mortality data of the first quarter of 2020 in May 2020 to analyse the impact of COVID-19 epidemic on the mortality of the Italian resident population.[10]

As enough data on patients were collected, studies were published on: how the infection presents and progresses clinically (Topic 1); prognostic risk factors associated with severity and mortality (Topic 5). Several publications focused on treatments and studies evaluating the safety and effectiveness of vaccines, including reviews of trial protocols (Topic 6). Finally, studies on the impact of COVID-19 on mental health can be identified to belong to Topic 10. Notably, the structural topic model found a correlation between topic 2 and 5; topic 5 and 1; and topic 1, 6, and 8. Moreover, a correlation was also found between topic 4 and 9, and topic 9 and 10; whereas topics 7 and 3 are standalone topic, meaning that they can be identified separately.

We want to stress that the figures we have reported represent a snapshot of the Italian scientific publishing on COVID-19, based on the documents indexed in PubMed and Scopus only. Thus, we have not identified all published materials. Even if underestimated, this dramatic increase in publications brings out several challenges and poses some questions. The pandemic publishing race with quickened times from data collection to publication raises concerns about the quality of much of this research and peer review. Nevertheless, in a time marked by unprecedented information explosion, it becomes of paramount importance to use innovative solutions to help researchers find the information they seek on the one hand. On the other hand, it is crucial how effectively disseminate a large volume of research, not overlooking information professionals needs to incorporate evidence and findings into their news stories and narratives.

Conflicts of interest: none declared.

REFERENCES

1. Brown A, Horton R. A planetary health perspective on COVID-19: a call for papers. Lancet Planet Health 2020;4(4):e125.
2. Chen Q, Allot A, Lu Z. LitCOVID: an open database of COVID-19 literature. Nucleic Acids Research 2021;49(D1):D1534-40.
3. Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: The COVID-19 Open Research Dataset. ArXiv 2020. doi: arXiv:2004.10706v2. Available from: https://arxiv.org/abs/2004.10706v2
4. Squazzoni F, Bravo G, Grimaldo F, García-Costa D, Farjam M, Mehmani B. Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals. PLoS One 2021;16(10):e0257919.
5. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet

2018;392(10159):1736-88.
6. Aviv-Reuven S, Rosenfeld A. Publication patterns' changes due to the COVID-19 pandemic: a longitudinal and short-term scientometric analysis. Scientometrics 2021;126(8):6761-84.
7. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003;3:993-1022.
8. Blei DM, Lafferty JD. A correlated topic model of Science. Ann Appl Stat;1(1):17-35.
9. Airoldi EM, Blei DM, Fienberg SE, Xing EP. Mixed membership analysis of high-throughput interaction studies: Relational data. arXiv 2007. doi: arXiv:07060294. Available from: http://arxiv.org/abs/0706.0294
10. ISTAT, ISS. Impatto dell'epidemia COVID-19 sulla mortalità totale della popolazione residente primo trimestre 2020. 04.05.2020. Available from: https://www.istat.it/it/files//2020/05/Rapporto_Istat_ISS.pdf