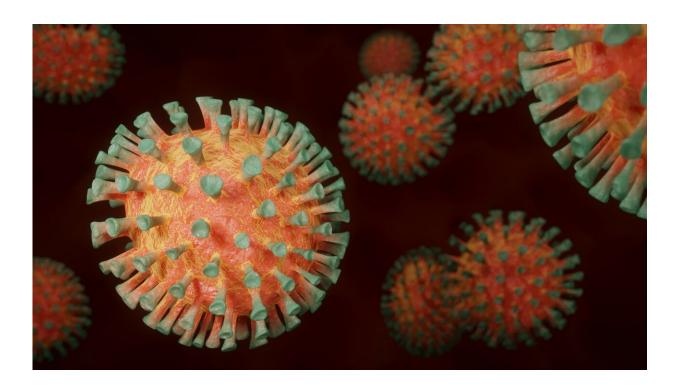# Analysis of COVID-19 publications identifies research gaps

September 17 2020

Credit: Pixabay/CC0 Public Domain

Since the start of the COVID-19 pandemic, scientific and medical journals have published over 100,000 studies on SARS-CoV-2. But according to data scientists who created a machine-learning tool to analyze the deluge of publications, basic lab-based studies on the microbiology of the virus, including research on its pathogenesis and mechanisms of viral transmission, are lacking. Their analysis appears

September 16 in the journal *Patterns*.

"In a crisis like this pandemic, we would expect research outside the lab to happen at a faster pace than lab research," says first author Anhvinh Doanvo a volunteer data scientist with the COVID-19 Dispersed Volunteer Research Network. "Nevertheless, the relative lack of lab-based studies seems to be unique to SARS-CoV-2, compared to other human coronaviruses. This shortage of lab-based research means that the scientific community may miss key aspects of the virus that could impact our ability to contain this pandemic and to counter future ones."

The investigators used research abstracts obtained from CORD-19 (COVID-19 Open Research Dataset). CORD-19 is updated daily and includes peer-reviewed studies from PubMed Central, as well as preprints from bioRxiv and medRxiv. At the time they conducted their first analysis at the end of May, the dataset included more than 137,000 studies. The analysis was later updated with data through July 31.

The team used two computational methods to analyze the data. The first was dimensionality reduction, which helps to find big patterns across many documents, such as abstracts from scientific studies, and to identify trends based on those patterns. The second method, topic modeling, allowed them to group the documents into different topics and to compare research on SARS-CoV-2 to research on other coronaviruses. Unlike previous studies that have focused only on keywords, both of these tools enabled them to review the full text of the abstracts.

"Broadly speaking, we found that the research community has produced a lot of work on the clinical manifestations of the virus, epidemiological models of its spread, and other work based on data collected from the field," says senior author Maimuna Majumder, a computational epidemiologist at Harvard Medical School and Boston Children's

Hospital's Computational Health Informatics Program.

The researchers also note that research has changed over time, with an acceleration in studies examining public health responses, clinical issues related to the virus, the societal impact of the outbreak, and how the disease spreads across populations, while reporting on the status of the outbreak has begun to plateau. "This is a positive development, as it indicates that the scientific community has transitioned from the role of a passive observer of the virus into a group studying ways to fight its spread," Majumder says.

"But basic microbiological research has been slow to pick up the pace, leaving potential knowledge gaps in its wake," Doanvo says. "It's possible that stronger resourcing in these time- and resource-intensive efforts would better enable the scientific community to respond quickly to this virus."

The researchers hope this analysis will help raise awareness about the importance of prioritizing lab-based studies on SARS-CoV-2 moving forward. They plan to conduct another analysis of scientific studies in about a year, using the tools they have already developed.

 More information: Anhvinh Doanvo et al, Machine Learning Maps Research Needs in COVID-19 Literature, *Patterns* (2020). DOI: 10.1016/j.patter.2020.100123

Provided by Cell Press