NLM Leverages Data, Text Mining to Sharpen COVID-19 Research Databases

Data-mining techniques will allow health researchers to sharpen COVID-19 literature and clinical trial database search results.

Melissa Harris
Mon, 05/11/2020 - 15:33

The National Library of Medicine is leveraging its database resources and artificial intelligence capabilities to rapidly provide COVID-19 literature and resources to researchers and scientists as the world races to understand and respond to the pandemic.

Photo Credit: xavierarnau/iStock
The White House in March tapped NLM, under the National Institutes of Health, to join a public-private partnership called the COVID-19 Open Research Dataset (CORD-19) to develop data-mining techniques that could help the science community answer critical questions pertaining to COVID-19. Leveraging its existing infrastructure and establishing processes for content submission, NLM has quickly brought access to COVID-19 literature and clinical trial content on its PubMed Central (PMC) and ClinicalTrials.gov databases.

“As of May 1, about 46,000 articles had been deposited by publishers to PMC or updated in PMC to have a license that allows for text and data-mining, of which more than 5,600 articles specifically focus on the current novel coronavirus,” said NLM National Center for Biotechnology Information Acting Director Stephen Sherry. “Some 49 publishers are now included in the PMC COVID-19 initiative.”

Within the first few weeks since launching the project, PMC saw significant COVID-19 download and data-sharing rates, said PMC Program Manager Kathryn Funk in an NIH webinar. As part of the project, Funk’s team worked to standardize submission data in a machine-readable format.

“The early results have been encouraging,” Funk said. “Articles in the Public Health Emergency Collection and PMC were retrieved more than 2 million times in the first two to three weeks of the initiative, and the CORD-19 dataset has been downloaded more than 75,000 times at this time. It’s our hope that through expanded access and machine learning, NLM will be able to help accelerate scientific research on COVID-19.”

NLM has also leveraged ClinicalTrials.gov’s existing infrastructure to scale up and provide quick access to information about trials related to COVID-19. Teams conducting trials around the world can submit standardized and structured information about their trials directly through an online submission portal called the Protocol Registration and Results system, where trial information is then posted to ClinicalTrials.gov within a couple of days of initial submission, Sherry said.

The data standardization and structure are critical to enabling AI technologies like machine learning and natural-language processing, which can help users more effectively mine and analyze the databases’ resources and literature to generate knowledge and support research that assist in responding to COVID-19, Sherry said.
“ClinicalTrials.gov also leverages NLM resources such as the biomedical vocabularies and standards integrated in the unified Medical Language System (UMLS) to support its search capabilities,” Sherry said, citing the database’s complete list of registered COVID-19 studies. “Users can filter the search results further by different study design characteristics, recruitment status, location information and other factors to identify trials of interest. All of these search capabilities are also available through the ClinicalTrials.gov API.”

Sherry likened the ClinicalTrials.gov infrastructure as an “information scaffold” for discovering information about clinical trials, as the platform applies unique identifiers called National Clinical Trial (NCT) numbers to each trial so that individuals can label and identify trials.

“As a result, different resources with information about particular trials can be linked and discovered through the use of unique NCT numbers, [such as] ClinicalTrials.gov records, press releases, journal articles, protocol document[s], informed consent forms, systematic reviews, reports, regulatory documents, individual participant-level data,” Sherry said.

Creating an open data repository ecosystem like ClinicalTrials.gov requires integrating different data contributors in a way that enable interoperability and usability of data, said NIH Director of Data Science Strategy Susan Gregurick, who helped establish the agency's data science office in 2018.

“NIH strongly encourages open-access, data-sharing repositories as your first go-to choice when you’re looking for a repository to share your data and your information,” Gregurick said during an agency webinar last month.

Although NLM had already pledged to modernize its databases, support data-driven science, collaborate with relevant stakeholders and build a future-ready workforce in its strategic plan, such as the multi-year effort to overall modernize ClinicalTrials.gov, COVID-19 has sparked a number of new data-backed initiatives and digital resources around COVID-19, said Sherry and Gregurick.

These are not just on PMC and ClinicalTrials.gov, but also on new platforms and resources, including:

-
LitCovid, a COVID-19-specific open-resource literature hub that curates and disseminates a constantly growing comprehensive collection of international research papers relevant to public health. “This resource builds on NLM research to develop new approaches to locating and indexing the literature related to COVID-19, including a text classification algorithm for screening and ranking relevant documents, topic modeling for suggesting relevant research categories and information extraction for obtaining geographic locations found in the abstract,” Sherry said.

- COVID-19 genetic sequence information additions to GenBank, the world’s largest genetic sequence database that released the first COVID-19 sequence to the public Jan. 12 and the first sequence collected in America in collaboration with the Centers for Disease Control and Prevention Jan. 25. “As of April 9, we have 579 SARS-CoV-2 sequences from 26 different countries publicly available,” Sherry said, adding that NLM has create a data hub on GenBank for individuals to search, retrieve and analyze COVID-19 sequences that have been submitted.
- The Sequence Read Archive, an 14-petabyte archive of high-throughput genetic sequence data that as of February became available on commercial cloud-computing platforms, which Sherry said significantly expanded the discovery potential of the data to help identify mutational patterns and inform drug and vaccine development.

View printer friendly version
COVID-19
coronavirus
data
data-mining
text-mining
machine learning
natural language processing
Artificial Intelligence
AI
NLM
National Library of Medicine