STAT

Hospital records hold valuable Covid-19 data. Making it usable is time-consuming work

By Rebecca Robbins @rebeccadrobbins

May 27, 2020

Reprints

To help scientists around the globe study Covid-19, researchers in Boston have shared genetic and other clinical data from thousands of patients with an international consortium. That data includes information from dozens of people with Covid-19, who had donated blood samples and opened up their medical records before the pandemic. *ALEX EDELMAN/AFP via Getty Images*

As tens of thousands of people worldwide test positive for Covid-19 every day, researchers are beginning to accumulate a trove of data from patients' medical records that will be critical to getting a better handle on the biology of the disease.

But even in the most advanced electronic health record systems, it's a painstaking process to turn the information in a Covid-19 patient's chart into a format that researchers can actually use.

That's playing out at Mass General Brigham, the Boston hospital network where researchers are toiling away to extract, clean, and check Covid-19 data so that the information can eventually be fed to researchers around the globe. Completing the task for a single patient can take eight hours. It's a reminder of just how much data wrangling must still be done by humans — even in an age in which medical research is increasingly done with the help of artificial intelligence and other sophisticated tech tools.

"I think people have no idea the amount of time that it takes," said Ann Woolley, an infectious disease physician who's leading one of the teams within the health system working with the Covid-19 patient data.

Related:

Pharma panics as Washington pushes to bring drug manufacturing back to the U.S.

At the biobank run by Mass General Brigham, which until recently was known as Partners HealthCare, researchers are packaging patients' data and sharing them with an international consortium of genetics researchers known as the Covid-19 Host Genetics Initiative. The Mass General Brigham effort is one of more than 150 programs worldwide that make up the consortium, which aims to unravel the unknowns about Covid-19, including why some patients become far sicker than others.

The team in Boston has already shared genetic and other clinical data from thousands of patients, including dozens with Covid-19, who had donated blood samples and opened up their medical records before the pandemic. But they're also working to incorporate data from new patients too, especially ones who've been hospitalized for Covid-19.

It could prove to be an important contribution.

"As the numbers mount, the way that a DNA biobank is tied into an EHR is extraordinarily powerful for teasing apart all sorts of questions about clinical phenomenology, pathophysiology, and eventually even treatment possibilities," said Robert Green, a medical geneticist and physician who's one of the leaders of the Mass General Brigham effort.

So far, the Mass General Brigham program has shared genetic data from 108 Covid-19 patients with the consortium, as well as data from more than 31,000 controls who are presumed not to be infected, but have not yet taken a Covid-19 test. Their information was already on file, because they had previously volunteered to have a blood sample collected and their data analyzed through the health system's biobank. A small fraction of them ended up with a Covid-19 diagnosis, flagged to researchers when the positive test showed up in their medical record.

The Mass General Brigham researchers expect to soon share data from many more patients. They plan to pull data from people already in the biobank who test positive for the virus in the coming days and months. They'll also share information from people not currently in the biobank who test positive for the virus at the health system's two hospitals — Massachusetts General Hospital and Brigham and Women's Hospital — and volunteer to provide a blood sample and their clinical data as part of a Covid-19 study being run in collaboration with the biobank.

Several hundred patients hospitalized for Covid-19 at the Brigham have already volunteered for the effort, according to Woolley, who's leading the work there. (Those and other new blood samples not already in the biobank have not yet been sequenced, but plans are in motion to do so.)

Woolley has assembled a team of eight research assistants who have spent weeks manually going through the charts of those patients in search of relevant information, which they then input into a software program. (Software later converts those inputs into a binary format — think 0s and 1s — that genetics researchers can more easily work with.) Those data cover patients' demographics, risk factors, exposures, medications, chest scans, lab test results, and co-infections. They also address essential questions: Did the patient require intensive care? How about intubation?

All told, the research assistants are on the hunt for some 800 possible data points, though they're never all relevant for a single patient. Often, the most helpful pieces of information about a patient are in what's known as unstructured text, such as a doctor's quickly typed notes, which need to be interpreted and formatted. Another layer of complexity: Patients who were

referred from another hospital must have their records from the other institution pulled in.

Related:

Two years of halting progress and high turnover preceded Atul Gawande's exit as Haven CEO

For a relatively simple case, scraping these data for a hospitalized Covid-19 patient could take a research assistant a minimum of three hours — not including extra time for quality control, Woolley said. But few cases are simple, and sometimes a clinician needs to get involved, such as in cases in which a patient experienced complications. In those situations, a medical resident, a fellow, or Woolley herself will go through the EHR data to assess exactly what happened, and then they'll double check each other's work.

All of those factors mean that scraping a single Covid-19 patient's data can equate to an entire day's worth of work, Woolley said. There are plenty of software tools on the market to automatically scrape EHR data into a format that's easier for genetics and artificial intelligence researchers to work with. But it's essential at this stage to have humans do the careful fact-checking and medical adjudication work, so that they can flag issues that could be addressed if the process is later scaled electronically, Woolley said.

The crucial human labor involved in the Mass General Brigham effort isn't just limited to scraping the EHR data. At the Broad Institute in Cambridge, Mass., research scientist Josep Maria Mercader has been cleaning and curating the genetic variants from patients whose samples had already been in the biobank, including by sorting patients in terms of their ancestry. The various data streams are later fed to postdoc Yen-Chen (Anne) Feng, who runs statistical analysis and formats the data in a way that's consistent with all the other programs in the international consortium.

The Mass General Brigham team will share its next batch of data with the consortium within the next couple weeks.

About the Author Reprints

Rebecca Robbins

San Francisco Correspondent

Rebecca covers the life sciences industry in the Bay Area. She is the co-author of the newsletter STAT Health Tech.

rebecca.robbins@statnews.com
@rebeccadrobbins