MARCO MANTOVANI/GETTY

Health-care workers check a patient's electronic health records on a COVID-19 ward in Cremona, Italy.

HEALTH DATA FOR ALL

Medical records can be tricky to access because of privacy and variability, but data-sharing efforts are unlocking their potential. **By Jyoti Madhusoodanan**

For the gastrointestinal condition known as ulcerative colitis, some physicians recommend using a particular drug twice a day, others, three times. But which protocol is the best way to help people with the condition to avoid surgery? Instead of launching a clinical trial, Peter Higgins, a gastroenterologist at the University of Michigan at Ann Arbor, examined the data.

Many health systems in the United States export clinical data from electronic health records (EHRs) into repositories known as health data warehouses for institutional use by researchers, Higgins says. Working with the University of Michigan's health informaticians, he identified and compared people on the two protocols. The scientists found that giving people the drug three times a day seemed to result in fewer operations (J. A. Berinstein *et al. Clin. Gastroenterol. Hepatol.* **19**, 2112–2120; 2021).

"A lot of good research is dropped because there's a huge learning curve to using these systems."

Such searches are complex because the underlying records are so variable, Higgins says. "It's a little bit of a needle in a haystack hunt," he explains, because the data are not standardized.

The variations in data formats, combined with regulations to protect patient privacy, make working with data warehouses challenging. Access to a repository is usually restricted to people within an institution, and international data protections can prove even more

daunting. "The data are just truly not interoperable across health systems," says Melissa Haendel, a data scientist at the University of Colorado Anschutz Medical Campus in Aurora.

Even for those trained in health informatics, learning how to work with such data is not trivial. "A lot of good research that could be done on the EHR is dropped because there's a huge learning curve to using these systems," says Charisse Madlock-Brown, a health-informatics researcher at the University of Tennessee Health Science Center in Memphis. Small institutions also often lack a health-informatics team that can assist biologists wanting to use these repositories, she says.

Data links

Spurred by the COVID-19 pandemic, researchers have begun to aggregate data from individual institutions in national repositories that are more accessible. In the United States, the National COVID Cohort Collaborative (N3C) is the largest patient-privacy-limited data set in the country's history, says Haendel, who co-leads the effort. Supported by the US National Institutes of Health, N3C encompasses data from more than 70 institutions and holds patient-level information for 13 million individuals. The data include EHRs, imaging scans and genomic sequences of viral variants, all of which are described using a common data model (E. R. Pfaff *et al. J. Am. Med. Inform. Assoc.* **29**, 609–618; 2022).

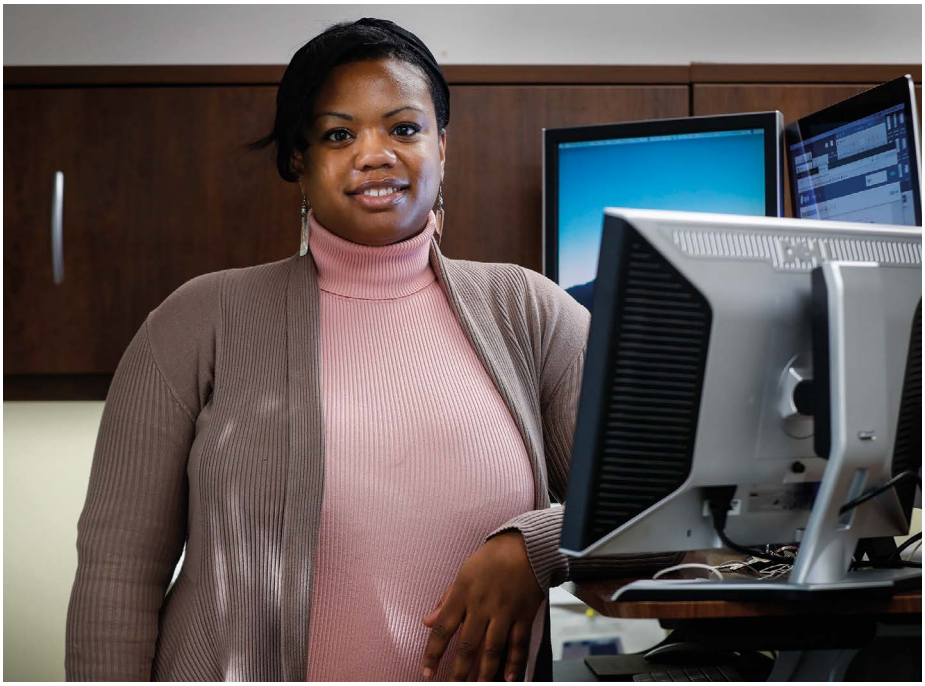
Equally important, Haendel says, is the collaboration's effort to link data across systems while preserving patient privacy. "If, for example, a patient went to a specialist at one hospital and a general-practice physician at another institute, we could connect those records to understand that person's health outcomes."

Likewise, the non-profit organization Health Data Research UK (HDR UK) in London launched its Innovation Gateway platform in 2020 to curate health data sets and a suite of analysis tools. "COVID-19 has been a good accelerator for this work," says Susheel Varma, chief technology officer at HDR UK. Such centralized repositories enable researchers to access a broader cohort of patient data.

Both HDR UK and N3C encourage researchers to work within the repository's digital workspace, where extra protections mean the data can be less anonymized to provide richer information for analysis, such as by including geographical information or dates. Given the need for speed with pandemic research, "we were able to have quite a permissive environment for people to use data", Haendel says of the N3C effort.

Researchers who want access to N3C data must sign data-use agreements at their home institutions and complete training on how to securely handle data from human participants.

Researchers at foreign institutions are able to access fully de-identified patient data from N3C, whereas citizen scientists can access only 'synthetic' data. (These are statistically similar to real patient information, but are computationally derived to protect privacy.) Researchers who wish to access any data limited by health privacy legislation, which include location information and important dates, require extra approvals from institutional review boards and the N3C's data-access committee, Haendel says. These privacy-limited data are restricted to US-based scientists for now.



MARK WEBER/*THE DAILY MEMPHIAN*

Researcher Charisse Madlock-Brown uses patient-level information from the US National COVID Cohort Collaborative data set to study the social determinants of health.

> **"COVID-19 has been a good accelerator for this work."**

The HDR UK effort also exercises jurisdictional control over certain kinds of data, Varma says. For international researchers, Varma recommends teaming up with UK-based research organizations. "It's not to prevent access but to contextualize the research, because the data are collected for national benefit," Varma says.

Quality control

Madlock-Brown uses N3C to study social determinants of health as risk factors for COVID-19 outcomes. Although the database is enormous, Madlock-Brown says, "a ton of data" does not necessarily mean the information is of high quality. To minimize errors caused by data bias and to ensure data quality, Madlock-Brown recommends working with informaticians who are familiar with the data sources to understand how the information is tabulated and organized — a consideration even at smaller data warehouses.

For example, Higgins points out that the diagnostic codes that describe a person's condition can be surprisingly inaccurate. When studying ulcerative colitis, he included fields for prescriptions and treatment duration in his query to rule out similar conditions such as ischaemic or infectious colitis. "You have to have a really good sense of what you want," he says. And researchers should check the results of a database query once they come back, Higgins adds. "It's really helpful to have, say, 20 patients who should be in the results and another set who should not be on the list," he says.

In a similar way to positive and negative controls in a bench experiment, these controls can help to refine a query to maximize the sensitivity and specificity of data analyses. "These are similar issues with all observational research," Madlock-Brown says.

The advantage of collaborative data warehouses such as N3C, Madlock-Brown says, is that biologists can team up with informaticians and others to understand these caveats before they start — even if their home institutions don't offer informatics support. "In that way, N3C is an equalizer," she says. "As long as your institution can get a data-use agreement, you have access to these resources. I can't think of another example in academia where you can access this much with no need to pay."

Jyoti Madhusoodanan is a science writer in Portland, Oregon.