Credit: Jennifer Kosig / DigitalVision Vectors / Getty

How patient data underpin COVID-19 research

Tracking the coronavirus SARS-CoV-2, developing treatments and testing vaccines have relied on access to the health data of millions of people. This massive trove could be used to understand other diseases, but not everyone is convinced.

Marion Renault

At the outset of the COVID-19 pandemic in early 2020, scientists and clinicians faced a daunting list of desperately needed answers. What are the symptoms of COVID-19? How can we detect the virus? How does it spread? How should we treat it and, someday, vaccinate against it?

The best means for settling these questions was to conduct research on an unprecedented scale and at an accelerated pace. "In times of a health crisis, things should be done differently," Tobias Kurth, an epidemiologist at Charité, a research and medical institute in Berlin, tells *Nature Medicine*. "If resources are put together, then science can move much, much faster." And it did.

Size is everything

In the United Kingdom, the RECOVERY Trial rapidly enrolled tens of thousands of patients at hospitals nationwide to test potential therapies for COVID-19. Since March 2020, RECOVERY has uncovered the life-saving potential of monoclonal antibody cocktails, the efficacy of the steroid dexamethasone and the interleukin antagonist tocilizumab, an immunosuppressant. RECOVERY has also demonstrated that certain drugs, including hydroxychloroquine, offer no benefit for seriously sick patients with COVID-19.

Rather than starting a clinical trial from scratch and wading through years of design and bureaucracy, the super-sized research project, with 41,000 participants enrolled so far, instead docked into the pre-existing architecture of the UK National Health Service (NHS). "By doing less, you can do more," says Martin Landray, a University of Oxford physician, epidemiologist and data scientist who co-led the RECOVERY Trial.

RECOVERY is just one example of the streamlined, real-time, large-scale research projects using patient health data that helped deliver answers—and, in turn, treatments, vaccines and virus-surveillance systems. For researchers, accessible patient datasets of this size was a long-time dream—made possible by the nightmare of COVID-19.

The pandemic spurred University of Edinburgh neurologist and epidemiologist Cathie Sudlow to tap into a novel dataset encompassing demographics, COVID-19 diagnoses, cardiovascular outcomes, death registries and hospital data of 54.4 million people. In all, the dataset represents more than 96% of the English population. "Answers to really important questions just

wouldn't be possible without a dataset of this size," she says.

Too much of a good thing

But some patient-privacy activists say that there is a fine line between appropriately facilitating public-health research in a global emergency and seizing the opportunity for overreaching data grabs. "[The public is] in a bit of a panic," says Phil Booth, founder of the advocacy group medConfidential, "and it has got our guard down."

Experts agree that the pandemic has thrown into relief the often blurry lines that govern the use of patients' data and shape the public's attitude toward this: the line between individual care and representative epidemiology; the line between routine medical care and consented research studies; and the line between the deeply personal and the profoundly public.

During this pandemic, the will of government agencies, healthcare providers, private industry and the public came together to transform lakes of patient data into usable streams of life-saving information. But what will happen to that body of data once there is no longer a public-health crisis? Landray says he hopes to keep seeing patients' data that are routinely collected during the course of medicine embedded into large, streamlined trials. "We've stretched a piece of elastic between where we are and where we were," he says. "We just have to decide which end to let go of."

Remove biases

Researchers say the advantages of these large patient databases are manifold and undeniable.

For one, clinical trials have historically over-enrolled white, male patients while under-representing women, people of color, immigrants and people who are too sick or cannot afford to volunteer as research participants. Massive data pooling allows passive, rather than active, participation in medical research. "If you're looking at the whole population," says Chris Carrigan, of DATA-CAN, a UK-wide cancer-research hub delivered by Health Data Research UK, the usual biases that plague clinical trials are not an issue, he says. "All the data [are] there." The result is a more representative dataset that produces findings that are more broadly applicable across demographics.

Large datasets also offer a better view of subgroups and rare conditions.

In a study of a condition that affects 1 in 10,000 people, for example, even a robust, 1,000,000-person dataset would contain a subpopulation of only 100 people with that condition. "The more people you can enroll, the more power we have to detect some signal," Kurth says.

The monumental will, coordination and resources necessary to pull together such large data collection did not happen with great frequency before 2020, when the pandemic hit. "The great enabler has, of course, been COVID," Sudlow says. "It's been everybody's problem," she adds, so everybody has wanted to be part of the solution.

Failures from fragmentation

In the UK alone, unprecedentedly large datasets helped scientists study who was at greatest risk of dying of COVID-19, the effects of post-COVID-19 syndrome on internal organs, and the potential adverse effects of vaccines. Real-time data from hospitals revealed that a majority of people with suspected cancer were not getting proper follow-up and that four of ten patients with cancer were not able to get treatment. The broad body of COVID-19 research in the UK is largely thanks to the NHS' national scope and informational unity—a global rarity. "There are very few countries in the world that can do [data linkages] currently, at this kind of scale," Sudlow says.

In the United States, for example, medical records are fragmented and siloed across a mix of public and private healthcare entities. The health system in the United States is also regulated by varying local, state and federal laws. For the past year and a half, this fragmentation has made it extraordinarily difficult to connect demographic information with vaccine registries maintained by states, SARS-CoV-2 genome sequencing conducted at academic centers, data collected by commercial pharmacies such as Walgreens and CVS, and patient records housed by for-profit, non-profit, public and private hospitals and clinics.

"We have a very fractured health system," says Vanderbilt University biostatistician Brad Malin. "So you're not getting a complete view of what's going on in the country. It's a bit of a free for all."

During the pandemic, the US National Institutes of Health built a centralized solution. The National COVID Cohort Collaborative, or 'N3C', has collected and curated medical records from 6.5 million patients so far, and grants access to the linked dataset on a case-by-case basis. So far, research from N3C has resulted in a dozen research articles that examined health disparities during the pandemic, investigated risk factors in patients with COVID-19 and cancer, liver disease or infection with HIV, and described a machine-learning model that predicts disease severity through the use of clinical data collected within 24 hours of hospital admission.

Researchers have been asking for this kind of centralized collaboration for a long time, Malin says. "It's not a paradigm shift," he says. "But it's being done on a scale that's never been seen before."

Other countries have also linked existing databases with COVID-19 surveillance to create new information ecosystems accessible to researchers.

SCIFI-PEARL, Sweden's new linked database, couples almost all of the country's known patients diagnosed with COVID-19 with a large control group for comparison. Some researchers have begun to connect SCIFI-PEARL with disease data for asthma, cardiovascular disease and chronic obstructive pulmonary disease, to sharpen its investigational power.

Belgium rapidly set up several national surveillance systems to complement hospital data, which helped researchers find that hospitalized patients with malignant tumors had a greater risk of dying within 30 days of a COVID-19 diagnosis. In France, a national database that collects discharge summaries for all admissions to public and private hospitals was used to demonstrate that COVID-19 posed a far greater risk for serious respiratory complications than influenza did. When Mexican scientists pooled data from 482,413 outpatients across 26 states, they found that logistics workers and delivery drivers were especially vulnerable to infection.

International collaboration

Many of these health databases transcend national borders.

Since March 2020, COVerAGE-DB, an open-access database, has gathered data on COVID-19 cases, deaths and testing from 108 countries across Europe, South America, North America, Africa, Asia and Australia. The database has contributed to at least 15 studies—including a UNICEF effort to monitor the pandemic's burden on children around the world, and a United Nations project focused on the elderly.

Observational Health Data Sciences and Informatics (OHDSI) is an international research database launched well before COVID-19, in 2014. OHDSI now hosts records for 600 million unique patients across 30 countries and has been used to study diabetes, adverse drug reactions, opioid addiction and lung disease. During the pandemic, researchers made use of its scale to study health complications in patients with COVID-19 and autoimmune diseases and in hospitalized children and adolescents. Another study analyzed the incidence of millions of adverse medical

events in eight countries, providing an international baseline against which to compare the rate of adverse events from COVID-19 vaccines.

Beyond COVID-19

Some researchers look enviously at the large datasets developed for COVID-19. There are hopes to generate the same volume of data from patients with other leading diseases, such as dementia, diabetes and cardiovascular disease—which altogether claimed 18.2 million lives, or one third of all deaths globally, in 2019. So far, COVID-19 has caused about 4.3 million deaths. "COVID is one of many pandemics," Sudlow says. "We have been sitting on pandemics for years and not using these kinds of resources."

Carrigan says isolated efforts to pool together datasets in cancer research have led to breakthroughs. By gathering and analyzing data leading up to diagnosis, during treatment, after remission and after death, scientists have learned far more about the progression of cancer and the effectiveness of screening and treatments. "Having all that [data] together revolutionized what you could see," Carrigan says. "That was fundamental."

The International Cancer Benchmarking Partnership combines data from Australia, Canada, Denmark, Ireland, New Zealand, Norway, Sweden and the UK to research cancer survival, incidence and mortality. The large pool of data offers almost unparalleled statistical strength. Recent research on survival in colon, ovarian and esophageal cancers, for example, benefitted from sample sizes of 264,000, 58,000 and 185,000, respectively.

The next big priorities for health data in cancer research include how to spot cancers earlier and how to address the reasons underlying delayed diagnoses, says Mark Lawler, a professor of digital health at Queen's University Belfast.

Heart disease, the leading cause of death globally, has no shortage of clinical research, but many studies are still underpowered. In 2019, when a group at Duke University reviewed the evidence behind more than 6,300 treatment recommendations, they found that less than 10% were based on evidence from multiple, large, randomized clinical trials.

Merging datasets could allow cardiovascular-disease researchers to build algorithms that can recommend targeted and personalized treatments, identify new uses for existing pharmaceuticals and surveil adverse reactions to newly approved drugs.

Within dementia research, linked databases combine information from neuroimaging, patient records and genomics to help improve early detection, before symptoms emerge. Some leaders in dementia research have noted that despite its increasing incidence and burden on society, "dementia receives a disproportionately low amount of funding compared to other disease areas"—all the more reason to pool existing data for research.

Landray says that rare health issues or ones whose evidence bases are "substantially weaker"—such as osteoarthritis, chronic respiratory conditions, mental-health issues, rare genetic diseases and perinatal care—would all benefit from large datasets.

'Disease X' also looms on the horizon, as Kurth points out: "We need to be prepared for the next pandemic, which will definitely come."

Nothing about me without me

In spring 2021, the UK NHS proposed a central medical history database comprising 61 million GP records—and caused an uproar.

Critics dismissed the hasty plan to scrape tens of millions of medical histories (including sensitive information on patients' mental and sexual health and criminal records) as mishandled, "disingenuous and misleading" and, at worst, potentially profiteering. "It is literally a grab for medical histories. It's never been done before," Booth says. "To do this in a rush in the middle of a pandemic—it seems to me to be the case of trying to get something while people are distracted."

In June, dozens of doctors in east London publicly urged their colleagues to refuse to hand over patient data for the database, which would be accessed by academics, health planners and commercial third parties alike.

Some doctors were concerned that the backlash could lead patients to reveal less information during consultations out of fear that sensitive information could eventually make its way, without their explicit knowledge, to researchers or private companies.

There are also fears that the controversy could also taint the data, and prevent the data from being useful for public health. "It will just be a toxic lake of data," Booth says. "People won't want to do research on that."

Missed opportunity

Responding to concerns, the UK government has delayed the planned database and extended the opt-out period. Opponents of the database have drawn comparisons to a 2016 debacle in which the NHS had to scrap another data-sharing initiative, Care.data, altogether after public outcry. "People are weary of exploitation, and weary of overhype and under-delivery," says Natalie Banner, who leads the Understanding Patient Data initiative at the Wellcome Trust, a medical-research charity in the UK. Banner worries about botching a prime opportunity for engaging the public in a conversation about patient data while there is unprecedented awareness, because of the pandemic.

Building transparent, well-protected and ethically managed databases could help strengthen "very fragile" public trust, Banner says. Trust is a currency that is easy to spend and difficult to earn back. "If we get this wrong again, we won't get research uses for another decade," Booth says.

Just because the public was willing to lend their health records for research during a global health crisis does not mean that there is an excess of public trust, or that attitudes toward data sharing have fundamentally changed, Banner and others say.

"Suddenly the idea of being able to collect and pool this data is much, much more tangible," Banner says. "But it's a mistake to assume there's a social license to continue." Landray agrees: "It's our job to address those concerns, not to dismiss them or pretend they don't exist."

Many health-data advocates say the path forward has to involve patients and the public as much as possible. Itzelle Medina-Perea, a researcher at the University of Sheffield, studies public attitudes toward data practices. Her research group has found that in the absence of transparency and engagement, "if people don't know what is happening with their data," Medina-Perea says, "they imagine the worst." The absence of transparency encourages suspicion, regardless of the potential benefits. And when the conversation is cut short by suspicion, then "the fantastic things we could do don't get heard about," Carrigan says, even though "virtually all of it is a good news story."

Failing to involve and engage the public with the collection and use of patient data could have a chilling effect on clinical research far beyond this pandemic's expiration. "We should not take any of the advances for granted," Sudlow says. "There is a risk of slipping back." ❐

Marion Renault
Freelance writer, Brooklyn, NY, USA.

Published online: 14 September 2021
https://doi.org/10.1038/s41591-021-01493-5