MARCO DI LAURO/GETTY

Health-care workers monitor a woman in a COVID-19 recovery gym in Genoa, Italy.

HOW COMMON IS LONG COVID? WHY THERE'S STILL NO ANSWER

Enormous databases do not necessarily allow scientists to solve the mysteries of long COVID.

By Heidi Ledford

Clinical epidemiologist Ziyad Al-Aly has access to a treasure trove that many researchers can only dream of: millions of sets of electronic medical records from the US Department of Veterans Affairs (VA), which provides health care for the country's military veterans.

With these data in hand, Al-Aly, who is based at the VA St Louis Healthcare System in Missouri, and his colleagues have undertaken the challenge of studying long COVID — a condition in which people experience symptoms months after an acute SARS-CoV-2 infection seems to have resolved — and recently published findings[1] that surprised some researchers. The team found that previous vaccination only reduces the risk of developing long COVID after infection by about 15%; other estimates[2] suggest that vaccines halve the risk.

It is the kind of whiplash result that people following long-COVID research have become accustomed to seeing, as data from various studies report discordant results. Differences in how the syndrome is defined, the kinds of data used to study it and how those data are analysed have left both the public and policymakers grappling with disparate answers to basic questions. How frequent is long COVID? And how does vaccination, reinfection or the latest SARS-CoV-2 variant affect the risk of developing the condition?

The answers to those questions can be used to develop COVID-19 policies, but the steady drip-drip of seesawing studies can also cause confusion, says Al-Aly. Having so much uncertainty doesn't engender a lot of trust, Al-Aly adds: "The public does not react very well to saying 'between 15% and 50%'."

Slippery definition

Part of the problem is the definition of long COVID, which has been linked to more than 200 symptoms, the severity of which can vary from inconvenient to debilitating. The syndrome can last for months or years, and has a distressing tendency to reappear, sometimes months after an apparent recovery.

So far, there is no agreement on how to define and diagnose long COVID. The World Health Organization's attempt at a consensus, published in 2021, has not proved popular, and studies continue to use a range of criteria to define the condition. Estimates of its prevalence can range from 5 to 50%.

A study of such a complex condition needs to be sufficiently large to reflect the range of symptoms and the possible impact of characteristics such as age. This is where analyses such as Al-Aly's offer a host of advantages: data from large health-care networks can provide enormous sample sizes. Al-Aly's study of long COVID after a 'breakthrough' infection — one that follows vaccination — included records from more than 13 million people. Although 90% of those people were men, that still left 1.3 million women in the analysis, Al-Aly notes, more than many other studies can muster.

Big-number benefits

These large numbers, as well as the types of data available in some health records, allow researchers to perform complicated statistical analyses to carefully match the demographics of people infected with the coronavirus to an uninfected control group, says Theo Vos, an epidemiologist at the Institute for Health Metrics and Evaluation at the University of Washington in Seattle, who has worked with a variety of data sources to study long COVID.

But there are also drawbacks. "People mistake the size of the study with its quality and its validity," says Walid Gellad, a physician who studies health policy at the University of Pittsburgh in Pennsylvania.

In particular, Gellad worries that studies that rely on electronic health records will be muddied by behavioural differences. For example, compared with someone who does not seek medical care for acute COVID-19, someone who does might be more likely to report long-COVID symptoms, he says.

Medical records and health insurance claims might not reflect a demographically diverse population, says computational epidemiologist Maimuna Majumder at Harvard Medical School in Boston, Massachusetts. This is particularly likely in the United States, she says, where health insurance coverage varies widely. "The number of data points considered is often so large that we mistakenly assume that these data must be representative," she says. "But this isn't necessarily the case."

She also wonders whether studying claims data could lead to an undercounting of cases of long COVID, because many people might not seek medical care for their condition.

Coding lessons

Another issue is how symptoms are recorded in the claims and electronic medical records. Doctors often record codes for several symptoms and conditions, but they rarely list a code for every symptom a patient is experiencing, says Vos, and the choice of codes for a given condition might vary from one doctor to the next. This could lead to differences in whether and how long COVID is reported. "Electronic health records have useful information in them, without a doubt," says Gellad, who

says that the VA study was particularly well designed. "But for answering the question of how common something is, they may not be the best."

Other methods also have their pitfalls. Some studies rely on self-reporting, such as the COVID Symptom Study app developed by King's College London and the data-science company ZOE, also in London. Data from the app showed that vaccination reduced people's risk of experiencing long COVID 28 days or more after an acute infection by about half[2]. But studies in which people voluntarily self-report their symptoms can be biased, because people who have symptoms are more likely to participate, says Gellad. And studies that rely on smartphone apps might not fully capture data from disadvantaged communities.

One particularly useful source of data has been the UK Office for National Statistics (ONS), says Nisreen Alwan, a public-health researcher at the University of Southampton, UK. In May, the ONS reported that the variant of SARS-CoV-2 that people are infected with can affect their risk of developing long COVID. Among double-vaccinated participants, those thought to have had COVID-19 caused by the Omicron BA.1 variant were roughly 50% less likely to develop long COVID symptoms four to eight weeks after infection than were those whose infections were probably caused by the Delta variant. This finding is in line with the results of an 18 June paper[3] based on ZOE data.

Seeking a common thread

Alwan, who has long COVID and has advocated for the collection of data on the condition, praises the ONS study design, which involved enrolling a group of people with careful attention to representing the UK population, and then following up with them to ask about their infection status and symptoms.

Other aspects of study design, such as whether a control group is used, can strongly affect results, says Alwan. But accounting for disparate methods and definitions need not stall research. "That's not something new," she says. "It's something that we had before COVID, for other conditions."

For Al-Aly, the discrepancies among study results are not surprising, nor are they damning. Epidemiologists often weave together evidence from multiple sources of data and methods of analysis, he says. Even if it is difficult to precisely quantify vaccination's effect on long-COVID risk, for example, researchers can look for trends. "You search for the common thread," Al-Aly says. "The common thread here is that vaccines are better than no vaccines."

1. Al-Aly, Z., Bowe, B. & Xie, Y. *Nature Med.* https://doi.org/10.1038/s41591-022-01840-0 (2022).
2. Antonelli, M. *et al. Lancet Infect. Dis.* **22**, 43–55 (2022).
3. Antonelli, M., Pujol, J. C., Spector, T. D., Ourselin, S. & Steves, C. J. *Lancet* **399**, 2263–2264 (2022).

MANY RESEARCHERS SAY THEY'LL SHARE DATA — BUT DON'T

Reasons include a lack of informed consent or ethics approval to share, and misplaced data.

By Clare Watson

Most biomedical and health researchers who declare their willingness to share the data behind journal articles do not respond to access requests or hand over the data when asked, reports a study.

Livia Puljak, who studies evidence-based medicine at the Catholic University of Croatia in Zagreb, and her colleagues analysed 3,556 biomedical- and health-science articles published in a month by 282 journals published by BMC. (BMC is part of Springer Nature, the publisher of *Nature*; *Nature*'s news team is editorially independent of its publisher.)

The team identified 381 articles with links to data stored in online repositories, and another 1,792 papers for which the authors indicated in statements that their data sets would be available on reasonable request. The remaining studies stated that their data were in the published manuscript and its supplements, or generated no data, so sharing did not apply.

But of the 1,792 manuscripts for which the authors stated they were willing to share their data, more than 90% of corresponding authors either declined or did not respond to requests for raw data (see 'Data-sharing behaviour'). Only 14%, or 254, of the contacted authors responded to e-mail requests for data, and a mere 6.7%, or 120 authors, actually handed over the data in a usable format. The study was published in the *Journal of Clinical Epidemiology* (M. Gabelica *et al. J. Clin. Epidemiol.* https://doi.org/h2q8; 2022).

Puljak was "flabbergasted" that so few researchers actually shared their data. "There is a gap between what people say and what people do," she says.

Data-availability statements are of little value because many of the data sets are never actually made accessible, says Valentin Danchev, a sociologist at the University of Essex in Colchester, UK.

Researchers who declined to supply data in Puljak's study gave varied reasons. Some had not received informed consent or ethics approval to share data; others had moved on from the project, had misplaced data or cited language hurdles when it came to translating qualitative data from interviews.

Aidan Tan, a paediatric physician and researcher in evidence-based medicine at the University of Sydney in Australia, says the study demonstrates that persistent barriers stop researchers sharing their data. His own research surveying leaders of clinical trials has found concerns about data privacy, participant confidentiality and data being misused in misleading secondary analyses (A. C. Tan *et al. Res. Synth. Methods* **12**, 641–657; 2021).

Tackling the problem

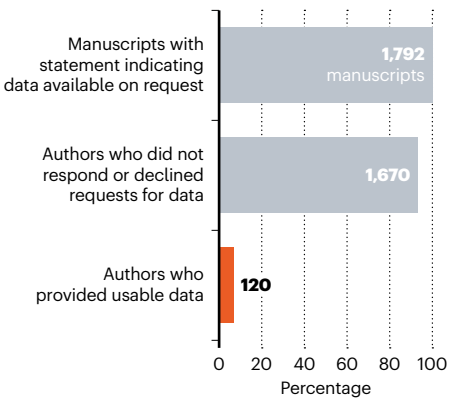
Rebecca Li, who is executive director of non-profit global data-sharing platform Vivli and is based in Cambridge, Massachusetts, surmises that many researchers don't fully understand what data sharing actually entails: that data underpinning manuscripts "should be ready, formatted and available for whoever asks", she says.

To encourage researchers to prepare their data, Li says, journals could make data-sharing statements more prescriptive. They could require authors to detail where they will share raw data, who will be able to access it, when and how.

Funders could also raise the bar for data sharing. The US National Institutes of Health, in an effort to curb wasteful, irreproducible research, will soon mandate that grant applicants include a data-management and sharing plan in their applications.

DATA-SHARING BEHAVIOUR

Of almost 1,800 manuscripts for which the authors stated they were willing to share their data, more than 90% of corresponding authors either declined or did not respond to requests for data. Only about 7% of authors actually handed over data.



SOURCE: LIVIA PULJAK *ET AL.*