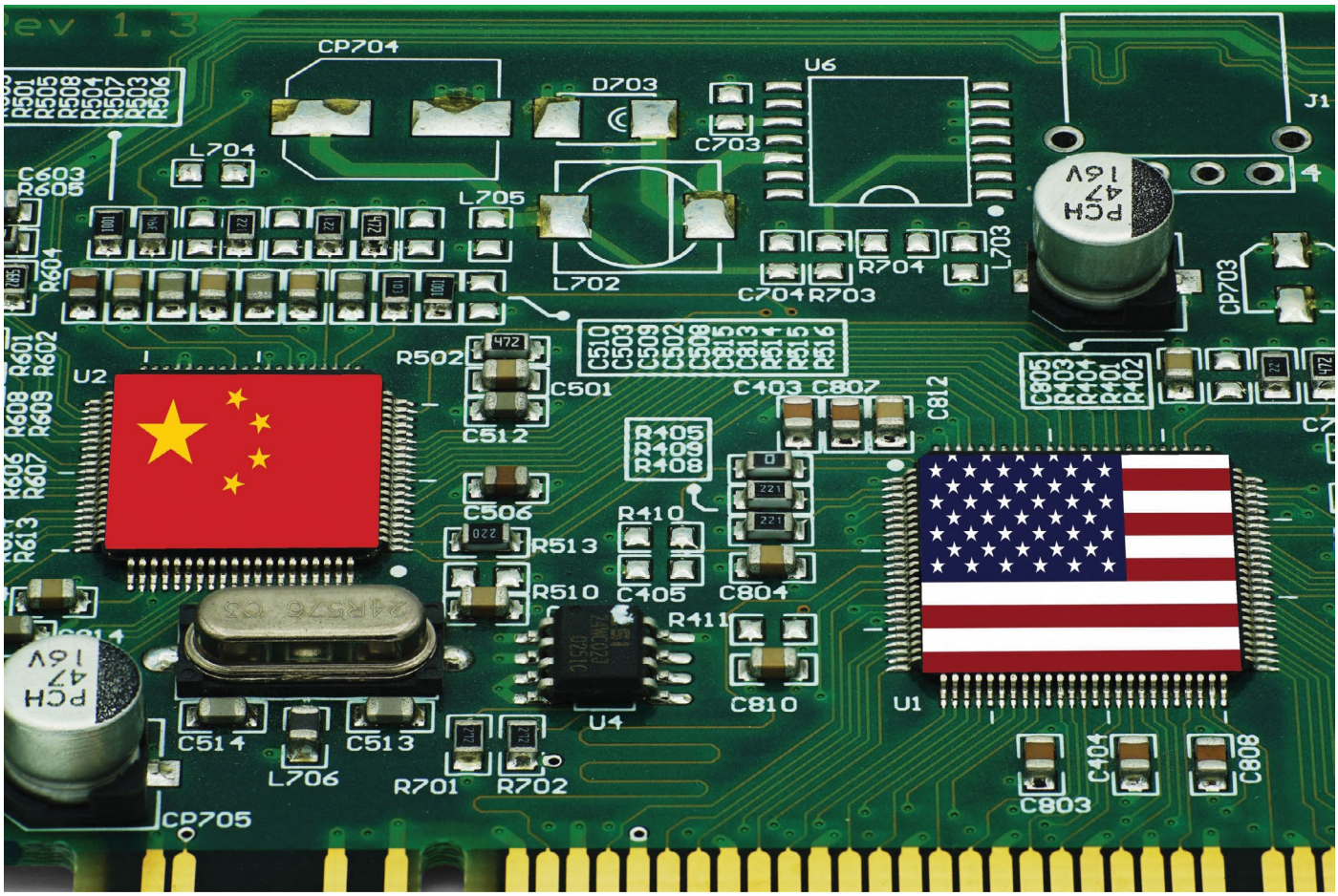
https://doi.org/10.1038/s41591-026-04378-7



WHO OWNS MY HEALTH DATA?

The geopolitics driving artificial intelligence superpowers is reshaping biomedical datasets, and who has access to them. **By Paul Webster**

When geneticist Jingyuan Fu heard that an artificial intelligence (AI) group in China had downloaded a large biomedical dataset her team built in Europe, she felt pride — and a jolt of unease.

"We spent millions on that dataset," says Fu, a professor of systems medicine at the University of Groningen in the Netherlands. "And the Chinese bought the whole thing for around €2,000."

The dataset grew out of a gut microbiome study launched in 2006. Over time, it expanded its scope: "We've been following 167,000 participants, about 10% of the population of the north province of the Netherlands, for 30 years," Fu says. "And we've gathered their transcriptomes, proteomics, metabolomics and different cellular molecular profiles."

In recent years, Fu's group, like many others, has also begun using such data as feedstock for artificial intelligence. The AI group in China that downloaded her dataset had the same goal. "The Chinese wanted all our data," Fu says. "And they also wanted our insights into how to mine it for AI development."

From her perspective, today's global scramble for biomedical data looks increasingly lopsided. "China has collected a huge amount of data," she says. "But their own data sharing and openness is very limited."

Fu's worry is not simply about who gets access (Fig. 1). It is about what happens to medicine when health data — built up over decades by patients, clinicians and publicly funded researchers — become ammunition in an AI arms race.

Farewell open science

Biomedical AI has surged on a simple bet: combine large, diverse datasets and AI models get better — with more accurate predictions and fewer blind spots. Alongside computing power and algorithm innovation, Fu emphasizes, AI is only as good as the data it learns from.

But after decades of policies pushing 'open science', governments are now promoting

CREDIT: RICHARD JONES/ALAMY

'data sovereignty' — the idea that sensitive datasets should remain under national control and foreign access should be conditional.

As nations and companies raise walls around health records, many researchers, such as Fu, fear that algorithms will be harder to validate across different populations, more prone to hidden bias and less likely to benefit the people who contributed their data in the first place.

Sovereign datasets

In Europe, nations are feeling pressure to keep the continent competitive in AI while staying inside strict privacy rules and reinforcing European data sovereignty. In November 2025, the European Commission launched RAISE, the Resource for AI Science in Europe — a virtual institute to coordinate AI resources, including scientific datasets, across member states. These resources will come largely from Horizon Europe, the EU's €93.5 billion research and innovation framework.

According to the Commission, databases such as Europe's Genomic Data Infrastructure and Cancer Image Europe will now be harnessed to serve planned public–private 'AI gigafactories' that will enable advanced AI modeling. Horizon Europe will channel €600 million to secure compute time for researchers.

But data sovereignty has a sharp edge. Fu says the mood in Brussels is defensive. She thinks that European Commission officials are embarrassed about having allowed Chinese AI developers to plunder European biomedical databases, even while China blocks foreign access to Chinese datasets. They are now belatedly closing international access to biomedical databases, after years of championing cross-border sharing, says Fu. "They now care more about local data control instead of the globalization of science," she worries.

The Commission declined interview requests with its health data officials but says that it supports the "mapping and mobilizing of health and medical data across the EU to advance AI in healthcare". In February, a spokesperson for the Commission said that despite some allowance for third-country access, "there are currently no partnerships involving the sharing of such data with China or the United States for AI development".

A Data Union strategy launched by the Commission last year aims to strengthen Europe's data sovereignty and includes an 'anti-leakage toolbox' and guidelines to assess "fair treatment of EU data abroad". But so far, the Commission has released no details about these measures and says that publications describing them will only be issued later this year.

Fig. 1 | Divided over datasets. Geneticist Jingyuan Fu at the University of Groningen in the Netherlands says that China has collected a huge amount of data, but their own data sharing and openness is very limited. Credit: Jiafei Wu.

"Europe is building an AI data ecosystem that is predominantly publicly governed, interoperable and regulated, even though 'public' does not mean open access," says Antonio Lavecchia, professor of medicinal chemistry at the University of Napoli in Italy.

"Data are available for AI development, but only through controlled, application-based mechanisms," Lavecchia explains. With respect to the interchange of healthcare and biomedical data between European and Chinese researchers, Lavecchia describes a "hybrid picture in which data sovereignty and protection laws restrict direct access, while scientific partnerships and shared frameworks create channels for meaningful exchange".

Europe's new emphasis on sovereignty is filtering into research funding. During a Horizon Europe call-for-funding webinar in late January, project managers explained that "Europe is at an economical and geopolitical conjuncture where we have to pay attention to what we do and who we collaborate with." Under some calls for funding, applicants are told they may be asked to show they can "prove that none of the applicants involved in the project is controlled by China".

For scientists like Fu running cohort studies and biobanks, the implication is that decisions around data access are no longer primarily scientific. They are geopolitical, and opaque.

"We have to report if we are collaborating with China or other countries on AI," Fu says about her university's rules under European Commission regulations. "But nobody tells us, what are the consequences? Why we should collaborate, or not collaborate? Nobody's able to give an answer."

Building Chinese data walls

In Beijing, Guanqiao Li, a professor at Tsinghua University who develops medical AI, has watched the access landscape tighten for Chinese researchers. "Every country seems to be very cautious for their genomic treasure," she says.

As of April 2025, the 2.5 petabytes of omics data in the US Cancer Genome Atlas Program database are now closed to Chinese researchers, and UK Biobank data, containing whole-genome and exome sequences for 500,000 people, is no longer internationally downloadable (Fig. 2). UK Biobank data must now be analyzed on the Biobank's own platform, which provides a cloud-based 'reading room' without allowing individual data downloads.

Li argues that having open data matters not just "because the data size is very important to train the AI", but because models trained on globally diverse populations are more relevant and reliable. She points to AI algorithms developed using UK Biobank data that explicitly validate models using external cohorts such as FinnGen or BioBank Japan. "This train–validate structure allows researchers to quantify performance shifts across ancestry groups and healthcare systems, which is where the value of international diversity becomes evident," she explains.

Similar validation strategies have been used in oncology, where data from the US-based Cancer Genome Atlas are often analyzed alongside or harmonized with the International Cancer Genome Consortium, enabling models to be trained or benchmarked using multi-country data rather than in a single national context. But access to these resources is increasingly constrained by governance and authorization requirements, Li says.

China's national cohorts and registries, such as the National Cancer Center, serve as 'national clinical infrastructures' that shape how AI tools are developed, validated and potentially translated into practice, Li says. Yet much of China's own health data remains difficult for outsiders to access — and sometimes risky even for Chinese scientists to discuss, she acknowledges.

Li pauses when asked about opening up Chinese data to foreign researchers. "I don't want

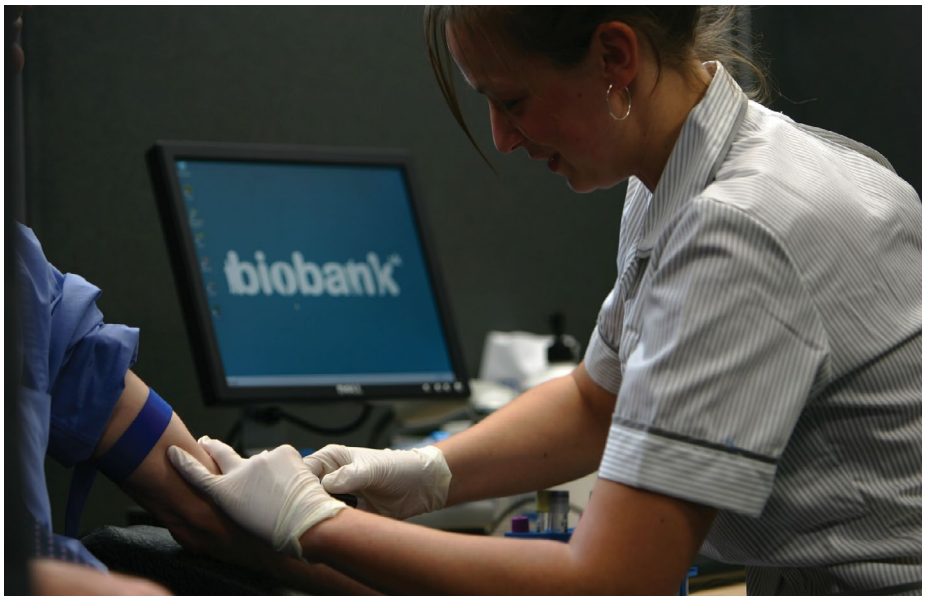


Fig. 2 | Open science under threat. Databases such as UK Biobank hold genome sequences for 500,000 people and were built on the promise of open science. New policies control who has access internationally. Credit: Christopher Furlong / Staff / Getty Images News.

to touch the red line of our government," she says. "Otherwise, I will be in trouble." Instead, she hopes for mechanisms that "could protect both of us", by making research collaboration possible without exposing study participants to privacy harm, or scientists involved in international collaborations to political risk, both in China and elsewhere.

An AI for an AI

In Washington, AI geopolitics are increasingly explicit.

Last June, while launching a national AI Action Plan, the USA announced a review of all clinical trials that send Americans' genetic data to China. Last September, the US National Institutes of Health issued new regulations for genomic data repositories and users aimed at "protecting Americans' sensitive personal health-related data from misuse by foreign adversaries" while enhancing "the privacy and autonomy of research participants".

In December, the US State Department launched its Pax Silica initiative, aimed at forming an international AI alliance that hedges against Chinese dominance. Although the initiative includes the UK and the Netherlands, so far it has sidestepped the European Commission, which it sees as favoring overly strict regulations on AI.

But US hostility to European AI regulations are a sideshow compared to the rampant anti-Chinese sentiment in Washington. At a Congressional hearing last year, Meta AI executive Alex Wang argued that China's data strategy is a competitive weapon. "As evident by China's investments, the country that wins on data will almost certainly win the AI race," he said, urging the USA to build a national reserve of government data for AI training.

The government is listening. Late last year, the White House unveiled its Genesis Mission to build an AI platform that harnesses federal scientific datasets to train models and accelerate discovery — framed explicitly as a bid for technological dominance. The Genesis AI models will have access to data from 17 national labs, including the Pentagon's health data system, and can form partnerships with scores of private companies, including IBM, OpenAI, Google and Microsoft.

In February, the US Department of Energy, which manages the Genesis Mission, placed innovation and US leadership in biotechnology among its top priorities. To some in industry, however, Washington's public-sector ambitions are not the main event.

Private data empires

Alex Zhavoronkov, CEO of the AI-driven drug-discovery company Insilico Medicine, describes the Genesis Mission as "a tiny bug on the wall" compared with private efforts to mine biomedical and healthcare data (Fig. 3). Zhavoronkov says that Google is on course to utterly dominate global AI development using healthcare data in the near term, "although the quality and quantity of Chinese biomedical data indicate that China may be the AI superpower of the future".

According to Zhavoronkov, China already holds the largest data repositories, with 1.4 billion people using the WeChat app, many of whom are already connected to hospital databases for data integration, analysis and even healthcare delivery. "China also runs the largest number of clinical trials in the world generating massive drug response and real-world-evidence datasets," he adds.

Outside of China, many of the most valuable health data are generated and held by hospitals, insurers, device makers, drug makers and data platform companies. For example, US-based electronic health records vendor Epic Systems Corporation manages records for over 300 million US patients and says that it has more than 150 AI features in development.

Such corporate AI juggernauts push the debate beyond borders and toward market power. Ashu Singhal, co-founder of Benchling, a data services company used by life-science researchers, notes that complex networks of corporate, academic and healthcare system data-gathering entities have formed. "Obviously there are things like the Worldwide Protein Data Bank in the USA and Europe that was really important for training protein models like AlphaFold," he says. "But now there's a newer crop of these data initiatives that are a combination of nonprofits and private companies."

The race is for scale and usability, Singhal says: "Everybody's working toward the same goal of 'How do you make larger and larger datasets that are actually usable by AI?'" But when these datasets and models sit behind corporate contracts, independent validation becomes harder — and the clinical community has less visibility into what AI the tools learned, and from whose data.

Restriction risks

For clinicians, biomedical researchers and the patients they serve, the consequences of data sequestration for non-scientific reasons, including geopolitics, are not abstract.

When sequestered datasets can't connect, and researchers cannot link genetic data from one country with clinical outcomes from another, says Alex Frangi, a University of Manchester computational medicine researcher, "they lose the statistical power required to identify rare genetic variants or subtle drug interactions". Frangi uses data from a mix of public and corporate databases to develop synthetic clinical trial methods.

Fig. 3 | Private powerplay. Alex Zhavoronkov, CEO of the AI-driven company Insilico Medicine, says that private efforts to mine biomedical and healthcare data for AI dwarf most government initiatives. Credit: Insilico.

AI models developed using sequestered datasets often 'overfit' to the specific demographics or clinical practices of their training environment, Frangi warns. "Without external, international validation, these biases are frequently only discovered after they have caused clinical harm," he explains.

As an example, Frangi says, many high-performing AI tools for melanoma detection show a precipitous drop in accuracy when applied to darker skin tones. "Because major datasets are often skewed toward light-skinned northern European or North American populations," says Frangi, "these tools can misclassify malignant lesions as benign in under-represented groups."

One approach to enabling research despite data sequestration is to move computation to the data rather than moving data to the researcher. UK Biobank, for example, has shifted analysis to a secure cloud-based Research Analysis Platform that generally does not allow researchers to download individual-level data. Other groups are experimenting with trusted research environments and federated learning, in which algorithms aggregate insights without exposing raw records.

One such model, says Singhal, is LillyTuneLab, which develops AI models trained on decades of pharma giant Eli Lilly's drug-discovery data. LillyTuneLab uses a federated data gathering and analysis mode, explains Singhal, in which AI developers combine internal company data with data from "outside their walls" that allows "third-party biotechs access to their models, as long as those third-party biotechs then contribute data back".

A braver new world

As China and the USA struggle for AI domination, scientists in other countries risk being sidelined. For biomedical scientists such as Fu in Groningen and Li in Shanghai, the danger is that medical AI becomes a set of national, and corporate, products trained in geopolitical isolation: powerful, but brittle and serving their political and commercial owners, with weaker external validation and hidden bias.

Ultimately, says Frangi, a balance must be struck between protecting national data and ensuring global inclusivity to make sure that synthetic data are "a tool for universal clinical progress rather than a source of new health inequalities".

And will patient interests get sidelined along the way? In an analysis of patient consent and UK Biobank published last year (G. Barn, *Med. Health Care Philos.* **28**, 533–547; 2025), University of Amsterdam researcher Gulzaar Barn warned that amid the fast-paced nature of AI-assisted genetic research, certain approved uses of UK Biobank data in the past by private companies, specifically insurance and direct-to-consumer genetic testing firms, arguably fell outside their original intent.

Naomi Allen, chief scientist for UK Biobank, says that the biobank was first created in 2012, during an era when "there was no such thing as a trusted research environment". She says that steps were subsequently taken to ensure that no data were supplied to the insurance industry or to any unapproved entities. UK Biobank acknowledges that many AI innovators now draw from its data, and Allen says, "We're in the process of reviewing our AI policy."

The promise that AI will improve medicine for everyone depends on something less glamorous than algorithms, Fu argues: it depends on rules that make sharing healthcare and biomedical data safe, reciprocal and worth it. The payoffs for medicine will be significant, she insists. "But to do it, we will need to get data much better integrated and harmonized across many different countries and world regions."

Economist Dianne Coyle, at the University of Cambridge, suggests that nations outside the American and Chinese AI superpowers act together to form an independent AI industrial model. They could follow the example of Europe's hugely successful Airbus project in the aerospace sector, which saw European countries banding together to develop numerous rival types of commercial aircraft.

Coyle says that she understands why smaller nations are fearful of the AI superpowers. "One of the things that is causing data barriers to go up is the sense that if the Chinese or American companies use that data, they're going to develop the products first, and they're going to get all the value," she says. "So, I can understand why countries are starting to put up barriers."

But Coyle frames the choice as moral as well as strategic, because when data are repurposed for geopolitical advantage or proprietary AI systems, questions about consent and benefit-sharing intensify. "We want all of our citizens to be able to benefit from it," Coyle says about the use of healthcare and biomedical data in powering AI development, "and the model of a few billionaires harvesting these data resources to make more billions is not very attractive".

Paul Webster is a freelance reporter based in Ontario, Canada.

Paul Webster
Nature Medicine.

Published online: 24 April 2026