Walking the tightrope between data sharing and data protection

With the increase in genomic data available and the advent of new technology for mining it, balancing the needs for data sharing and data protection becomes more challenging. Different disciplines must come together to find new solutions.

In the past 20 years, genomics has gained an ever-more-central role in biomedical research and healthcare. Genomic data has been used to advance diagnostics and to inform new strategies for patient stratification and screening, as well as to identify new targets and develop personalized therapies. The willingness to share genomic data, and the local and global initiatives to uphold this principle, have had a crucial role in this success, as recently exemplified by the rapid sharing of data during the early phases of the COVID-19 pandemic that enabled the development of vaccines and therapeutics in record time.

Genomic data has grown not only in importance but also in volume at an impressive pace. In the early 2000s, only two people had had their genome sequenced, whereas it is estimated that in 2020, more than 30 million people across the world had access to their genomic data, with 2 to 40 billion gigabytes of new data being generated every year. Nonetheless, limitations in translating genomic research into widely accessible therapies persist, as most of the genomic data available is for people of European ancestry, which hinders its benefit for the majority of the global population. Given the enormous amount of data available, the research community and regulators have been challenged by the need to balance the twin aims of making data accessible to researchers while at the same time protecting the privacy of study participants and patients — a far-from-trivial task.

Research has indeed shown that concerns around privacy and potential misuse of data are one of the main factors hindering participation of the public in genome-research studies. These concerns disproportionally affect the participation of under-represented communities and thus impact the diversity of the data collected and limit the benefit of genomic research globally. Worries around data security are not unfounded, as breaching the privacy of genomic data could expose sensitive information and potentially lead to loss of healthcare insurance, discrimination and stigmatization, or damage to family relationships. Another concern is that data, if not appropriately protected, could end up being used for applications that are not in line with the consent that participants provided in the first place.

Several technological safeguards have been brought forward to avoid re-identification issues, including data anonymization, de-identification and data aggregation. Nonetheless, making data truly anonymous is difficult. In a proof-of-concept study, researchers analyzing individual-level data were able to re-identify some participants of a study by inferring their surnames from short-tandem repeat sequences on the Y chromosome. Summary statistics from genome-wide association studies were also found to be not completely immune to privacy intrusions, with researchers finding that it may be possible to infer a person's participation in a genome-wide association study as part of a specific, potentially sensitive group.

Although these represent rare examples, there is reason to believe that the issue of genetic data privacy is only going to become more complex. We all share part of our DNA with relatives, and the more people have their DNA sequenced — as part of research studies or as a consequence of the popularity of direct-to-consumer genomic tests — the higher the risk of re-identification. The availability of larger genomic datasets, coupled with the ever-more-powerful applications of artificial intelligence, carries the risk of exacerbating existing vulnerabilities and presenting new societal consequences. What is considered 'safe' now might not be so in the future.

At the same time, as computational approaches become more sophisticated, it may be possible to leverage emerging technology to propose better ways of protecting genomic data. Recently proposed solutions borrowed from the fields of informatics and economics include encryption methods such as homomorphic encryption, which allows the computation of statistics without the need to decrypt genomic data. Other solutions include control of data access. For example, the National Health Service in the United Kingdom recently announced that it is transitioning to a 'Trusted Research Environment' model, whereby trusted researchers worldwide can access and work on data in an ad hoc virtual environment, without downloading the data. A similar concept has been adopted by the Global Alliance for Genomics and Health.

The evolving landscape of genomic research also presents new challenges for the regulatory and review bodies that approve and monitor studies that collect and use genetic information. In most countries, such studies are approved by review bodies, such as institutional review boards or research ethics committees. These bodies are in charge of the ethical oversight of research studies and are multi-disciplinary groups, although there is often no mandate for the inclusion of specific expertise, such as on data security.

Although large genomic consortia increasingly have entire teams working on data protection and have separate data-access committees, recent reports showed that often institutional review boards do not have dedicated expertise in the handling of big data, such as computational science, artificial intelligence, data ethics and privacy. The lack of specialized expertise represents a double-edged sword: on the one hand, this may lead to overlooking potential threats and vulnerabilities to data protection; on the other hand, lack of expert advice may mean that unjustified changes are requested or even that projects are not approved owing to perceived data-security threats. Both scenarios hinder the advancement of science.

The potential of genomic data to advance human health is enormous, but it can be tapped only if everyone feels safe taking part. Therefore, now is the time for the field to start thinking of how to best deal with emerging and future issues of data security and privacy in genomic research. Although the solution might not be a one-size-fits-all approach, it is key to involve different expertise in the process, probably from diverse backgrounds such as informatics, ethics and law, as well as to include patients and the public in these discussions. ❐

Published online: 18 May 2022
https://doi.org/10.1038/s41591-022-01852-w