WELLCOME LIBRARY, LONDON

Balancing privacy with public benefit

Maximizing access to research data will greatly benefit science, and users can help to establish universal principles on how to do it, says **Martin Bobrow**.

This week's publication in *Nature* of a second HeLa cancer-cell genome (see page 207) and the announcement from the US National Institutes of Health on how it will control who gets to use the sequence information (see page 141) highlight a growing issue in modern science: access to biomedical and health-related research data.

The amount of such data continues to grow at breakneck speed, generated by large epidemiological and cohort studies that track people's health over many years (for example, the UK Biobank project) and by studies that sequence the DNA of many individuals, such as the 1000 Genomes project. Researchers, funders and governments are becoming increasingly aware of the potential power of linking and co-analysing different data sets. Genomic data linked to large sets of patient records, for example, might reveal connections about disease that we would not otherwise discover, and data from the social sciences could add further value to these studies.

Maximizing access to data resources should increase the chances that scientists will make discoveries with medical benefits. As a result, most major research funders require grant recipients to make any large data sets they create available to other researchers. It is an ethical imperative that we seek to maximize the value of research data generated from human participants, particularly when using public funds.

In response to open-access policies, a trend is emerging of allowing legitimate researchers access to research data before publication. In making unpublished data available, however, two sets of interests need to be safeguarded. Most research participants expect privacy protection and do not want their genomes or health records to be readily identifiable. Furthermore, researchers who spend time, effort and ingenuity to generate, process and manage large research data sets expect to get appropriate credit. This also relates to emerging discussions about clinical trials: there is a need for more access to patient-level data (as highlighted by the AllTrials campaign), while respecting the terms of study participants' consent.

To navigate these issues, many large genome and longitudinal studies have set up specific data-access procedures, often overseen by committees. This is what the National Institutes of Health has done for the HeLa sequence. As the number of these data-access committees grows along with the links between data sets, a question arises: is such a piecemeal approach appropriate? The scientific and medical potential of data will only be realized if researchers are not stymied by myriad data-access mechanisms and by inconsistent ways of recording and describing data variables. So, does biomedical science need to establish and enforce common principles of governance?

There are reasons to be cautious: linking data is likely to increase the risk of individuals being identified. As more research based on linked data sets emerges, it will be extremely important to understand how these data are being used, to quantify the risks and to devise proportionate governance that allows innovative uses of data to flourish while protecting participant confidentiality as far as possible.

I chair the Expert Advisory Group on Data Access — a working group that has been set up to provide strategic advice on this issue to funders — and we need your help. We have already talked to those who produce and manage biomedical and social-science data. Now we want to hear from those who use the data, or who would like to use them in future.

What does the regulatory landscape look like for potential data users? Are we maximizing the value of the data, and if not, why not? How many data sets are out there, in what fields, governed by how many data-access committees, operating to what standards?

The remit of the working group is for UK-based funders, but our scope is international and we want to reach across both disciplinary and national borders. Still, so far we have found it extremely difficult to get an overview of the situation in the United Kingdom, let alone internationally. This is partly because of the proliferation of data in increasingly large, complex and heterogeneous data sets, but also because of the patchwork of regulations, standards and policies that govern the management of research data across the world.

To help fill in the gaps, we are conducting an online survey of users of research data, and we would value the input of *Nature* readers. If you use shared data in your research, wherever you are in the world, I urge you to participate (see go.nature.com/bmun1x).

We are interested in, for example, how controlled access to data affects your projects and how easy you find it to locate the right data sets. We know that some types of data are held in central repositories, but others are held, managed and formatted within local or institutional data-management systems that are known only to a small group of collaborators.

The challenge in data access is to achieve an appropriate balance. On the one hand, managers need to rigorously safeguard the interests of research participants and to apply serious sanctions against anyone who wilfully misuses their data. On the other hand, they also need to ensure that research data are accessible to legitimate researchers without undue costs and delay. All of this will be greatly helped if there is wide agreement on principles for structure, governance and use of shared data. ■

IT IS AN ETHICAL **IMPERATIVE** THAT WE SEEK TO **MAXIMIZE** THE VALUE OF RESEARCH DATA.

➲ **NATURE.COM**
Discuss this article online at:
go.nature.com/gp5hgu

Martin Bobrow *is an emeritus fellow at the University of Cambridge, UK.*
e-mail: mb238@cam.ac.uk