Moving beyond Bermuda: sharing data to build a medical information commons

Robert Cook-Deegan¹ and Amy L. McGuire²

¹Arizona State University, Tempe, Arizona 85287, USA; ²Baylor College of Medicine, Houston, Texas 77030, USA

The ubiquity of DNA sequencing and the advent of medical imaging, electronic health records, and “omics” technologies have produced a deluge of data. Making meaning of those data—creating scientific knowledge and useful clinical information—will vastly exceed the capacity of even the largest institutions. Data must be shared to achieve the promises of genomic science and precision medicine.

National and international public–private partnerships, consortia, and other initiatives are being formed to collect and share data on a large scale. Many aspire to build, and some predict that the profusion of data will contribute to the creation of a medical information commons—a networked environment in which diverse sources of health, medical, and genomic data on large populations become widely shared resources. The ideal of an individual-centered information commons was a leitmotif of the 2011 National Academies report on precision medicine (National Research Council 2011). Many of those goals are embedded in large-scale efforts around the world, including the Precision Medicine Initiative and “cancer moonshot” in the United States, creating an ecosystem of medical information commons (Collins and Varmus 2015; Biden 2016). For example, the National Cancer Institute of the NIH has established a Genomic Data Commons, and the NIH Commonfund is part of NIH’s Big Data to Knowledge Initiative. The Wellcome Trust, international science funders, and many national funders have long supported data structures that constitute a nascent scientific commons, which is now broadening into clinical applications. Success of these initiatives depends on policies and practices that promote data sharing, are consistent with international standards, address barriers to participation, and attend to the ethical, legal, and social issues that arise when sharing medical data about individuals.

Genomics has strong precedents for broad data sharing and open science, most notably the Bermuda Principles of 1996, now two decades old. However, the very success of the Human Genome Project (HGP) has created more complex challenges—well beyond Bermuda—that will have to be addressed to achieve the aspiration of a global information commons. The design principles of data access and transparency can help achieve this goal by preserving the value of open science while protecting the rights and interests of individuals who contribute their data as well as institutions that generate, use, and share those data.

The Bermuda precedent

In February 1996, representatives from the major DNA sequencing centers in five nations convened in Bermuda and agreed upon daily release of DNA sequence information by laboratories participating in the HGP. Prepublication, rapid disclosure was intended to solve very practical problems as the HGP shifted from mapping to sequencing: how to allocate the work and how to monitor the accuracy of sequencing at different laboratories (Reardon et al. 2016). The Bermuda Principles were also a commitment to open science, in the spirit of producing data that would be available to laboratories of all sizes throughout the world (Contreras 2011; Garcia-Sancho 2012). The Principles are thus rightly regarded as a touchstone for open science and a precedent for subsequent data-sharing policies (Arias et al. 2015).

The Bermuda Principles were primarily focused on enabling high-throughput DNA sequencing laboratories to assemble a human reference sequence—a global public good. Reaching consensus and ensuring compliance with the Bermuda Principles was no mean feat; indeed, the daily release rules initially conflicted with national policies to allow privileged data access to domestic companies in Germany and Japan and could have been interpreted to deviate from policies in the United States, France, and United Kingdom. However, the funders for over 90% of the initial sequencing under the HGP (the US National Institutes of Health and Department of Energy, and UK’s Wellcome Trust) supported the daily data-release rule. When Germany and Japan did not formally adopt the Principles, US project leaders sent letters that indicated noncompliance with the Bermuda Principles would lead to exclusion from the global genome-sequencing consortium (Reardon et al. 2016; KM Jones, RA Ankeny, and R Cook-Deegan, unpubl.). The Principles held, even if compliance could not be closely monitored and was never fully enforced. It took a concerted effort, but rapid and extensive data sharing was built into the DNA of the HGP. This mattered, as data sharing was one of the main features that distinguished the publicly funded HGP from the corporate genome-sequencing efforts of Celera and other firms, and it served as a precedent for sharing of many other kinds of data in biomedical research, from clinical data to genome-wide association data and beyond. The Bermuda Principles became an icon of effective open science.

As we enter the third decade since the Principles were enunciated, the legacy of Bermuda is evident in the amount and diversity of data proliferating around the globe, with explicit attention to sharing data. The Principles have had to be modified somewhat. The Wellcome Trust convened a meeting at Fort Lauderdale that broadened the open science ideal to other organisms but also acknowledged a difference between “community resource projects”...
and other research, for which the same rules for prepublication data release might not apply, but also cited the importance of attribution and fairness (Wellcome Trust 2003). The HapMap Consortium was aggressive in fostering data sharing and discouraging patents (The International HapMap Consortium 2003), but the ENCODE Project, because it was focused on functional elements, acknowledged that some discoveries might be the proper subject of patents (National Human Genome Research Institute 2000; Contreras 2011). A Toronto meeting broadened the principles of data sharing further to include “omics” data beyond genomic variation (The ENCODE Project Consortium 2004; Kellis et al. 2014). Modern data-sharing policies provide heightened protection for individuals whose DNA fuels research and genomic medicine, while continuing to embody the spirit of broadly sharing sequence data (Wellcome Trust Sanger Institute 2014; National Institutes of Health 2015). Yet the task of building a medical information commons in coming decades will be vastly more complicated than sharing DNA sequence contributing to a single human reference genome and sequencing model organisms. Establishing the Bermuda Principles was hard; building a medical information commons will be much harder.

Toto, we are not in Bermuda anymore

The Bermuda Principles applied to a relatively small number of laboratories in five countries, later joined by China in 1999. It was possible to get 50 people in a room to develop the rules with representation from each group that would be contributing data. Those laboratories sent data to DNA sequence databases in the US, Japan, and Europe that had already been exchanging their curated data for over a decade. The sources of DNA for the reference sequence had given informed consent for that purpose. Sequencing was performed on a set of reference samples, and the data were sequence reads assembled into contiguous sequences. It was assumed the data could be sufficiently de-identified so as to avoid issues with individual privacy and security. The work was funded by government and nonprofit research organizations focused on a shared scientific objective: a human reference sequence.

Today, the situation is quite different (see Table 1). Genomic data are highly complex, generated by public and private laboratories, for both research and clinical purposes, and they are most useful if they are further linked to many other kinds of data. Even the starting elements—sequence-based data—come in diverse formats at different levels of organization: microarray SNPs, raw sequence reads, assembled sequence data, variant call files, and large-scale insertions and deletions. In addition, the data are generated by different platform technologies and instrument suites. Moreover, data on sequence variants must be linked to many other kinds of data such as health records, exposures, and genealogy in more diverse contexts and by varied groups that are not all governed by the same rules, regulations, or incentives. The value of linked data and the inherent identifiability of sequence data make it impossible to guarantee privacy and raises new ethical and legal concerns about people from whom the data are derived.

A unitary set of Bermuda-like Principles to share sequence reads will simply not suffice. The prospect of building a medical information commons faces far more daunting challenges, which can only be addressed by understanding the major differences in how data are collected, used, and shared now as compared to 20 years ago (see Box 1).

| International Human Genome Sequencing Consortium (Bermuda 1996) | Global Alliance for Genomics and Health (2017 and beyond) | Key differences |
|---|---|---|
| **Type of data** | Assembled fragments of DNA sequence data contributing to an ultimate reference sequence | Data on genomic variants, clinical laboratory data, medical outcomes data, electronic medical records, exposure to environmental factors and toxins | Vastly broader array of data types; far more databases with more diverse formats |
| **Number of countries** | 5 | >70 | Far greater international diversity |
| **Number of individuals** | <50 | >800, at 400 institutions | More people and institutions |
| **Sources of data** | Major sequencing labs | Research laboratories of many types, clinical laboratories | Greater diversity |
| **Sharing structures** | Labs agree to public posting with links to a portal, ultimately flowing to GenBank, DDBJ, and EBI | Nascent efforts to build “beacons,” brokers, interoperable formats and standards | DNA sequence databases in US, Europe, and Japan had been sharing data systematically for a decade |
| **Organisms** | Four model organisms and human | Many organisms; diverse scientific communities | Human sequencing and data sharing developed largely from C. elegans framework |
| **Privacy, confidentiality, and informed consent** | Reference samples from a small number of donors who gave informed consent | Sequencing technology can enable re-identification, standards for informed consent changing, many new laws that require permission for sharing or exporting samples or data | Vastly more complex legal framework for sharing data about people |
| **Commercial landscape** | One major instrument manufacturer; Celera emerged as a private sequencing rival to the public HGP in 1998; and fewer than a dozen biotech and pharma firms doing intensive sequence-based analysis | Hundreds of companies in dozens of subsectors including sequencing, interpretation, clinical testing, instrument development, software, interpretive services, and direct-to-consumer genomic analysis | Far more diverse set of commercial firms with very different business models and data-sharing practices |
| **User communities** | Research tool primarily for researchers | Many tools for biology and medical research broadly; clinical use as well as scientific use; individual patients and consumers | Greater diversity of scientific users, clinicians and counselors, individual patient and consumer users |
The President envisioned medical information commons is unlikely to be a sin-
kinds of users. Forming a commons means finding rules to which
sanctions for violating them (Ostrom 2000). Experience with ge-
Advisors on Science and Technology 2014).

Commercial laboratories are highly diverse and some business models entail hoarding data.
Genomic technology has bred many kinds of companies of varying sizes, carrying out many different functions (from making instruments to developing software, test kits, or supply reagents for genomic analysis; from providing genetic testing services to interpreting or storing data to performing genetic counseling) (Curriatte et al. 2014). Firms differ not only in what they do and how they do it, but also in how they expect to make money. Some firms are committed to open science and data sharing. For example, some genomic testing services that spun out of nonprofit or government laboratories to do rigorous genetic testing, such as GeneDx, have a strong norm of data sharing and tolerate but do not relish gene patents. InVitae, Ambry, ARUP, Counsyl, and Pathway Genomics all contribute data to ClinVar, the freely available public genomic variant database maintained by the National Library of Medicine (Rehm et al. 2015). PrecisionFDA contains data from many commercial laboratories. The largest US laboratory testing companies, LabCorp and Quest, are collaborating with the Universal Mutation Database in Paris (Ray 2015), and the Global Alliance for Genomics and Health is sponsoring the BRCA Challenge and BRCA Exchange, an open science flagship project to foster interpretation of genetic variants associated with cancer risk (BRCA Challenge 2015). Other companies, however, are pursuing strategies that foresee permanent, proprietary databases. The salient example is Myriad Genetics, which premises its strategy on keeping the results of its genetic tests proprietary and binds users of its tests to maintain variant data as trade secrets. The value of a proprietary database depends in part on how good public databases become. Any firm that uses public databases but does not contribute to them is, by definition, a free rider and will have a proprietary advantage. This was true for Celera relative to the HGP, but it is true in spades for clinical data for which the pathologies of free-riding are not restricted to a research database but affect medical decisions. As public data sources “catch up” to proprietary ones, however, proprietary advantage dissipates, and eventually the cost of maintaining a private database will exceed its marginal value. This sets up a classic multiplayer prisoners dilemma scenario for policy (Ostrom 1990). The system as a whole will be more efficient and effective with a robust commons, but individual firms relying on proprietary databases will have incentives at the margin to contribute little to the commons and even to disparage it. A robust commons actually discourages the formation of proprietary data siloes, but absence of a commons invites a proliferation of siloes.

Genomic data must link to many other kinds of data. Genomic data will be linked to imaging and clinical data, medical records, and laboratory results beyond genomics, as well as family history and genealogy. Where there was one set of collaborating DNA sequence databases when the Bermuda meetings took place 1996–1998, the data in the medical information commons will be in many kinds of databases. Some will be public, some private, and some hybrids.

A commons is not a free-for-all
A medical information commons will have stakeholders of at least three types: (1) individuals whose data populate the commons; (2) researchers, clinical laboratories, and other for-profit and not-for-profit institutions that contribute data; and (3) many different kinds of users. Forming a commons means finding rules to which stakeholders—donors, data generators, and users—can agree. The envisioned medical information commons is unlikely to be a single central database, but rather a linked network of resources. As the President’s Council of Advisors on Science and Technology notes, interoperability will be key (President’s Council of Advisors on Science and Technology 2014).

A commons has rules, means of monitoring compliance, and sanctions for violating them (Ostrom 2000). Experience with genomic commons suggests that developing those capacities will take time and effort. The furore that greeted Longo and Drazen’s New England Journal of Medicine editorial, which characterized clinical researchers’ concerns about “research parasites” feeding on open science, suggests that clinical researchers have work to do convincing bioinformatics users and others who do not wish to impede the benefits of unforeseen uses and users to strengthen incentives for contributors (Longo and Drazen 2016). Conflict is common in building a commons; it can be messy and contentious. Open debate is part of the process and should be welcomed. Commons take work to build, and that work is starting, through organizations like the Global Alliance for Genomics and Health.

Data access and transparent analysis as design principles
The challenges of creating a medical information commons are daunting but not a counsel for despair. Design principles can guide
policy decisions for those funding science and paying for medical goods and services that can foster a learning system that builds a robust medical information commons. Three National Academies reports all point toward the importance of having access to data and the ability to reproduce the interpretation of the data, either for making scientific inferences (scientific replication) or for making clinical decisions (evidence-based medicine) (Cech et al. 2003; National Research Council 2011; Micheel et al. 2012). Two general principles stand out: data access and transparent analysis (DA-TA).

Data access

The philosophy and sociology of science do not always agree on details, but they do agree on some features that make scientific knowledge reliable, including the ability to replicate scientific work (Merton 1942; Popper 1959; Ziman 1978). If data are important for making scientific inference, they need to be shared so others can verify the findings, and those findings need to be subjected to "organized skepticism," in the phrase of Robert Merton (Merton 1942). Another, equally important rationale for data sharing is that some findings are simply not possible without expanding beyond the scale of individual institutions, or even countries. An analogous principle of data sharing to enable clinical validation underlies evidence-based medicine (Cochrane 1972; Eddy 2011), for example, when formulating clinical practice guidelines (Field and Lohr 1992).

Access to data is not just for science and medicine, however; it is also emerging as an individual right (Lunshof et al. 2014). Access to data about oneself is becoming a norm, both for those participating in research and in being able to get laboratory data and access to medical records from clinical care. Access to underlying data is distinct from access to "research results" interpreted by researchers, which may come with great uncertainty about their clinical or other significance. Access to data does not necessarily imply ownership, because the data may also be retained by the laboratory or institution that generated it. Rather, it is a right to get the underlying data and to transfer those data to others, rights to "get" and "send." Policies to codify such rights are beginning to be implemented. Until October 2014, for example, US laboratories certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA) were exempt from the requirement to share medical information with individuals under the Health Insurance Portability and Accountability Act (HIPAA). However, that laboratory exemption was removed, and individuals may now request the relevant record set of a laboratory test. Laboratories are obligated to provide the complete designated record set in the format specified by the requestor at a reasonable cost (DHHS Office of Civil Rights, http://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access/index.html).

Data access is a core principle for both moral and practical reasons, but it has not yet become a design principle for genomic and other databases and resources. Access is essential to respect the rights and interests of individuals, to comply with norms of scientific evidence, and to lay the foundation for evidence-based medicine, but it is not built into the system and will not be until and unless policy incentives are in place. A commons needs rules and ways to enforce them; this will fall to organizations like the Global Alliance or other forms of governance that may yet form.

Transparent analysis

Access to data is necessary but not sufficient. The canons of science also require that the methods used to interpret data as well as other forms of metadata needed to make scientific inferences be shared. Assessment of clinical validity rests on a bedrock of being able to verify and validate the interpretation of clinical data. Both scientific and clinical practice thus entail sharing data in sufficient detail to enable independent replication and verification.

The DA-TA principles have real consequences. They can guide policies to ensure that science and precision medicine rest on firm ground. They can be translated into policies for funding science. NIH, for example, has required data-sharing policies to be explicit in grants over $500,000 for over a decade. NIH broadened its data-sharing policy in 2014 (National Institutes of Health 2014). The International Committee of Medical Journal Editors has long since agreed on the need for data access and transparent analysis (International Committee of Medical Journal Editors 2013), but the principles also apply to medical goods and services that are not published. The American Medical Association Resolution 460.971 "encourages laboratories to place all clinical variants and the data that were used to assess the clinical significance of these results, into the public domain which would allow appropriate interpretation and surveillance for these variations that can impact the public's health" (American Medical Association 2013). The January 2017 statement of the American College of Medical Genetics and Genomics observes that "no single provider, laboratory, medical center, state, or even individual country will typically possess sufficient knowledge to deliver the best care for patients in need of care" (American College of Medical Genetics, Board of Directors 2017). Criteria for certification of laboratories by the College of American Pathologists, or under CLIA, should ensure compliance with this principle. Insurers and health plans that pay for laboratory tests should ensure that clinical inferences can be independently verified, which, in the case of genomic variants, necessarily implies deposit of data into public databases and sharing interpretive algorithms. Recent guidelines from the US Food and Drug Administration are addressed to "publicly accessible databases of human genetic variants," and the procedures for curation and proficiency testing presume sharing of data (Food and Drug Administration 2016).

Failure to implement policies that embody the DA-TA principles invites a "race to the bottom" of hoarding data for proprietary advantage in commercial medical services and in highly competitive academic research. Sharing costs money, takes time, and involves players outside the laboratory. It will always be cheaper to hoard than share. However, if all laboratories play by the same DA-TA rules, they can compete on other grounds, and the system will incrementally grow a medical information commons that all can draw upon to improve science and the practice of medicine.

Acknowledgments

Research was supported in part by grants from the National Human Genome Research Institute (R01 HG008918; P50 HG003391). R.C.-D. is a senior fellow at FasterCures, a Center of the Milken Institute.

References

American College of Medical Genetics and Genomics, Board of Directors. 2017. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet Med doi:10.1038/gim.2016.196. American Medical Association. 2013. D-460.971 Genome analysis and variant identification policy statement. https://searchpf.ama-assn.org/SearchML/searchDetails.action?uri=%2FAMADoc%2Fdirectives.xml-0-1601.xml. Arias JJ, Pham-Kanter G, Campbell EG. 2015. The growth and gaps of genetic data sharing policies in the United States. J Law Biosci 2: 56-68.

Biden J. 2016. Beginning cancer ‘moonshot’ at Penn Medicine. Penn Medicine Almanac 62: 1.
