
INTRODUCTION

Sharing Data in a Medical Information Commons

Robert Cook-Deegan, Mary A. Majumder, and Amy L. McGuire

In 2011, the U.S. National Academies published a report that called for an “Information Commons” and “Knowledge Network” to enhance understanding of disease and normal biology and to inform a learning health care system. The report became a major influence on President Obama’s 2015 State of the Union address proposing a “Precision Medicine Initiative,” which in turn gave rise to the current All of Us Research Program spearheaded by the U.S. National Institutes of Health. Two central ideas guiding the National Academies’ analysis were that rapid progress would depend on the continuous contribution of molecular, environmental, and health record/outcomes data linked to individuals, and that information generated from the data would need to be widely available for multiple uses. Although it is clear that the Information Commons and the Knowledge Network connecting its elements are inextricable, each is composed of discrete units with its own funding and support structures.

Even as U.S. efforts to implement the National Academies report were gathering steam, we began a project, “Building the Medical Information Commons (MIC),” to study the process. We somewhat naively initiated this project in search of the rules that would govern *the* MIC. The vision of a single, integrated, global MIC captures the goal of sharing and linking data so that it can be transformed into information, and ultimately knowledge, in order to advance sci-

ence, enhance clinical decision-making, and improve public health. However, our conception of an MIC evolved to accommodate both the data structures (plural) and the networks (also plural) connecting them. We no longer believe that it is realistic to expect a unitary commons with a single set of data-sharing practices or uniformly-defined community of contributors and users. Instead, the MIC is a collection of many different health-related commonuses (or common pool resources) that would benefit from the widespread adoption of a group of high-level but flexible principles.¹

Indeed, we have found that the notion of a single, integrated, global MIC is unrealistic for at least three reasons. First, it belies a necessary plurality of collectively-managed resources in many places for many purposes. As a landscape analysis of data-sharing initiatives revealed,² there are many sources of data, many users of data, and many research and health care institutions pursuing data-sharing functions that are only somewhat aligned. Some sites store data, while others curate data, aggregate data, broker access to data, communicate the meaning of data, and interpret the clinical significance of data.³ Disparate actors with different roles work to collect and manage the data and build the networks that make the data useful for biomedical research, clinical care, and public health.

Second, the complexities of moving masses of data make it unrealistic to expect a single, integrated, global

Robert Cook-Deegan, M.D., is a Professor in the School for the Future of Innovation in Society at Arizona State University. He is a physician and molecular biologist who turned to policy and then entered academe through Georgetown, Stanford, and Duke Universities before joining ASU. **Mary A. Majumder, J.D., Ph.D.**, is an Associate Professor of Medicine at the Center for Medical Ethics and Health Policy, Baylor College of Medicine. **Amy L. McGuire, J.D., Ph.D.**, is the Leon Jaworski Professor of Biomedical Ethics and Director of the Center for Medical Ethics and Health Policy at Baylor College of Medicine. Dr. McGuire serves on the program committee for the Greenwall Foundation Faculty Scholars Program in Bioethics and is immediate past president of the Association of Bioethics Program Directors.

MIC. Technologically, the size and scope of data make them both difficult and expensive to move. There are initiatives that still move data to a central repository, or mirror data among databases. For example, the three international human DNA sequence databases (GenBank, the European Nucleotide Archive, and the DNA Database of Japan) started sharing most of their data in the 1980s. Sequence data are copied and stored in each database.⁴ However, the future is mostly about the creation of federated systems where data and their analysis remain local because it is technologically easier and much cheaper to leave the data in place, do local analysis, and export only the parts of the data relevant to a scientific question or clinical inquiry.

Sociologically and ethically, medical and other personal data are widely regarded as highly sensitive and private. It will thus be important to keep data secure and to ensure that their use respects the terms of the consent or authorization (and cultural expectations) under which they were contributed. Economically, data are increasingly perceived as valuable sources for economic growth. Private firms are building data sets, and many countries have passed laws constraining the export of human genetic resources — often including data as well as samples.⁵ Some of these laws are intended in part to prevent “biocolonialism” — the extraction of biological resources including samples, DNA, and data from Indigenous and marginalized peoples by settler colonial forces.⁶ Global laws and norms are also intended to protect the rights and interests of citizens whose genetic and other personal data may be attractive to foreign researchers and companies.⁷

These technological, sociological/ethical, and economic factors do not, however, preclude the construction of a global network that uses many sources of data to advance understanding and to inform clinical decision-making. A federated system allows data to be left in place while permitting extraction of the information needed to address particular clinical, research, or public health questions. Further, certain kinds of valuable information at low or no risk of re-identification can be deposited in freely available global public databases. ClinVar, for example, contains information about genomic variants associated (or not) with disease, linked to information about the evidence base for variant classification.⁸ It does not contain individually identifying case-level data; those data remain in the data resources that feed into ClinVar.⁹

Third and finally, there is no international research infrastructure or funding mechanism to support a single, integrated, global MIC. Many of the component data sets were built initially through national

government-supported research or clinical testing initiatives. Most data are now generated by commercial firms, with aggregation and curation as ancillary functions of an ongoing business operation such as genetic testing, body imaging, or drug development. Some data resources (e.g., GenBank, ClinVar, LOVD, CFTR2, gnomAD) have become global public goods. However, truly global data resources that are funded through a stable international framework are rare. There are a few: the tripartite data-mirroring of DNA sequence data noted above, for example, is an unusual case of sustained international collaboration over four decades. Yet, even here, the funding sources are mainly national (USA and Japan) and regional (Europe). The organized efforts to create a human reference genome through the Human Genome Project, and many other organism-centered reference sequences, have been assembled from international collaborations. But again, these are funded through national funding streams pooled together; there are no large, stable funding structures that are truly international, and no international equivalent of the U.S. National Institutes of Health. Yet we find innumerable examples of global *use* of data for research, health care, or public health purposes. In sum, for the most part, data use is global but support of the infrastructure is national, or private. The absence of capacity for building robust public international infrastructure may not be a desirable feature, but it does appear to be a practical constraint for the foreseeable future.

The result is many work-arounds through ad hoc international collaborations, national efforts that have international dimensions, and public-private partnerships. Recent years have seen major large-scale efforts to develop standards and facilitate sharing of resources internationally, most notably the Global Alliance for Genomics and Health (GA4GH).¹⁰

It would be an exaggeration to call the landscape we have described *the* Information Commons and Knowledge Network envisioned in the 2011 National Academies’ report, but the existing resources could well evolve to serve many of its functions. The biggest point of divergence from the National Academies’ vision is that many data commons have emerged within the space we refer to as the MIC with no central administrative authority. Rather, the evolving MIC appears to be a combination of (a) aggregating and curating data generated by completed or ongoing research projects and consortia and establishing new health data-focused initiatives — usually supported by national governments, or public or private health care delivery organizations and technology companies — and (b) observing how well (or poorly) the various

models, pilots and major initiatives are meeting the needs of clinicians, researchers, and participants.

This special issue draws together the efforts of a multi-disciplinary research team, community advisory panels, and an international network of experts to think through how best to fulfill the promise of an MIC. An initial meeting of our advisory committee in March 2016 gave rise to eight principles based on case studies and a review of the literature.¹¹ We then proceeded to gather information through mixed empirical methods. In this special issue, we report the results of that research and the further reflection those results inspired. The opening paper by Amy L. McGuire and members of the research team and expert advisory committee represents the culmination of our different lines of inquiry and the pooling of the collective experience and wisdom of advisory committee members. It centers on two main findings related to the

that diversity contributes to a thriving, value-creating MIC aligned with Findable, Accessible, Interoperable and Re-usable (FAIR) data principles, and they also call attention to the ways in which some actors are challenging tradition research norms (for example, brokers that promise participants payment for their data). Next, Villanueva and colleagues describe the results of a review of the practices of data-sharing initiatives. They focus on transparency (i.e., accessibility of information to the public) related to practices in five domains: consent, privacy and security, data access, oversight, and participant engagement.¹⁴ They highlight areas of significant variation across initiatives reviewed, flagging some of these areas for further scrutiny, and make several recommendations with the aim of increasing transparency.

In parallel, members of the research team conducted interviews with expert stakeholders involved

This special issue draws together the efforts of a multi-disciplinary research team, community advisory panels, and an international network of experts to think through how best to fulfill the promise of an MIC. An initial meeting of our advisory committee in March 2016 gave rise to eight principles based on case studies and a review of the literature. We then proceeded to gather information through mixed empirical methods. In this special issue, we report the results of that research and the further reflection those results inspired.

sustainability of an MIC: (1) the case for making data resources truly oriented around participants is both ethical and pragmatic, and several dimensions of participant-centricity need attention and heightened visibility, including the participant role in decision-making, and (2) gaining and sustaining trust necessarily entails forming trustworthy institutions and practices that are committed to transparency, give participants access to their own data (and ensure that data are high-quality), invest in security, and understand the importance of accountability.¹²

One line of inquiry contributing to these findings consisted of a landscape analysis centered on systematically gathering information about and classifying data-sharing efforts and facilitators, compiling them in a database, and conducting a detailed review of a subset of efforts that both collect and distribute human genomic and other health-related data (“data-sharing initiatives”). Angela G. Villanueva and other members of the research team first characterize the broader data-sharing landscape, putting forward a typology of data-sharing efforts and facilitators.¹³ They conclude

in various aspects of data-sharing initiatives from diverse employment sectors (i.e., laboratory, academia, non-government organization, government, technology, and healthcare company). As described by Juli M. Bollinger and colleagues, part of each interview centered on defining an MIC, and identifying what might be done to narrow the gap between interviewees’ vision and aspirations of an MIC and current realities.¹⁵ Although there was full or near consensus on many points, including the desirability of pluralism and the importance of trustworthiness, views diverged concerning the priority of investing in making an MIC “clinical-grade.” Mary A. Majumder and colleagues address a different question that is critical to working out the meaning of participant-centricity: “What role should the people whose data populate the medical information commons play?”¹⁶ They found broad agreement on expanding the role of participants but also differences of emphasis between individual and collective forms of engagement. They discuss political aspects of engagement of participants as collectives, including issues of representation, and the signifi-

cance of efficiency concerns. Interviews also explored who “owns” data, who controls it, and how the medical and legal systems should frame the policy choices. Amy L. McGuire, Jessica Roberts, Sean Aas, and Barbara J. Evans discuss relevant findings and legal and ethical frameworks.¹⁷ They conclude that ownership may be a useful metaphor, but it is a poor focus for policy development related to an MIC.

Another line of inquiry consisted of convening of community advisory panels in three cities: Raleigh-Durham, NC, Austin, TX, and Oakland, CA. As described by Patricia A. Deverka, colleagues from the American Institutes for Research, and other members of the research team, public deliberations were anchored by hypothetical case studies capturing different possible MICs within the collection that makes up the over-arching MIC (including a case study based on the All of Us Research Program).¹⁸ They present recommendations regarding opt-in consent, transparent data policies, public representation on governing boards, and data security and privacy protection. They also discuss the importance of recognizing and accounting for high levels of public *mistrust* as the backdrop for efforts to build an MIC.

Finally, three papers round out this special issue, widening the lens to pick up important developments beyond the United States. Bollinger and colleagues use BRCA1/2 data sharing as a case study to capture the global data-sharing picture and concrete obstacles and incentives facing laboratories engaged in clinical genetic testing.¹⁹ They conclude that there is strong support for data sharing worldwide but also that the sharing norm is fragile, underlining the importance of addressing obstacles and attending to incentives. Tania Bubela and colleagues examine how MIC principles relate to the concept of a learning health system.²⁰ In particular, they use MIC initiatives in three Canadian provinces as case studies, and they discuss important lessons learned from each. Finally, Mark Phillips and Bartha M. Knoppers explore the tensions between the open science movement (and related efforts to build information commons) and the movement to strengthen data protection laws, with a focus on the European Union’s General Data Protection Regulation.²¹ They suggest that prior efforts to reconcile a public right of access to government documents and privacy protections offer hope that these tensions can be worked out.

We believe that, collectively, the papers contained in this special issue represent a body of work that can inform policy and practice in building an MIC. Like the deliberants in the community advisory panels, we are both hopeful and concerned. The promise of the MIC is great. Policies, especially if implemented

early while the commons are still under construction, can enable broad use leading to benefits that improve health and improve lives, while ensuring benefits are equitably distributed. We worry that this promise will be undercut by a superficial versus deep commitment to participant-centricity and trustworthiness, by failures to address obstacles and restructure incentives to support data sharing, and by laws and regulations that fall short of what is needed. The danger of steering off course is real when no one is at the helm. We hope this series of articles can initiate a wider conversation leading to a future that is in line with widely shared aspirations and that addresses the concerns.

Note

The authors report grants from the NIH during the conduct of the study, and Dr. McGuire reports personal fees from Geisinger Research, outside the submitted work

References

1. P.A. Deverka, M.A. Majumder, A.G. Villanueva, and M. Anderson et al., “Creating a Data Resource: What Will It Take to Build a Medical Information Commons?” *Genome Medicine* 9, no. 84 (2017): 1-5, available at <<https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0476-3>> (last visited January 3, 2019).
2. A.G. Villanueva, R. Cook-Deegan, B.A. Koenig, and P.A. Deverka et al., “Characterizing the Biomedical Data-Sharing Landscape,” *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 21-30; The sources of data include databases created for many different purposes. In genomics, most were created to collect data for researchers, but many genomic databases are now used in health care. Some databases, such as cancer registries, were established for public health surveillance, but are highly useful for both research and clinical care. Testing laboratories are a direct source of many data — for example, genomic testing laboratories, clinical laboratories, and imaging facilities. Moreover, as medical records are increasingly digitized, unstructured data are being transformed into forms that enable analysis of health outcomes and for other purposes. Also, many efforts are underway to integrate social determinants of health, genomic, imaging and laboratory data into electronic health records. The users of data include scientists hoping to understand biology or disease, but also health professionals helping individuals make decisions about health care interventions, counseling them about the meaning of data. Increasingly, individuals are using health and genomic data themselves.
3. See Villanueva, *supra* note 2; J.M. Bollinger, A. Sanka, L. Dolman, and R. Liao et al., “BRCA1/2 Variant Data-Sharing Practices,” *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 88-96 (see discussion of ClinVar and ClinGen in particular); M.A. Curnutte, K.L. Frumovitz, M.A. Majumder, and A.L. McGuire et al., “One Technology, Multiple Applications: Developing Context-Specific NGS Policy,” *Nature Biotechnology* 34, no. 5 (2016): 466-470, available at <<https://doi.org/10.1038/nbt.3545>> (last visited January 3, 2019).
4. National Center for Biotechnology Information, *GenBank Overview*, GenBank Website, available at <<https://www.ncbi.nlm.nih.gov/genbank/>> (last visited December 21, 2018); EMBL-EBBI, *European Nucleotide Archive*, ENA Website, available at <<https://www.ebi.ac.uk/ena/>> (last visited January 3, 2019), and DDBJ Center Website, available at <<https://www.ddbj.nig.ac.jp/index-e.html>> (last visited January 3, 2019); See also H. Stevens, *Life Out of Sequence: A Data-Driven History of Bioinformatics* (Chicago, IL: Uni-

- versity of Chicago Press; 2013) and S. Hilgartner, *Reordering Life: Knowledge and Control in the Genomics Revolution* (Cambridge, MA: MIT Press, 2017).
5. See M.A. Rothstein, B.M. Knoppers, and H.L. Harrell, "Comparative Approaches to Biobanks and Privacy," *Journal of Law, Medicine & Ethics* 44, no. 1 (2016): 161-172, available at <<https://doi.org/10.1177/1073110516644207>> (last visited January 3, 2019).
 6. N.A. Garrison, "Genomic Justice for Native Americans: Impact of the Havasupai Case on Genetic Research," *Science, Technology & Human Values* 38, no. 2 (2013): 201-223, available at <<https://doi.org/10.1177/0162243912470009>> (last visited January 3, 2019).
 7. See M.A. Majumder, R. Cook-Deegan, and A.L. McGuire, "Beyond Our Borders? Public Resistance to Global Genomic Data Sharing," *PLoS Biol* 14, no. 11 (2016): e2000206, available at <<https://doi.org/10.1371/journal.pbio.2000206>> (last visited January 3, 2019).
 8. H. Rehm, J.S. Berg, and S.E. Plon, "ClinGen and ClinVar – Enabling Genomics in Precision Medicine," *Human Mutation* 39, no. 11 (2018), available at <<https://doi.org/10.1002/humu.23654>> (last visited January 3, 2019); M.J. Landrum and B.L. Kattman, "ClinVar at Five Years: Delivering on the Promise," *Human Mutation* 39, no. 11 (2018), available at <<https://doi.org/10.1002/humu.23641>> (last visited January 3, 2019).
 9. M.J. Landrum, J.M. Lee, M. Benson, and G. Brown et al., "ClinVar: Public Archive of Interpretations of Clinically Relevant Variants," *Nucleic Acids Research* 44, no. 44 (Database issue, 2016): D862-868, available at <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702865/>> (last visited January 3, 2019).
 10. See, e.g., B.M. Knoppers and Y. Joly, "Introduction: The Why and Whither of Genomic Data Sharing," *Human Genetics* 137 (2018): 569-574, available at <<https://link.springer.com/article/10.1007/s00439-018-1923-y>> (last visited January 3, 2019).
 11. Deverka, *supra* note 1.
 12. A.L. McGuire, M.A. Majumder, A.G. Villanueva, and J. Bardill et al., "Importance of Participant-Centricity and Trust for a Sustainable Medical Information Commons," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 12-20.
 13. See Villanueva, *supra* note 2.
 14. A.G. Villanueva, R. Cook-Deegan, J.O. Robinson, and A.L. McGuire et al., "Genomic Data-Sharing Practices," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 31-40.
 15. J.M. Bollinger, P.D. Zuk, M.A. Majumder, and E. Versalovic et al., "What Is a Medical Information Commons," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 41-50.
 16. M.A. Majumder, J.M. Bollinger, A. G. Villanueva, P.A. Deverka, and B.A. Koenig, "The Role of Participants in a Medical Information Commons," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 51-61.
 17. A.L. McGuire, J. Roberts, S. Aas, and B.J. Evans, "Who Owns the Data in a Medical Information Commons?" *Journal of Law, Medicine & Ethics*, 47, no. 1 (2019): 62-69.
 18. P.A. Deverka, D. Gilmore, J. Richmond, and Z. Smith et al., "Hopeful and Concerned: Public Input on Building a Trustworthy Medical Information Commons," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 70-87.
 19. See Bollinger, *supra* note 3.
 20. T. Bubela, S.K. Genuis, N.Z. Janjua, and M. Krajden et al., "Medical Information Commons to Support Learning Healthcare Systems: Examples from Canada," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 97-105.
 21. M. Phillips and B.M. Knoppers, "Whose Commons? Data Protection as a Legal Limit of Open Science," *Journal of Law, Medicine & Ethics* 47, no. 1 (2019): 106-111.