Characterizing the Biomedical Data-Sharing Landscape

Angela G. Villanueva, Robert Cook-Deegan, Barbara A. Koenig, Patricia A. Deverka, Erika Versalovic, Amy L. McGuire, and Mary A. Majumder

Technologies such as next-generation sequencing have dramatically expanded capacity to generate genomic data at a reasonable cost, while advances in biomedical informatics have created new tools for linking and analyzing diverse data types from multiple sources. Further, many research-funding agencies now mandate that grantees share data. The National Institutes of Health's (NIH) Genomic Data Sharing (GDS) Policy, for example, requires NIH-funded research projects generating large-scale human genomic data to share those data via an NIH-designated data repository such as the Database of Genotypes and Phenotypes (dbGaP).[1] Another example is the Parent Project Muscular Dystrophy, a non-profit organization that requires applicants to propose a data-sharing plan and take into account an applicant's history of data sharing.[2]

The flow of data to and from different projects, institutions, and sectors is creating a medical information commons (MIC), a data-sharing ecosystem consisting of networked resources sharing diverse health-related data from multiple sources for research and clinical uses.[3] This concept aligns with the 2018 NIH Strategic Plan for Data Science, which uses the term "data ecosystem" to describe "a distributed, adaptive, open system with properties of self-organization, scalability and sustainability" and proposes to "modernize the biomedical research data ecosystem" by funding projects such as the NIH Data Commons.[4] Consistent with Elinor Ostrom's discussion of nested institutional arrangements, an MIC is both singular and plural and may describe the ecosystem as a whole or individual components contributing to the ecosystem.[5] Thus, resources like the NIH Data Commons with its associated institutional arrangements are MICs, and also

Angela G. Villanueva, M.P.H., *is a Research Associate at the Center for Medical Ethics and Health Policy at Baylor College of Medicine.* **Robert Cook-Deegan, M.D.,** *is a Professor in the School for the Future of Innovation in Society at Arizona State University.* **Barbara A. Koenig, Ph.D.,** *is Professor of Bioethics and Medical Anthropology, based at the Institute for Health & Aging, University of California, San Francisco.* **Patricia A. Deverka, M.D., M.S., M.B.E.,** *is Director, Value Evidence and Outcomes at Geisinger National Precision Health, where she focuses on demonstrating the value of genomic sequencing for health systems and policymakers.* **Erika Versalovic** *is a Ph.D. student in the philosophy department at the University of Washington and a neuroethics fellow with the Center for Neurotechnology in Seattle, WA.* **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.* **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.*

The Journal of Law, Medicine & Ethics, 47 (2019): 21-30. © 2019 The Author(s)
DOI: 10.1177/1073110519840481
https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

form part of the larger MIC that encompasses all such resources and arrangements.

Although many research funders incentivize data sharing, in practice, progress in making biomedical data broadly available to maximize its utility is often hampered by a broad range of technical, legal, cultural, normative, and policy challenges that include achieving interoperability, changing the standards for academic promotion, and addressing data privacy and security concerns. Addressing these challenges requires multi-stakeholder involvement.[6] To identify relevant stakeholders and advance understanding of the contributors to an MIC, we conducted a landscape analysis of existing data-sharing efforts and facilitators. Our work builds on typologies describing various aspects of data sharing that focused on biobanks, research consortia, or where data reside (e.g., degree of data centralization).[7] While these works are informative, we aimed to capture the biomedical data-sharing ecosystem with a wider scope. Understanding the components of an MIC ecosystem and how they interact, and identifying emerging trends that test existing norms (such as norms respecting the role of individuals from whom the data describe), is essential to fostering effective practices, policies and governance structures, guiding resource allocation, and promoting the overall sustainability of the MIC.

Understanding the components of an MIC ecosystem and how they interact, and identifying emerging trends that test existing norms (such as norms respecting the role of individuals from whom the data describe), is essential to fostering effective practices, policies and governance structures, guiding resource allocation, and promoting the overall sustainability of the MIC.

Methods

We conducted a landscape analysis to capture and characterize a broad range of data-sharing practices, focusing on efforts distributing DNA-derived data and related information and other actors facilitating such data sharing and utilization. Sharing genomic data for research and clinical uses is relatively new, and many proposals for improving the future of biomedical research and clinical care center on genomic data. Our review relied exclusively on publicly-available information.

Sampling

We employed a purposeful sampling approach to identify data-sharing efforts, and started by working on case studies selected by the project team in October 2015. Then from August 2016-January 2017, after refining a list of keywords, we conducted an online search by entering keywords into Google search to find data-sharing efforts that distribute DNA-derived data and other actors that facilitate the distribution and utilization of such data. Results containing all or some of the search terms were examined. Relevant efforts were also identified from expert input. Data-sharing efforts and facilitators originating outside the U.S. were included; however, our search was limited by the use of English-language terms. Efforts and facilitators reviewed were logged using Microsoft Excel and monitored regularly through August 2018. Efforts sharing data derived primarily from non-human organisms were not logged.

A total of 679 websites were identified and examined for inclusion (see Figure 1). Websites of efforts that did not share data or facilitate sharing (n=173), such as non-research, academic programs, were excluded. Since our focus was on the sharing of DNA-derived data, we excluded biobanks that primarily provide biological materials. We also excluded efforts sharing data through an external repository (n=49). For example, efforts that only share data by depositing it into dbGaP were excluded and are instead considered a type of data contributor in our analysis. Efforts that share data but not DNA-derived data (n=12) and efforts for which we could not confirm that DNA-derived data are shared based on the publicly available information (n=5) were excluded, in addition to inactive efforts (n=62) and initiatives that had not yet started to share DNA-derived data (n=13). A total of 327 initiatives were included in the analysis.

Classification and Enumeration

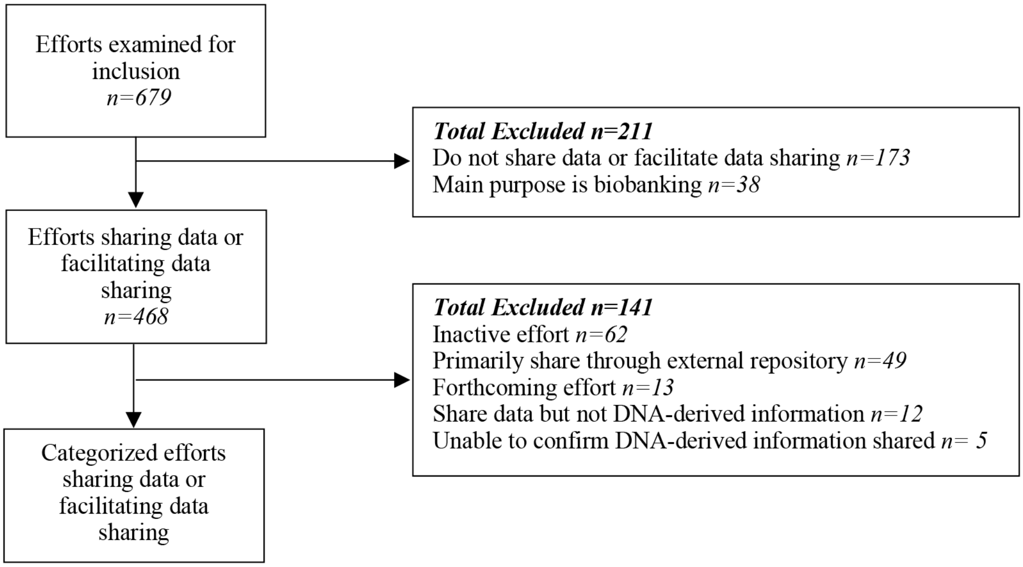
As we initially reviewed search results, the diverse mechanisms facilitating the flow of data became apparent. To capture the diversity and develop a useful

The Journal of Law, Medicine & Ethics, 47 (2019): 21-30. © 2019 The Author(s)

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

Figure 1

Selection Process of Efforts Sharing Data or Facilitating Data Sharing


Efforts examined for inclusion *n=679*
Total Excluded n=211
Do not share data or facilitate data sharing *n=173*
Main purpose is biobanking *n=38*
Efforts sharing data or facilitating data sharing *n=468*
Total Excluded n=141
Inactive effort *n=62*
Primarily share through external repository *n=49*
Forthcoming effort *n=13*
Share data but not DNA-derived information *n=12*
Unable to confirm DNA-derived information shared *n= 5*
Categorized efforts sharing data or facilitating data sharing

typology of the data-sharing landscape, the research team collaboratively categorized efforts based on their observed function, drawing on existing classification schemes.[8] The process of classification exposed inconsistency in how terms were used and the diversity of functions and practices among data-sharing efforts. While we noted how each data sharing effort described itself, we characterized them based on our analysis of their primary function in the data-sharing landscape.

Categories and code assignments were refined iteratively by the project team as the research progressed. In addition, for some of the categories described in this paper, we highlighted efforts fulfilling multiple functions that fit into more than one category. Two members of the research team coded the search results. Ambiguities were resolved by consulting additional members of the research team and reaching consensus on assignment to a single category based on primary function. Counts and percentages for each code were computed using Microsoft Excel. Our intent was to develop an empirically-grounded typology to present a snapshot of the existing data-sharing landscape, but it is not exhaustive. The reported relative frequency counts and percentages for the codes, therefore, reflect the landscape within the parameters of our sampling approach.

Results

Based on the review of the 327 efforts included in the final analysis, we identified and defined four discrete categories that capture components of the data-sharing ecosystem. These categories — data-sharing efforts, data-sharing facilitators, data sources, and end users — and their respective components are defined in Table 1 and described in more detail below.

Data-Sharing Efforts

Our review revealed that multiple efforts from the public and private sectors contribute to the ecosystem by collecting — either directly from individuals or from intermediaries — and distributing human data. These efforts varied, depending on how broadly they shared and whether they directly collected the shared data or simply aggregated data from other sources. Here we describe the types of data-sharing efforts we identified.

Controlled- and open-access data-sharing initiatives (n=50, 15%) collected biospecimens and data from individuals and shared them through open- or controlled-access portals. Controlled-access initiatives used a gatekeeping process that required a certain action, such as signing a data-sharing agreement, to access data. Some data-sharing efforts offered tiered access where some data were openly available and other data were controlled access. Individuals were usually recruited to share their data via these initiatives through: 1) general biomedical research; 2) specific research protocols; 3) disease-specific registries; and 4) open databases. In keeping with traditional research norms, these efforts typically operated under Institutional Review Board (IRB) or other ethics review body oversight, collected biospecimens

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

Table 1

The Data-Sharing Ecosystem: Components Defined

Category	Function	Example
1. Data-Sharing Efforts	Distribute data through established mechanisms in accordance with data-sharing policies and guidelines.	
Aggregators	Pool data from published studies, existing datasets, or from direct data submissions, and share an output, often through a browser.	UCSC Genome Browser
Closed consortia	Share data internally with two or more collaborators, each representing a different institution.	CHARGE
Controlled-access data-sharing initiatives	Require an action, such as submission and approved application and/or a data use agreement, to grant data access. Data may be shared through access tiers based on type of data.	Framingham Heart Study
Open-access data-sharing initiatives	Offer access to data with no action, such as creating an account or submitting an application, required.	openSNP
Repositories	Store data, deposited either voluntarily or in fulfillment of a funding agency mandate, for distribution and/or archival purposes.	dbGaP
Selective data-sharing initiatives	Provide some thematically linked data from a single or few sources.	FLOSSIES
2. Data-Sharing Facilitators	Enable the flow of data through mechanisms that do not involve the exchange of datasets.	
Brokers	Connect individuals to data by: Connecting researchers to data shared through a federated network of data sources where data initiatives maintain control over data access decisions but collaborate by contributing information to the broker	Beacon Network
	OR connecting researchers to individuals who may want to participate in research by sharing their data or enrolling in a study.	NuMe
Data analysis tool providers	Provide data analysis tools or a suite of tools to manage, analyze, and/or share genomic data, often offering a software application to visualize output.	Seven Bridges
Indexers	Catalog existing databases and provide a web-based query search engine yielding location of information in external databases.	CLINVITAE
Infrastructure providers	Offer technological infrastructure in the form of platforms to host a data-sharing project or platform-enabling technology developed to facilitate data sharing.	Sage Bionetworks
Policy and guideline developers	Develop and establish stakeholder-informed policies and guidelines promoting data sharing.	Global Alliance for Genomics and Health
3. Data Sources	Supply data into the ecosystem.	
Healthcare providers and health plans	Translate clinical encounters and related transactions into data that can be used for clinical or research purposes.	Hospitals, healthcare systems, clinical laboratories
Researchers	Collect and maintain data, or process and store biospecimens that are used to generate data that can distributed for further use.	Investigator-led research laboratories, pharmaceutical company research divisions, biorepositories
Publishers	Implement publication guidelines that require or encourage the distribution of datasets corresponding to manuscripts.	Basic science and clinical journals
4. End Users	Adhere to data access procedures and utilize the data.	
Citizen Scientists	Seek data for research purposes but typically lack professional training in the particular research field or an institutional affiliation.	Members of general public interested in genomics, patient-researchers
Clinicians	Seek data to inform clinical decision-making.	Physicians treating rare diseases
Researchers	Seek data under a research protocol, often with institutional oversight.	Investigators affiliated with academic institutions or pharmaceutical companies

The Journal of Law, Medicine & Ethics, 47 (2019): 21-30. © 2019 The Author(s)

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

and data, and solicited consent from participants to share data with other researchers either broadly or in keeping with a set of funder policies. Disease-specific registries typically recruited individuals to share data for general research, as opposed to a specific study protocol.

In a departure from traditional research norms, a number of data-sharing initiatives relied on individuals to contribute data received from direct-to-consumer (DTC) testing or smart-phone applications linked to wearable devices and often enrolled individuals through a website. Individuals were able to become end users, taking on the role of a researcher (i.e., citizen scientist).[9] Open databases, such as openSNP and Open Humans, enabled such uses since data were available for anyone to access. Open Humans also gave data contributors the option to upload a picture and a write-up about themselves, in addition to sharing data through direct, private communication with a data seeker. Open databases typically did not operate under IRB research oversight, did not fall under the purview of health data privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA), and did not require formal consent for participation. Individuals assumed all risks, including the risk of personal identification. As the openSNP website was noted to state, "There is zero privacy anyway, get over it."[10]

Selective data-sharing initiatives (n=9, 3%) provided data access but only to limited datasets, primarily sharing aggregated data analyses from one or a few thematically linked sources. For example, Fabulous Ladies Over Seventy (FLOSSIES) a public, web-based database indicated on their website sharing some demographic information along with allele frequencies of genes linked to breast cancer. The study population in the database biospecimens cancer-free women over the age of 70 whose biospecimens were collected as part of the Women's Health Initiative, and some data were made publicly available through the FLOSSIES genetic variant database.[11]

Closed consortia (n=77, 24%) consisted of multi-institutional research collaborations (often between industry and non-profit organizations) that primarily shared data internally, among consortium members. For example, the website for the company PatientsLikeMe had an online forum where patients connected with each other (researchers were also invited to access the forum). However, data sharing was noted to occur through research partnerships: "We do not participate in studies in which we simply provide data to researchers. We prefer to be active collaborators in the implementation of the research protocols for which patient data will be analyzed."[12] Most closed consortia provided limited or no information on the terms of their collaboration agreements on their websites; therefore, we have limited information about their data-sharing practices. An exception was the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Consortium, which posted on their website the consortium's data-sharing policy and expectations of its members. CHARGE noted that some data were shared with external researchers by depositing datasets from published reports into dbGaP.[13]

Aggregators (n=84, 26%) combined data from multiple datasets, analyze the data, and publicly share their analyses. While some aggregators solely pooled data from published studies and/or existing datasets, others also offered mechanisms for voluntary data submissions. Both types of aggregators shared summary statistics or metadata analyses, often through a public, online browser that facilitated interrogation of the data. The University of California Santa Cruz (UCSC) Genome Browser was classified as an aggregator because the data available were combined from multiple sources, such as the laboratories participating in the Human Genome Sequencing Consortium, and the UCSC Genome Browser provided a publicly and freely accessible visualization of the integrated, analyzed data.[14]

Repositories (n=20, 6%) served primarily to house datasets from multiple sources and engage in data curation to ensure datasets were adequately de-identified and met data submission criteria. Aggregators shared certain features with repositories, but repositories typically did not take the additional step of combining and analyzing the data. Repositories stored and shared individual datasets contributed voluntarily or in compliance with a funding agency mandate and offered data through controlled-, open-, or tiered-access (i.e., some data open, other data through controlled access with a gatekeeper). A few efforts did not fit neatly into a single category. For example, the NIH's GenBank DNA sequence database had both aggregator and repository features, which included archiving data, synchronizing data daily with other U.S. and international databases, offering a mechanism for researchers to submit data, and publicly sharing aggregated outputs.[15]

Data-Sharing Facilitators

The following sections describe efforts that facilitated the flow and use of data in an MIC ecosystem by connecting researchers and other third parties to existing data sources, providing critical infrastructure, and developing standards.

Brokers (n=16, 5%) connected individuals to specific kinds of data. Some brokers were structured as

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

federated networks, where data contributed by multiple efforts were funneled to a centralized database for the purposes of data interrogation, and the individual data contributors retained control over access to the datasets. We classified the Global Alliance for Genomics and Health (GA4GH) Beacon Network as such a broker because the Beacon resource had the ability to query the multiple participating databases to find out if any contained information on a particular genetic variant.[16] Taking the brokering function further was the Global Alzheimer's Association Interactive Network (GAAIN). GAAIN described its data resource as providing output of analyses on data from various U.S. and international research efforts, including the Canadian Longitudinal Study on Aging and the Japanese Alzheimer's Disease Neuroimaging Initiative, that GAAIN formatted based on standards established by the Clinical Data Interchange Standards Consortium for the purpose of data exploration and hypothesis development, as the contributing data initiatives themselves granted access to the original datasets.[17]

Other brokers connected researchers with individuals who may want to share data or participate in research. MyGene2 was classified as this type of broker. Through its website, MyGene2 indicated that clinicians and researchers who have contributed information on individuals with a rare condition are given access to phenotypic information for other similar individuals and connected to families of those affected, as well as to other clinicians and researchers.[18] MyGene2 also noted its participation in the GA4GH Beacon Network. We also classified we classified Portable Genomics' NuMe as a broker. NuMe, a for-profit enterprise, that offered a mechanism to connect researchers to data from NuMe customers. NuMe is noteworthy because the company indicated that customers would receive monetary compensation if their data were to be shared.[19] This contrasts with traditional research norms, according to which participants are not invited to share in any profits from the use of their data.

The repositories described above conveniently provided a "one-stop shop" for data from multiple sources. This abundance of data spurred the development of independent cataloging efforts, or *indexers* (n=9, 3%). Indexers functioned as a directory, typically offering a publicly-accessible, web-based query through which users could identify which existing resource, such as ClinVar and dbGaP, to access for information of interest. CLINVITAE, for example, a free and public resource developed by the genetic testing company Invitae Corp, directed users to public databases containing information on genetic variants of interest.[20]

Data analysis tool providers (n=33, 10%) offered discrete products to help researchers and others not specialized in data analysis to interpret genomic and health data. Their tools were intended to help researchers manage, analyze, and understand large datasets with simple, user-friendly software, often through data visualizations. An example is the company Seven Bridges, with its portfolio of products including the Seven Bridges Platform. The Seven Bridges Platform, available to U.S. and international research projects, was recorded as a data analysis tool provider because of its provision of bioinformatics analyses of large genomic datasets via Amazon Web services and Google Cloud.[21] The National Cancer Institute's Cancer Genomics Cloud has implemented the Seven Bridges Platform.[22]

The sharing and storage of vast amounts of data was supported by *infrastructure providers* (n=17, 5%), who provided a digital space to host data-sharing projects. Synapse Commons, developed by Sage Bionetworks, a non-profit organization, was one as such provider. Synapse has established general community guidelines concerning data privacy and security, while each research team using Synapse has the discretion to determine whether to offer data access to all or some registered users.[23] Other similar projects, such as the Platform for Engaging Everyone Responsibly (PEER) created through a partnership between a non-profit patient advocacy organization, Genetic Alliance, and the company Private Access, Inc., provided platform-enabling technology that can be used by research efforts. PEER has been recognized for its user-friendly interface designed to facilitate self-governance by giving individuals dynamic and granular control over data sharing.[24] The HEROIC Registry for individuals with hereditary cancer, which we classified as a data-sharing initiative, integrated the PEER platform to achieve its mission of providing a "patient-centric genetic database."[25]

Finally, data-sharing requirements imposed by governmental funding agencies, such as the NIH via its GDS Policy, and guidelines developed by professional organizations, such as the International Committee of Medical Journal Editors (ICMJE) proposal on the sharing of clinical trial data, are (or in the cases of the ICMJE proposal, could become) standard policies shaping and regulating the flow of biomedical research data.[26] Beyond these organizations, we identified work by *policy and guideline developers* (n=12, 4%), frequently international, who facilitated the development and establishment of stakeholder-informed policies and guidelines that promote data sharing. The GA4GH, Human Variome Project, and Public Population Project in Genomics and Society have developed

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

data-sharing guidelines informed by different global perspectives on challenges and best practices.[27]

Data Sources and End Users

While we did not attempt to quantify the data contributors and end users, we recognize their important role in sustaining the flow of data and creating value from data. As described in the sections above, the data-sharing ecosystem relied on four main contributors of data: 1) *individuals*, where data entered the ecosystem directly from the human beings the data describe, for example, through a DTC testing consumer's contribution of her data to an open-access data-sharing initiative, or a patient's use of her HIPAA access right to make a similar direct contribution of data from her medical record; 2) *healthcare providers and health plans*, where, for example, a healthcare system contributed data from medical records including DNA-derived data, or medical records were linked to DNA-derived data generated for research purposes, or a clinical laboratory contributed genetic testing results and related information, such as the basis for a variant call; 3) *researchers*, where data entered the ecosystem from datasets maintained by individual investigators who led laboratories or research organizations such as pharmaceutical companies and biorepositories that generated data from biospecimens; and 4) *publishers*, where datasets corresponding to manuscripts were made available to actors in the ecosystem, including aggregators and indexers.

From distributing datasets to brokering relationships with potential research participants, data-sharing takes many forms, creating opportunities for new actors such as technology companies and citizen science-friendly open-access data-sharing initiatives

The main end users of the data-sharing ecosystem were: 1) *researchers* accessing data for use in a scientific study who possessed the conventional bona fides for access to data maintained by controlled-access data-sharing initiatives including an advanced degree in a relevant scientific discipline, a research protocol, and an institutional affiliation entailing some degree of oversight (which could encompass IRB approval); 2) *clinicians* accessing data for patient care purposes; and 3) *citizen scientists*, here meaning members of the general public with an interest in engaging in genomic research, as well as researchers who would not be considered qualified to access data according to common controlled-access criteria (e.g., laboratory-based researchers interested in variant information but without a specific research protocol, researchers with limited resources outside institutional settings and lacking access to an oversight mechanism).[28]

Discussion

From distributing datasets to brokering relationships with potential research participants, data-sharing takes many forms, creating opportunities for new actors such as technology companies and citizen science-friendly open-access data-sharing initiatives. The purpose of our typology is not to provide new terminology for the field to adopt; rather, the purpose is to provide a tool to advance understanding of the different efforts and facilitators involved in creating an MIC as a whole and their functions. The interactions among the components described illustrate a data-sharing ecosystem that is complex and diverse; we believe each component is integral in order for an MIC as a whole to flourish. For example, many closed consortia were research partnerships involving multi-sectorial collaborators exchanging expertise and resources. Further, although closed consortia did not have institutional arrangements for direct sharing of their data with non-members, many shared data more broadly via NIH-designated repositories such as dbGaP (as required by the GDS Policy, in the case of NIH-funded consortia). Open databases captured data from individuals willing to share their data who may otherwise be unrepresented in databases if they did not participate in a research study or disease registry. Selective data-sharing efforts added utility to data that would otherwise lay fallow. Aggregators synthesized vast quantities of data circulating throughout the ecosystem, made data available efficiently with few or no restrictions, and often incorporated features that enhance utility for diverse end users. Reusing and repurposing existing data, plus allowing for new data to be submitted, maximizes the utility of existing resources, and reduces the costs associated with executing new research protocols to collect data.

The diversity of functions represented suggests that for the ecosystem to thrive and create value, just sharing data is insufficient; data resources must also be searchable and accessible by researchers and clinicians. Such features align with the FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles. The FAIR principles promote access and utilization

The Journal of Law, Medicine & Ethics, 47 (2019): 21-30. © 2019 The Author(s)

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

of existing electronic data, algorithms, and analytical tools.[29] Indexers helped researchers locate data in large databases. Brokers connected data seekers with data sources and potential research participants, and data analysis tools helped researchers and clinicians advance precision medicine goals while also facilitating the flow of data. The important role of the actors captured under the *data-sharing facilitators* category suggests the need for adequate resources, including funding, to support these facilitators of the broad distribution and efficient utilization of data in an MIC. Other important facilitators of the data-sharing ecosystem were the providers of technological infrastructure that supported the storage and distribution of data and the guideline developers who provided a policy roadmap to navigate the full array of data-sharing intricacies.

As described above, some of the data-sharing efforts and facilitators captured by our landscape analysis departed from traditional research norms, for example, with regard to the role of individuals whom the data describe. The traditional biomedical research paradigm is characterized by minimal involvement of these individuals throughout the research process. Investigators develop protocols covering practices such as recruitment and enrollment, informed consent, participant compensation, and data sharing under the oversight of IRBs and institutional officials. Study or registry personnel collect, store, and transmit data. The direct and active role of individuals in certain types of data-sharing initiatives — including direct contribution and personal control of data and unrestricted sharing through simple display interfaces — is striking. These kinds of developments may further advance a citizen science movement in biomedical research related to genomics, which many see as valuable.[30] The success of this movement depends, in part, upon adoption of technologies and related policies that make direct contribution easy (including the integration of smartphones or wearable digital devices as data collection and transmission tools) and do not data limit access to "qualified researchers" with institutional affiliations and resources. At the same time, many citizen science-friendly data-sharing initiatives are operating in a relative regulatory vacuum, outside the reach of HIPAA and federal research regulation, and some commentators argue that new guidelines and policies are needed to ensure that individuals contributing data to these initiatives are fully informed and protected from potential harms.[31]

Further, providing or arranging monetary compensation for individuals as the sources of data was a feature of some of the newer data-sharing facilitators, such as NuMe, captured in our analysis. In the traditional research paradigm, participants may be compensated for time, effort, and cooperation but typically are not paid for their data, and it is common for consent forms to contain language specifically disclaiming any plan to share profits that may result with participants.[32] If brokers who pay individuals to share their data succeed in attracting a large membership, a shift might occur, disrupting this traditional approach. Compensating individuals for data is a trend to observe, as it raises questions about data ownership, stewardship and control while, arguably, promoting fairness by acknowledging the value of a person's genetic information.[33] As this paper was being finalized, two companies — Luna DNA and Zenome — entered the data-sharing space promising to compensate individuals for their DNA.[34] Whether payment is a successful tactic has yet to be determined, but this novel approach points to the urgency of the need for further conversation regarding compensation.

Conclusion

Data sharing involves more than making datasets available. We captured how complex multi-stakeholder involvement will be, given the characteristics of the data-sharing ecosystem. Some policy- and norm-development projects will need representation from many or all of the categories of data-sharing efforts and facilitators presented above, as well as representation of data contributors and end users. Our analysis also revealed emerging trends related to citizen science and paying individuals for their data, challenging traditional norms. These trends are worthy of further monitoring and study. Recognition of the multiple actors facilitating the flow of data and its efficient use to create value is critical to understanding and ensuring the growth of a healthy data-sharing ecosystem and maintaining its long-term sustainability.

Note

Dr. McGuire reports personal fees from Geisinger Research, outside the submitted work. Erika Versalovic reports grants from The National Human Genome Research Institute Institute, during the conduct of the study

Acknowledgement

We are thankful to Jill Robinson for her editorial assistance. The following grants from the National Institutes of Health (NIH) supported the preparation of this article: R01HG008918 (AMG and RCD) and P20HG007243 (BAK). The views expressed in this article are those of the authors and do not necessarily reflect those of the NIH.

References

1. National Institutes of Health, "Genomic Data Sharing Policy," (August 2014), *available at* <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html> (last visited January 4, 2019).
2. Parent Project Muscular Dystrophy (PPMD), "Resources," PPMD Website, *available at* <https://www.parentprojectmd.

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

org/research/for-researchers-industry/resources/> (last visited November 16, 2018).

[3] 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 4, 2019).

[4] National Institutes of Health, *NIH Strategic Plan for Data Science* (2018) at 9 and 29, *available at* <https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf> (last visited January 4, 2019). The NIH Data Commons will integrate datasets so that they can be interrogated simultaneously.

[5] M.A. Majumder, P.D. Zuk, and A.L. McGuire, "Medical Information Commons," in B. Hudson, J. Rosenbloom, and D. Cole, eds., *Routledge Handbook of the Study of the Commons* (Taylor & Francis Group, Forthcoming 2019), *available at* <https://papers.ssrn.com/abstract=3131913> (last visited January 4, 2019).

[6] See Deverka, *supra* note 3.

[7] See G.E. Henderson, R.J. Cadigan, T.P. Edwards, and I. Conlon et al., "Characterizing Biobank Organizations in the U.S.: Results from a National Survey," *Genome Medicine* 5, no. 3 (2013), *available at* <https://genomemedicine.biomedcentral.com/articles/10.1186/gm407> (last visited January 4, 2019); M.D. Lim, "Consortium Sandbox: Building and Sharing Resources," *Science Translational Medicine* 6, no. 242 (2014): 242–246, *available at* <http://stm.sciencemag.org/content/6/242/242cm6.full> (last visited January 4, 2019); J.L. Contreras and J.H. Reichman, "Sharing by Design: Data and Decentralized Commons," *Science* 350, no. 6266 (2015): 1312–1314, *available at* <http://science.sciencemag.org/content/350/6266/1312> (last visited January 4, 2019); H.A. Piwowar, M.J. Becich, H. Bilofsky, and R.S. Crowley et al., "Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers," *PLOS Medicine* 5, no. 9 (2008): e183, *available at* <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050183> (last visited January 4, 2019).

[8] The Global Alliance for Genomics and Health, Data Sharing Lexicon (March 2016), *available at* <https://www.ga4gh.org/wp-content/uploads/GA4GH_Data_Sharing_Lexicon_Mar15.pdf> (last visited February 26, 2019).

[9] See Open Humans website, *available at* <https://www.openhumans.org/> (last visited September 24, 2018).

[10] openSNP, Sign Up website, *available at* <https://opensnp.org/signup> (last visited September 24, 2018).

[11] See FLOSSIES website, *available at* <https://whi.color.com/about> (last visited September 24, 2018).

[12] PatientsLikeMe, *Frequently Asked Questions about Research*, PatientsLikeMe website, *available at* <https://www.patientslikeme.com/research/faq#qr4> (last visited September 24, 2018).

[13] CHARGE Consortium, *CHARGE Results*, CHARGE Consortium website, *available at* <http://www.chargeconsortium.com/main/results> (last visited September 24, 2018).

[14] The Regents of the University of California, UCSC Genome Browser: Acknowledgments, Genome Browser Website, *available at* <https://genome.ucsc.edu/goldenPath/credits.html#human_credits> (last visited September 24 2018).

[15] National Center for Biotechnology Information, *GenBank Overview*, GenBank website, *available at* <https://www.ncbi.nlm.nih.gov/genbank/> (last visited September 24, 2018).

[16] The Global Alliance for Genomics and Health, "A Federated Ecosystem for Sharing Genomic, Clinical Data," *Science* 352, no. 6291 (2016): 1278–1280; Beacon Network, *About*, Beacon Website, *available at* <https://beacon-network.org/#/about> (last visited September 24, 2018).

[17] See the *Explore Data* webpage on the Global Alzheimer's Association Interactive Network website, *available at* <https://www.gaaindata.org/partners/online.html> (last visited September 24, 2018).

[18] University of Washington, *About*, MyGene2 Website, *available at* <https://mygene2.org/MyGene2/about> (last visited September 24, 2018).

[19] Portable Genomics, NuMe website, *available at* <http://nume.website/> (last visited September 24, 2018).

[20] Invitae, CLINVITAE website, *available at* <http://clinvitae.invitae.com/> (last visited September 24, 2018).

[21] Seven Bridges Genomics, The Seven Bridges Platform website, *available from* <https://www.sevenbridges.com/platform/> (last visited September 24, 2018).

[22] Seven Bridges, Cancer Genomics Cloud website, *available at* <http://www.cancergenomicscloud.org/> (last visited September 24, 2018).

[23] Sage Bionetworks, Synapse Commons Data Use Procedure, (October 26, 2015), *available at* <https://s3.amazonaws.com/static.synapse.org/governance/SynapseCommonsDataUseProcedure.pdf?v=4> (last visited February 26, 2019).

[24] R.H. Shelton, "Electronic Consent Channels: Preserving Patient Privacy Without Handcuffing Researchers," *Science Translational Medicine* 3, no. 69 (2011): 69cm4-69cm4, *available at* <http://stm.sciencemag.org/content/3/69/69cm4.full> (last visited January 7, 2019).

[25] AliveAndKickn, HEROIC Registry website, *available at* <https://aliveandkickn.org/heroic-registry-0> (last visited September 24, 2018).

[26] See National Institutes of Health, *supra* note 1 and D.B. Taichman, P. Sahni, A. Pinborg, and L. Peiperl et al., "Data Sharing Statements for Clinical Trials," *BMJ* 357 (2017): j2372, *available at* <https://doi.org/10.1136/bmj.j2372> (last visited January 7, 2019).

[27] B.M. Knoppers, "Framework for Responsible Sharing of genomic and health-related data," *HUGO Journal* 8, no. 1 (2014): 3, *available at* <https://thehugojournal.springeropen.com/articles/10.1186/s11568-014-0003-1> (last visited Janury 7, 2019); Human Variome Project, *Solutions*, Human Variome Project website, *available at* <http://www.humanvariomeproject.org/solutions/solutions.html> (last visited September 24, 2018); Public Population Project in Genomics and Society, *About P3G2*, P3G website, *available at* <http://p3g2.org/about-p3g2/> (last visited September 24, 2018).

[28] For a discussion on qualified researchers, see 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.

[29] M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg, and G. Appleton et al., "The FAIR Guiding Principles for Scientific Sata Management and Stewardship," *Scientific Data* 3 (2016), article160018, *available at* <https://www.nature.com/articles/sdata201618> (last visited January 7, 2019).

[30] See C.J. Guerrini, M.A. Majumder, M.J. Lewellyn, and A.L. McGuire, "Citizen Science, Public Policy," *Science* 361, no. 6398 (2018): 134-136, *available at* <http://science.sciencemag.org/content/361/6398/134.full> (last visited January 7, 2019).

[31] M.A. Rothstein, J.T. Wilbanks, and K.B. Brothers, "Citizen Science on Your Smartphone: An ELSI Research Agenda," *Journal of Law, Medicine & Ethics* 43, no. 4 (2015): 897–903, *available at* <https://onlinelibrary.wiley.com/doi/abs/10.1111/jlme.12327> (last visited January 7, 2019).

[32] The general norm has been supported by a range of justifications, including avoiding coercion and undue inducement, lack of value creation by individual participants, and ensuring the sustainability of the research enterprise. See E.A. Largent and H.F. Lynch, "Paying Research Participants: Regulatory Uncertainty, Conceptual Confusion, and a Path Forward," *Yale Journal of Health Policy, Law and Ethics* 17, no. 1 (2017): 61-141, *available at* <https://digitalcommons.law.yale.edu/yjhple/vol17/iss1/2/> (last visited January 7, 2019) (describing concerns about coercion and undue inducement related to the traditional norm of not paying for research partici-

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press

pation, and arguing that these concerns reflect confusion); R.D. Truog, A.S. Kesselheim, and S. Joffe, “Paying Patients for Their Tissue: The Legacy of Henrietta Lacks,” *Science* 337, no. 6090 (2012): 37-38, *available at* <http://science.sciencemag.org/content/337/6090/37> (last visited January 7, 2019) (review of other ethical justifications for the norm of not paying individuals, in the context of tissue).

33. J.L. Roberts, S. Pereira, and A.L. McGuire, “Should you Profit from your Genome?” *Nature Biotechnology* 35, no.1 (2017): 18–20, *available at* <https://www.nature.com/articles/nbt.3757> (last visited January 7, 2019).

34. Zenome, *About us*, Zenome website, *available at* <https://zenome.io/> (last visited September 24, 2018); LunaDNA website, *available at* <https://www.lunadna.com/> (last visited September 24, 2018).

https://doi.org/10.1177/1073110519840481 Published online by Cambridge University Press