Genomic Data Sharing

Case Studies, Challenges, and Opportunities for Precision Medicine

Edited by
Jennifer B. McCormick
Jyotishman Pathak

Genomic Data Sharing

Genomic Data Sharing

Case Studies, Challenges, and Opportunities for Precision Medicine

Edited by

Jennifer B. McCormick

Associate Professor, Department of Humanities, College of Medicine, Pennsylvania State University, Hershey PA, United States

Jyotishman Pathak

Professor of Medical Informatics, Professor of Psychiatry, Chief of the Division of Health Informatics, Vice Chair of the Department of Population Health Sciences, Weill Cornell Medicine, New York-Presbyterian Hospital, New York, United States

ACADEMIC PRESS

An imprint of Elsevier

elsevier.com/books-and-journals

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2023 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-12-819803-2

For Information on all Academic Press publications visit our website at https://www.elsevier.com/books-and-journals

Publisher: Stacy Masucci
Acquisitions Editor: Peter B. Linsley
Editorial Project Manager: Samantha Allard
Production Project Manager: Stalin Viswanathan
Cover Designer: Mark Rogers

Typeset by Aptara, New Delhi, India

Contents

Contributors

Shakuntala Baichoo, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Réduit, Mauritius

K.D. Blizinsky, Research Program, National Institutes of Health, Bethesda, MD, USA; Rush Alzheimer's Disease Center, Rush University, Chicago, IL, USA

S. Chandrasekharan, Research Program, National Institutes of Health, Bethesda, MD, USA

Melek Chaouch, Laboratory of Bioinformatics, Biomathematics and Biostatistics LR16IPT09, Institut Pasteur de Tunis, Tunis, Tunisia

Emile R. Chimusa, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle, Tyne and Wear, NE1 8ST, UK

Jorge L. Contreras, Department of Human Genetics, S.J. Quinney College of Law, University of Utah

Deanna Cross, Marshfield Clinic Research Foundation, Marshfield, WI, United States

Pieter De Smet, Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium

Faisal M. Fadlelmola, Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum, Al Khartoum, Al Sudan

Christopher Fields, High-Performance Computing in Biology (HPCBio), University of Illinois, Champaign, IL, United States

Robert R. Freimuth, Department of Artificial Intelligence and Informatics, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, United States

Elena Ghanaim, M.A., National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Anisah Ghoorah, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius

Zoe Gill, Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, Cape Town, South Africa; H3ABioNet, UCT Computational Biology Division, Institute of Infectious Disease and Molecular Medicine, University of Cape Town Health Sciences Campus, Cape Town, South Africa

Melissa Haendel, Center for Health AI, University of Colorado Anschutz Medical Campus

S. Jooma, Research Program, National Institutes of Health, Bethesda, MD, USA

Samar Kamal Kassim, Faculty of Medicine, Ain Shams University, Abbaseya, Cairo, Egypt

Melissa Konopko, GA4GH, Sanger Institute, Hinxton, CB10 1SD, United Kingdom

Judit Kumuthini, South African National Bioinformatics Institute, University of Western Cape. Belville, Cape Town, Republic of South Africa

Michael Lang, Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

Upendo Masamu, Sickle Cell Program, Muhimbili University of Health and Allied Sciences, Dar-es-salaam, Tanzania

Catherine A. McCarty, Marshfield Clinic Research Foundation, Marshfield, WI, United States

Jennifer B. McCormick, Associate Professor, Department of Humanities, College of Medicine, Pennsylvania State University, Hershey PA, United States

Robert P. Milius, National Marrow Donor Program/Be The Match, IT, Minneapolis, MN, United States

Nicola Mulder, Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, Cape Town, South Africa

Zahra Mungloo-Dilmohamud, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius

Mullai Murugan, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, United States

Liberata A. Mwita, Sickle Cell Program, Muhimbili University of Health and Allied Sciences, Dar-es-salaam, Tanzania

James C. O'Leary, Independent Scholar

Angela Page, Global Alliance for Genomics and Health Secretariat, Broad Institute of MIT and Harvard

Jyotishman Pathak, Professor of Medical Informatics, Professor of Psychiatry, Chief of the Division of Health Informatics, Vice Chair of the Department of Population Health Sciences, Weill Cornell Medicine, New York-Presbyterian Hospital, New York, United States

Dimitri Patrinos, Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

Fouzia Radouani, Chlamydiae and Mycoplasma Laboratory, Research Department, Institut Pasteur du Maroc, Casablanca, Morocco

Verena Ras, Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, South Africa; Department of Biodiversity and Conservation Biology, University of the Western Cape, Private Bag X17, Bellville, South Africa

J.A. Reusch, Research Program, National Institutes of Health, Bethesda, MD, USA

Laura Lyman Rodriguez, Ph.D., National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; The Patient-Centered Outcomes Research Institute (PCORI), Washington, DC, USA

Chaimae Samtal, Laboratory of Biotechnology, Environment, Agri-food and Health, Faculty of Sciences, Dhar El Mahraz–Sidi Mohammed Ben Abdellah University, Fez, Morocco

Raphael Z. Sangeda, Sickle Cell Program, Muhimbili University of Health and Allied Sciences, Dar-es-salaam, Tanzania

Dassen Sathan, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius

Milaine S.S. Tchamga, African Institute for Mathematical Sciences, Cape Town, South Africa

Mahsa Shabani, Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium

May Terry, MITRE Corporation, Bedford, MA, United States

Kimberly A. Thomson, Research Program, National Institutes of Health, Bethesda, MD, USA; New Jersey State Department of Health, Division of HIV, STD, and TB Services, Newark, NJ, USA

John Van Horn, Department of Psychology and School of Data Science, University of Virginia, Charlottesville, VA, United States

Gordon Wells, South African National Bioinformatics Institute, University of Western Cape. Belville, Cape Town, South Africa

Lyndon Zass, Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, South Africa

Ma'n H. Zawati, Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

CHAPTER 1

Introduction to the volume

Jennifer B. McCormick[a] **and Jyotishman Pathak**[b]

[a]Associate Professor, Department of Humanities, College of Medicine, Pennsylvania State University, Hershey PA, United States [b]Professor of Medical Informatics, Professor of Psychiatry, Chief of the Division of Health Informatics, Vice Chair of the Department of Population Health Sciences, Weill Cornell Medicine, New York-Presbyterian Hospital, New York, United States

Advances in digital technologies have enabled the collection, aggregation, use, and sharing of unprecedented amounts of data about humans and human activities. This vast and rapidly growing volume of data, often referred to as "big data" can be linked across and integrated into large, centralized databases to investigate the question in both basic and applied science as well as population science.[1] Definitions of "big data" vary with some scholars defining it as "a process of sifting and assimilating disparate data to obtain meaning"[2] while others define it by the three V's, that is, "volume, velocity, and variety"[3,4] or variability of data. Some of these data are from health-related sources such as electronic health records, clinical trial records, biorepositories, and genomic databases. In addition, publicly available data (e.g., national surveys) and commercial datasets (e.g., health insurance claims) also are being mined.[5] Data sources not typically considered health-related can similarly be used.[6] These include environmental data such as climate records and socioeconomic data such as consumers' use of loyalty cards. Social media, wearable fitness devices, and smartphone apps track individuals' daily activities and as such can provide insight into individuals' health.

For researchers, pooled datasets have multiple benefits. First and foremost, they permit investigators to access large sample sizes much larger than what they likely could ever create on their own, given time and resource limitations. Pooled datasets also facilitate linking data across sectors such as genomic data with environmental data.[7] Similarly, by accessing networked data repositories, researchers can repurpose others' data to develop new avenues of investigations perhaps overlooked in the initial study. For those researchers with limited funding, data sharing through pooled datasets provides a valuable resource otherwise inaccessible and unobtainable. Finally, studies have demonstrated the high impact of studies using shared datasets.[8]

The collection, aggregation, and analysis of genetic and genomic data are advancing a deeper understanding of disease etiology, health versus diseased state, disease progression, and

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00007-9
Copyright © 2023 Elsevier Inc. All rights reserved.

disease management. Large-scale, genomic, and genetic data sharing also holds significant promise for patients and research participants. Pooled genetic and genomic data from patients with complex conditions, for instance, can be analyzed to investigate disease causes and progression. The greater the amount of information, the greater the likelihood of advancing understanding. For individuals with rare diseases, combining or pooling datasets collected globally may be their best hope for understanding, diagnosing, and treating their conditions. Until the advent of big data, members of the rare disease community and the clinicians treating them were hamstrung in understanding their conditions because of sparse data and low prevalence.[9,10] Big data has not only advanced the understanding of rare diseases but also helped with the identification of treatment options. Genomic data sharing also can permit the tailoring of diagnostic and treatment decisions for individual patients.[11] Public databases such as The Cancer Genome Atlas allows for sequencing of cancer tumors that can result in identifying mutations. With that knowledge, clinicians can prescribe specific drug treatments to fight tumors.[12]

Large datasets, for instance, promote independent verification of published research findings by researchers not involved in the original investigation, thereby encouraging collaboration and reducing duplicative studies. Pooled datasets also hold the promise of accelerating research results into both new avenues of research. Furthermore, data sharing is key to respecting participants and patients who provide genomic data and who expect their contributions to be used to maximize knowledge and health benefits.[13]

It has been argued that, notwithstanding the benefits for researchers, the scientific community has a social responsibility—indeed, an "ethical imperative"—to share data.[14] Much of the research enterprise today is funded by the government and nonprofit foundations. Data shared, rather than data isolated, helps to ensure the best use of available—and limited—resources whether by expanding the impact of findings, increasing accountability of researchers, or identifying gaps in research. Genomic data also hold promise for advancing understanding of the health needs and inequities of underserved populations, who frequently have been overlooked in clinical trials. With data repositories, researchers can identify statistically significant results which can be particularly important for genomic changes in any given phenotype that might have relatively small effects. In short, data sharing assumes research findings are public goods.

To encourage the broad sharing of genomic data, many governments and funders have policies that foster sharing and provide repositories for storing the data. These policies largely evolved from the 1996 First International Strategy Meeting on Human Genome Sequencing that issued the "Bermuda Principles" with the aim of advancing the public benefits and scientific advancements of the Human Genome Project (https://www.genome.gov/human-genome-project/Timeline-of-Events). The following year (1997), the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) required investigators of large-scale projects to allow for sharing and access of data. NIH issued similar policies

beginning in 1999 and culminating in Data Sharing Policy (GWAS) in 2007 that applied to all NIH-funded projects. This policy has been revised twice (in 2008 and 2014) and applies to both human and non-human genomic data. Providers of human genomic data (e.g., publicly or privately funded repositories and data archives) fulfill their social contract with data donors when their shareable data conforms to FAIR (findable, accessible, interoperable, reusable) principles.

Despite the many advantages of genomic research and support from governmental entities for broad sharing of data; however, concerns exist about the volume of data being collected, used, and shared about us and our activities. Those concerns speak to fundamental differences between how big data research is conducted and overseen and how traditional research has historically been conducted and overseen. As mentioned previously, big data research draws upon publicly available information such as social media, wearable fitness devices, and geolocation data in addition to the more typical health-related data, such as genomic sequencing and medical records.

While this use of unconventional data sources enables innovative research, it also has upended the traditional model of oversight of health-related research. Institutional Review Boards, for instance, are ill-suited to review big data projects using publicly available information or projects that involve collaborations between public and private entities, and reuse of data for new and evolving purposes because they tend to lack the expertise.[15,16] Moreover, the traditional model of informed consent also has been upended. Unlike traditional consent documents developed for use with a specific research project or clinical trial, broad consent documents cover both immediate and unspecified future research. Broad consent, thus, maximizes the "future utility" of collected data.[17–19] However, by definition, donors of data are left in the dark about the nature of that future research. While numerous empirical studies have demonstrated public support for providing biological samples through broad consent,[20–25] studies have also shown that some participants want some control over how their data are used.[6,26–28]

The proliferation of unconventional sources of health data from web search engines to social media platforms also raises questions about the protection of individual privacy, considered a fundamental right. The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, which protects health data, was passed in a world without smartphones, mobile health apps, and wearable fitness and monitoring devices, and no changes have been made to the rule to account for these data sources Data generated by those devices as well as data tracking web searches and consumer purchases are not covered by HIPAA—although those data can reveal personal, private health information.[3,4,29,30] This includes genomic data from direct-to-consumer testing companies like 23andMe. Additionally, HIPAA's solution to protecting individuals through deidentification by removing some identifiers has been proven inadequate. Technological advances and multiple datasets can enable triangulation of data sources and re-identification of participants previously de-identified.[3,18,20] Furthermore, social

media platforms' Terms & Conditions do not clearly indicate that individuals' postings can yield private health data. Such privacy violations can have serious consequences. Individuals can face identity theft, employment discrimination, insurance denial as well as social stigmatization.

These last two issues suggest the need for better governance structures. While the era of genomic data sharing is evolving, the related governance practices and policies have not. Current practices rely heavily on individual consent, which as noted has limitations. Robust governance, on the other hand, includes various stakeholders in decisions about the use and access of collected information. Such inclusion is critical to maintaining the trust of the public. As a concept and a practice, governance is said to provide a way of addressing the ethical, regulatory, and policy challenges of research with personal information. More specifically, governance addresses how and why deidentified data are accessed and used by researchers, who makes those decisions, and how are these decisions made—all of which may not be known by patients who are providing the data. Thus, transparency is also critical to good governance.

This volume will tackle the issue of governance as well as a number of other issues. It also includes several case examples in genomic data sharing. In Chapter 2, Rodriguez and Ghanaim discuss the importance of sharing genomic data and provide examples of how sharing genomic data have enabled different areas of research. They close with a brief discussion of the need for responsible stewardship and the different challenges encountered. In Chapter 3, McCarty and Cross provide an overview of the Marshfield Clinical Personalized Medicine Research Project, a population-based biobank, discussing stakeholder engagement, technical procedures that facilitate genomic data sharing, and two pilot projects. This chapter provides a nice example of a local biobank, while Blizinsky et al. in Chapter 4 give us a broad overview of a national cohort study, the All of Us Research Project. In their chapter, they discuss the ethical, legal, regulatory, policy, and social considerations around genomic data sharing, and illustrate how those considerations have shaped the implementation of the data sharing and return of results for the program as well as explore future directions.

Page et al. discuss the impetus behind the creation of GA4GH and share the history of the organization through to the present day. In their chapter, they describe the GA4GH organizational and operational structures including challenges encountered. They conclude with recommendations that include three core principles they believe to have been of particular value to the success of GA4GH, and which other organizations may find useful as they seek to enable interoperability and sharing across an international community. In Chapter 6, Freimuth et al. provide the reader with case studies that show three examples of how early adoption and testing helped to advance the development of the Genomics Reporting FHIR Implementation Guide. Each case study highlights different types of stakeholders that had unique use cases, requirements, and outcomes. The authors describe how the Health Level 7 International (HL7)

standards organization, Electronic Medical Records and Genomics (eMERGE) Network, the Minimum Common Oncology Data Elements (mCODE) initiative engaged in the standards development process demonstrating how becoming an early adopter can yield significant benefits to both the standards developer and early adopter.

Another case study is shared by Kumuthini et al. in Chapter 7 where the authors discuss how the H3Africa consortium has achieved large-scale genomic data sharing based on the FAIR principles. In this chapter, the authors also discuss two widely adopted genomic data sharing practices: (1) data sharing via a public data repository, and (2) data sharing via a project-specific data commons. Each approach requires the development of robust data management plans that outline the data security and privacy policies, and adherence to standardized common data models, terminologies, and data access protocols. In Chapter 8, Kumuthini et al. discuss the state-of-the-art in standardization of "-omics" data including the development, dissemination, and adoption of data standards. Data standardization is critical to enable wide-scale reuse of data, yet despite the availability of data standards, their adoption is frequently a barrier. This chapter highlights some of the key barriers, and outlines strategies for increased adoption of -omics data standards in the community.

The final four chapters (Chapters 9–12) delve more deeply into the ethical, legal, policy, and social issues surrounding genetic and genomic data sharing. O'Leary provides an overview of public attitudes toward genomic data sharing, giving context to relevant case studies. The chapter also provides a framework for how to approach the issue of genomic data sharing in the age of precision medicine. This chapter is followed by a discussion of the European Union General Data Protection Regulation, or the GDPR. Smet and Shabani describe the possible legal grounds for genetic and genomic data sharing under the GDPR, discussing consent and the public interest. They also shed light on the challenges and uncertainties regarding the use of broad consent in the context data sharing under the GDPR. In the next chapter, Contreras describes copyright and patent law relevant to genetic and genomic data sharing as well as the Bermuda and Fort Lauderdale principles. He also discusses two case examples: *Assn. For Molecular Pathology v. Myriad Genetics* and *Mayo Collaborative Services v. Prometheus Laboratories, Inc.* Patrinos et al. round out the volume with a chapter on data governance. They use genomics as a case study for illustrating how data governance plays an essential role in precision medicine, describing three areas in which data governance is especially important: retrospective data integration, prospective data collection, and data access. They close with the argument that precision medicine scholars need to attend to and appreciate the role that data governance has in ensuring the benefits of precision medicine are realized.

Our goal with this volume is to provide the academic researcher and research administrator a succinct current overview accomplishments and challenges in data sharing from an interdisciplinary perspective, including IT, ELSI, and the researcher participating in large data sharing consortia. Clinician investigators, clinicians affiliated with academic medical

centers, policymakers, and regulators will also gain insight. As Francis Collins, MD, former Director of the US National Institutes of Health (NIH) noted during the 2017 AHSG Annual meeting plenary, "the era of sitting on one's own dataset, reanalyzing it over and over, is over" (paraphrased). He also said that we have a "transformative opportunity" with all the omic data and associated data being generated in the biomedical sciences, and the research community has a responsibility to make sure these datasets are sustainable, accessible, and used fairly by researchers who want to use them. Over the last decade, increasing emphasis has been placed on data sharing, with the goal of making datasets, in particular those generated with public funding, available to a wide range of researchers. Sharing data broadly has its challenges; however, it will no doubt facilitate discoveries and advancement in the biomedical and public health sciences and in medicine.

Acknowledgments

The authors thank Margaret Hopkins for her assistance.

References

1. McKeown A, Mourby M, Harrison P, et al. Ethical issues in consent for the resue of data in health data platforms. *Sci Eng Ethics*. 2021;27:9.
2. Seller R. Addressing benefits, risks and consent in next generation sequencing studies. *J Chin Res Biotech*. 2015;6:4.
3. Belani S, Tiaras GC, Moorkerjee N, et al. "I agree to disagree": Comparative ethical and legal analysis of Big Data and genomics for privacy, consent, and ownership. *Cureus*. 2021;13:2.
4. Price 2nd WN, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25:37–43 Epub 2019 Jan 7. PMID: 30617331; PMCID: PMC6376961. doi:10.1038/s41591-018-0272-7.
5. Schneble CO, Elger VBS, Shaw DM, et al. All out data will be health data one day: The need for universial data protection and comprehensive consent. *J Med Internet Res*. 2020;May 28;22(5):e16879 PMID: 32463372; PMCID: PMC7290498. doi:10.2196/16879.
6. Vayena E, Blasimme A. Health Research with Big Data: Time for Systemic Oversight. *J Law Med Ethics*. 2018 Mar;46(1):119–129 Epub 2018 Mar 27. PMID: 30034208; PMCID: PMC6052857. doi:10.1177/1073110518766026.
7. Byrd JB, Greene AC, Prasad DV, et al. Responsible, practical genomic data that accelerates research. *Nat Rev Genet*. 2020 Oct;21(10):615–629 Epub 2020 Jul 21. PMID: 32694666; PMCID: PMC7974070. doi:10.1038/s41576-020-0257-5.
8. Milham MP, Craddock RC, Son JJ, et al. Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun*. 2018;9:2818. https://doi.org/10.1038/s41467-018-04976-1.
9. Rubinstein YR, Robinson PN, Gahl WA, et al. The case for open science; rare diseases. *JAMIA Open*. 2020;Sep 11;3(3):472–486 PMID: 33426479; PMCID: PMC7660964. doi:10.1093/jamiaopen/ooaa030.
10. Scollen S, Page A, Wilson J. From the data on many, precision medicine for "One": the case for widespread genomic data sharing. *Biomed Hub*. 2017;2(Suppl 1):104–110 PMID: 31988941; PMCID: PMC6945905. doi:10.1159/000481682.
11. Low SK, Zembutsu H, Nakamura Y. Breast cancer: the translation of big genomic data to cancer precision medicine. *Cancer Sci*. 2018;109:497–506 Epub 2017 Dec 30. PMID: 29215763; PMCID: PMC5834810. doi:10.1111/cas.13463.
12. Adams JU. Genetics: big hopes for big data. *Nature*. 2015;527:S108–S109 PMID: 26580158. doi:10.1038/527S108a.

13. Banzi R, Canham S, Kuchinke W, Krleza-Jeric K, Demotes-Mainard J, Ohmann C. Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials*. 2019;20:169 PMID: 30876434; PMCID: PMC6420770. doi:10.1186/s13063-019-3253-3.
14. Carr D, Littler K. Sharing research data to improve public health: a funder perspective. *J Empir Res Hum Res Ethics*. 2015;10:314–316 PMID: 26297752; PMCID: PMC4547198. doi:10.1177/1556264615593485.
15. Ferretti A, Ienca M, Sheehan M, et al. Ethics review of big data research: What should stay and what should be reformed? *BMC Med Ethics*. 2021;22:51 PMID: 33931049; PMCID: PMC8085804. doi:10.1186/s12910-021-00616-4.
16. McKeown A, Mourby M, Harrison P, Walker S, Sheehan M, Singh I. Ethical issues in consent for the reuse of data in health data platforms. *Sci Eng Ethics*. 2021;27:9 PMID: 33538942; PMCID: PMC7862505. doi:10.1007/s11948-021-00282-0.
17. Howe Iii EG, Elenberg F. Ethical challenges posed by big data. *Innov Clin Neurosci*. 2020;17:24–30 PMID: 33898098; PMCID: PMC7819582.
18. Meller R. Addressing benefits, risks and consent in next generation sequencing studies. *J Clin Res Bioeth*. 2015;6:249 Epub 2015 Dec 14. PMID: 27375922; PMCID: PMC4930149. doi:10.4172/2155-9627.1000249.
19. Fisher CB, Layman DM. Genomics, big data, and broad consent: a new ethics frontier for prevention science. *Prev Sci*. 2018;19:871–879 PMID: 30145751; PMCID: PMC6182378. doi:10.1007/s11121-018-0944-z.
20. Garrison NA, Sathe NA, Antommaria AH, et al. A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. *Genet Med*. 2016;18:663–671 Epub 2015 Nov 19. PMID: 26583683; PMCID: PMC4873460. doi:10.1038/gim.2015.138.
21. Goodman D, Johnson CO, Bowen D, Smith M, Wenzel L, Edwards K K. De-identified genomic data sharing: the research participant perspective. *J Community Genet*. 2017;8:173–181 Epub 2017 Apr 5. PMID: 28382417; PMCID: PMC5496839. doi:10.1007/s12687-017-0300-1.
22. Kaufman DJ, Baker R, Milner LC, Devaney S, Hudson KL. A survey of US adults' opinions about conduct of a nationwide precision medicine initiative® cohort study of genes and environment. *PLoS One*. 2016;11 PMID: 27532667; PMCID: PMC4988644. doi:10.1371/journal.pone.0160461.
23. Sanderson SC, Brothers KB KB, Mercaldo ND ND, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. *Am J Hum Genet*. 2017;100:414–427 Epub 2017 Feb 9. PMID: 28190457; PMCID: PMC5339111. doi:10.1016/j.ajhg.2017.01.021.
24. Spruill IJ, Gibbs YC, Laken M, Williams T. Perceptions toward establishing a biobank and clinical data warehouse: voices from the community. *Clin Nurs Stud*. 2014;2 URL: http://dx.doi.org/10.5430/cns.v2n3p97. doi:10.5430/cns.v2n3p97.
25. Trinidad SB, Fullerton SM, Bares JM, Jarvik GP, Larson EB, Burke W. Genomic research and wide data sharing: views of prospective participants. *Genet Med*. 2010;12:486–495 PMID: 20535021; PMCID: PMC3045967. doi:10.1097/GIM.0b013e3181e38f9e.
26. Botkin JR, Rothwell E, Anderson R, Stark LA, Mitchell J. Public attitudes regarding the use of electronic health information and residual clinical tissues for research. *J Community Genet*. 2014;5:205–213 Epub 2013 Dec 5. PMID: 24307509; PMCID: PMC4059848. doi:10.1007/s12687-013-0175-8.
27. De Vries RG, Tomlinson T, Kim HM, et al. The moral concerns of biobank donors: the effect of non-welfare interests on willingness to donate. *Life Sci Soc Policy*. 2016;12:3 Epub 2016 Mar 11. PMID: 26968989; PMCID: PMC4788662. doi:10.1186/s40504-016-0036-4.
28. De Vries RG, Tomlinson T, Kim HM, Krenz C, Haggerty D, Ryan KA, Kim SY. Understanding the public's reservations about broad consent and study-by-study consent for donations to a biobank: results of a national survey. *PLoS One*. 2016;11.
29. Cohen IG, Mello MM. Big data, big tech, and protecting patient privacy. *JAMA*. 2019;322:1141–1142 PMID: 31397838. doi:10.1001/jama.2019.11365.
30. Schneble CO, Elger BS, Shaw DM. All our data will be health data one day: the need for universal data protection and comprehensive consent. *J Med Internet Res*. 2020;22:e16879 PMID: 32463372; PMCID: PMC7290498. doi:10.2196/16879.

CHAPTER 2

From public resources to improving health: How genomic data sharing empowers science and medicine

Laura Lyman Rodriguez, Ph.D.[a,b] **and Elena Ghanaim, M.A.**[a]

[a]*National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA*
[b]*The Patient-Centered Outcomes Research Institute (PCORI), Washington, DC, USA*

2.1 Introduction

Data sharing is an integral element of genomics research. It enables basic, translational, and applied research, promotes data quality, and maximizes the potential benefit achieved through data generated. Patients and research participants who agree to donate the data generated through their care or research participation recognize the value of data sharing for advancing science and health. Therefore, beyond the scientific opportunities that data sharing empowers, the research enterprise has a responsibility to use the information that these individuals contribute to the fullest appropriate extent. Moving forward, improved and more facile data sharing will be essential for realizing the potential of and equitable benefits of genomic medicine and precision health more broadly. As discussed in this chapter, and throughout this volume, the specific applications for data sharing have evolved since the initial sequencing of the human genome, but responsible genomic data sharing is and must remain a bedrock of the genomics field and central to the biomedical research enterprise as a whole.

Data on their own are simply information, but when properly curated and analyzed data create knowledge, and knowledge when applied to specific questions can lead to scientific or medical advances. Combining data (either more of the same data type or layering different data types) makes it feasible to realize outcomes that were not possible on the basis of the primary data itself, and perhaps to answer questions not envisioned when the data were first collected. For example, a 2019 study combined data from 17 Parkinson's disease datasets, analyzing more than 37 thousand cases and 1.4 million controls, discovering genetic risk signals that could not have been found with any one dataset.[1] As Adam Resnick, Director of Data Driven Discovery in Biomedicine (D3b) at Children's Hospital of Philadelphia, has noted, "data sets

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00006-7
Copyright © 2023 Elsevier Inc. All rights reserved.

and their connectivity, immediately impart new paths to knowledge not evident in the initial investigatory-specific cohort."[2] While genomic data sharing presents certain risks, use of data for many purposes or purposes not possible or evident at the time of data collection allows society to maximize the public benefit achieved through research.

Researchers, healthcare providers, health systems, and individual members of the public continue to generate and analyze genomic data in new ways and for expanded purposes, including predictive health, genealogy, or personal interest. The case studies and discussions elsewhere in this book will illustrate how the dynamics among traditional roles in research and medicine are changing and how data sharing is being adapted to serve the new models. The shifts in access to and uses of genomic data outside the research or healthcare context, for example, through direct-to-consumer health or ancestry services, are also reconfiguring the relationships between data generators and data users. The various evolutions in expectations now taking shape will need to be reflected in genomic data sharing policies and practices going forward.

2.2 The Human Genome Project set the paradigm for genomic data sharing

In 2003, an international group of researchers, funded by multiple governments and led in the United States by the National Human Genome Research Institute (NHGRI), completed the Human Genome Project (HGP) ahead of schedule and under budget. The conclusion of the project marked the first time that scientists had the full list of all three billion base pairs in a human genome sequence. HGP researchers used a team science approach to "divide and conquer" the project's goals (e.g., assigning regions of the genome to different teams to sequence and subsequently piecing them together). The so-called "Bermuda Principles" established a data sharing framework for the project and provided a core component of the HGP's success through its paradigm-shifting central tenets of rapid and broad prepublication data dissemination.[3] Concerns about efforts to patent segments of the human genome represented a partial impetus for the data sharing framework's development, but the timely and open data sharing expected under the framework also helped to maintain necessary connectivity and coordination across scientific teams and organize the large, complex project.[3] Cook-Deegan et al.[4] write that "such coordination was necessary for mapping the genome for the first time because it avoided duplicative efforts, ensured data quality, and allowed for verification."

During the initial stages of the HGP, researchers generated the genomes of five model organisms (yeast, bacteria, nematode worm, fruit fly, and mouse).[3] These subprojects adopted and began to implement data sharing practices based on the Bermuda Principles following their completion in 1996. Dedicated public genome databases already existed for some model organisms, in part due to the growth of collaborative communities dedicated to studying

them (e.g., the fruit fly, yeast, mouse, and, in particular, nematode worm communities).[3] The scale-up of such databases represented an explicit effort to create public resources that would facilitate genomics projects unrelated to the genome sequencing efforts themselves. Some termed these types of resource projects "hypothesis-generating" (in contrast to "hypothesis-driven"), since their aims included enabling future researchers to pose varied and innumerable questions. This capacity for stimulating broad use of HGP-generated data amplified the potential public value of producing the reference sequence itself.

The grand achievement of the HGP's completion accelerated the evolution and expansion of genomics as a field. It also marked the beginning of the genome era and the dissemination and application of genomic strategies to many other life science disciplines, including biomedical research. In the years that immediately followed the HGP, additional flagship projects, such as the Encyclopedia of DNA Elements (ENCODE), identified putative elements within the genome sequence, allowing scientists to interrogate their biological functions.[5] Likewise, the International HapMap Project and 1000 Genomes Project continued to increase the granularity with which the genome sequence could be assayed, and further refined the human DNA map by identifying and cataloging variation within and across human populations.[6,7] Importantly, these initiatives were also designed as community resource projects and made all data generated publicly accessible. Among the advances fueled by the availability of these community resource databases were genome-wide association studies (GWAS) with their innovative capacity to explore the breadth of the genome in a single assay. The extraordinary advances in genomic and genotyping technology that followed the completion of HGP enabled this experimental design to become a commonly used technique to identify potential genomic underpinnings of health and disease.[8,9]

In 2007, the National Institutes of Health's (NIH) National Center for Biotechnology Information (NCBI) launched the Database for Genotypes and Phenotypes (dbGaP) to support sharing of large-scale (and deidentified) genomic and phenotypic datasets produced by GWAS through a novel two-tiered data access model.[10] NIH rapidly incorporated dbGaP as a central vehicle to promote the sharing of GWAS data generated with NIH funds in order to capture the explosion of information and responsibly harness its power for public benefit and scientific advancement.[11] As the cost of DNA sequencing technologies continued to decrease,[12,13] the agency subsequently brought other types of large-scale genomic data (such as genome and exome data) under the data sharing policy.[14] Today, researchers from across the globe request access to data housed within dbGaP or similar data repositories, such as the European Genome-phenome Archive.[15] As of January 2020, dbGaP included over 2.4 million molecular assays across 1,418 studies, including over 200,000 exome sequences and nearly 100,000 genomes (Fig. 2.1).[16]

Data sharing on this scale enables researchers to examine the proverbial elephant from multiple vantage points concurrently. This ability to combine datasets is essential to getting a more

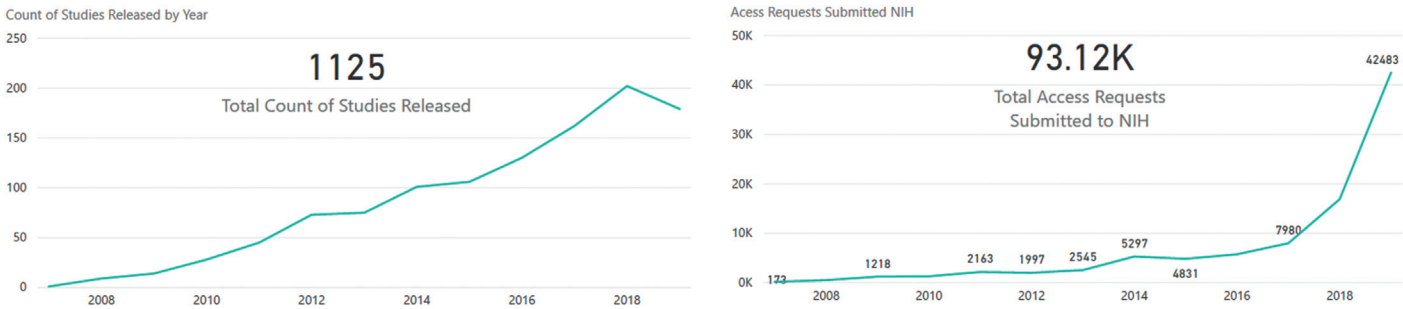

Figure 2.1
Studies available and access request statistics from dbGaP. December 31, 2019.
Figure credit: Jonathan Lawson, Broad Institute.

comprehensive view of complex biological relationships and increases the power of genomic studies. Courbier et al.[17] write:

> Sharing data - rather than data operated in isolation from others - is now recognised as one of the most important ways to ensure benefits for all, including patients, families, scientists, funders, health care providers and future users of the healthcare systems. The basic principal behind data sharing is that the scientific community should, wherever possible, pool their data to gain the maximum benefit from it; this would be, for example, combining two or more datasets from researchers working in the same area, to make one large dataset, which then becomes more statistically significant.

2.3 Genomic data sharing enables multiple areas of research

The core arguments for sharing genomic and associated clinical or phenotypic data, some of which have been discussed above, have been well-articulated by others and are synthesized in Box 2.1.[18–22] This chapter is focused on how genomic data sharing can fulfill the central promise of sequencing the human genome to advance human health. It is important to note that while the arguments below are applicable across biomedical disciplines, the specific context related to contributing research participants' or patients' data donation (e.g., informed consent, confidentiality) must always guide data sharing implementation.

BOX 2.1 Rationale for why genomic data sharing benefits science and society

Ethical/moral

- Fulfills obligations to research participants to use data generated through their participation.
- Maximizes potential for public benefit to be achieved from data.
- Honors the nature of medical research as a public good.
- Promotes trust in the research enterprise by increasing transparency.

Scientific/practical

- Data can be used to explore a wide range of secondary research questions beyond those that could be asked within any single study.
- Data may be combined across studies to power scientific discoveries not feasible in the context of a single dataset.
- Facilitates methods and tool development to advance research.
- Avoids redundant generation of data.
- Improves scientific quality/accuracy/reproducibility/validation of research findings.
- Increases dissemination of research results.
- Accelerates biomedical research by enabling all the items listed above.

Today, we recognize various avenues of basic and clinical research where data sharing is important. Each instance of sharing enables different insights, new tool development, or more

powerful analyses to be possible. The applications range from sharing within model organism communities to facilitate basic biological research to data sharing among rare disease clinicians to identify patients with similar conditions as their own in order to inform potential care options or identify new syndromes.

2.3.1 Research using model organisms

For basic research, the sequencing and deposition of model organism genomes into publicly available databases have continued to yield valuable insights into fundamental biology. Databases such as Flybase, where *Drosophila* researchers share genomic data of many kinds, "play an indispensable role" in facilitating research.[23] Flybase receives over 1.2 million page views per month and is considered a "storehouse of knowledge that is used daily."[23] Similarly, "[the Mouse Genome Informatics (MGI) repository] serves a strategic role for the scientific community in facilitating biomedical, experimental, and computational studies investigating the genetics and processes of diseases and enabling the development and testing of new disease models and therapeutic interventions."[24] These insights into basic biological mechanisms can be extrapolated to inform studies of human pathologies, which ultimately might help identify targets for new drugs and other therapies.[25]

One application of model organism data sharing that benefits biomedical research broadly is genome sequence analyses across species—known as comparative genomics—which can elucidate the connection between biological variation, evolution, and gene function.[26] The ability to compare DNA sequence data from a variety of organisms, including humans, via a common database or by creating a network of databases, allows researchers to look for clues about which genome elements are important for biological processes that contribute to health or disease. Conserved regions or common motifs across species signal preserved functions, whereas domains with extensive variation indicate places in the genome that can tolerate change without maleffect. In this way, genomic variation patterns across species can be a window into the natural history of adaptations that led to evolutions in form or function. The resulting hypotheses about the function of conserved elements can then be tested by manipulating the DNA sequence of an appropriate model organism to provide insights into human biology or pathology and advance health research. The Alliance of Genome Resources, formally established in 2016, is in the midst of modernizing existing databases to improve future data sharing and comparative genomic analyses across species.[27]

2.3.2 Research using human data

In order to inform the identification of variants associated with disease and stimulate translational research to support health applications going forward, there will be an on-going and acute need for broad data sharing to enable the aggregation of large volumes of human

genomic data. Sharing genomic data for secondary research, consistent with the informed consent of the research participants or patients who donated their data, permits investigators to dramatically increase a study's power and thereby discern even weak effects or extremely rare genome variants linked to traits or diseases of interest. Creating the capacity to interrogate sufficiently rich datasets to dissect out these kinds of associations through the aggregation of genomic and other data types (e.g., clinical information or environmental exposures) will become more important as we pursue the genetic and genomic bases of both rare and common diseases.

In rare diseases, exome and genome sequencing are increasingly being deployed in the clinic for patients with undiagnosed conditions, with some successes in the newborn screening scenario as well.[28,29] Approximately 80% of rare diseases have a genetic component.[30] The Undiagnosed Diseases Network, for example, uses a collaborative, cross-disciplinary research approach and genomics to investigate disorders, which are often extremely rare, in children and adults that have eluded diagnosis. To find the underlying genetic causes of a rare genetic disorder, where it is more likely that one or a small number of changes produce the disorder, researchers must compare a patient's genome to a high-quality reference genome to look for change(s) with potential functional consequence. Changes in these patients can be as discrete as just one nucleotide base alteration or a small difference in genome structure. As Brownstein et al.[31] write, "broad data sharing is at the core of achieving [the objectives of the UDN]."

In common disease research, data sharing powers analyses of sufficiently large numbers of individuals with and without disease to narrow in on the specific genomic changes that might contribute with relatively small effect to disease development or progression. Individual genes that contribute to common diseases generally have weaker effects and are but one component among a multitude of factors including one's environment and lifestyle, as well as synergistic gene-gene interactions and gene-environment interactions. Large-scale efforts enabled by the responsible data sharing of tens or hundreds of thousands of individual human genomes are generating breakthroughs and novel insights into the genomic underpinnings of common diseases, such as type 2 diabetes.[32] In 2019, an international forum called the International Common Disease Alliance was launched to accelerate the progress of using genomics to better understand and develop treatments for common disease.[33] Fundamental to its premise is the ability to bring together the scientific community to aggregate datasets and share findings in order to assess progress and develop the next steps to advance understanding and develop innovative medical interventions.

2.3.3 Technical analysis development

Finally, the manipulation and management of increasingly large datasets have and will continue to require targeted research efforts to develop novel computational tools and

methodologies, which then must be trained and tested on sufficiently sized genomic datasets. Database platforms that bring together analytic tools with robust data management infrastructures to support the compilation and data security needs for aggregating shareable datasets at the scale necessary are being developed and will be critical to future advances. One exciting opportunity to capitalize on the large, combined genomic datasets now able to be amassed through data sharing efforts is the use of artificial intelligence for a variety of applications, including to support the identification of clinically relevant variants, identify patients at higher-risk for disease based on their genomic profile, and even to improve the quality of genomic data.[34,35]

2.4 Putting data sharing into practice

Genomics-related professional societies, including the American College of Medical Genetics and Genomics and the American Society of Human Genetics, have articulated the importance of data sharing for medical benefits within their policy statements.[36,37] A culture of sharing data across research fields, not simply in genomics, will be required for ultimate success, however. Calls for this type of expansion of data sharing principles and practices have been put forward and policymakers are moving to create such expectations for other data types (e.g., imaging data, billing codes, free-text notes).[38–40]

Further, for the full benefit of genomic data sharing to be realized, simply sharing molecular data is not enough. Combining disparate datasets to increase the statistical power and potential for knowledge gain requires interoperability of the data, the infrastructures to support data sharing, responsible and robust data management, and capacities for dataset analysis. Going forward, all stakeholders should be more deliberate about adequate documentation of metadata, increased deposition of phenotype data when possible, integration of common ontologies, and the use of standard, machine-readable variables. Community-driven efforts, including international consortia such as the Global Alliance for Genomics and Health (GA4GH) and the Global Genomic Medicine Collaborative (G2MC) in the genomics research and clinical implementation domains respectively, are championing and working to address these needs.[41,42] Policy efforts are also capturing emerging standards for genomic and health data quality and building them into expectations to maximize the potential for data utility. An example in this space includes the integration of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles[43] into policy frameworks such as the Precision Medicine Initiative's Privacy and Trust Principles and Data Security Policy Framework to ensure the initiative's data sharing goals are achieved. Resourcing and managing the substantial costs associated with sustainable and FAIR data sharing practices and the infrastructure necessary to support it over time is and will be a substantive challenge. Research sponsors must continue to work through such challenges and the need to balance centralized and distributed responsibilities for these intrinsic data sharing costs.[44,45]

2.5 Data sharing will propel precision medicine

Precision medicine is an emerging area with substantial potential to achieve increased benefit through research and clinical data sharing. Unlike traditional medicine that treats disease based on interventions developed to work for the majority of people, precision medicine seeks to diagnose, and treat individuals and, ideally, to predict and then diminish the risk for disease onset, based on a person's unique biological constitution and clinical presentation.[46] Though precision medicine is focused on targeting medical care to an individual, it is powered by analyses of data from many people and populations. Its meaningful implementation will require far more data than are gathered via a conventional clinical trial because such study designs are simply not large enough to identify the small effects of individual factors that contribute to most health outcomes. As noted elsewhere in this chapter, the need to discern and contextualize the relatively weak signals commonly identified in precision medicine studies in order to realize health benefit for all necessitates that ancestrally diverse populations be engaged and included in research studies and their design.

At the leading edge of precision medicine is "genomic medicine," which, as the name implies, informs patient treatment strategies based on genome sequence information. This term has been defined by NHGRI[47] as:

> … an emerging medical discipline that involves using genomic information about an individual as part of their clinical care (e.g. for diagnostic or therapeutic decision-making) and the health outcomes and policy implications of that clinical use. Already, genomic medicine is making an impact in the fields of oncology, pharmacology, rare and undiagnosed diseases, and infectious disease.

The primary form of genomic medicine to date has been genetic, and more recently genomic, testing, which now ranges from the level of looking for a specific nucleotide change through to sequencing of a patient's entire genome. Some key types of genetic testing include: diagnostic testing, which looks for a known or suspected potential genetic cause of a patient's condition; predictive testing, which looks for specific genomic variation with known disease associations as a means to identify a person's risk of developing a condition; and pharmacogenomic testing, a form of predictive testing focused on how a person is likely to respond to certain medications based on their specific genomic variants.[48,49]

As described earlier, one instance where genetic and genomic testing is known to be extremely powerful is for the diagnosis of rare diseases. Sharing data in order to create large data resources allows researchers to assess with greater accuracy how rare or common a genetic variant of interest is in a given ancestral population. This knowledge is important for interpreting the clinical significance of a variant since common variants are substantially less likely to have a pathogenic effect.[20] Openly available resources such as the Genome Aggregation Database (gnomAD), which provides freely accessible summary data from

over 100,000 exomes and genomes (including ancestrally diverse populations) greatly enable this effort.[50,51] Rare disease research also benefits from efforts such as MatchMaker Exchange,[52] which links databases from around the world to enable the identification of matched genotypes from research participants and patients. This global accessibility of genomic information, which makes it possible for rare and ultra-rare disease patients to find analogous cases, is often considered essential for identifying and verifying causal variants.[53] In reflecting on his team's identification of a previously undescribed syndrome and discovery of the underlying genetic mutations responsible for its manifestation, Dr. Daniel Kastner, NIH Distinguished Investigator at NHGRI, highlighted the importance of data sharing saying, "This discovery underscores the tremendous power of combining astute clinical observation, state-of-the-art DNA sequencing, and the sharing of sequence data in large publicly-accessible databases."[54] Lastly, a benefit of data sharing that goes beyond the scientific and medical realm is the potential for personal utility and benefit to patients and their families because of the opportunity to identify and form social communities of those living with the same or similar conditions. Indeed, rare disease patients often expect and promote broad data sharing processes; their perspectives should be integral to policy discussions about how to achieve appropriate research participant and patient protections in research and health-related data sharing.

The case for predictive testing is increasingly being considered for determining risk for common diseases or predicting prognosis once the disease is present. However, it is only recently that researchers have had access to datasets large enough to study the complex, polygenic, and other contributions to target outcomes on a scale that might prove useful for anticipatory and even preventative individualized care applications. In 2018, Kathiresan and colleagues published seminal work on the development of polygenic risk scores for common diseases such as coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer.[55] The development and validation of polygenic risk scores to describe a person's relative risk for developing a common disease are exciting to many. However, it is important to note that before the approach is adopted for clinical care more broadly, there will need to be improvements to methods used in score generation, the ancestral diversity of the data used to generate the scores, as well as the development of clinical guidelines for healthcare providers.[56] These fundamental steps to advance the potential clinical application of polygenic risk scores will require responsible, broad genomic data sharing and the infrastructure and policies to support it.

Pharmacogenomics is another burgeoning area of genomic medicine that highlights the need for sustained data sharing. This genomic medicine application is expected to improve patient care by shifting from the current method of prescribing treatments primarily through "trial and error" to prescribing strategies based on an individual's likely response to a particular drug or their potential to have an adverse outcome. Pharmacogenomics is being implemented in a variety of areas, including HIV, cancer, and mental health.[57] Data sharing

will remain critical to the progress of pharmacogenomics because on-going data and experience collection will be necessary to inform and refine evidence development for clinically relevant associations between specific variants and drugs as well as differential sensitivities to drugs.

Community resource projects, such as the NCBI Clinical Variant Resource (ClinVar) and the NIH-funded Clinical Genome Resource (ClinGen), further demonstrate how central genomic data sharing is to genomic medicine. ClinVar is a data archive that aggregates information about individual genetic variants observed by organizations that are conducting genomic testing, the context within which they were observed, and the level of evidence available for any connection between a genomic variant and disease.[58] ClinVar is foundational for ClinGen, which builds upon the data archive to provide information about the type of evidence available for variants listed in ClinVar. ClinVar and ClinGen have quickly become essential and highly synergistic efforts that promote data sharing and using community standards for high-quality curation of genomic variants contributing to disease. These resources enable clinicians and genetic test developers to understand more deeply the potential clinical relevance of specific genome information in the context of a patient.[59] ClinGen and ClinVar, which registered its millionth variant in December 2019,[60] receive a number of submissions from the private sector, which is an indicator of how fundamental a resource the community views them to be for downstream clinical applications. Significantly, in December 2018, the Food and Drug Administration (FDA) recognized ClinGen as the first FDA-designated public genetic database. This formal FDA recognition allows test developers to rely on the database to support premarket submissions for future genetic tests.[61]

A different type of challenge for genomic medicine, and one that illustrates the important intersection of social and scientific issues in genomics, is the significant issue of underrepresentation of nonwhite communities in research. This imbalance prevents the development of a robust understanding of human variation and disease etiology across populations and impedes the pursuit of precision medicine for all patients.[62] It has been shown that most individuals whose data are included in genomic databases are of European descent, resulting in a much larger number of variants of uncertain significance for patients of non-European descent. This lack of diversity in the available data will affect the equity realized by genomics research[63] by limiting the utility of genetic test results for these patients.[64] It is, thus, imperative that the genomics community as a whole address this imbalance to realize the full potential and equitable benefit of genomic medicine across all populations,[63] and to do so by first and foremost being trustworthy and transparent stewards of participant and patient data. Data sharing could be an important element among the efforts to promote such equity by enabling broad dissemination of genomic datasets (with appropriate consent and full transparency) that include ancestrally diverse populations in order to maximize the potential for the data's inclusion in other research studies and clinical genomics resources. Communities that are invited to participate in research should be consulted (ideally directly involved) when

decisions are considered regarding if and how data from a project will be shared more broadly.[65] In addition, the genomics community should consider innovative governance strutures that reduce risks of so called "helicopter resaerch", stigmatization, or other harms.

2.6 Learning healthcare systems and data sharing

Genomics, by its very nature as the study of the entire DNA sequence of an organism, is the study of "big data." A single human genome, with its three billion individual data points, represents both a large volume[1] and a highly-complex variety of information to be interrogated.[66] GA4GH predicts that over the next several years, more than 60 million patients will have some form of genomic sequencing conducted by their healthcare system and notes that China alone plans to sequence 100 million human genomes by 2030.[67] As examples of big data analyses generating novel findings mount in a variety of health disciplines (e.g., public health, meta-analyses of genome-wide association studies, patient profile analytics), large healthcare systems are considering how to harness the data they collect and the workflows to capture patient information at the scale needed to enable active learning and support for precision medicine applications. As the early adopters of genomic medicine and other big data strategies launch new initiatives, they are probing how to effectively learn from patient data and develop options for the design and efficient implementation of evidence-based innovations into practice. This type of healthcare system—that is, one intentionally designed to learn from the data collected in the course of clinical workflows to continuously seek new understanding and to enhance patient care—is known as a learning healthcare system.[68,69]

Data sharing is one component of the "comprehensive implementation plan" needed to make precision medicine a reality in a learning healthcare system.[70] Genomic medicine, as an example, will require on-going data sharing within and across health systems, because there is much more to be learned about how variants might affect disease or health in different scenarios. Updated interpretations will need to be relayed to the clinical team and the patient in meaningful formats for their respective purposes.[71] Better data collection, a trained workforce, and improved health IT systems, among other infrastructure and resources, are also necessary to ensure the accurate and facile data sharing necessary to enable precision medicine applications.[70] To assess and realize the potential for cost savings from genomic medicine strategies, clinical systems will need to be able to capture and share health outcome data connected to the use of genomic data. For example, Verbelen et al.[72] write that "PGx-guided treatment can be cost-effective and even a cost-saving strategy. Having genetic information readily available in the clinical health record is a realistic future prospect and would

[1] To illustrate the size of a human genome, O'Driscoll et al. [66] estimate that the three billion base pairs are 100 GB of data, which is the same as 102,400 photos.

make genetic tests economically worthwhile." There are examples of various efforts to integrate (and improve the integration of) genomic data into electronic health records.[73–75] These are early and important steps to achieve the degree of data sharing necessary to implement genomic medicine successfully within healthcare systems.

While the changes needed to advance health data sharing are significant, it is also important to be aware that this evolution in the way that a healthcare system interacts with patient data and with learning activities modifies the typical roles of research and healthcare stakeholders and how they relate to one another. In a learning healthcare system, the delivery of clinical care can take on elements similar to the research enterprise, healthcare providers become data collectors, and patients could be considered to be taking on some of the roles more typical of research participants. Due to these shifting dynamics, some have suggested that there should be a new "social contract" created where patients, not just research participants, agree to have their data shared for the good of future generations.[76,77] Others frame it as rebalancing the ethical priorities of the research enterprise—increasing the importance of public benefit and transparency,[78] and moving away from overprotectionism.[79] This notion of an updated social contract relies on the premise that by studying the many, you will help the individual, and vice versa, by studying the outcomes of one, you might inform future care for many. While protection from potential harm must remain at the core of ethical research conduct, the increasingly frequent emphasis in the literature on the potential societal goods that data sharing generates suggests a move toward the recognition of how responsible data sharing practices can contribute to beneficence (by maximizing the possible benefits of research) in assessing the ethical foundations of genomic and other health data sharing.

2.7 Need for responsible data stewardship

Moving forward, the acceptability and success of the increased level of data sharing necessary to support precision medicine research and translation will rest heavily on responsible data stewardship to earn and sustain public trust. Multiple parties will be required to act as data stewards during the course of the research process, for example, those managing, governing, accessing, or using individual-level personal health and genomic data, and standards for conduct in these roles must be set at the highest levels. Data stewards in all sectors must continuously seek a responsible and respectful balance between providing patient and participant autonomy with regard to the boundaries for access to and use of their personal information and enabling research and health data to be used as broadly as is appropriate.

Full success will also require culture change regarding how researchers and other data stewards view data. There must be an increased awareness that data,even in its deidentified or derivative forms, are more than simply digital information; they represent a person who made a voluntary choice to share their information or biosample. Thus, data confidentiality, security,

and any parameters for use attached to an individual's data donation must be given the utmost attention, and the interests of the patient or research participant respected and protected. Finally, there must also be transparency in practice across data workflows from collection to use.

Due to the potential for reidentification of genome sequence information, and other sensitivities with this and associated data types, research sponsors and research community forums have developed frameworks to support responsible genomic data use. For example, the NIH Genomic Data User Code of Conduct lays out expectations for those utilizing controlled-access genomic data from an NIH-designated data repository. Chief among the expectations is that data only be used for the research purpose for which access was approved (consistent with any limitations established through the participants' consent), and that data users make no attempt to reidentify participants or groups within study populations.[80] GA4GH has outlined four foundational principles and ten core elements of responsible genomic and health data sharing (Fig. 2.2), and Knoppers et al. have called for an international code of conduct to guide genomic data sharing across the globe.[81,82]

2.8 Barriers to genomic data sharing

Though there is a longstanding tradition of data sharing in genomics, and many opportunities that data sharing will enable, important challenges remain to be addressed. Box 2.2 details a range of technical and social obstacles with the potential to slow progress in the field, many of which have been discussed in greater depth elsewhere in this chapter.[83–85] This list is not intended to be exhaustive, as there are additional factors that may impede genomic data sharing efforts.

BOX 2.2 Obstacles to data sharing

Lack of Incentives—Sharing data requires resources, including time, personnel, funds, etc. Data sharing activities need to be properly resourced, include a reward system for good data sharing practices (e.g., data sharing should be valued for career advancement considerations, an acknowledgment in publications, etc.), and disincentives for failing to share valuable data need to be in place.

Lack of Data and Metadata Standards—Data and metadata standards are needed to ensure that data are of good quality and clear provenance and can be combined with datasets including similar variables to increase the statistical power of future analyses. There is a natural tension intrinsic to assessing at what point and how (and by whom) data standards are most appropriately created and implemented.

Privacy Concerns—It is challenging to balance the risk of and public concern about reidentification and other privacy issues with the benefits of data sharing. Though it is useful to have

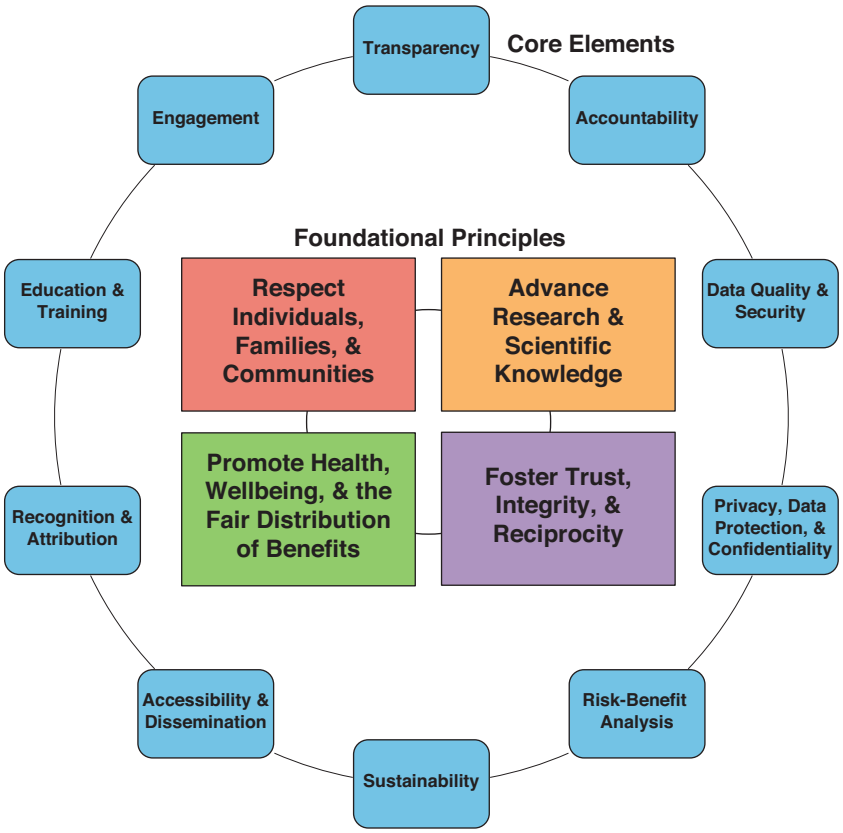

Figure 2.2
GA4GH framework for responsible sharing of genomic and health-related data.
Figure credit: Adapted from Knoppers.[81]

access to many information types about an individual to gain a more complete picture of the variables that contribute to health and disease (e.g., environmental, genomic, phenotypic, geographic, socioeconomic, etc.), such robust information can make it trivial to reidentify an individual. The need to address these issues is not limited to the research or clinical domains, but extends to those undertaking consumer/patient-driven data collection and sharing for health and nonhealth purposes.

Paternalistic Views—In an effort to protect research participants, some stakeholders within the research community may be overly conservative about with whom or how they believe it is appropriate to share genomic data. Policies and practices should align with community preferences and attitudes toward data sharing based upon data collected and applied to policy development through rigorous and transparent methods.

(*continued*)

BOX 2.2 Obstacles to data sharing —*cont'd*

Lack of a Data Sharing Culture—Due to traditional academic advancement models, researchers may be reluctant to share the data that they generate fully or in a timely way to ensure that they are able to conduct analyses to support the first (or exclusive) publication on the data they generated or developed for analysis.

Capacity Issues—As the cost of genome sequencing decreases, and the number of individuals who have had their genomes sequenced increases, there are growing infrastructure needs and costs for storing, securing, and computing on such large volumes of data.

Scalability Issues—As the number of controlled-access datasets grows, current models to adjudicate and oversee data access and use will not be able to meet demand. Innovative approaches to meet data sharing responsibilities and policy mandates through streamlined or automated methods must be devised and integrated such that the steps requiring manual (human) oversight are most judiciously implemented.

Data Inclusivity Issues—Lack of data from diverse populations within genomic data resources is known to affect the generalizability across populations of the results generated. Meaningful engagement and partnership with communities is needed to explore the principles related to building and sustaining trust with participants and patients to earn their willingness to contribute genomic information to data repositories for research or learning health approaches.

Sustainability Issues—It is unclear who should support the technical resources and infrastructures needed for data sharing over the long-term. It is challenging to distribute the balance of responsibilities between centralized and distributed funding for these foundational needs.

Addressing the technical barriers will be no small feat and will require the efforts of standards-generating organizations, genomic data scientists, and improvements to current infrastructure, among other developments. However, perhaps more daunting are the social issues, which may persist as long-standing barriers if general agreement on solutions or even paths forward cannot be reached. There are not universally appropriate answers in some instances, such as questions about how to balance the increasing risk of reidentification with the benefit of broad data sharing, how to incentivize researchers to share data, or how to provision and distribute the resource obligations required for the increasing data science infrastructure needs. What is clear is that these issues require substantial consideration and creative solutions to ensure that genomics research continues to advance our understanding of biological principles and other factors contributing to health and disease, achieves equitable access to and benefit from precision medicine for all, and, most importantly, fulfills the community's obligations to the research participants and patients who share their data in the hope of realizing benefit for future generations.

2.9 Conclusion

Genomics is at the leading edge of recognizing the benefits and addressing the challenges of implementing data sharing practices within biomedical research. However, dramatic decreases in genome sequencing technology costs, increasing computational capacity across sectors, and the resulting surge in access and relative affordability of genomic data generation and interpretation methodologies have led to mammoth increases in the numbers of and purposes for which genome sequence data are produced (both human and nonhuman). Furthermore, with the rapid shift from genomics data generation primarily in research settings to it increasingly emanating from the clinical arena, there is an explosion of information available to be explored to advance biological understanding and improve human health. To make the most of this extraordinary opportunity, the field must continue to move forward through deliberate and direct attention to the technical and social issues that could diminish the potential public benefit that data sharing offers.

Digital data, genomic data included, are not finite in that they can be used over and over again without being consumed. Thus, the potential for use is unbounded. However, practical and ethical limits on data use and data sharing do and should exist. While the technical challenges stem mostly from the capacity to manage, store, process, and analyze the volume of data,[66] critical considerations about if, when, and how data *should* be used, and not simply how many scientific questions can be asked, must be integral to any consideration of solutions or policy proposals for how to move forward. Discourse about genomic data sharing should incorporate deliberation of both benefits and risks to ensure appropriately balanced options are advanced and then evaluated against experience and through rigorous research and evaluation.

While sequencing the human genome was a momentous scientific milestone and achievement, it was but a single step prerequisite to applying genomic information to advance science and health. As both data stewards, research participants, or patients it is useful to remember the practical and philosophical reasons for why we share data. The need for data sharing will persist, or even become more pressing, into the future because of the sheer volume of genomic and other data needed to understand how genomic differences among individuals and populations contribute to variable risks and protections from disease, the aging process, and lifestyle and environmental influences on health. The case studies that follow this chapter will provide windows into how data sharing is advancing research and clinical care through more powerful, more efficient, or more inclusive strategies that capture scientific opportunity while also respecting and honoring the research participants and patients who contribute their data to improve human health.

References

1. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2019;18:1091–1102.
2. Resnick C's. *The Cancer Letter*. Big data will transition from research to the standard of care in the clinic; 2019.
3. Jones KM, Ankeny RA, Cook-Deegan R. The Bermuda Triangle: the pragmatics, policies, and principles for data sharing in the history of the Human Genome Project. *J Hist Biol*. 2018;51:693–805.
4. Cook-Deegan R, Ankeny RA, Jones KM. Sharing data to build a medical information commons: from Bermuda to the Global Alliance. *Annu Rev Genomics Hum Genet*. 2017;18:389–415.
5. Ecker JR, Bickmore WA, Pritchard JK, et al. Genomics: ENCODE explained. *Nature*. 2012;489:52–54.
6. Altshuler D, Donnelly P, Consortium IHM. A haplotype map of the human genome. *Nature*. 2005;437:1299.
7. Genomes Project Consortium, an integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56.
8. Genome-Wide Association Studies Fact Sheet. National Human Genome Research Institute. Accessed October 26, 2019. Updated 2015. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet.
9. Trevino V, Falciani F, Barrera-Saldaña HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med*. 2007;13:527–541.
10. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39:1181–1186.
11. National Institutes of Health, Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS), Department of Health and Human Services, Editor. 2007, Federal Register. p. 49290-49297.
12. Hayden EC. The $1,000 genome. *Nature*. 2014;507:294.
13. Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40: past, present and future. *Nature*. 2017;550:345–353.
14. National Institutes of Health, Final NIH genomic data sharing policy, Department of Health and Human Services, Editor. 2014, Federal Register. p. 51345–51354.
15. European genome-phenome archive. Accessed January 7, 2020. https://ega-archive.org/.
16. Summary Statistics of dbGaP Data. National Center for Biotechnology Information. Accessed October 31, 2019. https://www.ncbi.nlm.nih.gov/projects/gap/summaries/cgi-bin/molecularDataPieSummary.cgi.
17. Courbier S, Dimond R, Bros-Facer V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection-quantitative survey and recommendations. *Orphanet J Rare Dis*. 2019;14:175.
18. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. caBIG Data Sharing and Intellectual Capital Workspace. Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med*. 2008;5:e183. doi:10.1371/journal.pmed.0050183.
19. Knoppers BM, Harris JR, Tassé AM, et al. Towards a data sharing Code of Conduct for international genomic research. *Genome Med*. 2011;3:46.
20. Raza S, Hall A. Genomic medicine and data sharing. *Br Med Bull*. 2017;123(1):35–45. doi:10.1093/bmb/ldx024.
21. Downey AS, Olson S. *Sharing Clinical Research Data: Workshop Summary*. Washington, D.C., USA: National Academies Press; 2013.
22. NIH GDS Policy Overview. Office of Science Policy. Accessed November 5, 2019. https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy_Overview.pdf.
23. Bilder D, Irvine KD. Taking stock of the Drosophila research ecosystem. *Genetics*. 2017;206:1227–1236.

24. Eppig JT. Mouse genome informatics (MGI) resource: genetic, genomic, and biological knowledgebase for the laboratory mouse. *ILAR J*. 2017;58:17–41.
25. Simmons D. The use of animal models in studying genetic disease: transgenesis and induced mutation. *Nat Edu*. 2008;1:70.
26. Comparative genomics fact sheet. National Human Genome Research Institute. 2015. Accessed October 26, 2019. https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet.
27. Alliance of Genome Resources ConsortiumThe Alliance of Genome Resources: building a modern data ecosystem for model organism databases. *Genetics*. 2019;213:1189–1196.
28. Bodian DL, Klein E, Iyer RK, et al. Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. *Genet Med*. 2016;18:221–230.
29. Park KJ, Park S, Lee E, et al. A population-based genomic study of inherited metabolic diseases detected through newborn screening. *Ann Lab Med*. 2016;36:561–572.
30. RARE Facts. Global Genes. 2019. Accessed October 31, 2019. https://globalgenes.org/rare-facts/.
31. Brownstein CA, Holm IA, Ramoni R, et al. Data sharing in the undiagnosed diseases network. *Hum Mutat*. 2015;36:985–988.
32. Varshney A, Scott LJ, Welch RP, et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci*. 2017;114:2301–2306.
33. International Common Disease Alliance, From maps to mechanisms to medicine: using human genetics to propel the understanding and treatment of common diseases. Accessed January 7, 2020. https://www.icda.bio/sites/default/files/2019-09/ICDA%20While%20Paper.pdf.
34. Xu J, Yang P, Xue S, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Hum Genet*. 2019;138:109–124.
35. Rees, V. Uniting humans and data: the role of AI in genomics. Drug Target Review. 2019. Accessed January 7, 2019. https://www.drugtargetreview.com/article/47942/uniting-humans-and-data-the-role-of-ai-in-genomics/
36. ACMG Board of DirectorsLaboratory 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*. 2017;19:721–722.
37. American Society of Human GeneticsAdvancing research and privacy: achievements, challenges, and core principles. *Am J Hum Genet*. 2019;105:445–447.
38. National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: The National Academies Press; 2011:142.
39. Stebbins M. Expanding public access to the results of federally funded research. Obama White House Blog. February 22, 2013. Accessed January 7, 2020. https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.
40. National Institutes of Health, Request for public comments on a DRAFT NIH policy for data management and sharing and supplemental DRAFT guidance, Department of Health and Human Services, Editor. 2019, Federal Register. p. 60398–60402.
41. Global Alliance for Genomics & Health. (2017). *GA4GH connect: a 5-year strategic plan.* Accessed January 7, 2020. https://www.ga4gh.org/wp-content/uploads/GA4GH-Connect-A-5-year-Strategic-Plan.pdf.
42. Ginsburg GS. A global collaborative to advance genomic medicine. *Am J Hum Genet*. 2019;104:407–409.
43. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18.
44. Corpas M, Kovalevskaya NV, McMurray A, Nielsen FGG. A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Comput Biol*. 2018;14(3):e1005873. doi:10.1371/journal.pcbi.1005873.
45. Big Data @ NSF. National Science Foundation. Accessed January 7, 2020. https://www.nsf.gov/cise/bigdata/.
46. National Library of Medicine. What is precision medicine? Genetics Home Reference. April 2015. Accessed November 12, 2019. https://ghr.nlm.nih.gov/primer/precisionmedicine/definition.

47. Genomics and Medicine. National Human Genome Research Institute. April 2019. Accessed November 9, 2019. https://www.genome.gov/health/Genomics-and-Medicine.
48. National Library of Medicine. What are the types of genetics tests? Genetics Home Reference. October 2019. Accessed November 9, 2019. https://ghr.nlm.nih.gov/primer/testing/uses.
49. Pharmacogenomics FAQs. National Human Genome Research Institute. May 2016. Accessed November 9, 2019. https://www.genome.gov/FAQ/Pharmacogenomics.
50. Francioli, L and D MacArthur. gnomAD v3.0. MacArthur Lab. October 16, 2019. Accessed January 7, 2020. https://macarthurlab.org/2019/10/16/gnomad-v3-0/.
51. Broad Institute. About gnomAD. gnomAD browser. Accessed November 19, 2019. https://gnomad.broadinstitute.org/about.
52. Philippakis AA, Azzariti DR, Beltran S, et al. The matchmaker exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36:915–921.
53. Keener, AB. Exome sequencing helps crack rare disease diagnosis. The Scientist Magazine. May 1, 2018. Accessed November 9, 2019. www.the-scientist.com/features/exome-sequencing-helps-crack-rare-disease-diagnosis-64277.
54. Ganguly, P. Researchers discover a new autoinflammatory disease called CRIA syndrome. National Human Genome Research Institute. December 23, 2019. Accessed January 7, 2020. https://www.genome.gov/news/news-release/NHGRI-Researchers-discover-a-new-autoinflammatory-disease-called-CRIA-syndrome.
55. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50:1219–1224.
56. Sugrue LP, Desikan RS. What are polygenic scores and why are they important? *JAMA*. 2019;321:1820–1821.
57. Roden DM, McLeod HL, Relling MV, et al. Pharmacogenomics. *Lancet*. 2019;394:521–532.
58. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1):D862–D868.
59. Rehm HL, Berg JS, Brooks LD, et al. ClinGen—the clinical genome resource. *N Engl J Med*. 2015;372:2235–2242.
60. National Center for Biotechnology Information. ClinVar Celebrates 1 Million Submissions. National Library of Medicine. December 20, 2019. Accessed January 26, 2020. https://ncbiinsights.ncbi.nlm.nih.gov/2019/12/20/clinvar-celebrates-million-submissions/.
61. Food and Drug Administration. FDA takes new action to advance the development of reliable and beneficial genetic tests that can improve patient care. December 4, 2018. Accessed January 7, 2020. https://www.fda.gov/news-events/press-announcements/fda-takes-new-action-advance-development-reliable-and-beneficial-genetic-tests-can-improve-patient.
62. Hindorff LA, et al. Prioritizing diversity in human genomics research. *Nat Rev Genet*. 2018;19:175.
63. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nat News*. 2016;538:161.
64. Hoffman-Andrews L. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *J Law Biosci*. 2017;4:648.
65. May T, Bogar S, Spellecy R, Kabasenche W, Craig J, Dick D. Community-Based Participatory Research and its Potential Role in Supporting Diversity in Genomic Science. *J Health Care Poor Underserved*. 2021;32(3):1208–1224. doi:10.1353/hpu.2021.0127.
66. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. *J Biomed Inform*. 2013;46:774–781.
67. Stark Z, et al. Integrating genomics into healthcare: a global responsibility. *Am J Hum Genet*. 2019;104:13–20.
68. Roundtable on Value & Science-Driven Health Care. The Learning Health System and its Innovation Collaboratives. Institute of Medicine of the National Academies. 2011. Accessed January 26, 2020. http://www.nationalacademies.org/hmd/Activities/Quality/~/media/Files/Activity%20Files/Quality/VSRT/Core%20Documents/ForEDistrib.pdf.
69. McGinnis JM, Saunders R, Stuckhardt L, McGinnis JM, Committee on the Learning Health Care System

in America; Institute of Medicine, eds. *Best Care at Lower Cost: the Path to Continuously Learning Health Care in America*. Washington, D.C., USA: National Academies Press; 2013.
70. Chanfreau-Coffinier C, Peredo J, Russell MM, et al. A logic model for precision medicine implementation informed by stakeholder views and implementation science. *Genet Med*. 2019;21:1139–1154.
71. The Office of the National Coordinator for Health Information Technology (ONC). Precis Med. Healthit.gov. 2019. Accessed December 3, 2019. https://www.healthit.gov/topic/scientific-initiatives/precision-medicine.
72. Verbelen M, Weale ME, Lewis CM. Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet? *Pharmacog J*. 2017;17:395–402.
73. Ohno-Machado L, et al. Genomics and electronic health record systems. *Hum Mol Genet*. 2018;27(R1):R48–R55.
74. The Office of the National Coordinator for Health Information Technology (ONC). *Sync for Genes.* Healthit.gov. 2017. Accessed December 3, 2019. https://www.healthit.gov/topic/sync-genes.
75. Gottesman O, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet Med*. 2013;15:761–771.
76. Nimita L, Carol I. The modern social contract between the patient, the healthcare provider, and digital medicine. *J Socialomics*. 2014;3:105.
77. Mills, P and Miller J. Why we need a new social contract for data in healthcare. World Economic Forum. March 21, 2019. Accessed December 3, 2019. https://www.weforum.org/agenda/2019/03/why-we-need-a-new-social-contract-for-data-in-healthcare/.
78. Deverka PA, et al. Hopeful and concerned: public input on building a trustworthy medical information commons. *J Law Med Ethics*. 2019;47:70–87.
79. Meslin EM, Cho MK. Research ethics in the era of personalized medicine: updating science's contract with society. *Public Health Genom*. 2010;13:378–384.
80. National Institutes of Health. Genomic data user code of conduct. Office of Science Policy. Accessed December 3, 2019. https://osp.od.nih.gov/wpcontent/uploads/Genomic_Data_User_Code_of_Conduct.pdf.
81. Knoppers BM. Framework for responsible sharing of genomic and health-related data. *The HUGO J*. 2014;8:3.
82. Phillips M, Molnár-Gábor F, Korbel JO, et al. Genomics: Data sharing needs an international code of conduct. *Nature*. 2014;578(7793):31–33. doi:10.1038/d41586-020-00082-9.
83. Wellcome. Sharing research data to improve public health: full joint statement by funders of health research. Accessed December 3, 2019. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm.
84. McGuire AL, et al. Who owns the data in a medical information commons? *J Law Med Ethics*. 2019;47:62–69.
85. Cook-Deegan R, Majumder MA, McGuire AL. Introduction: sharing data in a medical information commons. *J Law Med Ethics*. 2019;47:7–11.

CHAPTER 3

Biobank case example: Marshfield clinic

Catherine A. McCarty and Deanna Cross
Marshfield Clinic Research Foundation, Marshfield, WI, United States

Genomic data sharing for precision medicine research has two primary components: (1) stakeholder engagement to ensure ethical guidelines and requirements are met while ensuring buy-in from stakeholders, and (2) technical procedures to facilitate data sharing with investigators from other institutions who have potentially used different techniques for specimen and data management. We will describe how these two components were handled in the Marshfield Clinic Personalized Medicine Research Project (PMRP). The PMRP is a population-based biobank with more than 20,00 participants and DNA, plasma, and serum samples linked to electronic health records (EHRs).[1] The overall aim of the PMRP was to facilitate genomics research that would ultimately improve healthcare delivery. It was one of the original five biobanks comprising the NHGRI-funded eMERGE (**E**lectronic **Me**dical **R**ecords and **Ge**nomics) network, which allowed researchers the opportunity to test the engagement and technical procedures that had been implemented to facilitate collaboration with researchers external to Marshfield Clinic.[2,3]

3.1 Stakeholder engagement

3.1.1 External stakeholders

The PMRP investigators convened three external advisory groups to provide guidance to investigators about the planning and implementation of the project: a Scientific Advisory Board (SAB), an Ethics and Security Advisory Board (ESAB), and a Community Advisory Group (CAG).[1,4–5]

The **SAB** met twice during the planning phase for PMRP, and a final time approximately one year after enrollment had started and more than 10,000 participants had been enrolled. The SAB included seven members with expertise in computational biology, pharmacogenetics, statistical genetics, and molecular epidemiology. PMRP investigators sought external guidance in part so that the policies and procedures would support collaboration with external

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00001-8
Copyright © 2023 Elsevier Inc. All rights reserved.

investigators, knowing that the power of the biobank would be increased through combination and collaboration with other study cohorts. During the planning meetings, the SAB members stressed the need to develop a rigorous, transparent process to determine access to the limited supply of biological samples in the biobank. Data and tissue access guidelines were developed in response to this recommendation (https://www.marshfieldresearch.org/cpmr/pmrp/access). The SAB members felt that a properly curated biobank linked to clinical data from a comprehensive EHR would garner interest from many scientists, both internal and external, to facilitate genomic discoveries. They suggested the development of cell lines to preserve the nonrenewable resource of biological samples. This recommendation could not be taken up due to limited resources. SAB members advised on potential proof of concept pilot projects that could be done quickly to test biobank infrastructure.

The **ESAB** comprised nine members with national expertise in bioethics, law, and spirituality. They met twice during the planning phases for PMRP and provided guidance to the Marshfield Clinic Institutional Review Board. One of their initial discussions was whether children should be included in the biobank. They felt that the project was a minimal risk, thus it might be considered unethical to NOT include children. Feasibility of identifying, locating, and reconsenting minors when they become legal adults was discussed. Ultimately, the decision not to enroll children was made on scientific grounds, not legal. A population-based cohort design is not efficient for children where diseases of study interest are not common.

The **CAG** met the most frequently of all the external advisory groups. The group initially comprised 15 members who represented a cross-section of the community, including a real estate agent, a teacher, a journalist, a farmer, patient advocate, a politician, a volunteer firefighter, and a retired person. During the planning phases, they reviewed all recruitment documents and provided suggestions on groups and individuals to speak to about the project. After recruitment had started, they provided guidance on newsletter content and study priorities. When the ESAB raised concerns about potential protocol changes that they thought required reconsent, the CAG requested a meeting with the ESAB to discuss the issue and let them know that they felt that study resources should be used to advance discovery rather than used to recontact and reconsent all participants.

3.1.2 Internal stakeholders

3.1.2.1 Scientific planning team

A scientific planning team comprising 17 Marshfield Clinic scientists met monthly during the project planning phase to discuss study logistics, and funding and prioritization of pilot projects using the biobank. Acting on recommendations from the external SAB, the internal scientific planning team selected two initial pilot proof concept projects: a pharmacogenetics study of Coumadin[6] that was conducted concurrently with recruitment to the biobank, and a genetic epidemiology study of APOE and Alzheimer's disease.[7]

Marshfield Clinic providers and employees were consulted through formal focus group discussions and provided with ongoing educational opportunities, including grand rounds, and an annual precision medicine conference for primary care providers. The ESAB felt strongly that recruitment to the biobank for employees of Marshfield Clinic should happen through their places of residence, not their work site, to avoid potential coercion.

3.2 Technical procedures to facilitate genomic data sharing with collaborators

Technical procedures that facilitate data sharing with both internal and external investigators can broadly be broken into four different phases. The initial phase includes gathering phenotypic data and sample processing. This phase is dominated by procedures that ensure that any data generated is of high quality and that all have been through a rigorous quality and identification check. The second set of technical procedures occurs when data and any sample is returned to the institution; samples and data need to be rechecked for identification and quality and any generated data needs to be incorporated into a data repository. The third phase of data and sample handling that needs rigorous technical processes in place is for finalizing a storage dataset that may include linking the phenotypic and genotypic data in a static data capture phase. Finally, procedures need to be created that allow new investigators to access data collected previously and to build on this work for future studies. Here, we discuss each phase of technical procedures that were implemented within the PMRP study and provide insight into key take-aways and lessons learned.

3.3 Phase 1—Sample identification, phenotyping, and quality controls

3.3.1 Phenotype data quality controls

In the case of the PMRP cohort, phenotypic data were collected from a dynamic medical record repository which meant that for each project phenotypic data would be collected at the point in time of the project. In order to ensure accuracy of the phenotypic data a number of processes and procedures were developed. One standard practice was to have a rigorous data dictionary, each element that was collected from the medical record was required to have a defined strategy and metadata table. In many instances, the metadata table included confidence intervals for phenotypic accuracy. To ensure the accuracy of any electronic data abstraction for phenotyping a subset of the abstracted data was reviewed by an expert personal (such as a physician or trained research coordinator) to ensure the phenotypic data was accurate. If the data did not accurately capture the phenotypic state, the investigator was requesting, an iterative process of data abstraction and data validation was performed until the phenotypic elements met quality control standards. During each iteration, the metadata for each element was documented and any program created to abstract from the medical record was cataloged

to ensure that any analyst provided the same data elements for a project. After phenotypic data, quality control standards were met the data from the medical record database was collected and captured as a completed dataset.

3.3.2 Sample data quality controls

To ensure the sample data were accurate within the PMRP sample repository a number of quality control and identification methods were developed. Initially, samples were all processed manually, however, during the course of the PMRP project this was modified so that a semiautomated process was in place. Standard laboratory protocols were followed such as bar coding samples and documenting the placement and utilization of both initial sample aliquots as well as all daughter aliquots of the samples. Quality data such as DNA quality and quantity were collected and linked to the bar codes to signal when samples needed to be replenished or placed in quarantine due to a lack of quality. Furthermore, a DNA fingerprint of each sample was created that could be used for individual sample identification.[8] This fingerprint included a mixture of polymorphisms that were present on the two most common GWAS platforms at the time Illumina and Affymetrix. In addition, the samples included a sex marker that was utilized to help ensure that phenotypic and genotypic data were attributed to the correct sample. Any sample with discordant sex identification was placed in quarantine and was not released for utilization until either a new sample was collected for that participant or sample misidentification was corrected manually based on other information such as utilizing relatives to identify sample DNA. Extensive pedigrees were created within the PMRP database to track families and relatedness within the participant population. Samples were also placed in quarantine and flagged for redraw if the fingerprint panel did not produce results or produced poor quality genotyping calls.

3.4 Phase 2—Data integration and sample return

Phenotype Data integration—Because the data from the medical records was frozen at the time of retrieval it was important to collect information such as the date of medical record abstraction. Each phenotypic dataset was stored as a separate file. However, all programs, algorithms, and logic that were utilized to build the phenotypic dataset was cataloged and stored separately in a library that could be used by new investigators after a period of embargo or with principal investigator permission. Any new variables created by investigator(s) were also cataloged and required to be returned to the PMRP data repository with appropriate metadata. Because most phenotypic data were generated in house. this phase of the phenotypic integration did not typically yield much new information for storage and integration.

Genotypic data integration—Because a single sample could be genotyped multiple times on numerous platforms genomic data that was returned went through a number of data capture

and quality assessments. Upon data return, sample identification was verified utilizing the unique SNP panel identifier, if there were unexpected results such as a discordant calls between genotyping, the confidence of the genotype call is investigated and samples are checked for potential mislabeling. Ancillary information such as genotype platform and call confidence were also collected for all external genotyping. Raw data files from the sequencing platform were also returned and archived within the PMRP data repository. Each time genotyping was performed the dataset was stored, however, based on all genotyping platforms a most probable genotype is also compiled for any loci that were genotyped on multiple platforms.

3.5 Phase 3—Finalizing datasets

An integrated dataset with both the collected phenotypic and genotypic data was stored within the PMRP data repository. This was essential as individuals may choose to leave the PMRP population at any time. Individuals who left the PMRP cohort have all samples destroyed and were not utilized in any future data analysis but the information on the individual was kept in the archival integrated dataset for research data integrity purposes. These finalized datasets were not available for interrogation going forward.

3.6 Phase 4—Data access

One of the requirements for PMRP was that all genotyping and other sample analysis that was performed would be available to investigators in the future after an embargo period. This meant that the genotyping that was performed on a sample could be utilized for alternative studies. Investigators determined whether or not they wished to utilize the most probable genotype for the sample or the genotyping from a particular platform or previous study. Previously collected phenotypic data could also be utilized by implanting logical algorithms that were created for the purpose of phenotyping, and specifying the appropriate time period.

3.6.1 Pilot genomic data sharing projects with participants

As mentioned previously, the written PMRP consent form explicitly stated that genomic data would not be shared with participants. In the context of the eMERGE II project, PMRP investigators undertook two pilot projects that involved sharing of genomic data with study participants. Both studies required new study recruitment to allow data sharing.

3.6.1.1 AMD pilot study

For this single-site study, 100 optometry patients aged 50 years and older with a family history of age-related macular degeneration (AMD) were recruited to a study of presymptomatic testing of genomic markers associated with increased risk of AMD.[9] Risk scores were calculated and the study optometrist presented the information to the participants. The risk

scores from the external lab were scanned into the medical record and a note was dictated related to the consultation visit. There was a high level of enthusiasm for this study among eligible patients, in part because of the genomic data sharing, and in part because of a trusting relationship with their eye care provider.

3.6.1.2 PGx pilot study

The eMERGE-PGx project is a multicenter pilot study of preemptive pharmacogenomics testing.[10] Genomic data sharing includes the sharing of clinical data with patients and providers, deposition of data in EHRs, and sharing of data with other scientists in the eMERGE network. As with other multicenter eMERGE projects, standardization of phenotypes and genotypes was essential to combine data across sites. Genotype standardization was supported through core laboratory facilities.

3.7 Summary

The PMRP research team was one of the first in the US to develop a population-based study to facilitate genomic discoveries with the long-term aim of delivery of precision medicine in a clinical setting. Genomic data sharing was essential to the success of the project. Rarer conditions required combining data with investigators from other sites to increase sample size, and external collaboration was necessary for the validation of genomic discoveries. Stakeholder engagement throughout the study provided feedback to develop processes to allow genomic data sharing (Table 3.1).

Table 3.1: Lessons learned in the development of a biobank to facilitate genomic data sharing.

- Stakeholder engagement at all stages of a project is vital and will improve the project, as well as acceptance and use by all interested parties, including participants, scientists, and clinicians.
- Regular communication with study participants is essential.
- Technology is changing rapidly. Investigators need to plan ahead.
- Bioinformatics can be a rate-limiting step for genomic analyses and return of results, and requires adequate resources.
- Create a process to ensure accurate sample and data identification for each step of the process from identification through data return.
- Consider storing many aliquots of plasma and serum because of the volume needed for some assays and the sensitivity of some assays to freeze/thaw cycles. Serial samples can be very useful for biomarker studies.
- Plan for return of results to study participants, starting with informed consent from participants and handling samples in a CLIA-certified lab, and including incorporation of results into EHRs and necessary decision support for providers and patients.
- Standardization of phenotypic data from electronic health records to facilitate multisite collaboration is more challenging and time-consuming than standardization of genomic data.

References

1. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Pers Med.*. 2005;2(1):49–79.
2. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genom.*. 2011;4:13.
3. Gottesman O, Kuivaniemi H, Tromp G, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet Med.* 15:761-771.
4. McCarty CA, Garber A, Reeser JC, Fist NC. Study newsletters, community and advisory boards, and focus group discussions provide ongoing feedback for a large biobank. *Am J Med Genet*. 2011;155:737–741.
5. McCarty CA, Chapman-Stone D, Derfus T, Giampietro PF, Fost NMarshfield Clinic PMRP Community Advisory Group. Community consultation and communication for a population-based DNA biobank: The Marshfield Clinic Personalized Medicine Research Project. *Am J Med Genet*. 2008;146A:3026–3033.
6. Caldwell MC, Wilke RA, Berg RL, et al. Impact of age, CYP2C9 genotype and concomitant medication on the rate of rise for prothrombin time during the first 30 days of warfarin therapy. *Clin Med Res*. 2005;3(4):207–213.
7. Ghebranious N, Mukesh B, Giampietro PF, et al. A pilot study of gene/gene and gene/environment interactions in Alzheimer Disease. *Clin Med Res*. 2011;9:17–25.
8. Cross DS, Ivacic LC, Stefanski EL, McCarty CA. Population based allele frequencies of disease associated polymorphisms in the Personalized Medicine Research Project. *BMC Genet*. 2010;11:51. doi:10.1186/1471-2156-11-51.
9. McCarty CA, Fuchs MJ, Lamb A, Conway P. How do patients respond to genetic testing for age-related macular degeneration? *Optom Vis Sci*. 2018;95:166–170.
10. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, et al. Design and anticipated outcomes of the eMERGE-PGx Project: a multicenter pilot for preemptive pharmacogenomics in electronic health records. *Clin Pharmacol Therapeut*. 2014;96:482–489.

CHAPTER 4

Multidirectional genetic and genomic data sharing in the All of Us *research program*

K.D. Blizinsky[a,b], S. Chandrasekharan[a], S. Jooma[a], J.A. Reusch[a] and Kimberly A. Thomson[a,c]

[a]Research Program, National Institutes of Health, Bethesda, MD, USA [b]Rush Alzheimer's Disease Center, Rush University, Chicago, IL, USA [c]New Jersey State Department of Health, Division of HIV, STD, and TB Services, Newark, NJ, USA

4.1 Introduction

Health is not wholly effected by a uniform set of conditions. Rather, what is "healthy" is whatever conditions promote and/or sustain normal functions in the complex system that is the human body. In some cases, these conditions may be reasonably generalizable across all peoples or across large, circumscribed populations. In such cases, public health approaches can be effectively applied to promote health and wellbeing.[1] However, in other cases, the variables that promote health may be more personalized. Likewise, when bodies are "unhealthy," or when there is dysfunction in a particular aspect of the system, the steps that must be taken to return that part of the system to proper functioning may not be the same for everybody, even when the underlying dysfunction is the same. This uniqueness is a product of our genomes, our environment, and our lived experiences, which together inform our individual state of being. It is in these situations of uniqueness where precision medicine can prove beneficial.[2,3] Precision medicine can be conceptualized as the practice of delivering the right treatment to the right person at the right time, and informed by, and in tandem with public health approaches, it is a formidable tool to improve health outcomes for all.[81]

In order to realize the power of precision medicine, though, there must first come biomedical research that can inform clinical approaches, and the antecedent to this research must be resources that furnish studies with the data necessary to answer the research questions relevant to that clinical progress. Enter the *All of Us* research program, an ambitious effort by the US government to facilitate precision medicine research.[6] The program aims to build a comprehensive, longitudinal dataset from one million or more individuals, offering deep

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00010-9
Copyright © 2023 Elsevier Inc. All rights reserved.

Table 4.1: *All of Us* research resources: current and planned.

Participant surveys[a]	Basic information ("The Basics")
	Lifestyle
	Overall health
	Personal medical history
	Healthcare access & utilization
	Family medical history
	COVID-19 participant experience
	Mental health
	Social determinants of health
Electronic personal health information	Medical records
	Pharmacy data
	Vision and dental records
Physical measurements	Blood pressure
	Pulse
	Body mass index
	Height
	Hip circumference
	Waist circumference
	Weight
Biospecimens	Blood
	Saliva
	Urine
	DNA (from blood and/or saliva)
	RNA
Omics	Whole genome sequence
	Genotyping arrays
	Other omics
Assays	COVID serology[b]
Mobile/wearable tech	Fitbit data
	Data from Other mHealth Devices
Geospatial/environmental data	
Data Linked from external sources	

Items in italics denote potential future activities.

[a] All current participant survey materials are available on the Research Hub Survey Explorer (https://www.researchallofus.org/data-tools/survey-explorer/).

[b] Not available for all participants. Please see Althoff, K.N. et al. (2022). Antibodies to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in All of Us Research Program Participants, 2 January to 18 March 2020. Clin Infect Dis. 2022 Mar 1;74(4):584-590. doi: 10.1093/cid/ciab519 for details.

biological, clinical, social, environmental, and other data of utility to scientists from a wide variety of biomedical disciplines (Table 4.1).

But critical to the success of this unprecedented endeavor is a dedication to diversity. A crucial part of improving health outcomes for everyone is addressing health inequities. Precision medicine can be an incisive tool to promote health equity. But to reap the entelechial benefits broadly, the research at its fons et origo must be able to both study participants whose circumstances place them at greater risk of experiencing health inequities and ask the

relevant questions of the data to make meaningful progress toward health equity. Many populations and communities that are subject to health inequities are historically underrepresented in biomedical research.[7,8] These trends also persist in genomic research. As late as 2016, only 19% of Genome-Wide Association Study populations consisted of individuals of non-European descent.[9,10] Research has repeatedly demonstrated that certain variants appear at different frequencies in different ancestral populations, conferring differential relative risk in populations with dissimilar ancestry.[11,12] Similar patterns of underrepresentation exist when examined through the lens of social factors like race and ethnicity, despite the fact that such social factors are known to contribute to disease risk and health outcomes. Clinical research overall has generally been conducted in overwhelmingly white, often male populations from higher socioeconomic strata,[13] and the paucity of inclusive data resources with respect to a number of key demographic dimensions—including ancestry, race, ethnicity, socioeconomic status, age, sex assigned at birth, gender, sexual orientation, educational attainment, geography, and more—is likely a major contributor to health inequities.[11,12]

All of Us has embraced, as one of its core values,[14] the goal of developing a cohort that reflects the diversity of the US population, with specific efforts to enfranchise and engage underrepresented populations over the long term.[15] However, the demographic dimensions that define underrepresentation do not exist in isolation. Underrepresentation is intersectional and multifaceted, and any successful strategy for meaningful inclusion that aims to reverse damaging historical trends must recognize that in the end, each person has a unique constellation of factors that bring to bear on their research participation.

As an important part of its data compilation efforts, the *All of Us* research program generates and curates genetic and genomic data.[16,17] *All of Us* has charged itself with sharing this data both with the research community and with the participants themselves. In this chapter, we will discuss the considerations, preparatory activities, and guiding concepts that undergird this multidirectional data sharing, as well as the implementation strategies employed. We also share the lessons learned from the program to date that may help to shape the future of both the program and precision medicine research.

4.2 Sharing data with researchers

Genetic and genomic data are mainstays of research across the biomedical sciences. Their superfluity reaches far beyond the bounds of research, as well, arising in fora from police investigations[18] to popular culture.[19,20] However, their ubiquity may lead one to falsely assume that as a collective, society has appropriately understood the full utility and significance of genetic and genomic data. These data are clearly sensitive: they are collectively unique to an individual and therefore harbor risks inherent in that uniqueness. And while in some forms, genetic and genomic data may arguably be deidentified, they can never truly be anonymized.[21,22]

Ergo, sharing genetic and genomic data with a wider audience, even when limited to legitimate research use, presents a set of questions[23] that revolve around questions of trust.[83] How much should the program trust people who want to access the data? How much should the participants trust the program to act on their behalf? How does any resource the program might build maintain an appropriate balance of protectiveness and permissiveness in the face of rapid social and technological changes that may alter dynamics of trust between participants, government, researchers, and the public?

4.2.1 Relevant considerations

The program must address the above questions in order to fulfill its mission. First, if the program is to engender trust between itself and its participants, particularly those from populations, groups, and communities that are currently and/or have historically been subject to discriminatory practices and abuse,[24] it must foremost consider its obligation to protect participants' interests. One major factor of this protection must be privacy and the legal, regulatory, and policy frameworks that dictate the privacy protections the program can extend to participants and their data. The Common Rule (45 C.F.R. 46) and NIH Genomic Data Sharing Policy[25] dictate the specific requirements for, and conditions under which, data—including genomic data—may be shared and used for research purposes, while operation under Certificates of Confidentiality (42 U.S.C. 241(d)) limits disclosures and uses of sensitive data gathered for research. In some contexts, other laws and regulations—such as HIPAA (Pub. L. No. 104-191) and 42 CFR Part 2—may apply. Any model of data access and use *All of Us* employs must be shaped by these legal, regulatory, and policy guardrails.

Another aspect of protection that *All of Us* must consider is its responsibility, within the realm of beneficence, to prevent intentional and unintentional misuse of the data. Data misuse may affect participants at an individual level,[26,27] but it may also have repercussions for groups and communities. For example, genetic and genomic research may fall prey to genetic determinism, leading to the overestimation of the impact of genes on biomedical and behavioral outcomes.[28] Genetic determinism is particularly detrimental when paired with implicit or explicit social biases.[29] When results from these studies are parsed using new or extant socially defined categorical groups, they may be proffered as evidence to justify novel or existing prejudices, lending the veneer of scientific credibility to bigotry and social stigma.

The principle of justice charges us to ensure the harms and benefits of research are fairly distributed. Many of the demographic groups and communities the *All of Us* seeks to engage are not only underrepresented in research but are also subject to institutional and structural discrimination.[30,31] As destructive as stigmatizing research can be in and of itself, there is an added risk when the resultant stigma compounds an existing stigma burden. Thus in order to fulfill its duties to just and beneficent research, the program must take appreciable steps toward the prevention of data misuse, and stigmatizing research in particular.

There are also specific considerations for engaging with Tribal and American Indian/Alaska Native (AI/AN) populations. Indigenous peoples in the United States have been subjected to tremendous injustices at the hands of nonindigenous peoples from the beginnings of European colonialism to the present day. These violations, both generally and those perpetrated under the auspices of scientific research, have left some tribes with concerns about taking part in research studies, particularly those leveraging genetic and genomic data.[32,33] If *All of Us* is to meaningfully engage with Tribal Nations and AI/AN individuals and communities, the program must recognize and account for these injustices in its approach and commit to efforts to protect AI/AN participants and communities from additional stigma and harm. The program must also recognize that in engaging with Tribal Nations, it is engaging with sovereign nations. As such, interactions are more than an interface between community groups and the program; they are government-to-government relations and should adhere to official processes, such as those prescribed by the HHS Tribal Consultation Policy.[34]

But perhaps most importantly, the program must also consider that ultimately, participants themselves are best suited to determining what is and is not in their own best interests. Fully recognizing participant autonomy is best achieved through partnership with participants.[35] In addition to providing participants visibility into the inner workings of the program, the program must be transparent about the steps it takes to protect participants and their data, and the ways in which those data are and have been used, so that prospective and current participants are able to make informed decisions about their participation. Partnership also means actively integrating participants' perspectives into program governance and decision-making. Participants must be involved in determining, among other things: the acceptable uses of participant data; and the principles for, and conditions of, the return of individual research results.

All of Us also has an obligation to make sure participants' data are used for the purposes for which they have been so generously donated. While maximizing participant privacy and data security—as well as addressing the ethical principles of beneficence, justice, and autonomy—the program must also maximize data utility, which means it must find the appropriate balance between these sometimes opposing forces.

4.2.2 Guiding concepts for sharing data with researchers

Data access and use, including access and use of genetic and genomic data in the *All of Us* research program, are framed by three guiding concepts.

4.2.2.1 Primacy of participant protection

The nature of the *All of Us* participant population and data being collected dictate that participant protection must be paramount. This should include not only participant privacy and data security, but also preventing potential misuse of the data that could result in stigma and discrimination at the individual, community, and population levels.

4.2.2.2 Broad accessibility and utility

However, when considering data access and use, the program should consider data—unlike participants' time, effort, or biospecimens—a nonscarce resource. Data are not exhausted with use. Furthermore, the more data are used, the more findings can be drawn from them, and the more potential benefit those data can ultimately provide, both to participants and society. In addition, the more the data resources can support a wide range of researchers from diverse backgrounds and with multifarious research interests, the more likely the findings resulting from use of the data will translate to greater equity in the benefits reaped from those findings. Thus, the program can promote the derivation of maximum, and most equitable, benefit from the data by ensuring the broadest appropriate access to them. This should include the empowerment of researchers outside the "characteristic data user" communities of well-resourced academic researchers and researchers affiliated with the biotech and pharmaceutical industries.

4.2.2.3 Iterative reappraisal

While initially, the program must err on the side of stringency with respect to data protections, in order to ensure participant protections are upheld, the program should also continuously reevaluate technological development, as well as patterns of user behavior. The field of genomics, especially, is rapidly changing, and programmatic reassessments should anticipate and match that cadence. In these reassessments, the program must seek to remove unnecessary barriers to access, in order to support broad access and use. No group of data users should have privileged access to *All of Us* data resources based on anything other than the protection of participants and participant data.

Thus, the below discussion of implementation should be viewed as a first incarnation of the *All of Us* data resources, with the expectation that this strategy will grow and change with time.

4.2.3 Implementation

The *All of Us* research program collects and shares health, environmental, and genetic and genomic data about its participants and applies the above-described concepts through its approach to data sharing and privacy protection. The policies and implementation strategies detailed below have been informed by foundational program efforts, such as the Precision Medicine Initiative Privacy and Trust Principles,[84] and subject matter experts and interested parties, including attendees of the *All of Us* Research Program's 2019 Ethical, Legal, and Social Implications (ELSI) Research Priorities Workshop.[36] They also continue to be shaped by valuable input from *All of Us* governance bodies and participant ambassadors.

In addition to constructing a general approach for genetic and genomic data sharing, *All of Us* undertook a separate process, through formal Tribal Consultation, to understand ways to appropriately share data from self-identified AI/AN individuals, as well as respectful ways to partner with Tribal Nations for the enrollment of individuals with tribal affiliations. To date, *All of Us* has not recruited participants on tribal land, nor will the program do so without express permission from Tribal Nations. However, individuals who self-identify as AI/AN have enrolled in the program on their own. In the interests of shaping AI/AN engagement around Tribal input, *All of Us* placed a moratorium on processing self-identified AI/AN participants' biosamples and sharing their data until completion of the Tribal Consultation and the publication of the accompanying final report. The report was published and widely disseminated in March 2021.[37,38] Based on the outcomes of the consultation, *All of Us* extended the moratorium to the end of September 2021, providing self-identified AI/AN participants the time to review the report and talk to their tribal leaders before making a decision about their continued participation. If self-identified AI/AN participants chose to withdraw before the September 30th deadline, their data and biosamples were removed from all program repositories and will never be shared with researchers. For those who chose to remain enrolled, their data and biosamples will proceed along the standard *All of Us* pipeline for inclusion in the Research Hub and ultimate use by researchers. The program will continue to solicit input from tribes on important issues pertaining to the program. Moreover, the program will work to establish partnerships with individual tribes for tribally-driven participation pilots, which will present opportunities to further address unique data-sharing concerns.

Following from the guiding concepts for the *All of Us* approach to data sharing, the program makes data available through three tiers in the *All of Us* Research Hub: the Public Tier, the Registered Tier, and the Controlled Tier (Fig. 4.1). The three tiers undergo different levels of data transformation, and the steps required to gain access to each tier are commensurate with the risks to privacy posed by the sensitivity of the data that each tier contains. No direct identifiers, such as name or address, are available in any tier, and participant identities cannot be readily ascertained or associated with the information. As such, no research with data from any of the currently available tier would constitute nonhuman subjects research under the Common Rule. *All of Us* may create additional data tiers at a later date, but as of yet, no additional tiers are planned.

The Public Tier, available through the program's public Data Browser,[39] consists of only aggregate data and summary statistics about the participant population, and all data are rounded into bins of 20 to ensure a negligible risk of reidentification. As the name implies, the Public Tier is publicly available and does not require a log-in or special access steps. Individuals can run basic analyses using these data, and researchers wishing to access the program's more restricted data tiers can visit the Data Browser to learn about the data available through *All of Us,* prior to embarking on the registration process, and investigate how they might use the data to pursue their research questions.

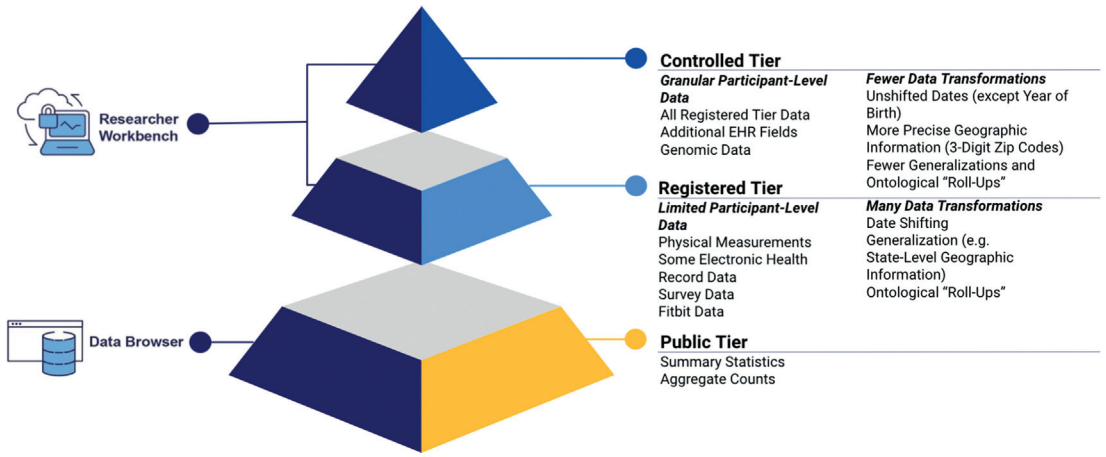

Figure 4.1

The *All of Us* Research Hub. *All of Us* shares participant data through three tiers: the Public Tier, accessed through the Data Browser, and the Registered and Controlled tiers, available through the Researcher Workbench. As shown, the Public Tier contains aggregate counts and summary statistics, while the Registered and Controlled Tiers contain individual-level participant data, with the Registered Tier containing fewer data types and being subject to a greater level of data transformation than the Controlled Tier to maintain a lower risk of reidentification.

Moreover, the Data Browser provides a way for members of the public, citizen and community scientists, and *All of Us* participants to explore the data without registering for Workbench access.

The Public Tier provides a snapshot of the different data types that *All of Us* has curated to date, including information from participants' electronic health records, such as health conditions, medications, procedures, and lab results; participants' survey data, including survey data about their health, lifestyle, medical history, family health history, health care access, and experiences during the COVID pandemic;[40] participants' physical measurements; genomic variant data; and data from wearable devices, like Fitbits. *All of Us* will continue to populate the Data Browser as it collects or generates new data types.

The Registered and Controlled Tiers together constitute the Researcher Workbench (Fig. 4.1), the program's secure, cloud-based analysis platform developed by the Data and Research Center (DRC).[17] Both tiers require researchers to take a series of steps to gain access[41]. However, access is predicated on trust in the researcher, rather than foreknowledge of the proposed research. This model is referred to as a "data passport" model, and it does not require researchers to seek approval for each project they wish to carry out. Instead, researchers receive their "passport" after successful completion of the required application and approval steps, after which they are able to carry out projects without any preapproval of their studies.

Researchers are required to renew their data passports annually by reviewing and updating their research profile and project workspaces, retaking the required training, and resigning their agreements, all discussed in greater detail below.

The Registered Tier, which became available in May 2020, contains participant-level data subject to a number of privacy-preserving transformations to obfuscate participant identities without compromising data utility. Such transformations include removal of direct identifiers, suppression of fields that may contain identifying information (e.g., free text fields from electronic health records), date-shifting (shifting dates in a participant record forward or backward a set number of days), and generalization (combining categories—e.g., race and ethnicity categories, gender identity categories, etc.—to create larger, less granular groups). Some data have been omitted from the Registered Tier, as the transformations necessary to accommodate inclusion are either impossible or would sufficiently diminish the utility of the data such as to render them functionally useless; at least initially, genetic and genomic data are not included in the Registered Tier.

The Controlled Tier, launched in spring 2022, also contains participant-level data, but data has been subjected to fewer transformations than the data in the Registered Tier. In addition, the Controlled Tier contains data generated from whole genome sequencing and genotyping arrays. The data on the whole are more granular and sensitive than data available through the Registered Tier.

The DRC's privacy experts conduct risk assessments to determine which data should be made available in each tier. These risk assessments are based on empirical analysis of reidentification risks associated with each data element, both alone and in conjunction with other data elements. Specifically, the empirical analysis estimates the chance that the values encountered in the combination of fields (e.g., age, sex, US state of residence) would uniquely distinguish an individual within the resource. This method also takes into account other information that is likely to be available to an attacker using external data, including sources of identifiable data, and simulates threats based on different degrees of knowledge that different types of attackers may possess about individuals.[42,43]

Of course, the data passport model, while lowering barriers to accessing data and conducting research, presents its own unique set of risks and relies heavily on trust vested in the researchers using the platform. In order to foster this trust and trustworthiness, *All of Us* relies on four key components: education, accountability, support, and oversight. In this section, we will expound upon how each component supports the feasibility of the data passport model in achieving its ambitious goals of broad access to data while preserving participant privacy.

4.2.3.1 Education

The program operates under the assumption that the vast majority of *All of Us* data users are and will be well-intentioned individuals who have set out to conduct research to advance

human health. Nevertheless, the history of research in the United States is rife with abuses and misconduct–both intentional and unintentional–that have harmed individuals and communities, especially members of the underrepresented groups *All of Us* hopes to engage in the program. Genetic and genomic research, in particular, has been at the root of many of these instances in the recent past. To ensure historical mistakes are not repeated, educating prospective and current data users is critical to minimizing the chance of unintentional misconduct, noncompliance, and stigmatizing or otherwise unethical research.

The program's primary vehicle for educating data users is the program's Responsible Conduct of Research Training (RCR). The RCR Training content was specifically created for *All of Us*, though much of it is generalizable, and it differs from extant research ethics trainings in its focus on secondaruse of human research data, rather than human subjects research itself. Increasingly, the biomedical research community is recognizing that secondary research is not conducted in a vacuum and can have profound implications for research participants and their communities.[44,45]

The training is organized as a series of modules that cover a range of topics. Topical modules in the original RCR training are required for all applicants for data access, and include:

- An introduction to the aims of the program and its core values.
- A contextualization of how the principles of ethical research, as laid out by the Belmont Principles[46] and codified through the Common Rule , can be extended to research with participant data in a nonhuman subjects environment.
- The importance of participant diversity to research equity.
- An overview of the process by which stigma is generated, the potential relationship of research to the creation and perpetuation of stigma, and how to prevent stigmatizing research.
- A review of the specific expectations outlined in the program's Data User Code of Conduct (DUCC; described below) and the Researcher Workbench oversight mechanisms.
- Privacy and security measures in place at the program-level and the additional responsibilities of the data users to protect participants and data privacy and security.

Recently, the program has expanded upon these base modules. This expansion includes:

- Additional information on the prevention of stigmatizing research.
- Avoidance of the biologization of social factors, like race and ethnicity.
- The importance of social determinants to health and wellbeing.
- Prevention of group harm.
- Techniques to address implicit biases and build cultural competence and humility.
- An overview of social responsibility in research, extending from the development of the research question to the communication of research findings.

Most of this new content is intended for all prospective data users. However, given the increased granularity and sensitivity of data in the Controlled Tier, applicants who wish to access this tier will have to complete additional training, consisting of training modules to orient applicants to the content and risks of the Controlled Tier, reinforce methods to prevent stigmatizing research, and surface issues related to the joint use of genetic/genomic and social variables.

After prospective data users complete the training, they are required to pass an assessment that measures their ability to apply the information and techniques covered in the training. They must pass with a score of 80% or higher to be able to complete the Researcher Workbench registration process for either Registered or Controlled tier access.

4.2.3.2 Accountability

In addition to education, *All of Us* employs both individual and institutional measures of accountability to hold data users to the specific expectations of the program. This includes a Data Use and Registration Agreement (DURA), an agreement between the program and the institution with which the applicant is affiliated. Master DURAs cover all institutionally affiliated data users, or institutions may opt to establish DURAs for institutionally affiliated data users on an individual basis. In addition, there is an agreement between the individual data user and the program called the Data User Code of Conduct (DUCC).[47] The DUCC is organized as a set of "will" and "will not" statements clearly outlining the program's expectations for individuals working within the Registered or Controlled Tiers. These include a requirement to provide meaningful and accurate descriptions of each research project established and prohibitions on attempting to reidentify research participants or their relatives and on row-level data exfiltration, among others.

Furthermore, the DUCC contains a clause requiring data users to comply with any policies promulgated by the program governing data use. The policies provide the program an avenue by which to elaborate upon the expectations in place for data users, as well as to delve into the rationale for these expectations. This clause of the DUCC, holding data users accountable to program policies generally, provides the program with the flexibility to modify policies and implement new ones as needed based on program evolution and patterns of behaviors and data use within the Researcher Workbench.

The program currently has policies to promote ethical conduct in research,[48] prevent stigmatizing research,[50] and ensure the responsible dissemination of the outcomes of research—in terms of protecting participant privacy,[49] ensuring participant contributions are acknowledged in publications, and promoting the rapid and public accessibility of research findings.[50]

While there are no DUCC clauses or data use policies specific to genetic and genomic data, the Controlled Tier RCR content characterizes what the program considers to be unacceptable

uses of genetic and genomic data, and links these unacceptable practices—participant reidentification, potentially stigmatizing research, data exfiltration, etc.—to expectations under the DUCC and associated data use policies.

4.2.3.3 Support

While individuals using the Researcher Workbench are required to understand and apply myriad concepts for the responsible use of data, they do not have to do so without assistance. In addition to technical support that data users can obtain through the User Support Hub in the Researcher Workbench, they may also reach out to the program's Resource Access Board (RAB) for assistance. The RAB oversees compliance with the DUCC and comprises members from across the consortium with expertise in biomedical research; ethical, legal, and social issues; legal, regulatory, and policy issues; privacy and security; and health equity. In addition, RAB membership includes participant ambassadors and other participant advocates who can bring the perspectives of individuals, groups, and communities—particularly those who are underrepresented in biomedical research—to RAB deliberations. The RAB also has the flexibility to consult with community and subject matter experts and to include *ad hoc* members as needed for its appropriate oversight.

When data users begin a new project and start a new workspace, they are required to complete a form that outlines the purpose of their research, their approach, anticipated outcomes, and their populations of interest, among other details. At any point in the instantiation and maintenance of a workspace, data users are given the opportunity to request a RAB workspace review to ensure compliance with the DUCC, particularly if they are concerned about inadvertently conducting research that could stigmatize participants or participant communities. If a data user requests a review, the RAB reviews the relevant workspace materials and provides advice to data users on how to amend their approach to help avoid potential stigma and group harm or otherwise bring their project into compliance.

4.2.3.4 Oversight

Finally, although projects are not preapproved, *All of Us* has oversight mechanisms in place to detect and take action against DUCC or policy noncompliance. The RAB is primarily responsible for this oversight, and projects are brought to the RAB's attention in one of three ways. One way, as noted above, is that data users are encouraged to request preemptive RAB review of their workspaces , especially when they have concerns about the potential for their research study to create or promote stigma.

In addition, all workspace descriptions are posted publicly on the *All of Us* Research Projects Directory;[51] members of the public may read through the descriptions and flag any workspaces that they believe violate the provisions of the DUCC. All flagged descriptions are directed to the RAB for review.

Third, the program conducts audits of workspaces, both random and targeted, and the DRC supports this auditing through the use of automated monitoring measures, such as alerts for the exfiltration of large amounts of data.

If the RAB determines that a violation has occurred, it goes on to consider the intentionality, scale, and scope of the potential violation, as well as history of the workspace owner and contributors, before recommending corrective action or penalties. Penalties may include suspending an individual's Researcher Workbench account, posting their name on a publicly available list of violators, notifying funding agencies like NIH of misconduct, or, in extreme cases, potential financial or legal repercussions. Depending on the scope and severity of the potential violation, the RAB may also take preventative measures during its review, such as suspending a user's access or suggesting mitigative actions like additional user training. Data users have the ability to appeal decisions made by the RAB up to two times if they feel that the determination is unwarranted.[52] All appeals will be assigned to a review team with appropriate expertise, which may include *All of Us* staff members and program leadership. The review team may consult with consortium subject matter experts as needed.

4.2.4 Lessons learned and future directions

To date, the *All of Us* data resources are predominantly used by tenure-track researchers at academic research institutions (Fig. 4.2A), mainly male, nonhispanic, and white (Fig. 4.2B), and all users are affiliated with institutions within the United States. While the proportion of early career researchers is to be applauded, these data point to a severe under-engagement of researchers both outside of academia and from diverse racial, ethnic, national, and sexual and gender demographic groups.

At present, access to *All of Us* data is limited such that access by international data users, data users affiliated with commercial entities, and citizen and community scientists is largely prohibited. The program intentionally limited access in these ways during the early days of the Researcher Workbench in order to give the program time to evaluate operational fidelity and make necessary adjustments to facilitate safe and intuitive access. However, broad access and use are foundational values of the program. As the program continues to develop, it will phase in a series of changes to the current access model to pilot strategies that would allow access to a wider range of researchers, including researchers from industry, international researchers, and citizen and community scientists. This will likely include changes to aspects of the data user application and approval pathway. It will also include outreach efforts to help introduce the *All of Us* data resources to researchers who may not otherwise be aware of them. Additionally, the program may explore new models of partnerships with academic and nonprofit research organizations to enable responsible and meaningful access to citizen scientists and community-based researchers.

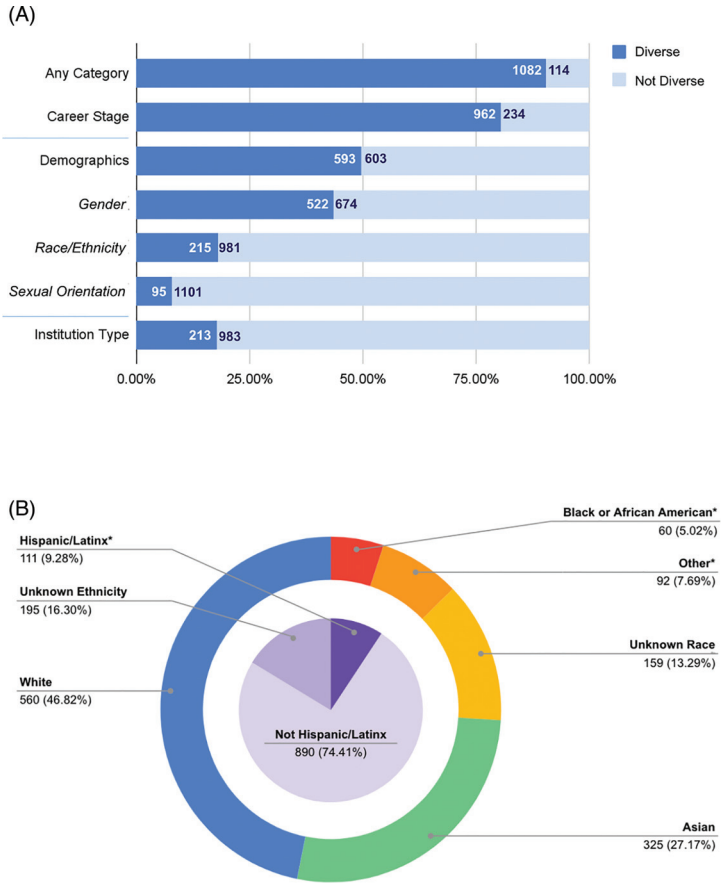

Figure 4.2

Composition of the Researcher Workbench data user population. As part of their profiles, authorized data users of the Researcher Workbench are asked to report basic professional and demographic information. Below describes the total authorized user breakdown by diversity categories (A) and race and ethnicity (B). A | Proportion of total ($n = 1196$ at time of publication) authorized users who self-identify with an underrepresented diversity designation according to the following: "any category" denotes authorized users responding "yes" to one or more designations contributing to workforce diversity in any category; "career stage" denotes authorized users who are not mid- or late-career tenure track faculty or senior researchers; "demographics" denotes authorized users who self-identify as belonging to one or more underrepresented designations for race/ethnicity (see below), sex assigned at birth (intersex or other), gender (see below), sexual orientation (see below), and disability status (physical and/or cognitive disability); "gender" denotes authorized users who self-identify as woman, nonbinary, transgender, or other; "race/ethnicity" denotes authorized users who self-identify as American Indian/Alaska Native, Black/African American, Native Hawaiian/Pacific Islander, Multiracial, or other race and/or hispanic/latinx; "sexual orientation" denotes authorized users who self-identify as LGBTQ+; "institution type" denotes authorized users from minority-serving institutions, smaller colleges, community colleges, foundations, and other organizations that do not receive the bulk of federal research dollars. B | Breakdown of authorized users' self-identified race and ethnicity; * denotes underrepresented subcategory.

Table 4.2: Resource Access Board reviews.

Review type	Current state of disposition	
Researcher	14	Reviews in progress
Requested	28	Workspaces requiring changes
($n = 70$)	56	Reviews completed
Publicly	1	Reviews in progress
Requested	10	Workspaces requiring changes
($n = 19$)	18	Completed
Program	9	Reviews in progress
Initiated	23	Workspaces requiring changes
($n = 45$)	36	Reviews completed

The Resource Access Board (RAB) conducts reviews based on requests that are initiated: by authorized users themselves (*Researcher Requested*); by members of the public viewing publicly displayed project descriptions (*Publicly Requested*); or upon request by the DRC, which monitors workspace description completion and supports users, and through routine program audits (*Program Initiated*). We have broken down the total number of RAB reviews requested at time of publication ($n = 134$) by type of request and further by whether the review is in progress or completed and the number of such reviews requiring changes to the workspace under review.

The program also continues to learn from its oversight experiences. It has been heartening to see that more than half of the reviews conducted by the RAB were initiated by requests from users themselves (Table 4.2). This suggests that when given the opportunity, researchers will avail themselves of support in ways that may help prevent participant and social harm. These reviews have also highlighted the need for more resources to assist data users in developing "meaningful" workspace descriptions that balance both needs for public transparency and sufficient scientific detail for conducting fair and detailed reviews, which has led to changes in the RCR Training. Going forward, the addition of new data types and the widening range of eligible data users will no doubt bring familiar and new challenges to ensuring safe and socially responsible use of *All of Us* data resources. Through continuous evaluation and learning from its own processes, the RAB will continue to identify important technical, consultation, and process improvement needs in order to enhance robust oversight of data use and assistance to data users.

In addition, the program will be examining ways to support researchers with different methodological and technical needs. At present, the Workbench provides some minimal support for researchers' methodological questions, but the program sees this as a ripe area for development. Especially as the population of researchers grows and diversifies, and the amount and complexity of the *All of Us* data increases, there is the foreseeable benefit to providing standardized "plug-and-play" code for routine analytical processes, as well as developing solutions to enable broader, but ethically and socially sound, use of more complex techniques like artificial intelligence and machine learning. Furthermore, the Workbench's reliance on R and python excludes many scientists with legitimate research interests that could be served by use of the *All of Us* data resources. The program is in the process of soliciting input from researcher groups across the spectrum to help ensure future iterations of the Workbench

include tools that enable these researchers to use the *All of Us* data resources effectively. While some solutions to these challenges will likely come directly from the program, *All of Us* is also looking at ways to encourage community building amongst the user and other interested parties to facilitate the sharing of knowledge, expertise, and other resources.

Researcher demographic diversity is another critical area of improvement for the program, and one that is not as easily addressed. On this issue, the program must struggle against centuries of institutionalized and structural prejudices that give rise to overwhelming demographic homogeneity amongst researchers, particularly in academia. Diversity on project teams has been shown to encourage more equitable team output.[53] Research teams are no exception, and studies demonstrate the importance of diversity in increasing the quality[54,55] and applicability[56] of scientific research. If the program truly wishes to improve health equity through the use of the *All of Us* data resources, it must take meaningful steps toward increasing diversity among data users.

All of Us takes this responsibility seriously, and while the initiative is still in its relative infancy, the program is committed to making meaningful headway in the effort to diversify the research community. At this stage, the program is reaching out to minority-serving institutions, such as Historically Black Colleges and Universities and Hispanic Serving Institutions, to introduce staff and students to the Researcher Workbench. The program also works with its community partners regionally and nationally to reach out to diverse research groups, including researchers from racial, ethnic, and sexual and gender minority groups. While the program cannot fully address the issue of researcher diversity on its own, *All of Us* hopes to be a part of a greater movement that not only ushers in a more inclusive age in research but also sees the positive effects of this inclusivity borne out in the increased equity of benefits derived from the practice of biomedical research.

4.3 Returning genetic and genomic results to participants

Much empirical research and scholarship over the last decade have focused on the ethical, legal, social, and practical issues surrounding the return of genetic and genomic research results.[57,58] These issues emerge from questions around which results to return, whether and what potential benefits and harms are associated with learning one's genetic and genomic information outside the clinical context; whether there is a right to, or not to, know such information; how to ensure informed consent; and how to protect participant privacy.[59,60] These questions, and many others, also present ethical tensions that require balancing of respect for persons/participant autonomy, beneficence/nonmaleficence, and justice considerations.[61–64]

4.3.1 Relevant considerations

Personal information comes with potential risks and possibility of harm to the individual and often, especially with genetic and genomic information, to a participant's family or blood

relatives. These risks may be psychosocial, ranging from temporary feelings of anxiety to long-term distress. Risks may also impact familial relations, family structure, or self-identity. Much of the relevant research thus far has focused on anxiety and stress upon learning of results. In some cases, these have been reported as being minimal, and often lessen over time.[65,66] A recent study by Mighton et al.[67] quotes participant responses to learning secondary findings as ranging from helping them "*have a good quality of life*" to "[it] *could ruin* [their] *quality of life.*" Risks may also be physical in nature, particularly when they arise due to unnecessary, unwarranted, or premature actions based on research results, such as adopting new health-related behaviors or interventions or discontinuing preventive measures without medical guidance.[68] These risks may be significantly reduced when the information returned is clearly actionable and of known medical significance.

Other risks may be economic in nature, such as the potential for increased medical spending, necessary changes in employment, or resources needed for support, such as disability and/or life insurance. Genetic and genomic information also poses privacy and confidentiality risks for individuals and their families, as when disease risks are uncovered, depending on with whom they share such information. Moreover, there are potential risks related to how genetic and genomic information may be used to discriminate against individuals. While the legal protections afforded by the Genetic Information Nondiscrimination Act (GINA; Pub. L. No. 110-233) prevent most cases of employment and health insurance discrimination based on "genetic information," life, disability, and long-term care insurers may still raise premiums and/or deny coverage in most localities.

For participants to make informed decisions about whether they want all or certain types of genetic or genomic results, participants must be clearly informed of all reasonably foreseeable risks and benefits of receiving such information. In general, respecting the individual's choice about whether and which results to receive is considered essential.[69] The research participant's choice(s) will depend on how they weigh the benefits and risks of knowing or not knowing this information, and will likely vary significantly among participants, given each individual's unique combination of beliefs, personalities, cognitive styles, values, and life circumstances.[66] For participants, this is without doubt a complex decision to make. Learning one's genetic information may have multiple social, emotional, and psychological implications. *All of Us* must balance beneficence and nonmaleficence in such return; there are also varying perspectives regarding what constitutes "harm" or "benefit," and these variations must inform the program's responsibilities and duties. Respect for persons requires that researchers be careful and deliberate in structuring informed consent processes such that processes allow individuals to make a truly informed and voluntary decision, free of coercion, that is aligned with their best interests and values. This translates into different models for informed consent, with most commonly used approaches requiring either "opt-in" or "opt-out" at the time of recruitment to the research study. As detailed in the next section, *All of Us* has developed both a separate consent process and a supporting opt-in approach to facilitate informed and contextualized decision-making about whether and which results participants opt to receive.

The notion of value is also complex and may depend on individual, familial, sociocultural, and other factors for each participant. Respecting the autonomy of participants to make voluntary informed choices about whether or not they want their research results back also means respecting the different ways in which value is construed, and how participants ascribe it to such results. In practice, this means that information provided during the consent process must not overly emphasize the value of knowing this information, while recognizing that learning such results have different and/or multiple values for participants depending on their individual, familial, and communal circumstances and other lived experiences.

Still, a common concern raised in the early planning stages for return of health-related genetic and genomic results was, "What about participants with limited or no access to genetics health care professionals?" *All of Us* strives to engage and enroll participants from populations that have historically been underrepresented in biomedical research, many of them also being medically underserved, making this a likely scenario. This and other questions such as, "What is the utility of results for participants with limited or no health insurance?" raise considerations of justice and whether potential benefits of receiving such information will be equitable. To help set realistic expectations, consent forms and related materials must clearly articulate that *All of Us* is a research program that does not provide medical care. Indeed, as a research program, *All of Us* would not be able to address systemic issues related to access to healthcare, including those due to systemic discrimination and inequality. Nonetheless, the program must remain cognizant of such concerns in approaches developed for implementing return of results, including considerations for whether and what resources and support systems it should create, to address such concerns and not further exacerbate inequities.

Lastly, participants' needs with respect to the manner in which genetic and genomic results are communicated may vary. As the 2018 National Academies of Science, Engineering, and Medicine report states, "… the process of communication is important to promote understanding of the meaning, potential uses, and limitations of the information."[58] Communication needs include considerations of participant literacy, including health and genomic literacy; accessibility of information, such as primary language, disability, and other factors; access to healthcare; need for emotional support; and cultural humility. Thus responsible return of results extends well beyond obtaining consent and respecting participant choice. It must take into account the life context of study participants, particularly when there is diversity of participants across demographic variables and inclusion of participants from groups that have limited resources and/or may be otherwise disenfranchised.

4.3.2 Guiding concepts for the return of genetic and genomic results

In its implementation, the return of genetic and genomic results by the *All of Us* research program is guided by the following concepts.

4.3.2.1 Accuracy

Despite the fact that results produced by the program are research results, and not intended for clinical use, results returned to participants should meet the highest of standards regarding accuracy. Specifically, any research results related to health outcomes must be analytically valid, as ensured by leveraging existing standards and best practices, and ultimately as deemed by the Food and Drug Administration. Strategies to ensure analytic validity include the use of College of American Pathologists/Clinical Laboratory Improvement Amendments (CAP/CLIA) certified genomics laboratories and validation of informative hereditary disease risk (HDR) findings with orthogonal methods prior to return to participants.

4.3.2.2 Clarity

As mentioned, there is a chance of harm if participants are unable to understand what their results mean, particularly when it comes to health-related research results. It is imperative that results are returned to participants in ways that minimize these harms. Thus *All of Us* must ensure results, their interpretations, and their limitations, especially with regards to the need for clinical confirmation, are presented clearly and comprehensively. Because genetic and health literacy may vary dramatically within the *All of Us* participant cohort, the program must anticipate a wide range of needs when it comes to clear communication of results.

4.3.2.3 Choice

As a research program that is not providing medical care, *All of Us* arguably does not have the same duty of care as entities involved in clinical practice.[70,71] As such, the full exercise of participant autonomy suggests the program should value a participant's right to refuse to receive results, regardless of the type and nature of those results.

Furthermore, different types of results present different types of potential harms, and each participant may have different risk tolerances with regard to these potential harms. They also may be of different perceived utility or value to each participant. In order that participants are able to make the choices that best suit their preferences and circumstances, different types of results should be separated from one another. Participants should be given the opportunity to accept or decline each type of result independently, and must be able to revoke their consent to receive results at any time. Similarly participants must also have the right to change their mind and provide consent to receive results, at any point, as changing life circumstances, such as their personal or family members' health status(es), may impact their decision about knowing their results.

4.3.2.4 Support

The potential for misunderstanding genetic and genomic results is high, as are the potential harms of these misunderstandings, particularly when it comes to health-related results. While high standards for clarity, as discussed previously, is a key feature of the *All of Us* model for

returning results, the program must also include additional support for participants. Such programmatic support should include dynamic, interactive strategies to ensure participant comprehension, such as genetic counseling, and should span the return pathway, from decision-making to downstream use, and take into consideration differences in accessibility needs.

4.3.2.5 Cultural humility

With a participant population as diverse as the *All of Us* participant cohort, the program must anticipate a variety of sensibilities and cultural needs when it comes to genetics and genomics. Results must be returned in ways that accommodate for these cultural differences and set clear boundaries around interpretation that protect participants' traditions, beliefs, and lived experiences while maintaining exacting standards of accuracy. Moreover, the program must ensure that materials used in the return of results do not create or reinforce prejudices against individuals, groups, or communities. As appropriate, the program should seek input from interested parties to inform the strategies and materials used in the return of genetic and genomic results.

4.3.3 Implementation

The program's implementation approaches for returning genetic and genomic results have been, and continue to be, informed by feedback from subject matter experts and interested parties, including those from the 2017 *Return of Genetic Results* workshop and 2019 *Ethical, Legal, and Social Implications (ELSI) Research Priorities* workshop;[72,73] the Food and Drug Administration; multiple internal governance bodies; and participant ambassadors. In addition, return protocols and materials underwent rigorous testing for comprehension and usability before implementation. Nevertheless, the program understands that the current iteration of the return process will necessarily change with time, in response to changes in information and technology, social and cultural context, and participant needs.

4.3.3.1 Types of results

All of Us will offer to return two main categories of genetic and genomic results to participants: (1) nonhealth-related; and (2) health-related. Currently, the nonhealth-related results offered are genetically-influenced traits, such as bitter taste perception, and genetic ancestry. Initially, four traits are offered (bitter taste perception, cilantro preference, earwax type, lactose intolerance), with plans to add additional traits at regular intervals to encourage continued program engagement. Genetic ancestry results may be updated over time as increased evidence in this field allows for refinement of current results.

For health-related genetic results, the program will initially offer pharmacogenetic (PGx) results and hereditary disease risk (HDR) results. As of the time of this publication, the program's PGx report will return results on seven genes: CYP2C19, DPYD, G6PD, NUDT15,

SLCO1B1, TPMT, and UGT1A1. The gene–phenotype–drug combinations the program will return appear in FDA-approved drug product labeling[74] or the FDA Table of Pharmacogenetic Associations,[75] or have a recommendation for alternative medication or dosing modification within a CPIC guideline.[76] Similarly, the genes analyzed and Pathogenic (P) and Likely Pathogenic (LP) variants returned for the HDR report are those with medical actionability for the associated condition(s), based on the ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing.[77] The analysis and return specifications for both the PGx and HDR results will evolve in an evidence-based fashion, as appropriate. For HDR results, the program plans to continue to follow ACMG recommendations as they are updated. However, the process of updating the program's IDE and analysis pipeline will necessitate some degree of delay between release of new ACMG recommendations and implementation by the program. Participants who have already received PGx or HDR results will receive an update if their results have a significant change in interpretation or new genes and variants are added to the results report. Additionally, the program will continue to explore the possibility of offering additional types of health-related genetic and genomic results in the future, as evidence builds to support the return of such results.

4.3.3.2 Informed consent and decision-making

As a longitudinal program that will continually add components throughout its lifetime, *All of Us* has built a modular electronic consent process (Fig. 4.3).[78] The Consent to Get DNA Results is a separate and subsequent consent experience from the primary consent, known as the Consent to Join the *All of Us* Research Program. During the Consent to Get DNA Results process, participants receive a general overview of DNA, variants, and limitations of our current knowledge; review potential risks and benefits of learning their genetic and genomic results; and are introduced to resources, like *All of Us* genetic counselors. *All of Us* genetic counselors and support agents at the Genetic Counselling Resource (GCR) Call Center are available to all participants to answer questions about choosing to receive genetic and genomic results from the program, as well as about the specific results participants receive. Subsequent to the receipt of results, participants may include their family members in their phone calls with genetic counselors or the GCR Call Center, and health-related results reports will have a phone number for healthcare providers to directly contact the GCR Call Center, if they wish.

As described above, *All of Us* offers several different types of genetic results to participants, each with their own potential implications, risks, and benefits. The program may also offer additional types of genetic and genomic results in the future. To allow for this, the program has designed a dynamic, just-in-time education and opt-in process for each type of genetic or genomic result, termed "informing loops" (Fig. 4.4). Each informing loop provides tailored information and allows the participant to decide if they want to receive that type of result or not, independently of other types of genetic or genomic results. Participants who have signed

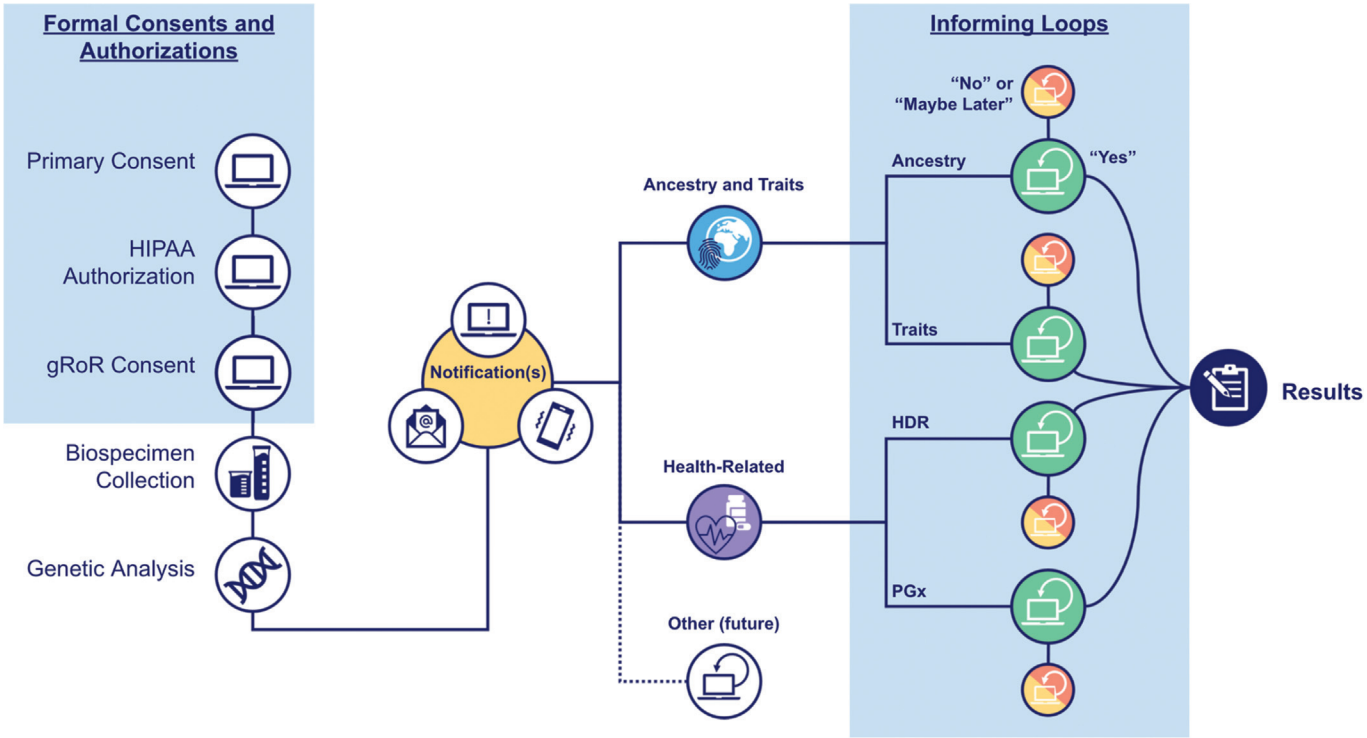

Figure 4.3

The All of Us modular consent process. Participants consent to engage in the program through a modular process that allows each participant granular control over their participant journey. Among the formal informed consent documents is the consent for the return of genetic and genomic results (Consent to Get DNA Results). Participants who opt to receive their results by consenting to return are eligible to make a "yes," "no," or "maybe later" decision for each type of genetic or genomic result through a dedicated "informing loop." These informing loops provide just-in-time information to aid in decision-making, and allow participants to make decisions about the receipt of each type of result independently. Participants can change their mind at any time about any part of the consent, authorization or informing loop process.

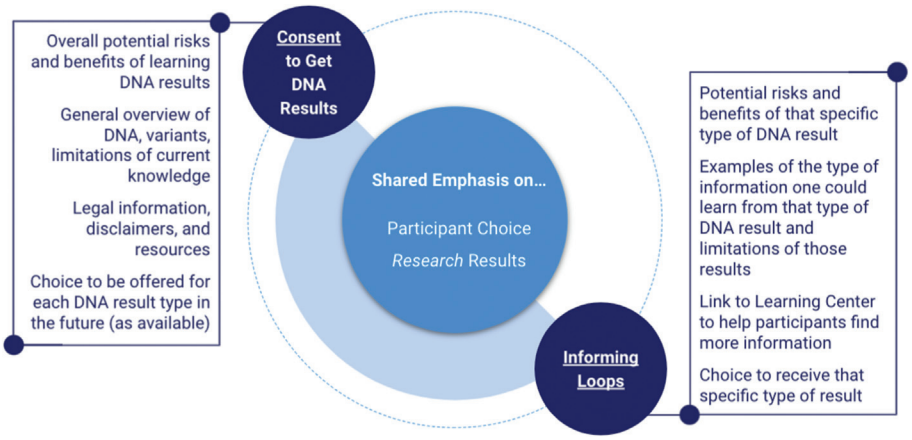

Figure 4.4
Informed consent versus informing loop decision points. *All of Us* employs two primary methods of registering participant choice with regards to the return of genetic and genomic results: formal consent (Consent to Get DNA Results) and just-in-time "informing loops." These two methods are used in different ways and at different junctures, as detailed in the figure.

the Consent to Get DNA Results are alerted each time a new type of genetic or genomic result can be generated for them. From that alert or the next time they log into their *All of Us* account, the participant will be directed to enter into the informing loop for that result type, review a few screens of information, and decide if they want the program to generate that type of result for them or not. As with consent, participants may change their mind at any time with respect to the receipt of different types of genetic and genomic results.

The informing loops also link out to the *All of Us* Learning Center, which serves as an additional educational resource for participants seeking more information and broader context. The Learning Center's genetic and genomic content provides a high-level overview of genetics and genomics, genetic and genomic research, genetic testing, genetic risk, and pharmacogenetics, as well as definitions for key terms. All participant-facing content created by the program is written at or below a 7th-grade reading level, and many sections are accompanied by illustrations or links to additional educational resources. Furthermore, the informing loops for health-related results link to the current list of genes the program will analyze for that particular result type. This allows the lists of genes to be updated as needed throughout the lifetime of the program and gives participants an accurate account of what will be tested at the time they make their decision about that result type.

4.3.3.3 Result generation and return

Once a participant has said "Yes" to a particular type of genetic or genomic result, results can be generated. This approach was specifically chosen for two reasons. First, the program

does not want to possess personalized genetic or genomic results that the participant does not want. This is both as a layer of privacy protection—that is, to not have a named result report on file unless and until it is requested—and to avoid the potential ethical quandary of having knowledge of informative pathogenic (P) or likely pathogenic (LP) HDR results that may not be delivered to a participant, depending on consent status. Second, this approach promotes voluntary decision-making. Knowing results are already available may bias participants to consent to return of results due to common cognitive biases triggered in decision-making, such as choosing the "default status" or loss-aversion.[79]

Results are derived from analysis of either array-based genotyping data (genetic ancestry and traits) or whole genome sequencing data (health-related results) that the program generates for every participant as part of its genetic and genomic research data pipeline. Genetic ancestry and trait results are automatically generated and available for viewing nearly instantaneously in the participant's *All of Us* account portal after completion of the associated informing loop. The health-related results, though, take more time to generate and return. Participant samples with a P or LP HDR result will undergo orthogonal testing by a medically established method for confirmation before report generation. Final reports for all health-related results are reviewed—and P/LP HDR results are signed out—by a board-certified laboratory geneticist or molecular pathologist at one of *All of Us*' clinical validation laboratories. These reports are then returned to participants as research results, which would require clinical validation for use in the course of medical care.

Participants can access their PGx and nonpathogenic HDR results through their *All of Us* account portal. Participants will receive a notification that their results are ready and will be prompted to view them by logging into their account. Upon initial selection of the new results, they will first view a few preresults screens. These will remind participants of some key information about the type of results they are about to see. Then they will be able to view their results in a dynamic web environment, as well as print or download a PDF version of their results report.

All of Us will employ a higher-touch method for returning P or LP HDR results. Participants with these results will be invited to schedule a phone-based appointment with an *All of Us* genetic counselor to review their results, free of charge. Genetic counseling appointments and the availability of the GCR Call Center to all participants are essential aspects of responsible return of genetic and genomic results to *All of Us* participants. The GCR Call Center is staffed with genetics experts and counselors fluent in English and Spanish. However, genetic counseling appointments can be scheduled in over 200 languages with the help of a medical translating service. This allows participants to receive and discuss these important and complex results in the language with which they are most comfortable.

During the initial appointment to review P/LP HDR results, the genetic counselor will "release" those results for viewing and download in the participant's *All of Us* account portal

and walk the participant through them. While it is suggested that the participant follow along on the device of their choice, it is not required. Participants can receive their results purely over the phone without requiring internet or device access. The genetic counselor will help to contextualize what the participant's results may mean for them; assist them with sharing this information with family, if they wish; and discuss next steps, such as sharing the results with their health care provider. The participant can also request their results be faxed directly to their health care provider. If the participant does not have a health care provider, the genetic counselor can help connect them with local resources. Participants can also have a family member or close friend join their appointment call with them. In the event that a participant exhibits high levels of emotional distress, the genetic counselor will work to connect them with a mental health care provider and transition the participant to their care through a warm handoff whenever possible.

4.3.4 Lessons learned and future directions

One of the challenges of developing the genetic ancestry and traits informing loops and results content was that *All of Us* is a federally funded scientific research program, and these types of results are often associated with direct-to-consumer genetics companies. Given this backdrop, the program felt that it was critical to strike the right balance for returning these results in the context of engaging and interesting information, while also being accurate about their limitations. For example, results for genetic ancestry may be more precise for those with European genetic ancestry because they are based on comparisons with publicly available datasets that have more data from people with European genetic ancestry. As discussed earlier, the focus on diversity by *All of Us* is in response to, and hopes to address, these persistent gaps.

Additionally, it was critical that the program return scientifically accurate information in a way that also clarifies that genetic and genomic results are only part of a person's identity, and may not align with family stories or religious and cultural beliefs about ancestry. For example, *All of Us* reports note that genetic ancestry results indicate where one's ancestors "might have lived hundreds of years ago," rather than definitively stating their ancestors originated from a certain region or regions. Also, they clearly articulate that the socially constructed categories of race and ethnicity are not determined by DNA. These harken back to the program's guiding concept of cultural humility for the return of genetic and genomic results. This was an area the program tried to approach thoughtfully, engaging in conversations with a variety of subject matter experts and multiple rounds of review.

A main driver in offering return of health-related genetic and genomic results to *All of Us* participants is the return of value to them. However, as noted above, these results are secondary findings from a research pipeline and therefore require clinical confirmation testing before they can be used in patient care. The program has taken great care to provide clear

disclaimers in all health-related return of results materials, starting with the Consent to Get DNA Results through the results reports themselves, as well as all supporting educational materials. To assist health care providers with whom participants share their research results, results reports contain a technical section that provides more detailed information about the analysis performed and directions for contacting the GCR Call Center with any questions.

The program has also realized that an important part of promoting health equity is supporting the transition to care for participants who may not have ready access to clinical genetic testing. The program does not aspire to further blur the line between research and clinical care by dispensing clinical services within the confines of the program. So to provide follow-up confirmation testing for participants who have a P or LP HDR result, the program has partnered with a third-party provider of physician-ordered clinical genetic testing. *All of Us* will connect eligible participants with the service and cover the cost of the clinical confirmation testing. The confirmation testing will not be carried out as part of *All of Us* research, but rather separate from it, under a consent process between the participant and the testing company. Participants may have the option to disclose their confirmation testing results back to *All of Us* in the future, but they will not be obligated to do so in order to receive the testing. While the program cannot provide follow-up medical care based on these results, this will give participants the results they need to engage in discussions with their health care provider about changes to their care.

Finally, the program will engage in continuous evaluation of its return of genetic and genomic results processes. As the *All of Us* cohort grows and the program evolves, the program will identify potential challenges, seek out input, and refine its processes to ensure the program continues to return results responsibly. To this end, the program is performing consultations with genomics and ELSI subject matter experts, and its participant ambassadors, to develop the goals and approaches for such continuous evaluation and process improvement.

4.4 Concluding remarks

The *All of Us* research program is one of several cohort programs marking a dramatic shift in research agendas away from pursuing one-size-fits-all solutions toward an emphasis on solutions tailored to the unique needs of individuals and communities. In addition to informing solutions to unmet clinical needs, these precision medicine and precision public health approaches can begin to combat health inequities in ways previously unimagined. Yet to do so, these approaches mandate new data sources, sources that both empower historically underrepresented populations and provide researchers with the data necessary to address health inequities.

Genetic and genomic data will be central to these efforts, as it will help researchers define the endogenous biological underpinnings of human existence, differentiate between

biological and environmental root causes for health and illness, and inform mechanisms for gene-environment interaction. We are also in a period of ascendency in the interest in, and social value of, genetic and genomic results beyond the research community. However, these data bring with them a host of complicating factors for research and nonresearch use.

As central tenets, the *All of Us* research program has committed itself to sharing genetic and genomic information with the research community and with participants. It has established means of doing both that endeavor to meet the needs of the recipients in ways that are responsive to a variety of appropriate applications and audiences. However, the program does not consider itself to be static: as times change, so must the program's approaches. The considerations and implementation strategies discussed in this chapter should be understood as the initial stages of a hopefully long and fruitful process, one that helps bring society from this moment into a better, more equitable, more informed, and empowered future. To meet the challenges of the future, *All of Us* must continue to listen to input from the participant, research, and other interested parties; constantly consider the social context in which the program exists; and revise its operations accordingly.

References

1. Teutsch SM, Fielding JE. Rediscovering the core of public health. *Annu Rev Public Health*. 2013;34:287–299. doi:10.1146/annurev-publhealth-031912-114433.
2. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC, USA: National Academies Press (US); 2011.
3. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793–795. doi:10.1056/NEJMp1500523.
4. Centers for Disease Control and Prevention. *Precision Public Health and Precision Medicine: Two Peas in a Pod | Blogs | CDC*. Genomics and Precision Health; 2021. https://blogs.cdc.gov/genomics/2015/03/02/precision-public/. [Accessed 29 August 2022].
5. Velmovitsky PE, Bevilacqua T, Alencar P, Cowan D, Morita PP. Convergence of precision medicine and public health into precision public health: toward a big data perspective. *Front Public Health*. 2021;9. doi:10.3389/fpubh.2021.561873.
6. Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman EAll of Us Research Program Investigators. The "All of Us" Research Program. *N Engl J Med*. 2019;381:668–676. doi:10.1056/NEJMsr1809937.
7. Konkel L. Racial and ethnic disparities in research studies: the challenge of creating more diverse cohorts. *Environ Health Perspect*. 2015;123:A297–A302. doi:10.1289/ehp.123-A297.
8. Oh SS, Galanter J, Thakur N, et al. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Med*. 2015;12. doi:10.1371/journal.pmed.1001918.
9. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538:161–164. doi:10.1038/538161a.
10. Mills MC, Rahal C. The GWAS diversity monitor tracks diversity by disease in real time. *Nat Genet*. 2020;52:242–243. doi:10.1038/s41588-020-0580-y.
11. Hindorff LA, Bonham VL, Brody LC, et al. Prioritizing diversity in human genomics research. *Nat Rev Genet*. 2018;19:175–185. doi:10.1038/nrg.2017.89.

12. Bentley AR, Callier SL, Rotimi CN. Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genomic Med*. 2020;5:5. doi:10.1038/s41525-019-0111-x.
13. Knepper TC, McLeod HL. When will clinical trials finally reflect diversity? *Nature*. 2018;557:157–159. doi:10.1038/d41586-018-05049-5.
14. All of Us Research Program. (2021). Core values. The All of Us Research Program|National Institutes of Health. https://allofus.nih.gov/about/core-values. [Accessed 29 August 2022].
15. Mapes BM, Foster CS, Kusnoor SV, et al. Diversity and inclusion for the All of Us research program: a scoping review. *PLoS One*. 2020;15. doi:10.1371/journal.pone.0234962.
16. All of Us Research Program. (2018). The all of us research program operational protocol. The All of Us Research Program|National Institutes of Health. https://allofus.nih.gov/sites/default/files/aou_operational_protocol_v1.7_mar_2018.pdf. [Accessed 29 August 2022].
17. Ramirez AH, Gebo KA, Harris PA. Progress with the all of us research program: opening access for researchers. *JAMA*. 2021;325:2441–2442. doi:10.1001/jama.2021.7702.
18. Zhang, S (2018). How a genealogy website led to the alleged golden state killer. The Atlantic. https://www.theatlantic.com/science/archive/2018/04/golden-state-killer-east-area-rapist-dna-genealogy/559070/. [Accessed 29 August 2022].
19. Zhang S (2018). Your DNA is not your culture. The Atlantic Magazine. https://www.theatlantic.com/science/archive/2018/09/your-dna-is-not-your-culture/571150/. [Accessed 29 August 2022].
20. Shafer, E (2019). Here are the lyrics to Lizzo's "Truth Hurts."Billboard. https://www.billboard.com/articles/news/lyrics/8517815/lizzo-truth-hurts-lyrics. [Accessed 29 August 2022].
21. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339:321–324. doi:10.1126/science.1229566.
22. Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. *Science*. 2018;362:690–694. doi:10.1126/science.aau4832.
23. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. *Nat Rev Genet*. 2020;21:615–629. doi:10.1038/s41576-020-0257-5.
24. Scherr CL, Ramesh S, Marshall-Fricker C, Perera MA. A review of African Americans' beliefs and attitudes about genomic studies: opportunities for message design. *Front Genet*. 2019;10:548. doi:10.3389/fgene.2019.00548.
25. Office of Science Policy, National Institutes of Health. (2017). NIH Genomic Data Sharing Policy. NIH GDS Policies. https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf. [Accessed 29 August 2022].
26. Metcalf J, Crawford K. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data Soc*. 2016. doi:10.1177/2053951716650211.
27. Metcalf, J, Keller, EF, & Boyd, D (2016). Perspectives on big data, ethics, and society. Council for Big Data, Ethics, and Society. https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/. [Accessed 29 August 2022].
28. Sabatello M, Juengst E. Genomic essentialism: its provenance and trajectory as an anticipatory ethical concern. *Hastings Cent Rep*. 2019;49(Suppl 1(Suppl 1)):S10–S18. doi:10.1002/hast.1012.
29. Wensley D, King M. Scientific responsibility for the dissemination and interpretation of genetic research: lessons from the "warrior gene" controversy. *J Med Ethics*. 2008;34:507–509. doi:10.1136/jme.2006.019596.
30. Hatzenbuehler ML, Pachankis JE. Stigma and minority stress as social determinants of health among lesbian, gay, bisexual, and transgender youth: research evidence and clinical implications. *Pediatr Clin North Am*. 2016;63:985–997. doi:10.1016/j.pcl.2016.07.003.
31. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet (London, England)*. 2017;389:1453–1463. doi:10.1016/S0140-6736(17)30569-X.

32. Garrison NA, Barton KS, Porter KM, Mai T, Burke W, Carroll SR. Access and management: indigenous perspectives on genomic data sharing. *Ethn Dis*. 2019;29(Suppl 3):659–668. doi:10.18865/ed.29.S3.659.
33. Garrison NA, Hudson M, Ballantyne LL, et al. Genomic research through an indigenous lens: understanding the expectations. *Annu Rev Genomics Hum Genet*. 2019;20:495–517. doi:10.1146/annurev-genom-083118-015434.
34. U.S. Department of Health and Human Services (2010). U.S. Department of Health and Human Services Tribal Consultation Policy. https://www.hhs.gov/sites/default/files/iea/tribal/tribalconsultation/hhs-consultation-policy.pdf. [Accessed 29 August 2022].
35. Anderson N, Bragg C, Hartzler A, Edwards K. Participant-centric initiatives: tools to facilitate engagement in research. *Appl Trans Genomics*. 2012;1:25–29. doi:10.1016/j.atg.2012.07.001.
36. All of Us Research Program. (2020f). All of us publication and presentation policy. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Publication_and_Presentation_508.pdf. [Accessed 29 August 2022].
37. All of Us Research Program. (2021a). All of Us research program tribal consultation final report. National Institutes of Health (NIH) — All of Us. https://allofus.nih.gov/all-us-research-program-tribal-consultation-final-report. [Accessed 29 August 2022].
38. All of Us Research Program. (2021b). information for American Indians and Alaska Natives. https://www.joinallofus.org/information-for-american-indians-and-alaska-natives. [Accessed 29 August 2022].
39. All of Us Research Program. (2021c). All of Us public data browser. All of Us Research Hub. https://databrowser.researchallofus.org/. [Accessed 29 August 2022].
40. All of Us Research Program. (2021d). Survey explorer. All of Us Research Hub. https://www.researchallofus.org/data-tools/survey-explorer/. [Accessed 29 August 2022].
41. All of Us Research Program. (2021e). Apply. All of Us Research Hub. https://www.researchallofus.org/apply/. [Accessed 29 August 2022].
42. All of Us Research Program. (2020). All of Us Data Access Framework. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Data_Access_Framework_508.pdf. [Accessed 29 August 2022].
43. Xia W, Liu Y, Wan Z, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. *J Am Med Inform Assoc*. 2021;28:744–752. doi:10.1093/jamia/ocaa327.
44. Zook M, Barocas S, Boyd D, et al. Ten simple rules for responsible big data research. *PLoS Comput Biol*. 2017;13. doi:10.1371/journal.pcbi.1005399.
45. Raymond N. Safeguards for human studies can't cope with big data. *Nature*. 2019;568:277. doi:10.1038/d41586-019-01164-z.
46. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, DC: Department of Health, Education and Welfare; 1978 *https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.*
47. All of Us Research Program. (2021f). The All of Us Data User Code of Conduct. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/12/AoU_Data_User_Code_of_Conduct_508.pdf. [Accessed 29 August 2022].
48. All of Us Research Program. (2020a). Ethical, legal, social implications in the All of Us Research Program: learnings, vision and approach for current and emergent issues. The All of Us Research Program | National Institutes of Health. https://allofus.nih.gov/sites/default/files/ELSI_White_Paper.pdf. [Accessed 29 August 2022].
49. All of Us Research Program. (2020c). All of Us Policy on the Ethical Conduct of Research. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Ethical_Principles_508.pdf. [Accessed 29 August 2022].

50. All of Us Research Program. (2020d). All of Us Policy on Stigmatizing Research. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Stigmatizing_Research_508.pdf. [Accessed 29 August 2022].
51. All of Us Research Program. (2021g). Research Projects Directory. All of Us Research Hub. https://www.researchallofus.org/research-projects-directory/. [Accessed 29 August 2022].
52. All of Us Research Program. (2020e). All of Us Data and Statistics Dissemination Policy. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Data_and_Statistics_Dissemination_508.pdf. [Accessed 29 August 2022].
53. Ferryman, K, & Pitcan, M (2018). Fairness in precision medicine. Data & Society. https://datasociety.net/library/fairness-in-precision-medicine/. [Accessed 29 August 2022].
54. Guterl F. Diversity in science: why it is essential for excellence. *Sci Am*. 2014;311:38–40.
55. Nielsen MW, Alegria S, Börjeson L. Opinion: gender diversity leads to better science. *Proc Nat Acad Sci USA*. 2017;114:1740–1742. doi:10.1073/pnas.1700616114.
56. McEligot AJ, Behseta S, Cuajungco MP, Van Horn JD, Toga AW. Wrangling big data through diversity, research education and partnerships. *Calif J Health Promot*. 2015;13:vi–ix.
57. Kaufman DJ, Baker R, Milner LC, Devaney S, Hudson KL. A survey of U.S adults' opinions about conduct of a nationwide precision medicine initiative® cohort study of genes and environment. *PLoS One*. 2016;11. doi:10.1371/journal.pone.0160461.
58. National Academies of Sciences, Engineering, and Medicine. *Returning Individual Research Results to Participants: Guidance for a New Research Paradigm*. Washington, DC, USA: The National Academies Press; 2018.
59. Knoppers BM, Deschênes M, Zawati MH, Tassé AM. Population studies: return of research results and incidental findings Policy Statement. *Eur J Hum Genet*. 2013;21:245–247. doi:10.1038/ejhg.2012.152.
60. World Medical AssociationWorld Medical Association Declaration of Helsinki ethical principles for medical research involving human subjects. *J Am Med Assn*. 2013;310:2191–2194. doi:10.1001/jama.2013.281053.
61. Henderson GE, Wolf SM, Kuczynski KJ, et al. The challenge of informed consent and return of results in translational genomics: empirical analysis and recommendations. *J Law Med Ethics*. 2014;42:344–355. doi:10.1111/jlme.12151.
62. Christenhusz GM, Devriendt K, Van Esch H, Dierickx K. Ethical signposts for clinical geneticists in secondary variant and incidental finding disclosure discussions. *Med Health Care Philos*. 2015;18:361–370. doi:10.1007/s11019-014-9611-8.
63. Horton R, Lucassen A. Consent and autonomy in the genomics era. *Curr Genet Med Rep*. 2019;7:85–91. doi:10.1007/s40142-019-00164-9.
64. Halverson C, Bland ST, Leppig KA, et al. Ethical conflicts in translational genetic research: lessons learned from the eMERGE-III experience. *Genet Med*. 2020;22:1667–1672. doi:10.1038/s41436-020-0863-9.
65. Stewart K, Wesselius A, Schreurs M, Schols A, Zeegers MP. Behavioural changes, sharing behaviour and psychological responses after receiving direct-to-consumer genetic test results: a systematic review and meta-analysis. *J Community Genet*. 2018;9:1–18. doi:10.1007/s12687-017-0310-z.
66. Wade CH. What is the psychosocial impact of providing genetic and genomic health information to individuals? An overview of systematic reviews. *Hastings Cent Rep*. 2019;49(Suppl 1):S88–S96. doi:10.1002/hast.1021.
67. Mighton C, Carlsson L, Clausen M, et al. Quality of life drives patients' preferences for secondary findings from genomic sequencing. *Eur J Hum Genet*. 2020;28:1178–1186. doi:10.1038/s41431-020-0640-x.
68. Beskow LM, Hammack CM, Brelsford KM. Thought leader perspectives on benefits and harms in precision medicine research. *PLoS One*. 2018;13. doi:10.1371/journal.pone.0207842.
69. Knoppers BM, Zawati MH, Sénécal K. Return of genetic testing results in the era of whole-genome sequencing. *Nat Rev Genet*. 2015;16:553–559. doi:10.1038/nrg3960.
70. Beskow LM, Burke W. Offering individual genetic research results: context matters. *Sci Transl Med*. 2010;2 38cm20. doi:10.1126/scitranslmed.3000952.

71. Prince AE, Conley JM, Davis AM, Lázaro-Muñoz G, Cadigan RJ. Automatic placement of genomic research results in medical records: do researchers have a duty? should participants have a choice? *J Law Med Ethics*. 2015;43:827–842. doi:10.1111/jlme.12323.
72. Ray, T (2017). Precision medicine project mulls how to return genetic test results to 1M participants. GenomeWeb. https://www.genomeweb.com/sequencing/precision-medicine-project-mulls-how-return-genetic-test-results-1m-participants#.YRQqldNKiem. [Accessed 29 August 2022].
73. All of Us Research Program. (2020h). Ethical, legal, and social implications in the All of Us research program: learnings, vision, and approach for current and emergent issues. https://allofus.nih.gov/sites/default/files/ELSI_White_Paper.pdf. [Accessed 29 August 2022].
74. National Center for Toxicological Research. (2021). FDA label: full-text search of drug product labeling. U.S. Food and Drug Administration. https://www.fda.gov/science-research/bioinformatics-tools/fdalabel-full-text-search-drug-product-labeling. [Accessed 29 August 2022].
75. Center for Devices and Radiological Health. (2021). Table of pharmacogenetic associations. U.S. Food and Drug Administration. https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations. [Accessed 29 August 2022].
76. Clinical Pharmacogenetics Implementation Consortium. (2021). Guidelines. https://cpicpgx.org/guidelines/. [Accessed 29 August 2022].
77. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017;19:249–255. doi:10.1038/gim.2016.190.
78. Doerr M, Grayson S, Moore S, Suver C, Wilbanks J, Wagner J. Implementing a universal informed consent process for the All of Us Research Program. *Pac Symp Biocomput*. 2019;24:427–438.
79. Cho I, Bates DW. Behavioral economics interventions in clinical decision support systems. *Yearbook Med Inform*. 2018;27:114–121. doi:10.1055/s-0038-1641221.
80. All of Us Research Program. (2021b). The All of Us data user code of conduct. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/12/AoU_Data_User_Code_of_Conduct_508.pdf. [Accessed 29 August 2022].
81. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. *Am J Prev Med*. 2016;50:398–401. doi:10.1016/j.amepre.2015.08.031.
82. All of Us Research Program. (2020g). All of Us data user appeals policy. All of Us Research Hub. https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_User_Appeals_508.pdf. [Accessed 29 August 2022].
83. Milne R, Morley KI, Howard H, et al. Trust in genomic data sharing among members of the general public in the UK, USA, Canada and Australia. *Hum Genet*. 2019;138:1237–1246. doi:10.1007/s00439-019-02062-0.
84. The White House Precision Medicine Initiative Interagency Working Group. *Precision Medicine Initiative Privacy and Trust Principles.* https://allofus.nih.gov/protecting-data-and-privacy/precision-medicine-initiative-privacy-and-trust-principles.

CHAPTER 5

A community approach to standards development: The Global Alliance for Genomics and Health (GA4GH)

Angela Page[a], **Melissa Haendel**[b] **and Robert R. Freimuth**[c]

[a]Global Alliance for Genomics and Health Secretariat, Broad Institute of MIT and Harvard [b]Center for Health AI, University of Colorado Anschutz Medical Campus [c]Department of Artificial Intelligence and Informatics, Center for Individualized Medicine, Mayo Clinic

5.1 Introduction

The Global Alliance for Genomics and Health (GA4GH) is an international standards development organization (SDO) focused on advancing human health and medicine through genomic data sharing and interoperability. Founded in 2013, the organization has evolved over time and has adapted to unexpected challenges in ways that we believe would be of interest to the broad biomedical community. In this chapter, we present our experiences to support others wishing to share data through global, community-driven standards. We focus on three themes that have emerged over the years that have been critical to the organization's success: (1) community needs must drive development; (2) agility is necessary to create global equity and opportunity; (3) developing an idea into a widely adopted standard requires multiple levels of consensus (e.g., broad agreement that a standard is needed, identification of a common scope and definition of the problem, and codevelopment of a specification while considering multiple potential and often disparate approaches). By applying these principles, GA4GH has demonstrated how an adaptable approach to organizational structure allows for iterative change and swift progress.

5.2 The rationale for and promise of an international alliance (2012–2014)

Following the completion of the first draft of the Human Genome in 2001 and the subsequent million-fold decrease in the cost of sequencing, a massive influx of genomic data was on the

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00011-0
Copyright © 2023 Elsevier Inc. All rights reserved.
For Robert Freimuth part of the contribution, Mayo retains copyright on the drawings, questionnaires, and survey instruments contained within the contribution.

horizon by the early 2010s. At this time, a group of leaders in the genomics domain, led by David Altshuler (then Deputy Director of the Broad Institute of MIT and Harvard), began to recognize that in order make the most use of these data, and realize the promise of genomic medicine established over the previous decade, incomparably large and more diverse datasets would be required. That meant the data would need to be shared across international and regulatory borders. Doing so would provide researchers with the statistical power that can only come from large cohorts and population-based studies, and would ensure that discoveries from research would have the potential to translate into healthcare more equitably. But such widespread harmonization and sharing was not yet the norm in the biomedical community, which would need to come together around a common set of technical practices and regulatory principles in order to achieve its goals.

With these motivations front-of-mind, and inspired by the examples of the development of the Internet and the World Wide Web, wherein technical harmonization enabled open sharing and collaboration and led to unprecedented scientific advances, Altshuler worked with Peter Goodhand of the Ontario Institute of Cancer Research and six other collaborators to plan a meeting of the world's leading experts in genomics and medicine. The goal of the meeting was to identify the infrastructure and standards that would be needed to share and integrate genomic data in a secure, controlled, and interoperable manner. They drafted a white paper that outlined the opportunities and challenges of sharing genomic and related health data, a document that was ultimately shared with more than 70 colleagues across the globe, and later released publicly (https://www.ga4gh.org/wp-content/uploads/White-Paper-June-3-final.pdf).

On January 28, 2013, 50 members of that group met to discuss how the community could work together to advance genomic medicine through data sharing, unlocking discovery while simultaneously respecting patient autonomy and their right to privacy. The group concluded that a multicomponent approach would be needed to meet the disparate needs of the patient, research, and clinical communities.

The first of these components would be a "global alliance of international partners, entrusted with the mission of enabling rapid progress in biomedicine." This collaborative group would create and maintain "interoperability of technical standards for managing and sharing sequence data in clinical samples." This would be accomplished by engaging stakeholders across sectors to encourage responsible and voluntary sharing of data and of methods. In addition, the group realized that technology platforms based on open standards would need to be designed to enable secure storage, responsible data access, participant-centric consent, and reproducible analysis. Lastly, operating entities would need to create platforms that would implement the standards to provide services and aggregate data for users. In doing so, these groups would commit to shared principles, spark innovation, and advance research, application, and practice.

In June 2013, 6 months after the initial meeting, the anticipated global alliance of international partners was launched when 73 organizations signed a letter of intent to formalize their commitment to this endeavor. The Global Alliance for Genomics and Health was born.

In that same month, an article published in *Science Magazine* titled "A Human Right to Science" caught the attention of Bartha Knoppers, a legal scholar based at McGill University and a founding member of GA4GH.[1] Knoppers had been studying bioethics and law for more than 20 years, but this framing of science as a human right was new. Article 27 of the 1948 Universal Declaration of Human Rights stated that every individual in the world had the right to both (1) "share in scientific advancement and its benefits" (including to freely engage in the responsible scientific inquiry), and (2) "the protection of the moral and material interests resulting from any scientific…production of which [a person] is the author." The purpose of the editorial was to explore how the scientific community could activate this human right, which had been adopted into the International Covenant on Economic, Social, and Cultural Rights in 1966 but which was otherwise rarely discussed in the literature and even less present in the minds of policymakers.

This was a "Eureka moment" for Knoppers, who was simultaneously working with Kazuto Kato of Keio University to establish one of four founding GA4GH working groups—the one that would focus on regulation and ethics for genomic data sharing. Knoppers and Kato began asking how GA4GH could use Article 27 and "catapult [data sharing] into a new way of thinking."[2] Instead of focusing on protecting against the risks and potential harms that sharing might reveal, GA4GH would focus on the benefits and the value proposition. Knoppers began working with Edward Dove, now at the University of Edinburgh, and an international group of colleagues on the seminal document that would underpin all GA4GH work going forward:

> [The *Framework for Responsible Sharing of Genomic and Health-Related Data*] is guided by the human rights of privacy, non-discrimination and procedural fairness. At the same time, it considers all human rights principles relevant, complementary and interrelated, founded as they are on respect for human dignity. Since science proceeds only with the broad support of society, respect for all persons is a primary driver underlying all other derived principles.[3]

The *Framework* established a set of foundational principles that would hold researchers and research data systems accountable while stating that data producers and users have a *responsibility* to access and share genomic and health-related data across the translation continuum—a necessary action to enable the human right of everyone to share in the benefits of genomic science.

5.3 Convening the community (2014–2017)

Since its earliest days, GA4GH has had a global operational structure to match its global scope, with expert guidance and leadership provided by members of the international community and operational support from a dedicated, globally distributed staff team. The GA4GH leadership consists of an international Steering Committee that is reflective of the diverse perspectives, backgrounds, and geography of the GA4GH community.

By April 2014, over 175 organizations across academia, healthcare, and industry had signed onto the mission and principles outlined in the *Framework* and joined GA4GH as organizational members. Along with the Regulatory and Ethics Working Group led by Knoppers and Kato, three other founding working groups had been established to accelerate progress in the areas of genomic data (the Data Working Group, initially led by David Haussler of UCSC and Richard Durbin of the Wellcome Sanger Institute), clinical data (the Clinical Working Group, led by Kathryn North of the Murdoch Children's Research Institute, and Charles Sawyers of Memorial Sloan Kettering Cancer Center), and security and privacy (the Security Working Group, led by Dixie Baker of Martin, Blanck, and Associates, and Paul Flicek of EMBL-EBI). Together, this community had collectively identified and launched three demonstration projects to highlight the value that responsibly sharing genomic data could have on the scientific and health communities.

The first of these, matchmaker exchange, was launched prior to the GA4GH but was brought under its wing as an exemplar of the rare disease use case in which finding just one additional patient from many datasets—the proverbial needle hidden in a field of haystacks—can be sufficient to diagnose an otherwise intractable phenotype. Next, the beacon project aimed to show the world how sharing even a minimal amount of information—simply the presence or absence of an allele in a dataset—could yield enormous value for researchers around the world who could otherwise find themselves locked into regional data silos. Finally, BRCA Challenge was launched as a gene-specific open sharing initiative intended to confront the risk of commercial ownership of data while simultaneously developing a template for openly sharing information about other genes in the future.

Three years into the lifetime of the organization, it had become clear that more structure was needed. The data deluge that had previously been a twinkle in the eyes of GA4GH founders was now becoming a reality, and while each of the demonstration projects was an individual success that operated as a self-sustaining entity long after its initial launch and with hundreds to thousands of data points that demonstrated its utility, they did not fully reflect the broad needs of the community.

After the development of matchmaker exchange and beacon and the success in achieving their initial goals, their standards needs had been met. In addition, with its focus on aggregating publicly accessible data, the BRCA Challenge did not face some of the most pressing

concerns of those wishing to harness patient data for research or to transfer data across diverse regulatory jurisdictions.

At the same time, the initial working groups seemed to be operating, ironically, in their own silos, developing solutions often in search of problems. The real-world needs being outlined in the Clinical Working Group were not driving the development of standards in the Data Working Group, and the outputs of the latter were not being evaluated in real-world contexts.

In 2016, the GA4GH steering committee nominated its third chair, Ewan Birney, who was Director of EMBL-EBI and is now also Deputy Director of EMBL. As Birney said in an interview, "There was lots of good work and interesting work happening inside of GA4GH at the time. But [we at EBI] noticed that there wasn't a connection to the incredibly practical questions that many of the people doing genomics in research, and in health care delivery, were facing….We realized…that we were either going to have to create a competitor to GA4GH, or we were going to have to make GA4GH very, very practically work for a large number of groups. We chose the latter."[4]

In December 2016, Birney announced the launch of a strategic planning effort that would be nothing short of an organizational revolution. Each working group, task team (subgroups focused on specific projects within a working group), and demonstration project was asked to outline their "Vision for 2022," including what they saw, from their respective corners, as the ideal state of the field in 5 years, and the technical requirements needed to achieve that state. Simultaneously, each group was asked to draft a five-year roadmap of projected milestones and accomplishments, along with any dependencies with other areas of the community.

To complete these requests, the groups consulted with the broader community, gathering inputs and priorities from large-scale data centers as well as high-level directional input from international leaders in genomics. By February 2017, the data had been collected and submitted to the GA4GH secretariat, which consolidated it into a draft strategic plan (https://www.ga4gh.org/wp-content/uploads/GA4GH-Connect-A-5-year-Strategic-Plan.pdf). In May 2017, Birney invited the most active contributors within the GA4GH community to the Wellcome Genome Campus in Hinxton, UK, where EBI is based, for a 2-day working meeting to iron out the wrinkles in that draft and ratify it as the new foundational vision. The group also worked out a new matrix organizational structure that would come to be hailed by many in the community as the disruptive key to GA4GH's future success.

5.4 GA4GH connect (2017–2019)

The GA4GH matrix structure (Fig. 5.1) was developed as a mechanism to directly link the needs of users to the standards development activities of the technical community. It was structured around a global desire to meet shared challenges, with users serving as the "verticals" (driver projects), providing insights into the needs of the broad stakeholder community.

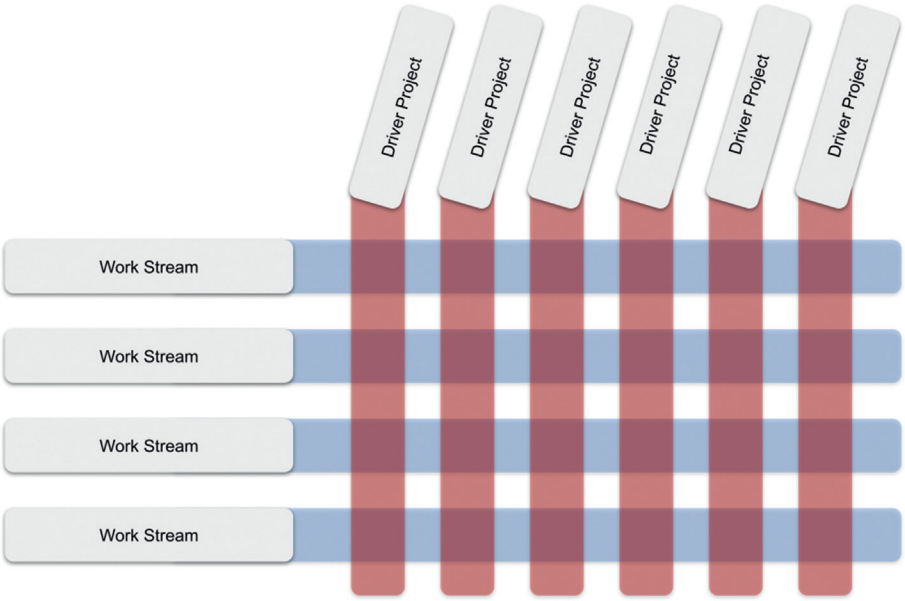

Figure 5.1
The GA4GH matrix structure brings together users (driver projects) with developers (work streams) to ensure that all deliverables are addressing real world needs.

These intersect with the "horizontal" development teams (work streams), which include expert contributors from across sectors and nations. In this way, GA4GH brings together an otherwise unlikely community of users and technical experts to advance "FAIR" science (findable, accessible, interoperable, reproducible).

The four initial working groups took on a new life as eight work streams (Fig. 5.2), each focused on a specific area identified as critical to the success of international genomic data sharing. The groups that set out those needs would be called driver projects, the first fifteen of which would be announced in October 2017 at the fifth annual GA4GH Plenary Meeting.

The new structure mandated that driver projects commit to not only providing use cases and requirements but also participating in the development, adoption, and testing of standards. While GA4GH remains an open and inclusive community in which anyone—regardless of domain, sector, nation, or affiliation—is invited to participate, driver projects are distinct from other groups in three ways. They must have:

1. A broad global remit or reach.
2. The potential to positively impact human health and contribute to a culture of data sharing.
3. The capacity to contribute to GA4GH standards development and policy framing.

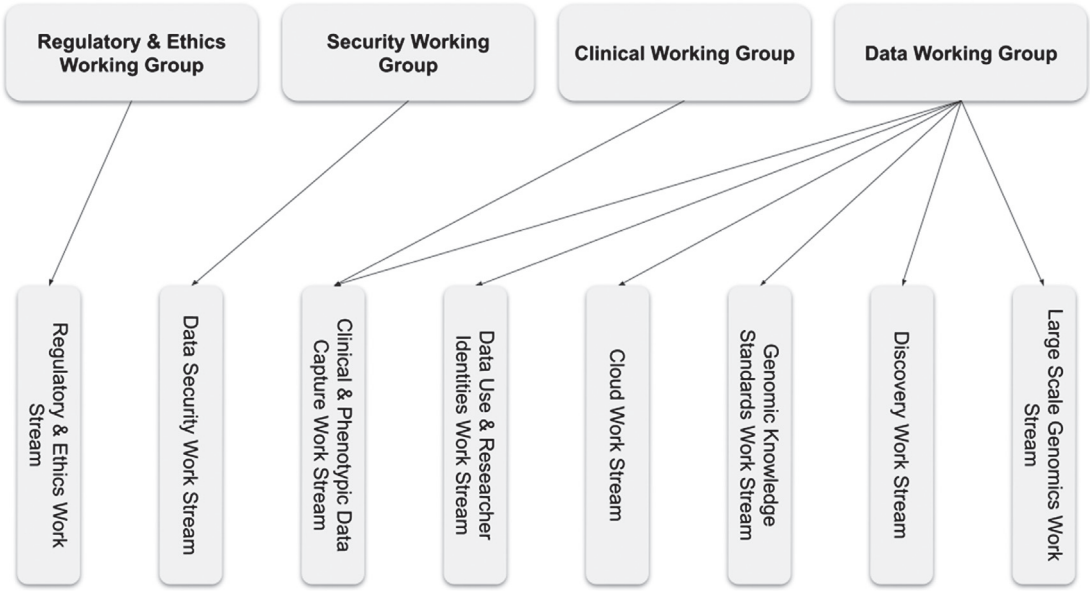

Figure 5.2
In 2017, the four original GA4GH working groups split into eight work streams—six technical and two foundational, with a remit to review all work that comes out of the organization.

Furthermore, driver projects are expected to contribute two full-time equivalents toward the active development of standards and policy framing within at least two work streams. Additionally, all driver projects must have a representative on each of the two foundational groups that review all GA4GH deliverables: the regulatory and ethics and data security work streams.

While common in some fields (e.g., information technology), this matrix organizational model is unusual in an SDO, bringing together a vast array of stakeholders to both provide input on and directly contribute to the development of the standards and frameworks they need. Through this approach, GA4GH leverages capacity well beyond its direct funding resources by connecting with data stewards, bioinformaticians, and software engineers around the globe who can contribute to GA4GH efforts while also advancing their own local priorities. The matrix structure encourages meaningful collaboration across a diverse spectrum of contributors, from different commercial sectors, scientific domains, jurisdictions, and institutions.

This new approach within GA4GH launched a flurry of activity over the following 2 years, during which the community developed and released over two dozen standards, which have since been implemented by over 40 organizations worldwide. Much of this success is attributed to the fact that the work streams directly engage the stakeholders at every stage of the standards development process, which also provides a tremendous opportunity for academia

and industry to work together pre- and post-competitively. An early motto within the community, attributable to founding GA4GH vice-chair and DWG co-chair David Haussler, was to "collaborate on interoperability and compete on implementations."

The driver project community (Fig. 5.3) has truly embraced that collaboration mandate. As a result of a second open call for new driver projects in 2018, the group currently consists of 24 distinct external projects focused on a broad set of scientific and clinical topics, from national initiatives to cancer knowledgebases to rare disease data sharing platforms. Some are focused on discovery research, others on building electronic infrastructure. Geographical scope varies as well: some driver projects are focused on generating data within a single nation while others are attempting to share data and knowledge across national borders. Despite their differences, this extensive group has come together to catalyze the outputs of GA4GH by both seeding our work streams with dedicated technical and policy experts as well as piloting work stream deliverables in real-world scenarios. While the diversity of driver projects spans the spectrum of use cases from clinical implementation to basic research, they are not—and could never be—an exhaustive representation of the entire genomics community. Instead, we strive for representation within every major class of genomics to include as many broadly relevant use cases as possible.

Despite its success, GA4GH continues to identify (and respond to) governance challenges in both a structural sense as well as a technical sense. For instance, while the organization strives to engage a truly global community, English-speaking countries still dominate the majority of the activity. There is also significant variability among the needs and offerings of driver projects, since the priorities and solutions that receive the most attention are often heavily influenced by the funding streams underlying the projects. Other challenges include limited interoperability between the work streams (and the standards they produce) and a lack of clarity around the minimum standards requirements needed for responsible international data sharing in the context of globally variable resource availability.

In keeping with its iterative, scientific approach to organizational operations, the GA4GH community rallied to address these challenges. Discussions around landscape analysis, improved requirements analyses, and approaches to work with existing industry standards all began to emerge as the bulk of the 2017 roadmap had been achieved.

5.5 Gap analysis (2019–2021)

In response to these discussions and with an informal sense of the issues at play, GA4GH launched a gap analysis in 2019 to confirm its internal suspicions and identify targeted responses from the community. The gap analysis consisted of both a bottom-up approach, led by the work streams, as well as a top-down approach led by GA4GH Vice-Chair Heidi Rehm of the Broad Institute of MIT and Harvard and Andrew Morris of Health Data Research, UK, senior leaders in the field of genomics—one internal to Ga4GH (Rehm) and one external

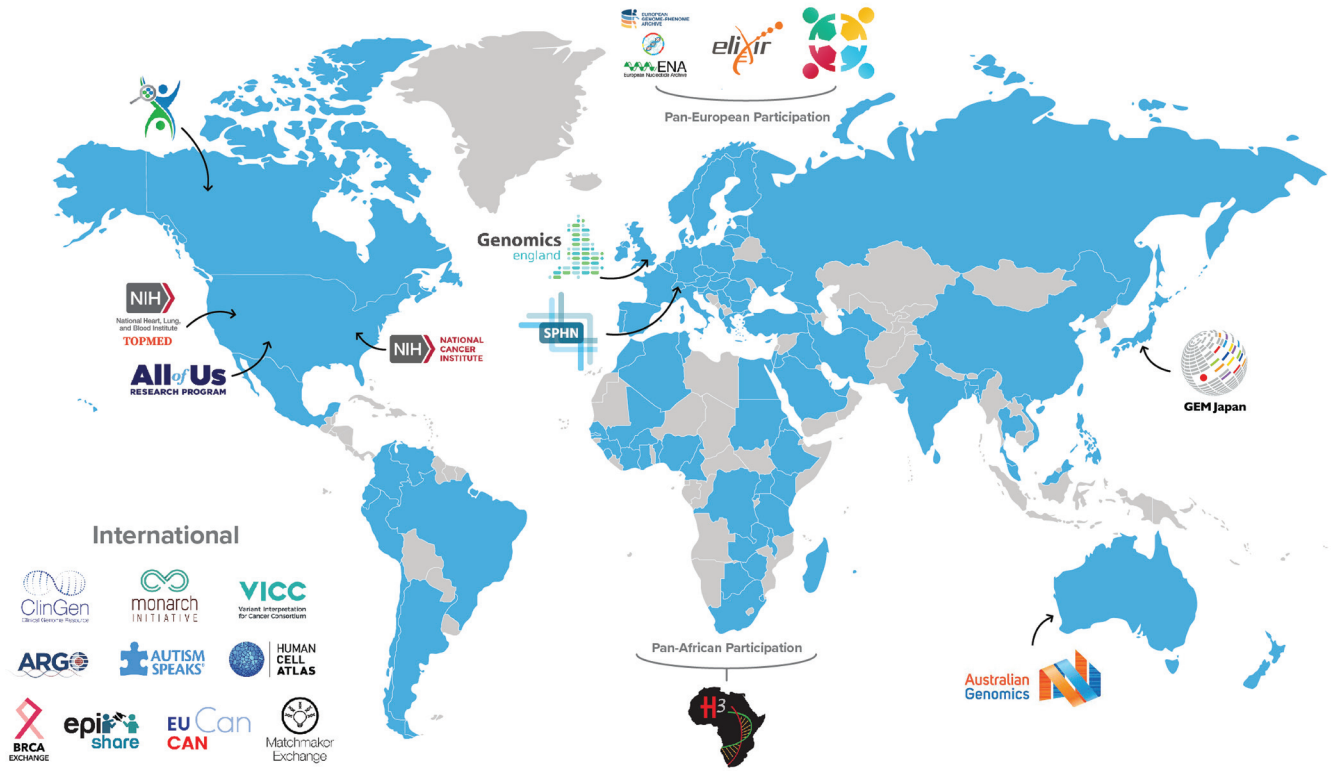

Figure 5.3
The GA4GH driver project community provides a broad international reach through manageable set of touchpoints.

(Morris). In addition, the GA4GH secretariat issued an open survey, in which all members of the genomics and health community were invited to submit their thoughts on GA4GH's work to date and future vision.

In the bottom-up process, the work streams engaged members of the community—hailing from within driver projects and elsewhere—during their regular video conferences. This activity led to the identification of 30 new deliverables that would be needed to advance genomic medicine; the collection was released in a revised product roadmap in early 2020.[5]

The top-down approach aimed to (1) identify gaps that may not be identified through the work stream roadmap process, (2) identify new opportunities for work that had not previously been on the GA4GH radar, (3) improve internal and external interoperability and collaboration, and (4) identify new strategies to drive uptake and usage of existing standards. To achieve these goals, Rehm and Morris chaired more than a dozen video conferences with driver project champions and work stream leads (see Fig. 5.4 for organizational chart and role descriptions) to discuss topics such as whether the GA4GH vision developed in 2017 was still appropriate given the current landscape, the most promising areas of impact in the health data field, opportunities to improve the GA4GH vision and shape the organization's future, and bottlenecks in standardization across the community. These discussions unearthed a host of concerns across 10 thematic areas:

1. Optimizing the pace and scope of standards development.
2. Developing a suite of truly interoperable standards.
3. Supporting and driving implementation and adoption.
4. Documenting, maintaining, and versioning standards over time.
5. Aligning activities across work streams.
6. Engaging a globally representative community of contributors.
7. Creating mechanisms to track implementation metrics and other key performance indicators.
8. Gaining clarity on the scope and intention of "openness" in standards development.
9. Engaging with external standards organizations.
10. Clarifying the GA4GH definition of the federation and outlining its optimal use cases.

The GA4GH steering committee reviewed these themes at a meeting in early 2020 and voted to focus on three specific areas over the coming years: (1) improve interoperability and alignment with external standards and between GA4GH standards, (2) improve implementation support for technical standards, and (3) engage more closely with healthcare and clinical standards. Future meetings outlined a strategy for achieving each of these "community imperatives," which was released as a new strategic plan alongside the product roadmap developed by the work streams. For each imperative, GA4GH outlined the significance of the problem, defined what success would look like, and identified potential solutions and tactics to achieve that success.

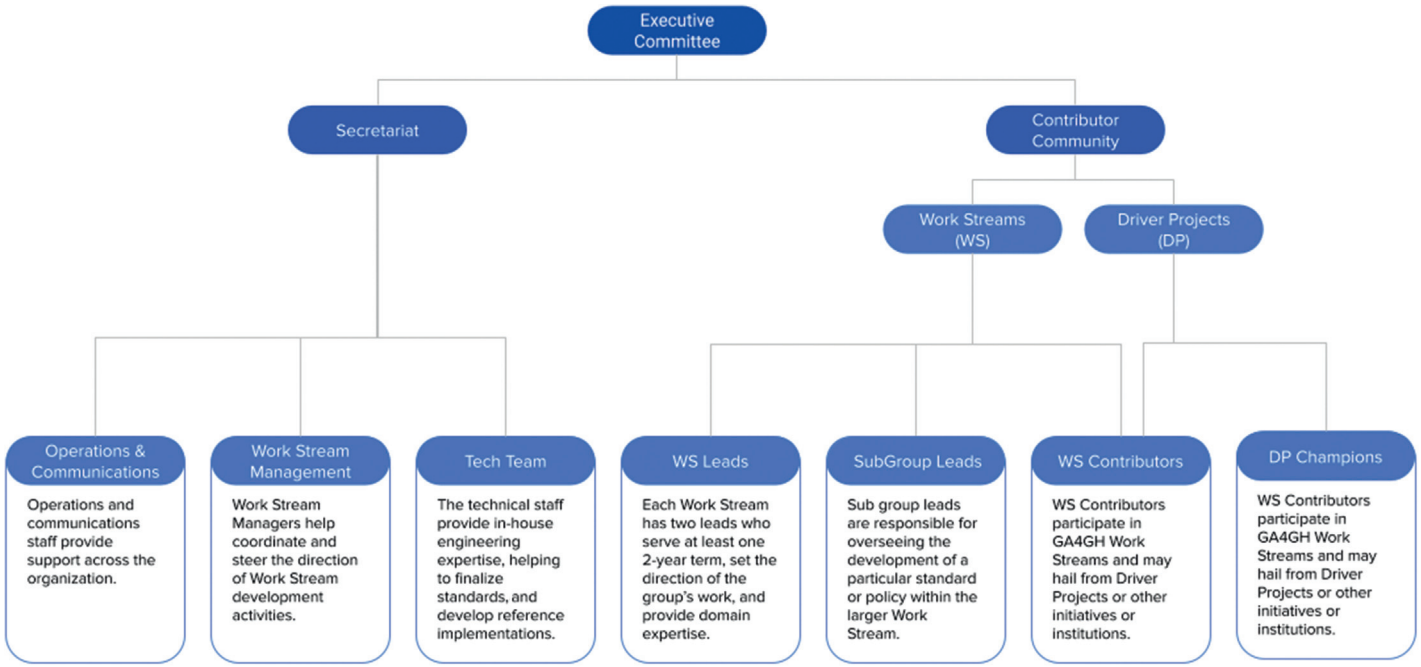

Figure 5.4

The broad GA4GH community, including both core secretariat and external contributors, is overseen by an executive committee. Work stream leads and driver project champions make up the current standards approval body, the GA4GH steering committee.

5.5.1 Technical alignment

In the area of **technical alignment**, the GA4GH community defined success as a suite of truly interoperable standards that would enable solutions encompassing multiple components of a pipeline across multiple platforms to support a variety of use cases. However, the distributed nature of the GA4GH meant that the resources were not being developed as a library of integrated standards that could be readily deployed together. Issues such as schema format incompatibility, identifier management, and metadata were incompatible across artifacts. Two existing activities rose to the top as key to achieving technical alignment, on which GA4GH leadership began placing greater emphasis: the technical alignment subcommittee (TASC) and the Federated Analysis Systems Project (FASP). TASC provided a space for early identification of opportunities for cross-functional alignment and provided the foundation for a cross-work stream governance body, but it lacked the authority to make and enforce decisions across the organization. FASP was an exciting demonstration of connectivity across multiple initiatives, but it required significant oversight and coordination across product teams to be successful. Both TASC and FASP arose in response to identified needs and were successful in beginning to address those needs, but to more completely meet our goals for technical alignment additional support was necessary. With sustainable funding (see Section 5.7), GA4GH has recently hired its first Chief Standards Officer as well as a number of dedicated software developers. These additions have made it possible for us to augment FASP and TASC and grow the role of those initiatives (as well as other similar efforts of the organization) to address the evolving challenges related to technical alignment across GA4GH products, which would be difficult to do without dedicated support.

5.5.2 Implementation support

In the area of **implementation support**, GA4GH defined success as the rapid implementation and adoption of its standards by driver projects and the expanding community, with subsequent uptake and broad interoperability across the globe. The FASP was identified as a key stepping stone to success, as were providing easy access to consistent documentation across deliverables and both in-person and virtual training workshops. In addition, core funding was secured to support a staff technical team to supplement the efforts of the broader contributor community and expedite implementation, documentation, testing, and adoption. Equitably allocating, coordinating, or otherwise prioritizing work across work streams remained a surmountable challenge as the team began its expansion in late 2021.

5.5.3 Clinical engagement

A key factor driving GA4GH's success is that genomics is a dynamic, rapidly growing domain, with significant support, demand, and enthusiasm from the international scientific

and healthcare communities. This urgency has helped ensure GA4GH has the resources and accountability it needs to follow through on its goals. As more and more national and international clinical genomic initiatives emerge, we expect this pressure to increase. GA4GH has already lowered barriers to data and information sharing with and between higher and lower resource settings, particularly in the area of genomic research. As we move forward, we must create and nurture partnerships with healthcare technologies and programs, so that genomics is brought into the mainstream of care not just for the wealthy few, but for the entire global population. We must be able to efficiently and effectively respond to the needs of clinical stakeholders and develop standards that support them.

The challenges of moving data into (and out of) health care are many, particularly when it comes to protected personal health information. To address some of these challenges, GA4GH sees promise in the federated approach of "data visiting" rather than data sharing, in instances where exchanging data is not possible or preferable. In this model, analyses are undertaken on the data where they reside and only aggregate results are returned to the researcher. Regardless of which mechanism is implemented, to work in the context of healthcare it is critical for the industry to adopt the required standards.

Conversely, the phenopackets standard was developed to enable deidentified case-level phenotype data sharing for computational use. Such individual case information is challenging to obtain both clinically and in the literature, yet is needed computationally for systems such as the matchmaker exchange and for diagnostic tools. GA4GH has begun to align the phenopackets standard for clinical data exchange with the Fast Healthcare Interoperability Resources developed by the healthcare SDO, Health Level Seven (HL7). Phenopackets is also now an International Organization for Standardization (ISO) standard, and is used globally for biobanking and rare disease data exchange.

GA4GH has helped establish a consortium of SDOs, the Cross Standards Development Organization (xSDO), with representatives from GA4GH, HL7, and ISO. The xSDO strives to coordinate the alignment of disparate but overlapping products, minimize the unnecessary proliferation of redundant standards, and avoid the development of semantically- and syntactically-conflicting standards. Through these goals, the xSDO aims to drive large-scale interoperability and consistency across the adopter community. While there is broad abstract enthusiasm across this consortium, success has been hampered by technical, process, and—most significantly—cultural differences across organizations. We collectively recognize the challenges in achieving this next level of coordination and technical harmonization, which will require dedicated resources, intentional effort, and strong organizational commitment.

GA4GH also developed a structure that was originally focused entirely on driving clinical engagement. By convening national genomic data initiatives that often live in the context of a national healthcare system, the Genomics in Health Implementation Forum (GHIF) provides a logical mechanism for advancing genomics into clinical care. GHIF's primary aim is to

empower knowledge exchange and collaboration through the implementation of GA4GH standards among global genomics initiatives as they pursue the common goal of advancing human health. While it has evolved to include research projects, international consortia, and more, it remains the primary mechanism by which GA4GH promotes the uptake of standards within a clinical context.

5.6 Beyond GA4GH connect (2021 and beyond)

In 2021, with more than two dozen GA4GH standards now available in the areas of data discovery, data access and researcher authentication, variation representation, cloud-based workflow interoperability, and more, the community began to put a larger emphasis on uptake and adoption. Development roadmaps are now primarily focused on alignment and integration between standards as well as federated interoperability across institutions and jurisdictions, rather than on the development of entirely new standards.

Prior to this point, GA4GH primarily focused its efforts on overcoming the challenges of enabling interoperability within new initiatives built on a "cloud-first" framework. However, an additional challenge is bringing existing high-performance computing (HPC) infrastructures into the federated interoperability network envisioned by this community. So while many GA4GH solutions were developed with cloud environments in mind, we must also ensure they can support interoperability with traditional HPC environments. Having specified many of the necessary elements for a minimal viable protocol for federation, the growing GA4GH developer staff team and other dedicated contributors began creating a set of Starter Kit Implementations in 2021 to ensure these existing environments were compatible with the wider GA4GH network. These resources took the form of the minimal implementations needed to facilitate interoperability among large-scale systems. At the same time, these resources served as off-the-shelf implementations that could be easily deployed within HPC and cloud environments alike. The Starter Kit was intended to provide a lightweight mechanism for HPC systems to achieve GA4GH compatibility, while also providing an on-ramp for entirely new data-sharing initiatives looking to build interoperability into their infrastructures from the ground up. As of late 2021, Starter Kit implementations had been initiated for the Passport, refget, Service Registry, Data Repository Service, and Workflow Execution Service standards, with pilot testing having commenced within EMBL-EBI and the BRCA Exchange.

In the early days of GA4GH connect, the driver project mechanism served as an appropriate means for representing the needs of the broader community without requiring direct interaction with each. Now, as the standards become available and optimized for widespread use, GA4GH must create new strategies for interacting with as much of the community as possible. Building on the driver project model of leveraging broad representation through a minimum number of touchpoints, the organization has begun to focus its engagement efforts on external consortia (rather than on individual projects) as well as GA4GH-hosted

"implementation communities" with GHIF that are organized around domain-specific use cases and convene experts from around the globe. In doing so, GA4GH hopes to access a scalable mechanism for incorporating many voices into its work without the need to add individual touchpoints. Furthermore, by providing more opportunities to engage beyond the driver project model, more initiatives can contribute to and influence GA4GH standards and technology development without needing to navigate or commit to work streams. The approach also strives to increase global membership and involvement in GA4GH, and aid its efforts to encourage diverse input and ensure that standards fit needs across multiple initiatives and national healthcare systems.

5.7 A novel approach to funding and support

Partly to address these new challenges, GA4GH has also begun to expand its funding model, which has always been unique. The organization's first two years were supported entirely by the three founding Host Institutions—the Broad Institute of MIT and Harvard in the United States, the Ontario Institute for Cancer Research in Canada, and the Wellcome Sanger Institute in the UK. These organizations committed in-kind resources such as staff and human resources support, meeting spaces, and more. Furthermore, these organizations served as the recipients of funding made on behalf of GA4GH. For example, a large portion of GA4GH core support was secured through supplements to larger grants such as The Clinical Genome Resource (ClinGen) and the Big Data to Knowledge (BD2K) program. At the same time, sponsorship funding from commercial companies supplemented these resources and provided flexibility in the organization's spending strategies.

Another element of the GA4GH funding model that sets it apart from other similar organizations is the very matrix structure upon which its development efforts are built. Driver projects and others could demonstrably allocate resources for hands-on technical development in the context of the GA4GH work streams, meaning that standards development—which is typically abstract and difficult to fund—becomes a concrete deliverable that projects and individual contributors alike can list on grant applications. This also means that grantees are accountable to the timelines set out by GA4GH roadmaps and they become dependent on the community to deliver on those expectations. This drives further success because it is the contributors themselves who make up that community.

In 2020, GA4GH made a pivotal step in its sustainability strategy by establishing a not-for-profit organization, GA4GH, Inc., in Canada. Prior to this point, GA4GH was an unincorporated entity entirely dependent on its host institutions. As a legal entity that can hold assets and enter into contracts, GA4GH Inc. enables the organization to develop a legal framework that lays out a robust approach to intellectual property, ensures freedom to operate, and maintains the ability to provide open standards. It also has the ability to enter into binding contractual agreements with other organizations, accept sponsorship funding from private

sector entities, and potentially disburse funds in pursuit of GA4GH's purpose, none of which were straightforward in the host institution context.

The not-for-profit came alongside several new funding agreements that would support GA4GH over the ensuing 5 years. Collectively, these agreements provided GA4GH with a $3M annual budget distributed across eight funding agencies in three countries. Each agreement was developed within the local funding schemes and came out of international discussions with the global collective, Heads of International Research Organizations.

Because it provides a legal entity that others can directly engage, GA4GH Inc. serves as a keystone for the organization that can better support the larger international structure of funders, Host Institutions, in-kind contributors, funded contributors, and core staff. An emerging strategy within this foundation is the concept of an "Assigned Expert." Assigned Experts are software engineers, bioinformaticians, and regulatory/ethics experts who have been assigned by a funder or employer to dedicate at least 30% of their job function to GA4GH standards development, working closely with the core technical staff to advance standards and policies over the finish line. The mechanism is being developed with the explicit intention of enabling lower-resource countries to have a significant voice in the international conversation.

Challenges in GA4GH's multifaceted funding approach arise when fully- and partially-funded contributors work alongside volunteers, which can lead to an imbalance if funded voices are allowed to dominate because they are better resourced and therefore can dedicate greater effort. The Assigned Experts model is one potential response to this challenge, by placing those individuals in close proximity to core activities.

The GA4GH secretariat is a professional staff team that is distributed across the four host institutions and supports the executive leadership, steering committee, and work streams, and increasingly participates in standards and policy development. Since the organization's inception, this lightweight group has been led by Peter Goodhand, founding Executive Director and now CEO. The team includes full-time dedicated staff who support project coordination and management and enable the broader work of a vast community of additional contributors outside of the core staff. In the early days of the organization, Working Group Coordinators were responsible for organizing activities and driving work forward, however, they did not have sufficient authority within the contributor teams to steer work toward institutional goals. In 2017 with the rollout of GA4GH Connect, this role was promoted to a management level and the individuals that held it (now called Work Stream Managers) became partners with the Work Stream Leads, rather than mere facilitators. With the procurement of sustainable funding through the Core Funders Forum, GA4GH was able to significantly expand the secretariat beginning in 2020 to include a Chief Standards Officer who oversees the technical team and Work Stream Managers, and a Director who oversees the operations and communications teams.

5.8 Three recommendations

From our experiences over GA4GH's evolution, we have identified three core principles that have been of particular value to our success, and which other organizations may find useful as they seek to enable interoperability and sharing across an international community. Below, we describe these activities and provide some details on each.

5.8.1 Community needs should drive development

From its earliest days, GA4GH has been driven by community needs. Research leaders recognized that broad alignment, interoperability, and collaboration would be necessary to fully realize the potential of huge volumes of genomic data to advance human health and medicine. In response to this need, the initial founders of GA4GH came together as a group to listen to one another and attempt to reach a better understanding of the needs and opportunities of their own international community. After launch, the GA4GH demonstration projects served as a powerful force to convince the wider world of the value of genomic data sharing, and as the needs of those individual projects evolved, GA4GH evolved as well. The Demonstration Projects broke off into distinct, self-sustaining efforts that now join a larger group of GA4GH driver projects that are collectively defining new use cases that drive standards development. These projects are only a small handful of initiatives among a global landscape of hundreds more, but they were selected to represent that community in both breadth and depth. As genomics further evolves toward a domain that is increasingly driven by healthcare, GA4GH will again need to respond to new needs and adapt its mechanisms for ingesting and responding to them.

While driven by broad needs in the areas of regulation and ethics, genomic and clinical data exchange, and privacy and security, the early GA4GH working groups were disconnected from one another. The needs of the clinical community did not sufficiently interface with the development efforts of the data working group. These groups served an important purpose in laying the foundation for GA4GH activities and for surfacing gaps in our process, and it was in direct response to these gaps that the GA4GH work stream and driver project matrix was born. Today, we strive to ensure that every activity within the organization is explicitly matched to a user need, and preferably one that is shared by a broad cross-section of the global community. This revision to our approach required large-scale operational bandwidth, but the value-added has justified that investment many times over.

5.8.2 Create global equity and opportunity to ensure fit-for-purpose development

Historically, genomics has been largely dominated by highly funded scientific powerhouses in the English-speaking world. While an international effort, the Human Genome Project (HGP) was led primarily by the United States and the UK, with funding from the National Institutes

of Health in the United States and the Medical Research Council and Wellcome Trust in the UK. However, genomics is relevant to all of us, and scientific discovery and application to healthcare will be maximized when it includes representation from all of humanity. Over the last twenty years, robust genomics efforts have been built in Estonia, South Africa, Brazil, Japan, India, and China, which represent opportunities to broaden participation and representation.

Building on the idea of letting community needs drive work, GA4GH believes that standards and tools must be developed with the widest possible stakeholder community at the table. As a direct descendant of the HGP, GA4GH has struggled to engage on a truly global level and we recognize that this limitation holds us back. In 2019, we began a deliberate effort to explore the challenges and opportunities in the areas of equity, diversity, and inclusion, with a particular interest in diversity metrics that may lie outside the standard norms. We recognize that many levels of diversity—including geographic, linguistic, ethnic, sector, and ancestry—are all important to achieving a suite of interoperable standards that are fit for purpose across the globe. What works in one country may not be effective in another, and only with truly interoperable standards will we be able to achieve cross-pollination between those communities. Standard development can serve to bring communities together, and therefore standards must be developed with broad participation from as many stakeholders as possible. With increased sustainable funding, GA4GH is now, perhaps belatedly, developing mechanisms to meet this critical need.

5.8.3 Strive for consensus and intentional decision-making

It is not easy to obtain consensus about difficult challenges in a field that is growing as rapidly as genomics, among stakeholders with widely disparate and sometimes competing priorities. Effective standards development requires consensus at multiple decision points, and only through the common implementation and broad adoption of standards by the community will interoperability of genomic data for research and healthcare be achieved. Standards development is, in many ways, as much of a social process as a technical effort.

First, there must be general agreement that a standard is necessary through the identification of common gaps and requirements. Stakeholders must then define the problem and set the scope. Choices also must be made regarding the approach, overall design, and technologies to be used, all of which are obvious explicit decision points that are required for standard development. Decisions can also be made implicitly, however, and because they can potentially impact the success of the standard it is important that decision points are recognized and decisions are made intentionally.

For example, within GA4GH at least two driver projects must commit to codeveloping each standard. This requires compromise regarding the problem to be addressed, scope of work, and technical approach, as mentioned above. However, because driver projects have a stated

objective to not only develop but also implement a standard, the line between a *standard* and an *implementation* can begin to blur. An implemented standard is a product that is designed to meet specific use cases within a defined organizational environment and electronic architecture. A standard is a specification that guides implementations and that can be adopted by other stakeholders in different environments, with different use cases, and with different system architectures. Within GA4GH, development teams must find a balance between the implementation that serves as a minimum viable product for a driver project, and a generalizable standard specification that supports future extensibility. Developing robust, enduring standards that gracefully evolve to support advancements in the domain is incredibly difficult and time-consuming, so it is often tempting to focus on the short-term, but setting the scope too narrowly and focusing on only immediate needs could require major, breaking changes to the standard in the future when it must support new data types and use cases.

5.9 Conclusion

GA4GH achieved its current status as the international standards community for genomics largely because of an ability to recognize not only our successes but also our failures, to see beyond the horizon, and to respond to evolving opportunities and shifting priorities. We have reflected on our process and structure at regular time points throughout our history, and have formalized this reflection process into a biannual gap analysis and strategic review. Rather than simply highlighting our achievements, however, these efforts are focused on identifying gaps and shortcomings and turning those findings into opportunities to advance and become a better organization that grows through iterative adaptation. This has not been by accident—the individuals that make up our community are passionate about scientific implementation that meets the needs of the broadest possible user community, and it is natural to apply the scientific review process to evaluate our own work. We care about standards because of the enormous benefit we know they will have on human health. Collectively, we have participated in thousands of hours of hair-splitting conversations about the representation of genetic variation and what information to include in the header of a message. To many, these discussions might feel excessively granular beyond the point of value, but GA4GH contributors will not rest until they land on an effective solution to a common need. It is no wonder we treat our organization with similar scrutiny and refuse to let it flounder under the pressure of history or design.

Acknowledgments

The authors would like to thank the current GA4GH executive committee, Ewan Birney, Peter Goodhand, Kathryn North, and Heidi Rehm, for their invaluable support and perspective. We also thank the broad GA4GH community for their collective contributions to the success of the organization and its efforts. RRF was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R35HG011899; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Chapman A, Wyndham J. Human rights. A human right to science. *Science*. 2013;340:1291. https://doi.org/10.1126/science.1233319.
2. OmicsXchange podcast episode 1: Science as a human right: an interview with dr. bartha maria knoppers n.d. https://www.ga4gh.org/news/omicsxchange-podcast-episode-1-science-as-a-human-right-an-interview-with-dr-bartha-maria-knoppers/ (accessed May 17, 2022).
3. Knoppers BM. Framework for responsible sharing of genomic and health-related data. *Hugo J*. 2014;8. doi:10.1186/s11568-014-0003-1.
4. OmicsXchange podcast episode 2: 7th anniversary of GA4GH: an interview with peter goodhand and ewan birney n.d. https://www.ga4gh.org/news/omicsxchange-podcast-episode-2/ (accessed May 17, 2022).
5. 2020-2021 roadmap part I n.d. https://www.ga4gh.org/how-we-work/2020-2021-roadmap/2020-2021-roadmap-part-i/ (accessed May 17, 2022).

CHAPTER 6

Clinical genomic data on FHIR®: Case studies in the development and adoption of the Genomics Reporting Implementation Guide

Robert R. Freimuth[a], Robert P. Milius[b], Mullai Murugan[c] and May Terry[d]

[a]Department of Artificial Intelligence and Informatics, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, United States [b]National Marrow Donor Program/Be The Match, IT, Minneapolis, MN, United States [c]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, United States [d]MITRE Corporation, Bedford, MA, United States

6.1 Background

6.1.1 Health Level 7 (HL7)

Health Level Seven International (HL7) is an American National Standards Institute (ANSI)-accredited standards developing organization (SDO) that is dedicated to developing standards to support the exchange of electronic health information, including that which supports clinical practice.[1] Members of HL7 hail from more than 50 countries and represent many different types of stakeholders, including healthcare providers, government agencies and policy groups, payers, vendors, and researchers. Since its inception in 1987, HL7 has developed numerous standards that are widely adopted within the United States and worldwide, including Fast Healthcare Interoperability Resources (FHIR®).

HL7 standards are developed by work groups that are responsible for defining, developing, and maintaining the standard specifications. Currently, HL7 supports more than 35 work groups, each of which "owns" a conceptual subdomain.[2] Examples of work groups include Clinical Decision Support, Clinical Genomics (CG), Orders & Observations, Patient Administration, Pharmacy, and Vocabulary. Some work groups, such as FHIR Infrastructure (FHIR-I) and Structured Documents, cut across conceptual domains.

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00009-2
Copyright © 2023 Elsevier Inc. All rights reserved.
For Robert Freimuth part of the contribution, Mayo retains copyright on the drawings, questionnaires, and survey instruments contained within the contribution.

The process of developing a standard is intentionally rigorous. HL7 defines five aspects of the standards development process [3]:

1. Foster consensus.
2. Ensure content is fit for purpose.
3. Ensure content is implementable.
4. Establish an appropriate implementer community.
5. Ensure ongoing maintenance of the standard.

Throughout the development process, work groups must follow a formal decision-making process to ensure consensus, openness, and balance of interest.[4] In addition, all candidate standards must undergo a formal balloting process through which feedback on the specification is gathered and reconciled prior to publication of the specification as a normative standard.[5] The processes that govern these activities conform to the requirements of the ANSI, which provides accreditation to HL7 as a SDO. The formality of the standards development processes ensures the resulting specifications were developed openly and fairly, and were vetted by a wide variety of stakeholders prior to publication. Those processes, however, impose a significant impact on the pace at which a standard can be developed. This includes any change to the specification, from minor corrections to major feature enhancements.

6.1.2 HL7 Clinical Genomics

The CG Work Group develops and maintains standards that enable the semantically meaningful exchange of clinical genomic data and family health history, thereby supporting precision medicine.[6] The exchange of those data, when linked to relevant clinical information, can help researchers identify and understand genomic factors that contribute to human disease and health. The membership of the CG Work Group includes stakeholders from a variety of organizations, including academic medical centers, genomic testing laboratories, genomic data repositories, and genomic knowledge bases.

The CG Work Group documents use cases for the clinical genomics domain, which are used to guide the development of standards that focus on clinical result reporting. In addition, because translational research has blurred the line between research and clinical results, the CG Work Group seeks to develop standards that not only support those developed by other HL7 work groups, but also data exchange standards developed outside of HL7. As a result, the CG Work Group attempts to both prospectively and retrospectively harmonize with standards developed by organizations such as the Global Alliance for Genomics and Health (GA4GH) and the Clinical Pharmacogenetics Implementation Consortium, and domain-based communities such as those supporting histocompatibility testing and molecular oncology.

The CG Work Group is currently developing the Genomics Reporting FHIR Implementation Guide (IG). The scope of the Genomics Reporting IG mirrors that of the CG Work Group itself and is intended to cover all aspects of human genomic reporting, including:

- All assay technologies (e.g., sequencing, array).
- All types of genomic variation, from simple variants to complex structural variants.
- Known and novel variations.
- Clinical interpretation and relevance of variations to disease pathology.
- Mosaicism and transplant scenarios.
- Mitochondrial genome and microbiome.

The scope of the Genomics Reporting IG is enormous, rapid advances in clinical sequencing technology create new use cases regularly, and the work group has a stated guiding principle to maximize computability by representing results as discrete, structured values. Given the combination of these factors and the fact that the base FHIR specification itself continues to evolve, it is not a surprise that the Genomics Reporting IG has undergone incremental improvements over time. The first draft of the Genomics Reporting IG (Standard for Trial Use 1, or STU1) was first balloted in April 2018. An update to that specification was balloted again in December 2018, and after feedback was addressed STU1 was published in November 2019. The Genomics Reporting IG was balloted again in May 2021, and STU2 of the IG was published in May 2022. Throughout that time, the CG Work Group continued to develop the IG by making structural changes to the specification and by clarifying ambiguities in response to use cases from stakeholders and feedback from early adopters.

6.2 Case studies: implementation of HL7 FHIR

The early adoption and testing of draft standards are an essential part of the standards development process. Implementing standards within a particular electronic environment and for specific use cases often requires a more detailed interpretation of the draft specification than is usually documented within the standard, and the choices that are made about how to represent individual data elements according to the specification can reveal gaps and limitations that were not previously evident. Feedback provided by early adopters is critical to identify and address issues in the draft specification, and therefore it is an important part of the maturation of a draft standard into a robust and generalizable specification.

The HL7 Clinical Genomics Reporting Implementation Guide has been implemented and tested by several key stakeholders throughout the development of the specification. The case studies described below summarize the experience of different organizations, each of which had a unique driving use case and engaged with the HL7 CG Work Group in a different way. The lessons learned from these early adopters provided critical feedback on the ability of the draft Implementation Guide to represent diverse genomic data types, its ability to be

implemented technically, and insight into ways communication and coordination can be improved between early adopters and the owner of a standard that must follow the specific decision making processes and procedures.

6.2.1 Exchanging HLA data for histocompatibility and immunogenetics

6.2.1.1 National Marrow Donor Program (NMDP)

The National Marrow Donor Program (NMDP) is a nonprofit organization dedicated to creating an opportunity for all patients to receive the hematopoietic cell (blood stem cell) transplant they need, when they need it. NMDP operates the Be The Match Registry®, the largest and most diverse donor registry in the world. Our partnerships with international and cooperative registries provide doctors with access to more than 39 million potential donors and more than 800,000 cord blood units worldwide.

NMDP is the central hub of a network of over 500 partners, including donor centers, collection centers, apheresis centers, transplant centers, testing labs, recruitment groups, member cord blood banks, international donor centers and cooperative registries, with immunogenetic data being exchanged between partners. Similarly, the Center for International Blood and Marrow Transplant Research (CIBMTR), a collaboration between NMDP and the Medical College of Wisconsin, collects transplant data including biomarker data for outcomes research.

6.2.1.2 Human leukocyte antigen (HLA)

Human leukocyte antigen (HLA) is the most important biomarker for predicting outcomes of hematopoietic cell transplants, and therefore it is critical to represent and exchange HLA data accurately. Each year, NMDP receives HLA genotypings for over 500,000 new potential donors for the Be The Match Registry.

HLA is one of the most highly variable gene systems in the human genome. Current donor/patient matching strategies target diploid pairs of five of the most polymorphic HLA genes, each with thousands of possible alleles, with hundreds of new alleles being discovered every year. The process is complex but largely involves matching HLA names for each allele, which are represented using a nomenclature system that links to reference sequences for each allele. As more HLA alleles are being discovered, it has become clear that the existing nomenclature system does not capture all the nuances of the alleles and that exchanging the actual molecular sequence data of donor and patient HLA loci is important to understand and predict transplant outcomes.

6.2.1.3 HLA reporting

With the advent of next-generation sequence technologies, the HLA research community recognized the importance of reporting molecular data in addition to reporting the asserted

allele names based on the conventional nomenclature system. This recognition resulted in a series of community-driven HLA Data Standard Hackathons (DaSH) that were designed to develop a standard for exchanging HLA sequence data, which produced at least three important products:

1. **Minimum Information for Reporting Immunogenomic next-generation sequencing (NGS) Genotyping (MIRING)**: MIRING outlined a set of principles for reporting HLA typing results. The minimum information required for reporting HLA results includes the assigned allele names (based on the existing nomenclature system), but it also encourages reporting molecular sequence data, detailed description of novel polymorphisms identified, and the test methodology. [7] The principles outlined by MIRING represented an important first step in the identification of critical data elements, as determined by experts in the domain.
2. **Histoimmunogenetics Markup Language (HML)**: An XML-based implementation of the principles of MIRING.[8] The definition of the HML specification was a necessary step toward the technical implementation of MIRING, and it formed a foundation that informed the future development of the HLA Reporting FHIR Implementation Guide.
3. **Genotype List String (GL String)**: The GL String is a grammar that describes the ambiguity that exists when HLA sequence data are interpreted into alleles and genotypes, a process that exposes limitations in sequencing technology when applied to a highly variable region like the HLA locus. The GL String format uses a hierarchical set of operators to describe the relationships between alleles, lists of possible alleles, phased alleles, genotypes, lists of possible genotypes, and multilocus unphased genotypes, without losing typing information or increasing typing ambiguity. [9] Therefore, GL String captures information about the *interpretation* of complex sequence data that is normally not represented in typical genomic results.

While HML works well to represent sequence data related to HLA, it is not used outside of the HLA community. Furthermore, the ability to integrate HML reports into clinical information systems, including electronic medical records (EMRs), is not widely supported even in healthcare systems that have or are affiliated with transplant centers. This critical gap led us to explore HL7 FHIR, a modern standard that was being adopted rapidly and widely throughout the healthcare community, as a way for exchanging HLA data using a technology that was native to a wide array of clinical systems.

6.2.1.4 HLA Reporting FHIR Implementation Guide

NMDP participated in Phases 1 through 3 of a program sponsored by the Office of the National Coordinator (ONC) called Sync for Genes. The overall goal of NMDP's participation was to develop an FHIR Implementation Guide that supports the complex data in HLA genotyping reports and that follows the principles of MIRING.

The initial steps toward developing an FHIR IG for HLA required mapping the data elements defined in HML to elements in the FHIR STU3 specification and the early work of the HL7 CG Work Group. The results of that analysis were a strategy for reporting HLA genotypes using FHIR and the subsequent development of a software tool to perform the data transformation.

During that work, the FHIR base specification matured to the R4 release, and the CG Work Group developed an early prototype of the Genomics Reporting Implementation Guide. Therefore, effort was made to update the tooling that converted HML to FHIR so that it aligned with current versions of the FHIR standard. The development of that tooling ultimately allowed our partner transplant centers to send HLA reports to NMDP in a familiar format (HML), and for NMDP to develop local infrastructure to consume HL7 FHIR data.

The achievements made by NMDP's participation in Phases 1 and 2 of the Sync for Genes program represented important steps in demonstrating the feasibility of using FHIR for HLA reporting, but implementation was limited to NMDP internal systems. In order to facilitate the adoption of FHIR for HLA reporting by other organizations, the development of a formal FHIR IG was necessary. Therefore, in Sync for Genes Phase 3, NMDP focused on two goals:

1. Develop the HLA Reporting FHIR Implementation Guide by leveraging the recently published HL7 FHIR Genomics Reporting Implementation Guide (STU1), which would be constrained by NMDP/CIBMTR rules for creating FHIR bundles and which would be designed to support reporting for HLA genotyping.
2. Use the HLA Reporting FHIR Implementation Guide to inform the development of a tool that could be used by an HLA testing lab, enabling the lab to create a FHIR-based HLA genotyping report directly from commonly used HLA analysis software without having to go through an HML intermediate.

While the Genomics Reporting Implementation Guide (STU1) served as a useful base, the development of the HLA Reporting FHIR Implementation Guide (http://fhir.nmdp.org/ig/hla-reporting) required significant work to tailor the specification to HLA data types, including the generation of several profiles, extensions, value sets, code systems, and examples of use. Core components of the HLA Reporting FHIR Implementation Guide that had to be created are listed in the table below.

FHIR Entity	Artifact
Profile	HLA Summary Report
Profile	HLA Genotype Observation
Profile	Allele Observation
Profile	Molecular Sequence
Extension	HLA Genotype Summary
Value Set	HLA Gene Name
Code System	HLA HGNC GeneID
Code System	Genotype List (GL) String Code

It is not surprising that the development of a domain-specific FHIR IG would require custom artifacts such as those listed above. While some of the new artifacts, such as the profile for allele observation or an expanded code system for HGNC Gene ID, would potentially be reusable in other contexts, others were highly specific to the HLA domain. Regardless, the development of these artifacts exercised the ability to use FHIR to represent HLA data, either natively or through extensions, and we identified several limitations that required coordination with the appropriate committee within HL7 and/or workarounds. For example, we discovered that the GL String format could not be easily represented in FHIR without loss of the code system information of the HLA nomenclature. To address this, we developed the GL String Code, a grammar that described a gene family namespace, a version of the nomenclature, and the GL String itself. The GL String Code (http://glstring.org) was implemented as a Code System in the HLA Reporting FHIR Implementation Guide.

The second goal was the development of a tool that could be used by an HLA testing lab for reporting genotype results in FHIR format. To develop this prototype, NMDP partnered with an HLA typing lab (Versiti, https://www.versiti.org) that used a popular HLA analysis software package (GenDx NGSEngine®). The software produced an XML-formatted output file (TARR) that contained detailed HLA test results, including full genotyping data, molecular sequence information, and GL Strings that represented any allelic and genotypic ambiguity. In this phase of the project, we developed a tool that converted the XML file to a FHIR bundle that conformed to the HLA Reporting FHIR Implementation Guide. The FHIR transaction Bundle contained instances of several FHIR resources and profiles that together represented the HLA test results:

- A single HLA Summary (an FHIR profile based on Genomics Report) for a single subject.
- HLA Genotype (an FHIR profile based on Genotype Observation), one per HLA gene.
- HLA Allele (an FHIR profile based on Haplotype Observation), one per allele.
- HLA Molecular Sequence (an FHIR profile based on Molecular Sequence), one per sequence reported.
- A single Provenance resource (to describe the input file(s) and software used to generate data).
- A single Device resource (to describe the conversion software and version).

The tool that converted the TARR file to FHIR (TARR2FHIR, available at https://github.com/nmdp-bioinformatics/tarr2fhir) provided a web-based graphical user interface as well as a REST API that could be invoked using tools such as Postman, cURL, or a programming/scripting language.

The implementation and testing of the TARR2FHIR tool identified several challenges. For example, while the HLA Reporting FHIR Implementation Guide was designed to support the expression of novel genomic variants, the TARR2FHIR tool could not support reporting novel variants because variants were defined in the TARR file through coordinates based on proprietary multi-sequence alignment of a set of HLA alleles. We also found that the original

TARR file lacked some of the metadata that was needed for a complete interpretation of the reported results, including:

- The nature of the subject being tested, which could be a donor, recipient, or a cord blood unit. For outcomes research, recipients also must be associated with a particular donor or cord blood unit.
- The namespace of the sample ID, which could have origins in a number of systems. For example, we need to know if the sample ID in the TARR file is a CIBMTR Research ID (CRID), a global donor registry ID (GRID), or from some other defined namespace and repository.
- The reporting organization. The TARR file does not contain any information about the organization sending this data, or on whose behalf.

It is not surprising that the TARR file lacks some of these data because the HLA analysis software used to generate the file is not intended to be a full a laboratory information management system. This is an example of a gap that is not due to a limitation in either the standard specification or the system generating the source data, and must be addressed through the use of external systems that can augment the tooling.

These two challenges, the inability to express novel variants and gaps in clinically important metadata, were not directly related to limitations in FHIR and instead were examples of the type of issues that must be addressed by clinical software and systems that may not have been designed to handle highly detailed and nuanced genomic data. These findings also underscore the importance of implementation and testing standards by early adopters.

6.2.1.5 Lessons learned and conclusions

This case study serves as an example of the successful adoption and pilot testing of an emerging standard to support a specialized domain. There were three important perspectives represented in this case study. First, the NMDP was already deeply engaged with the HL7 CG Work Group and team members had considerable knowledge of and technical experience with the FHIR standard prior to the initiation of this project; thus, NMDP represented a stakeholder that was an "insider" that functioned as both a developer and adopter of the standard. Second, this case study illustrates the process by which a domain-specific, community-developed standard can be used to seed the development of a more formal standard that was based on FHIR, thereby reducing barriers to implementation by vendors of clinical software. Lastly, this project demonstrated the importance of engaging a testing laboratory, which represented an industry point of view, in the process of standard implementation and pilot testing.

There were several lessons learned from this project:

- Constraints facilitate the mapping between two specifications. It is difficult to create tooling for converting between data formats that are not strictly defined.

- Implementation guides and profiling are critical to validation. HL7 FHIR validation tooling is invaluable during development and implementation, providing assurance that the implemented software conforms to the standard specification.
- Engaging with the community is invaluable. Open dialog between the NMDP implementation team and the owners of the FHIR standard enabled the team to verify understanding, to discuss approaches to address gaps, and to provide feedback on the standard specification.
- The Genomics Reporting Implementation Guide was very complex and early discussions with the HL7 CG Work Group helped to lower the learning curve even among those that already had extensive knowledge of FHIR.
- The community-driven hackathons, such as DaSH and the HL7 FHIR Connectathons, helped to vet the standards and tools that were developed by this project. For example, during a DaSH event, we discovered our early implementation of GL String as an FHIR observation was fundamentally wrong. The GL String code was developed as a solution after direct dialog between experts in the HL7 FHIR community and the owners of the HLA nomenclature.

Ultimately, we believe that these standards-based endeavors will make it possible for us to achieve our vision of exchanging patient/donor immunogenetic data with consent directly with EMRs, typing labs, and other healthcare and research systems. The case study described above is an important step toward that goal.

6.2.2 Electronic medical records and genomics (eMERGE) network

6.2.2.1 Background

In September 2007, the National Human Genome Research Institute of the National Institutes of Health (NIH) launched the Electronic Medical Records and Genomics (eMERGE) Network. The eMERGE network is a consortium of medical research institutions within the United States that are working to advance genomic discovery and genomic medicine implementation.[10] A primary focus of the eMERGE network is the integration of genomic data into clinical systems, including electronic health records (EHRs). As such, each phase of the eMERGE network includes an evaluation of opportunities to deliver genomic results through EHRs to providers, thereby making those data available for consideration in clinical decision-making.

6.2.2.2 eMERGE III

Phase III of the eMERGE network began in September 2015 and completed in March 2020. Phase III included 11 study sites, a coordinating center, and two centralized sequencing and genotyping facilities (CSGs): Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) and the Broad Institute (BI)/Partners Laboratory for Molecular Medicine

(LMM). Each CSG sequenced about half of the specimens from study participants and generated clinical reports. The results were utilized for research and discovery, and the clinical reports were disseminated to the study sites for delivery to providers through clinical decision support and to return results to participants.

The eMERGE network had a federated structure: each study site had unique clinical systems and policies, and the CSGs used different laboratory systems and report-generating pipelines. Nonetheless, the network had to share and reliably integrate data from each of the multiple sources, and therefore the need for a common structured representation of genetic test results was identified as a key deliverable for the Network.

The timeline of the eMERGE III study was very tight, and the network did not have time to evaluate, implement, and test new technologies for exchanging genomic results. To meet study deadlines and ensure that the electronic infrastructure that was needed for communicating clinical results from the CSGs to study sites was implemented in time, the network decided to use a custom XML message format to represent genomic results. The CSGs generated both PDF clinical reports and XML files containing structured data, which were returned to the study sites for integration into clinical systems and workflows.

Recognizing that the custom XML format was not a scalable solution, and as an effort to support the advancement of standards for genomic data, the network undertook a parallel activity. At the time, HL7 FHIR was rapidly emerging as the next generation of standards for healthcare, and the CG Work Group (CG WG) was developing an FHIR Genomics Reporting Implementation Guide (IG).[11] Although the Genomics Reporting IG was not yet finalized and would not be implementable within production clinical systems, including the EHRs at study sites, the Network formulated an objective to evaluate the nascent Genomics Reporting IG with the intent of advancing the standard and contributing to its further development.

6.2.2.3 Evaluation of the FHIR Genomics Reporting Implementation Guide

The evaluation of the FHIR Genomics Reporting IG included two components: a formal analysis of the Genomics Reporting IG and the development of a specification to support the eMERGE study, and a pilot implementation designed to validate the eMERGE FHIR specification and demonstrate its potential utility within clinical systems. The participating organizations included the two CSGs, which jointly developed the eMERGE FHIR Specification, as well as study site Northwestern University (NU) and nonclinical affiliate Johns Hopkins University (JHU), which implemented and validated the eMERGE FHIR specification.[12] To demonstrate data transfer, the BCM-HGSC, in its role as a diagnostic laboratory, generated a sample set of clinical reports using the eMERGE FHIR specification, which were sent to NU or JHU (as appropriate, depending on the participant) for EHR integration and clinical decision support (see Fig. 6.1).

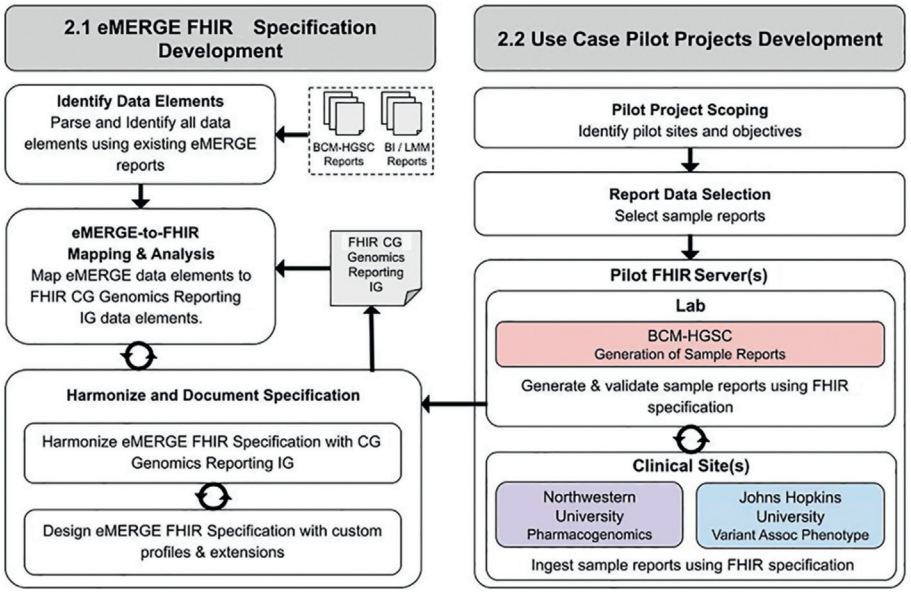

Figure 6.1
eMERGE FHIR specification development and implementation use cases.[12]
Reprinted from "Genomic considerations for FHIR; eMERGE implementation lessons", *J Biomed Inform*. 2021 Jun;118:103795, with permission from Elsevier.

The principal goals of this pilot project were:

- Evaluate the ability of the FHIR Genomics Reporting IG to support eMERGE clinical genomic reporting use cases, including a detailed gap analysis.
- Create an FHIR specification, based on the Genomics Reporting IG and supporting eMERGE reporting requirements, for use in pilot implementation and testing.
- Provide feedback from the evaluation and pilot implementation to the CG WG to help advance the maturity of the Genomics Reporting IG.

A detailed description of the methods and results of this study has been published.[12] In summary, the evaluation of the FHIR Genomics Reporting IG began with a thorough review of the current version of the IG, which involved reading documentation and engaging with the CG WG to clarify ambiguities and verify the accuracy of understanding. Next, each data element in the eMERGE narrative reports (in PDF format) produced by the two CSGs was mapped to a corresponding element in FHIR, performed in close partnership with members of the CG WG. That process identified a diverse set of structural and content variations, which were documented in detail as a gap analysis between the data contained in the eMERGE clinical genomic reports and the data elements supported by the FHIR Genomics Reporting IG. The results of the evaluation and gap analysis enabled the development of an FHIR specification,

based on the Genomics Reporting IG, designed specifically to support the eMERGE III clinical genomic reports. The eMERGE FHIR specification was successfully used by BCM-HGSC, NU, and JHU as a core part of the pilot implementations.

The steady progress and direction of the project were largely dependent on engagement with members of the CG WG and experts from the larger FHIR community. Communication was supported by several channels, including informal exchanges through the FHIR Zulip Chat Board,[13] which enabled rapid responses to simple questions. Complex questions were addressed through live discussion with members of the CG WG during regularly scheduled meetings, where decisions were captured as part of the meeting minutes. Technical changes and enhancements to the Genomics Reporting IG were formally logged in the HL7 issue tracking system. As part of its formal decision-making process, the CG WG regularly triages logged issues, discusses and debates potential approaches to address them, votes on proposed resolutions, and finally implements approved changes within the specification. This project resulted in the identification of more than 20 significant issues, most of which were resolved during the project period, as documented on the eMERGE FHIR specification website.[14] Feedback from this project to the CG WG resulted in several refinements to the Genomics Reporting IG.

To aid future implementers, the network published its experiences and lessons learned through a manuscript[12] and on the eMERGE FHIR specification website. The design and development of the eMERGE FHIR specification is documented at https://emerge-fhir-spec.readthedocs.io/. Sample reports are available on GitHub at https://github.com/emerge-ehri/fhir-specification. Pilot implementation code is available at https://github.com/emerge-ehri.

6.2.2.4 Lessons learned and conclusions

This case study is an example of a rare but highly beneficial interaction between a standards development organization and representatives of a large consortia, which included 11 study sites and two clinical laboratories, each of which had a unique clinical electronic environment and system architecture. Interactions with a pre-established *network* of both data senders and receivers, which had slightly different requirements but which were centered around a common use case, encouraged a deeper, and more focused evaluation of the specification. In addition, because the members of the consortia were relative newcomers to FHIR and engaged with the CG WG from the perspective of standard implementers and adopters, the evaluation included both technical and non-technical (e.g., phrasing, definitions) aspects of the specification. The experience of the eMERGE team underscored the need for robust, comprehensive documentation that includes examples that illustrate the usage of complex specifications like the Genomics Reporting IG. As of this writing, the CG WG has systematically and methodically responded to this type of feedback, which is critical not only for the development of the standard but also to ensure its accurate implementation, which is necessary to achieve interoperability.

The importance of direct engagement and collaboration between the eMERGE team and members of the CG WG and larger FHIR community cannot be overstated. The timely progression and successful completion of the project was due in part to the ability to ask detailed questions about the intent and use of the FHIR specification and receive authoritative feedback relatively quickly. This was especially important since the iterative harmonization process and ultimate implementation required that the members of the eMERGE team, who already had a comprehensive understanding of genomic testing data and clinical reports, had to rapidly gain a working knowledge of the HL7 FHIR specification and the Genomics Reporting IG in order to function as an early adopter during the eMERGE funding period.

Timelines impacted implementation decisions in other ways as well. In particular, the Genomics Reporting IG was classified as trial use[15] and was under active development, which meant the specification changed substantially throughout the project period, but the eMERGE team could not wait for the Genomics Reporting IG to stabilize before performing its evaluation. Similarly, the eMERGE team had to implement unofficial extensions or workarounds for features that could not be resolved by the CG WG in time for the eMERGE pilot implementation.[16] It is not uncommon for implementation/adoption timelines to be incompatible with the formal process of standards development, where issue resolution, balloting, and release cycles are often much longer than those for implementation projects.

Despite these challenges and others, this project and the Network's participation in the CG WG emphasizes the importance for early adopters from cross-platform groups, including research consortia programs, medical institutions, diagnostic laboratories, and EHR system vendors, to engage in the standards development process to help advance the maturity of draft specifications and prepare them for mainstream use.

6.2.3 Minimum common oncology data elements (mCODE)

6.2.3.1 Background

According to the National Cancer Institute, 38.5% of men and women will be diagnosed with cancer at some point during their lifetimes. In 2018, there were 18.1 million new cases and 9.5 million cancer-related deaths worldwide.[17] While these numbers are staggering, advances in cancer genomics research have accelerated our understanding of molecular pathways and subsequent biological processes in the genesis of certain cancers. This has led to significant improvements in how we diagnose, treat, and monitor the cancer patient's journey.

Research-quality data from all cancer patients would enable higher quality health outcomes; however, there exists a lack of consensus on oncology data models, technologies, and methods to capture that data. Subsequently, lack of high-quality structured and coded impeded the leveraging of real-world data to inform clinical decisions in cancer diagnosis, treatment, and monitoring.[18] To address this problem, the American Society of Clinical Oncology in

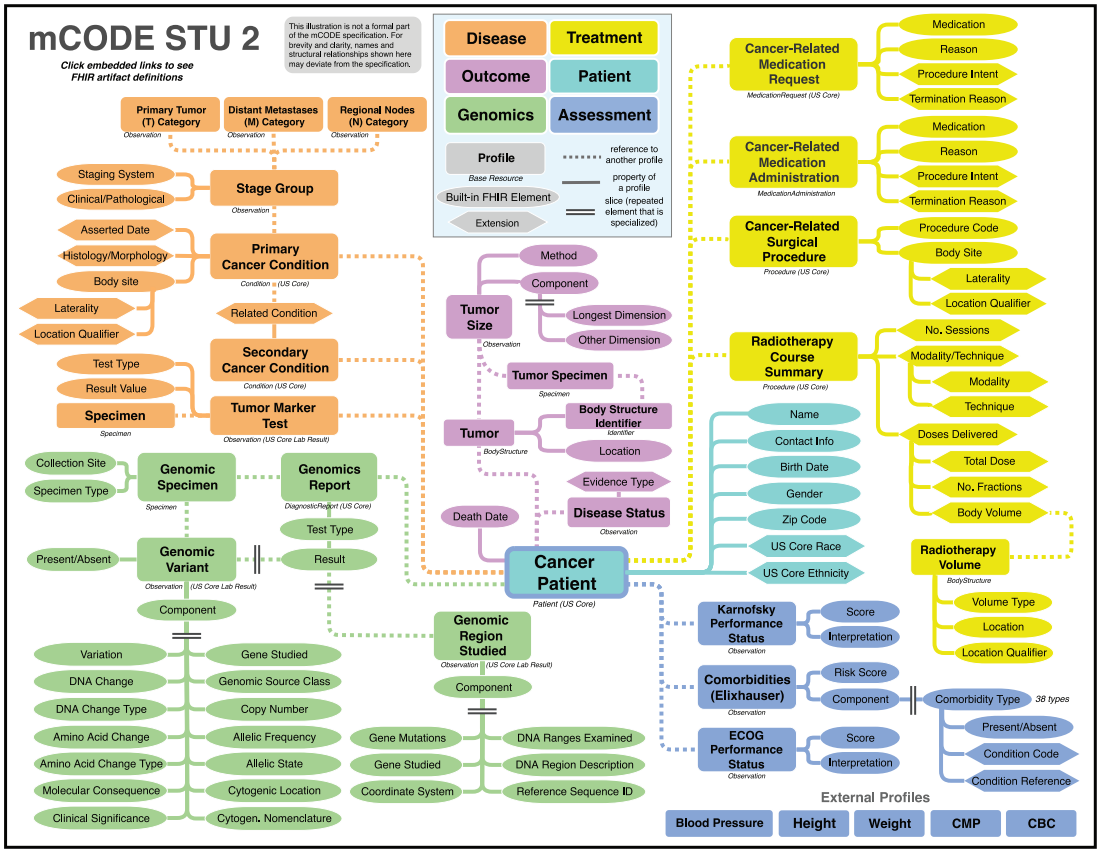

Figure 6.2
mCODE conceptual model.

collaboration with the MITRE Corporation, established the Minimal Common Oncology Data Elements (mCODE) initiative, intended to assemble a core set of structured data elements for oncology EHRs. mCODE is a step toward capturing research-quality data from the treatment of all cancer patients by allowing for easier methods of data exchange between health systems.

6.2.3.2 Minimum Common Oncology Data Elements (mCODE) and FHIR

The mCODE information model was developed through a multidisciplinary community that included oncologists, informaticians, researchers, and experts in terminologies and standards.[19] Two initial use cases drove the initial scope and information model: (1) comparative effective analysis and support for cooperative and shared provider-patient decision making, and (2) comparative effective analysis with next-generation sequencing.

The mCODE conceptual model (Fig. 6.2) depicts data elements organized into six domains that minimally and actionably represent the cancer patient journey: patient demographics and

overall status, disease characterization, the assessment of cancer-relevant comorbidities and performance statuses, and medication, surgical, and radiation therapy treatments. Genomics reporting elements were categorized into its own domain because of the complexity of cancer genomics data. The information model leveraged the standards development work from the HL7 CG Working Group (CG WG) by aligning with HL7 Genomics Reporting FHIR Implementation Guide (IG) where possible. The mCODE information model adopted a subset of the data elements from the HL7 Genomics Reporting IG to maintain mCODE's intent to preserve a minimal and actionable scope. The mCODE information model was ultimately used to develop the mCODE FHIR IG.

Overall, the FHIR IG developed by mCODE was successfully aligned to the Genomics Reporting FHIR IG developed by the HL7 CG WG. There were challenges in achieving complete harmonization, however, primarily in defining the scope of a minimal set of genomics elements. In this regard, the HL7 Genomics Reporting IG aims to be a comprehensive framework that supports structured representations for genomics data, including any primary and supporting data that could be included in a clinical report. The mCODE community, however, emphasizes the definition of a minimal set of data elements and is focused on including just enough minimally actionable data to enable the discussion and collaborative decision-making between the cancer patient and their managing provider.

The mutual relationship between the mCODE community and the HL7 CG WG resulted in changes within both IGs. Guidance from the CG WG resulted in improvements to how mCODE represented functional annotation and clinical significance. Conversely, mCODE influenced changes made to the Genomics Reporting IG, which initially contained multiple "must-support" elements that mCODE's minimal use case did not support; those constraints were subsequently relaxed for the STU2 release of the Genomics Reporting IG. Other areas of collaboration between the groups included the need for a more simple representation of tumor markers and a representation of fusion genes.

The beneficial collaboration between the mCODE community and the HL7 CG WG bridged multidisciplinary communities of both research and provider organizations that jointly collaborated to develop a representation of genomics data that could be implemented by both provider and research-based organizations. The interaction between the HL7 CG WG and other genomics organizations, such as the Global Alliance for Genomics and Health (GA4GH) and federally funded initiatives like the eMERGE network (NIH) and the Sync for Genes program (Office of the National Coordinator (ONC) for Health Information Technology), also helped to expand the potential scope of alignment with other efforts.

6.2.3.3 Future directions: broadening harmonization

The mCODE community is continuing to advance the standardization of genomic data exchange in multiple ways. CodeX (Common Oncology Data Elements eXtensions) is a member-driven HL7 FHIR Accelerator, building a community to accelerate interoperable data

modeling and applications leading to step-change improvements in cancer patient care and research.[20] CodeX members further promote cancer data interoperability and build upon the use of mCODE's minimal and actional elements by applying them to new use cases driven by existing and new applications. The genomics elements in mCODE are being considered to support CodeX use cases such as clinical trial matching. At the federal level, a core set of genomics elements for a minimal variant representation is included in version 3 of the United States Core Data for Interoperability (USCDI), the set of health data classes and constituent data elements developed by the ONC for nationwide, interoperable health information exchange.[21]

Finally, with the eventual vision of providing a standards-based approach to building an oncology learning health system that enables a synergistic exchange between genomics research data and clinical decision support, mCODE has been noted as an oncology use case in the HL7 and Observational Health Data Sciences and Informatics (OHDSI) partnership to align the representation of genomic variants in the FHIR and Observational Medical Outcomes Partnership (OMOP) common data model. The initiative, while still early in development, may further support research in genotype-to-phenotype discovery and pharmacogenomics.

6.2.3.4 Lessons learned

The development of the mCODE minimal and actionable FHIR-based model has been successful, with over 40 organizations that have implemented parts of the mCODE specification since its first publication in March 2018. There is great interest in the use of mCODE genomic elements given their relevance for supporting precision medicine at the point of care, but they are still in the early stages of adoption.

One reason that the genomic elements have experienced slower adoption is that genomic data are still largely sent from genomics reference labs to EHRs as unstructured documents, mostly in PDF format, rather than as discrete, structured, and standardized data. Another reason is that stakeholder requirements vary as to what is considered "minimum and necessary" genomic data to ensure clinical validity and actionability. For example, cancer research requires extensive detail about the processing, annotation, and interpretation that led to the results that are ultimately sent to a provider. On the other hand, most oncologists value this information but do not necessarily need it available within the EHR. Obtaining consensus regarding the identification and prioritization of genomic data elements to include in standard specifications, therefore, is an ongoing and challenging process.

It is important to note that mCODE is not instantiated as a single document containing patient information, but as an aggregation of important data elements captured as a snapshot in time during the course of a cancer patient's journey. CG is an important aspect of this journey, especially for cancer types for which a genomic test is indicated. An example of this is illustrated in Fig. 6.3, which shows the mCODE Genomics Report and TumorMarker

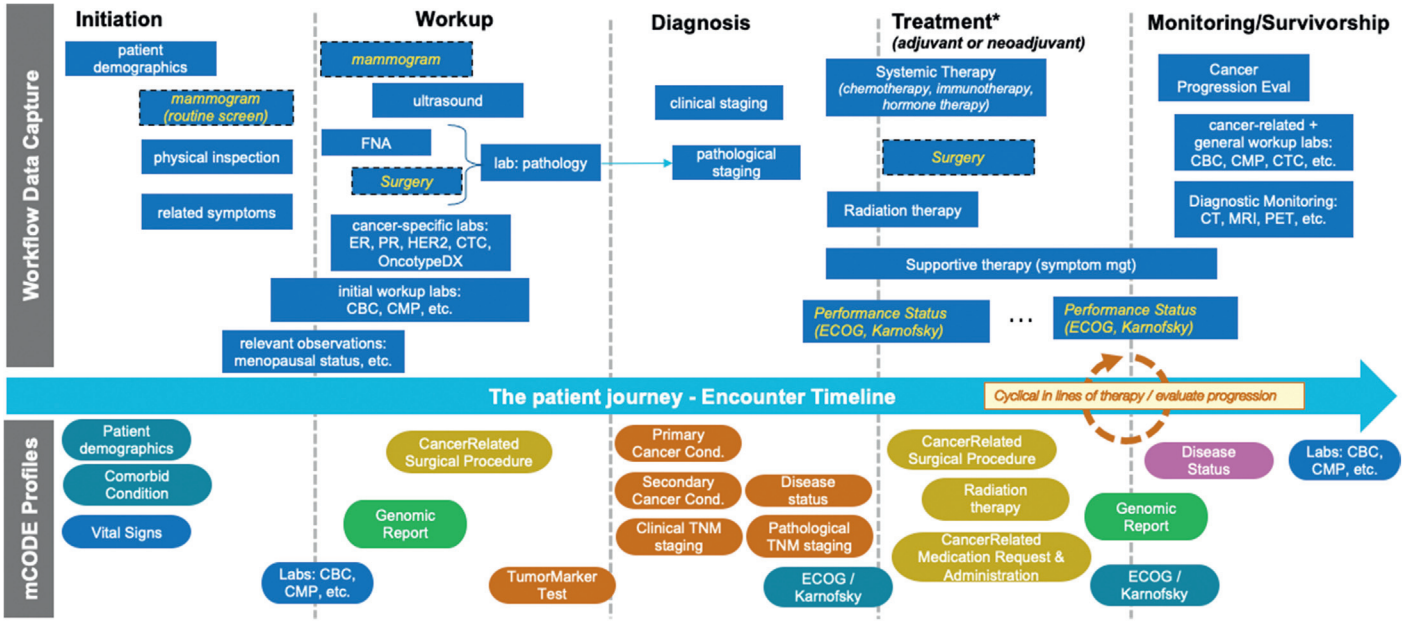

Figure 6.3
Example of mCODE snapshots throughout a cancer patient journey.

Test FHIR profiles, which could be included in an mCODE-conformant FHIR bundle in the Workup stage of the patient journey. That information could be collected and reported again during the monitoring phase, or following the completion of a line of therapy to support the effectiveness of a targeted therapy drug or whether the tumor itself has mutated. Thus, the ability to assemble an arbitrary set of data that is relevant at any given point in time means that the structure of the underlying standards must support a larger degree of optionality in how they are implemented.

While the initial scope for mCODE emphasized decision-making at the point of care, it should be noted that the applicability of genomic data could expand as the mCODE community incorporates care teams that include a greater portion of the cancer patient journey. For example, molecular pathologists and a molecular tumor board would have a need to further explore the clinical validity associated with the genomics test, and the detailed evidence that pertains to particular genetic changes. These use cases are potential future areas of consideration for the CodeX communities.

In spite of the challenges outlined above, and to achieve broader harmonization, the mCODE community continues to work with standards development organizations such as the HL7 CG Work Group, CodeX FHIR Accelerator, and OHDSI communities. The collaboration among these communities is invaluable in developing consensus and adoption for genomic data exchange.

6.3 Conclusion

The development of the Genomics Reporting FHIR Implementation Guide is an example of how difficult it is to develop rigorous standards that support the exchange of complex data in a rapidly evolving domain, and the use of those data for clinical care puts extra emphasis on the importance of semantic accuracy rather than simply syntactic data representation. To maintain progress, effort must be focused on high-priority use cases presented by champions that remain engaged through the standards development process.

It is extremely important for stakeholders to participate in the process of standards development by contributing use cases and examples that guide development, by reviewing draft specifications and providing feedback, and by testing the standard through trial implementations. It is through these processes that a draft standard can be refined into a more mature specification that promotes interoperability.

The three case studies presented in this chapter illustrate how early adopters, each of which had unique use cases, requirements, and desired outcomes, engaged with the CG Work Group and contributed feedback that advanced the development of the Genomics Reporting FHIR Implementation Guide. In parallel, that direct engagement provided significant benefit to the early adopters through education, validation of the implementation of the draft specification,

and changes to the specification. The case studies described above can serve as models for other potential adopters and may encourage them to become early contributors to the standards development process.

Acknowledgments

The authors thank Kevin Power and Luke Rasmussen for their assistance in proofreading the content of this chapter. RRF was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R35HG011899; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Approved for Public Release, Distribution Unlimited 22-3633. ©2022 The MITRE Corporation. All Rights Reserved. May Terry's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

References

1. Health level seven international - Homepage | HL7 international [Internet]. (cited May 18, 2022): https://www.hl7.org/index.cfm.
2. HL7 work groups & projects - HL7 - confluence [Internet]. (cited 2022 May 18). https://confluence.hl7.org/pages/viewpage.action?pageId=4489802.
3. Understanding the standards process - HL7 - confluence [Internet]. (cited May 18, 2022): https://confluence.hl7.org/display/HL7/Understanding+the+Standards+Process.
4. HL7 - Decision Making practices documents | HL7 international [Internet]. (cited May 18, 2022): http://www.hl7.org/participate/decisionmaking.cfm.
5. HL7 balloting - HL7 - confluence [Internet]. (cited May 18, 2022) : https://confluence.hl7.org/display/HL7/HL7+Balloting.
6. WorkGroup home - Clinical Genomics - Confluence [internet]. (cited May 19, 2022): https://confluence.hl7.org/display/CGW/WorkGroup+Home.
7. Mack SJ, Milius RP, Gifford BD, et al. Minimum information for reporting next generation sequence genotyping (MIRING): guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Hum Immunol*. 2015;76:954–962.
8. Milius RP, Heuer M, Valiga D, et al. Histoimmunogenetics markup language 1.0: reporting next generation sequencing-based HLA and KIR genotyping. *Hum Immunol*. 2015;76:963–974.
9. Milius RP, Mack SJ, Hollenbach JA, et al. Genotype list string: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens*. 2013;82(2):106–112.
10. Electronic medical records and genomics (eMERGE) network [Internet]. (cited 2022 May 9, 2022): https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE.
11. Genomics reporting implementation guide [Internet]. (cited May 9, 2022): http://hl7.org/fhir/uv/genomics-reporting/index.html.
12. Murugan M, Babb LJ, Overby Taylor C, et al. Genomic considerations for FHIR®; eMERGE implementation lessons. *J Biomed Inform*. 2021;118.
13. (242) recent topics - FHIR Community - Zulip [internet]. [cited May 9, 2022): https://chat.fhir.org/#narrow/stream/189875-genomics-.2F.20eMerge.20Pilot.
14. Issues & resolutions — emerge-fhir-spec 1.0 documentation [Internet]. (cited May 9, 2022): https://emerge-fhir-spec.readthedocs.io/en/1.0/issues_and_resolutions.html.
15. Versions - FHIR v4.0.1 [Internet]. (cited May 9, 2022): https://www.hl7.org/fhir/versions.html.
16. Discussion — emerge-fhir-spec 1.0 documentation [Internet]. (cited May 9, 2022): https://emerge-fhir-spec.readthedocs.io/en/1.0/discussion.html.

17. Cancer statistics - NCI [internet]. (cited May 9, 2022): https://www.cancer.gov/about-cancer/understanding/statistics.
18. Bertagnolli MM, Anderson B, Norsworthy K, et al. Status update on data required to build a learning health system. *J Clin Oncol*. 2020;38:1602–1607.
19. Osterman TJ, Terry M, Miller RS. Improving cancer data interoperability: the promise of the minimal common oncology data elements (mcode) initiative. *JCO Clin Cancer Inform*. 2020;4:993–1001.
20. CodeX | HL7 international [Internet]. (cited May 9, 2022): https://www.hl7.org/codex/.
21. United states core data for interoperability (USCDI) | interoperability standards advisory (ISA) [Internet]. (cited May 9, 2022): https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi.

CHAPTER 7

Genomics data sharing

Judit Kumuthini[a], Lyndon Zass[b], Melek Chaouch[c], Faisal M. Fadlelmola[d], Nicola Mulder[b], Fouzia Radouani[e], Verena Ras[b,f], Chaimae Samtal[g], Milaine S. S. Tchamga[h], Dassen Sathan[i], Anisah Ghoorah[i], Raphael Z. Sangeda[j], Liberata A. Mwita[j], Upendo Masamu[j], Samar Kamal Kassim[k], Zoe Gill[b], Zahra Mungloo-Dilmohamud[i] and Gordon Wells[a]

[a]South African National Bioinformatics Institute, University of Western Cape, Belville, Cape Town, Republic of South Africa [b]Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, South Africa [c]Laboratory of Bioinformatics, Biomathematics and Biostatistics LR16IPT09, Institut Pasteur de Tunis, Tunis, Tunisia [d]Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum, Khartoum, Sudan [e]Chlamydiae and Mycoplasma Laboratory, Research Department, Institut Pasteur du Maroc, Casablanca, Morocco [f]Department of Biodiversity and Conservation Biology, University of the Western Cape, Private Bag X17, Bellville, South Africa [g]Laboratory of Biotechnology, Environment, Agri-food and Health, Faculty of Sciences, Dhar El Mahraz–Sidi Mohammed Ben Abdellah University, Fez, Morocco [h]African Institute for Mathematical Sciences, Cape Town, South Africa [i]Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius [j]Sickle Cell Program, Muhimbili University of Health and Allied Sciences, Dar-es-salaam, Tanzania [k]Faculty of Medicine, Ain Shams University, Abbaseya, Cairo, Egypt

7.1 Introduction

Recent advances in biological sequencing, computing technologies, and the internet of things have enabled the scientific community to focus their efforts on large-scale data-generating projects.[1] Such projects generate big datasets which can enable scientific discovery through the use of management infrastructures, or platforms that support simultaneous access to large datasets by multiple researchers globally.[2] In the past two decades, there has been a huge increase in the number of sensitive human genomic and health datasets being made available to researchers around the world. This has been associated with concerns related to the ethics of responsible research conduct and data sharing, as well as how these practices are collectively organized and monitored.[3,4] These concerns include escalating research costs, particularly for low and middle-income countries (LMICs), as well as questions related to research quality and transparency.[5] Social concerns such as lack of public trust and diversity in scientific

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00003-1
Copyright © 2023 Elsevier Inc. All rights reserved.

communities, insufficient community engagement, and ethical concerns associated with research practices, reusability, reproducibility, and resource limitations have also been raised.[6] These concerns have sparked an increase in open science collaborations, as well as a reduction in restrictive intellectual property (IP) rights.[7] These movements have also led to the increased use and development of platforms that enable sharing of genomic and phenotypic data in order to promote genomics research, share research resources, and maximizing the utility and value of existing datasets.[8]

The future of genomics research relies on the sharing of health, environment, and genomic data to facilitate large-scale biological data analyses and support the clinical interpretation of genomic variation.[9,10] With scientific and policy support, the field of data sharing has advanced rapidly, and efforts are ongoing to develop technical, ethical, and legal solutions to connect genomic databases and make them more accessible for clinical and research purposes.[2,11–13] In addition, funding bodies are increasingly requesting data sharing policies to accompany funding proposals and making data sharing a requirement of support for all funded projects, including hypothesis-driven projects, which primarily focus on specific research questions rather than the generation of data to be used by others.[2,13] Examples include the Wellcome Trust's "Policy on data management and sharing" and National Institutes of Health's (NIH) "Genomic data sharing" policies. Such policies also extend to international initiatives and organizational bodies such as the Global Alliance for Genomics and Health (GA4GH) and the Organisation for Economic Co-operation and Development, amongst others.[14]

Despite its slow rate of adoption worldwide, according to the latest report made by the World Health Organisation for "Developing global norms for sharing data and results during public health emergencies" sharing data is endorsed to become a default procedure in the future.[13] In this chapter, we discuss current data sharing practices and models, take a closer look at the specific model employed by the Human Heredity and Health in Africa (H3Africa) consortium, and discuss the challenges and considerations associated with data sharing in modern biological research.

7.2 Current practices

The overall process of data sharing is a combination of several characteristics, considerations, and requirements that need to be carefully explored.[15] Among these are those components that are unique to the type of data being shared. These characteristics determine where you can and, in some cases, where you should share data. Characteristics include the **funding resource** that enabled the generation of data (funding sources may require that the generated data should be stored for third-party access) and the **type(s) of data** generated.[15] In this case, type(s) do not only include the formats associated with the data but also the field in which the data was generated, whether it is specific to a particular disease or whether it was

produced from a specific population of interest. Each of these types may influence the method through which data is shared at a more global level. Type of data is a particularly important characteristic when it comes to genomics data sharing, as there are both data archives as well as data commons that are specific to the sharing of genomics data.[9] The considerations refer to those components that are absolutely crucial to the actual process of biological data sharing. These include consent, privacy and security, data access, oversight, participant engagement,[15] and lifetime maintenance. The type of **consent** garnered at the beginning of a research study determines if data collected and generated during the study can be shared, to what extent it can be shared, and the requirements or restrictions that need to be adhered to in relation to data sharing. The Data Use Ontology was created to provide a framework to describe the restrictions and requirements as it pertains to the sharing and secondary use of a particular dataset.[16] If researchers use specimens and data obtained without the prior intention of submitting it to a repository or database, then approval from an Institutional Review Board alone is insufficient for sharing data. Instead, researchers are encouraged to recontact the participants to request whether their genomic and clinical data can be shared. Recontacting research participants present practical difficulties, particularly when data came from past studies, contact information may have been lost or changed since or a participant could be missing or ill. To prevent such concerns, data sharing should be taken into account at the beginning of the project proposal stage. Some researchers are proposing an effective approach to adopt an e-consent system, which enables dynamic consent, whereby study subjects can monitor how research develops and their data are used and can at any time express whether to continue with a study or not and allow researchers to share their data.[17]

Privacy and security are two of the most important considerations of data sharing, particularly so when working with genomic data generated from human subjects.[15,18,19] The Health Insurance Portability and Accountability Act was passed by the United States government in 1996, primarily to modernize the flow of healthcare information, however, it has since been extended and covers the protection of human data not just at healthcare level, but at research and business levels too.[18] Many countries have since employed or are in the process of employing similar versions of these laws.[20] As it pertains to human data, privacy and security are not just important because it is mandated by law, but are also ethically crucial, as these data can be subject to discrimination and exploitation. The value of genomics analyses, in particular, has increased in recent times, as we move toward an era of precision medicine, especially in the field of rare diseases and cancer.[21–23] Therefore, the necessity of protection when sharing data generated from both patients and their family members has been widely discussed.[24] Indeed, sharing these data collectively can raise ethical tension between the value of datasets and the rights of participants, as well as how this pertains to the risk of reidentification. In this context, Takashima et al., 2018 highlighted the importance of engaging all stakeholders, including researchers, primary data users, data submitters, database operators, and review boards to shape a project's data sharing policy and practices. They emphasized

that further research on the opinions and attitudes of the public be conducted to continuously review current rules and practices, as well as new techniques as they are developed and as social values change.[24,25]

Data access policies/guidelines determine who can access a given set of data.[15] In modern biological research, we may want as many individuals as possible to access the data, in an effort to be fair and transparent, and ensure maximum reuse. However, data protection is crucial to prevent the exploitation of disenfranchised or disadvantaged communities and to ensure the reliability of sharing practices. Data access should be governed by consent.[17] While consent may determine how data can be shared and used, data access determines who can access the data and how the access is enabled or actioned.

Lastly, **oversight** and **participant engagement** cover how the practice of data sharing is monitored and fed back to the original data sources, respectively.[15,25] Oversight is crucial to maintain and ensure the considerations previously discussed and is the primary manner in which breaches, both intentional and unintentional can be rectified.[25] Ultimately, biological data generated from human subjects, particularly genomic and clinical, are produced in order to improve the health and healthcare of the participants from which data were generated. Therefore, data users should always bear this in mind, and feedback to participants (in terms of health benefit or education) should be a crucial component of all genomics data-sharing practices.

Notably, in addition to the discussed characteristics and considerations, findable, accessible, interoperable, and reusable (FAIR) Data Principles (https://fairsharing.org/) are equally relevant to data sharing in order to ensure that shared data are FAIR.[26] The principles are relevant to all stakeholders in the current digital ecosystem, facilitating FAIR data management and stewardship.

Hereafter, we will discuss two ways through which data are commonly shared:

1) Data sharing through a data archive or public data repository.
2) Data sharing through a data commons which can be project-, application-, or domain-specific.

We will look at the benefits and considerations of both methods. In addition, specifically with regards to sharing via a data archive or repository, we will discuss how this method is employed by the H3Africa (www.h3africa.org) consortium.

7.3 Case study: H3Africa model

The H3Africa consortium is a continent-wide initiative that aims to enable and facilitate innovative research into the genetic determinants of diseases affecting African populations.[27–29]

Table 7.1: H3Africa datasets made available through EGA (June 2020).

Dataset name	EGA accession	Dataset description	Technology type
H3AFRICA ACCME	EGAD00001005310	49 Samples from Nigeria	Illumina HiSeq 2500
H3AFRICA AWI-GEN	EGAD00001004448	60 Samples from Burkina Faso and Ghana	Illumina HiSeq 2500
H3AFRICA CAFGEN	EGAD00001004533	48 Samples from Botswana	Illumina HiSeq 2500
H3AFRICA ELSI	EGAD00001004316	50 Samples from Cameroon	Illumina HiSeq 2500
H3AFRICA MALSIC	EGAD00001004557	50 Samples from Benin	Illumina HiSeq 2500
H3AFRICA NEEDI	EGAD00001004334	50 Samples from Mali	Illumina HiSeq 2500
H3AFRICA TRYPANOGEN1	EGAD00001004393	26 Samples from Cameroon	Illumina HiSeq 2500
H3AFRICA TRYPANOGEN2	EGAD00001004220	41 Samples from Zambia	Illumina HiSeq 2500
H3AChipDesign Phenotype Dataset	EGAD00001005310	Phenotypic data for 348 samples	NA

The consortium is primarily funded by the National Institutes of Health (NIH, USA) and the Wellcome Trust (UK) through the Alliance for Accelerating Excellence in Science in Africa (AESA). A key goal of the initiative is not only to fund scientific research but to develop research infrastructure and capacity across the African continent.[27–29] An important component of the H3Africa initiative is sharing the genomic data produced by the consortium. For example, Table 7.1 lists a subset of H3Africa projects that have generated data and made it available to other researchers through data archiving. The number of projects to share generated data is expected to increase in the future. Hereafter, we discuss important components of sharing data through a data archive, and how these components are being implemented in LMICs by H3Africa.

7.3.1 Data archive

Data archives provide the infrastructural support that allows researchers to easily share and enable long-term data preservation. In contrast to databases, which can be primary (submitted data without annotation) and secondary (well-annotated) in nature and easily searchable and retrievable, data archives are employed for secure long-term storage where data is not expected to be retrieved often, and access for data is by request for whole datasets.[30] Data archiving, that is, submitting data to a data repository, serves a critical function to ensure the long-term availability of a particular dataset(s) and fulfilling contractual agreements with funders.[4,31] These platforms enable increased collaboration and discovery and facilitate the transfer of knowledge amongst researchers globally.[4,31] It is, therefore, necessary that data management and sharing plans be developed in conjunction with a particular archive or database in mind to ensure the availability and traceability of data in the future.

When submitting data to a data archive, the data must typically comply with criteria and standards set out by policies, procedures or guidelines specific to the selected archive. Different archives may have different requirements upon submission, depending on the data type submitted.

Some of these requirements include:

1) Standard file formats: The data must be organized in such a way that makes it reusable. Ideally, data needs to be in a standard format for a particular data type.[3,32,33] For example, primary human sequence data should typically be submitted in CRAM format to the European Genome-Phenome Archive (EGA) (https://www.ebi.ac.uk/ega/home) while paired-end reads should be submitted in FASTq format to comply with community standards.[34]
2) Secondary results: Typically, research projects only share raw data produced during the project. However, these studies often produce secondary results which researchers are encouraged to share along with their primary data, should it be available.[35,36]
3) Metadata: it is crucial that metadata accompany the submitted primary and secondary data, in order to describe the features and provenance of the data.[3,37,38] As an example, if next-generation sequencing data are submitted as BAM files or variants are provided in VCF files, it is important to annotate the data with information about the source, reference genome information, and ideally the software tools used, along with version numbers.[26]

Some of the biggest limitations for data archiving in the majority of LMICs are the inability to transfer data via a stable transfer channel, coupled with the lack of local infrastructure and technical support, as well as the lack of data curation capacity to ensure the data are well described.[4]

7.3.1.1 H3Africa data archive

The H3Africa consortium tasked the H3Africa Bioinformatics Network (H3ABioNet) with securely archiving the genomic and phenotypic research data generated by the various projects within the consortium. H3ABioNet was also tasked with preparing the data for submission to selected public repositories, in this instance the EGA and ENA.[39] A secure data archive located at the University of Cape Town, South Africa [39] is used for data storage and transfer to repositories. The archive prepares and encrypts data for submission and contains a Dashboard to easily monitor the progress of these various steps.[39] Submission is preceded by the completion of a data submission request form, which collects the data's associated metadata—including the study details, platform details, and core phenotypes. In order to prevent incomplete metadata submission, the request form uses a controlled vocabulary for completion. Data types allowed include whole genome- and exome sequencing data, genotyping array data, and microbiome sequencing data. Once the data transfer is complete, submission is verified to ensure no data was lost during the transfer. In addition to archiving these data in

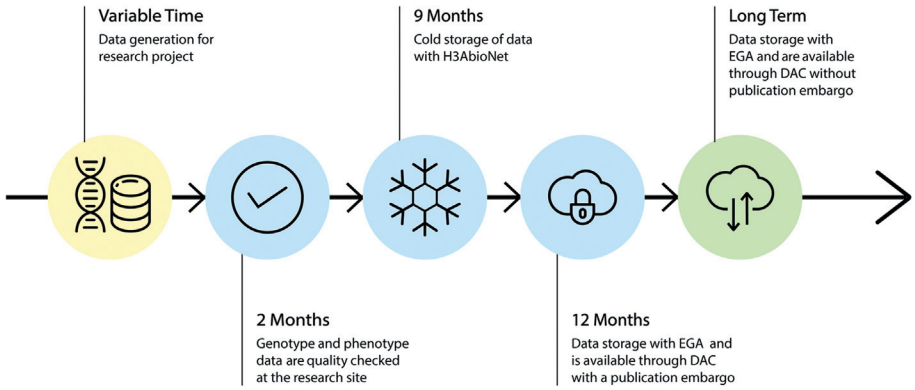

Figure 7.1
Timeline associated with data submission to H3Africa data archive.

Africa, H3ABioNet works alongside the data providers to submit the data to a publicly accessible repository, where controlled access is provided to the international research community. Data are submitted to H3ABioNet for archival purposes after it has been through the relevant quality control processes according to the project's standard operating procedures and policies and thereafter is under a moratorium for nine months before the data has to be submitted to public repositories. This includes the encrypted dataset along with metadata previously listed.

The data generated from the H3Africa projects are maintained in cold data storage in the archive for an extended period. Cold data storage refers to the storage of inactive data that is rarely used or accessed. H3ABioNet supports users with data submission, maintains the security of the data, and facilitates the data transfer,[39] with the goal of ensuring efficient and responsible data flow between the researchers, the archive and public repositories. Data are secured via encryption ahead of conducting any data transfers, via a built in security and fault recovery application on the transfer platform. The H3Africa Data Archive physical infrastructure comprises three main components, a landing area server, a vault server, and cold storage. The submitted data remain encrypted, except in the highly secure vault server, where it may be decrypted for validation purposes.[39] Several components were considered during the development of the archive, including; electrical supply, access to equipment for uninterruptible power supplies, access to electrical generator hardware, IT resources (human and infrastructure), and data backup infrastructure.[39] Currently, data are continuously being received from the H3Africa consortium and submitted to selected public repositories. The flow of data is illustrated in Fig. 7.1.

7.3.1.2 Data submission to EGA/ENA

The EGA and European Nucleotide Archive (ENA), developed and maintained by EMBL-EBI (and Centre For Genomic Regulation (CRG) in the case of the EGA), represent two

genomics data repositories based in Europe. The EGA and ENA serve as a space for the permanent archiving and sharing of all types of genetic and associated phenotypic data. The EGA focuses specifically on the archiving of genome and phenome data produced from humans,[40] whereas the ENA extends the focus to genomic and metagenomic data from all organisms.[41] Notably, the ENA mirrors data with both the US and Japan, making it widely accessible and faster rates. The EGA has strict access protocols to ensure data are securely managed, stored appropriately, and released only to approved researchers. They also have a strict set of submission guidelines to ensure data are submitted in standard, interoperable, reusable formats (https://www.ebi.ac.uk/ega/submission/). The existing access protocols and submission guidelines made the EGA an ideal archive of choice to distribute the data generated by H3Africa to the wider scientific community. The H3Africa consortium deposits their data in an EGA deposit box. Data are encrypted using the EGA encryption key prior to transfer in order to maintain data security. The associated study and datasets are provided with unique EGA accession IDs. Once in EGA, data are available on request; however, any associated publications are prohibited for a specified time period as explained hereafter.

7.3.2 Data sharing, access and release policy

Although there is a critical need for sharing collected and generated data, research participants and researchers often feel concerned about data access and ownership, particularly in LMICs due to resource limitations and disease burden.[42–44] Therefore, developing a data sharing, access and release policy is a crucial component of data sharing. This is particularly important when practicing data sharing in LMICs, where national data protection policies may not yet be implemented. The policies, both institutional and national, should address the ethical concerns of all stakeholders and endeavor to be transparent in the rules and practices as they relate to sharing and access to data.

Due to a lack of expertise, available research funds and infrastructure, many institutions particularly those in LMICs struggle to develop practical policies. Waithira et al. (2019) suggested four main components that should form part of a data-sharing policy, which broadly follow international standards. The first of these is a **good Data Management Plan (DMP)**. The DMP should outline the institution or project's stance on data collection, storage, and curation and should allow for project-specific needs. Guidelines on what constitutes a good DMP plan have previously been described.[3,45] The data sharing policy forms an important part of the DMP, therefore, **various models** of making data available to secondary users should be considered, and the method(s) best suited to a project's institutional aims should be adopted. Some models involve depositing data into external repositories such as figshare (https://figshare.com/) or the EGA, or online open-access platforms through journal publication. The policy should also **provide criteria** that align with the specific requirements of the funders, consent, by laws and regulation, for example, embargo periods for data release to the

public that institutions may need to adhere to.[42] Lastly, but most importantly, various **consent models** should also be considered when undertaking the research. The consent model chosen for primary research depends largely on the aims of the research; however, it is important to explore all models for secondary use/sharing, adhering to national policies, and funders requirements.[46,47] There are definite advantages and disadvantages to each, for example, broad consent for data sharing in addition to the consent for the actual study may deter users from consenting to data sharing.

The drafting of the H3Africa data sharing, access and release policy was a volunteer-driven, consortia-wide effort, spearheaded by a data sharing, access and release committee. An important consideration when developing the H3Africa policy was ensuring and enabling African benefit from the shared data. For data (raw and metadata) and biospecimen (sample) access, H3Africa wanted to ensure that the consortium had enough time to exploit the generated and collected data and that their research was not duplicated. Therefore, researchers are given nine months to submit their data to EGA after sufficient quality control processing according to employed operating procedures, and subsequent 12-month publication embargoes are placed on the shared data. This allows H3Africa researchers to complete their projects and publish results before wider access is provided. In addition, data requesters are encouraged to propose methods by which they will include African collaboration when accessing the shared data. Though this is not mandatory for data use, it is an important step to encourage more collaboration and capacity development from the more developed world. The H3Africa data sharing, access and release policy focuses on the requirements that need to be adhered to when requesting data, and how such requests are processed. Requests are reviewed by a Data Access Committee (DAC), or DBAC, in the case of H3Africa as the committee reviews access to biospecimens too. More info regarding the H3Africa data sharing policy can be found on the H3Africa website: https://h3africa.org/index.php/consortium/consortium-documents/.

7.3.3 Data access committee

A DAC is a group who are responsible for managing and reviewing access requests to data that have been collected or generated by individual research institutions or consortia.[25,48] A number of criteria, policies and procedures are used when reviewing data access requests submitted by researchers. Examples of these criteria include ethical and legal issues such as consent, privacy and confidentiality, qualification of the applicants, and scientific feasibility of the submitted requests.[25,48,49] Repositories operate in different ways, for example, for the EGA, DAC's operate locally in a decentralized style with data submitters being responsible for assigning their own DACs, the same mechanism is used by Database of Genotypes and Phenotypes (dbGAP) (a repository sponsored by (NIH). The advantage of a decentralized style is that data providers can control the terms of secondary data usage that are being specified in a data access agreement (DAA). But this means the existence of multiple DAAs which

force data requesters to adhere to varieties of agreements if they request access to multiple datasets.[48] With the centralized system, users do not have to adhere to multiple DAAs, but it has been suggested that special attention should be paid by those opting for that model since the interests and concerns of researchers differ between studies.[48] To maintain fairness in reviewing requests, the DAC should have nonconflicted representatives with a range of expertise from both scientific (e.g. genomics and data experts) as well as nonscientific stakeholders (e.g., patient advocates and ethical and legal expertise).[50,51] Furthermore, a DAC should be independent of the data providers in order to ensure impartiality when making decisions on data access requests.[48]

The H3Africa Data and Biospecimen Access Committee (DBAC) is composed of nine members from various areas of expertise such as scientists, biobanking, ethics, legal, data experts, and a patient advocate. The H3Africa steering committee and funders performed the selection of the DBAC members and members were appointed for an initial term of 3 years. Three members typically remain when the term ends, while the rest are replaced to enable continuity. No conflicts of interest should exist between these members and any of the H3Africa projects. Should a conflict of interest arise from any member of the DBAC, that member/s will automatically disqualify from any proceeding discussion and voting on the request they are in conflict with. A researcher will be required to submit a "Data and/or Biospecimen Access Request" form in order to gain access to data or biospecimens, which will be reviewed by the DBAC to ensure that they complied with the terms in informed consent and any limitation stipulated by the submitting investigator/institution for each study. The requestor should receive the result of an application within 30–60 days of submission. Apart from criteria such as ethics and legal grounds, the decision about the request will also be evaluated by checking the scientific merit of the request, institutional or researcher capacity, potential for research to be published, and the requirement to collaborate with African investigators (only for biospecimens).

7.3.4 H3Africa catalog

The H3Africa catalog was developed in order to facilitate access to the data produced by the H3Africa consortium, with the ultimate objective to facilitate further research that could benefit study participants.[29] Information in the H3Africa catalog is derived from the H3Africa Archive hosted by H3ABioNet and the three biorepositories, Integrated Biorepository of H3Africa Uganda (IBRH3AU) based in Uganda, Institute of Human Virology Nigeria (IHVN) based in Nigeria and Clinical Laboratory Services (CLS) based in South Africa. The catalog is intended to be a space where secondary users can explore metadata about the data and biospecimens that are available for request and to search for information about H3Africa studies.[29] The catalog provides a user-friendly web interface where users can perform simple queries, or register to be able to filter the data from the database based on criteria such as

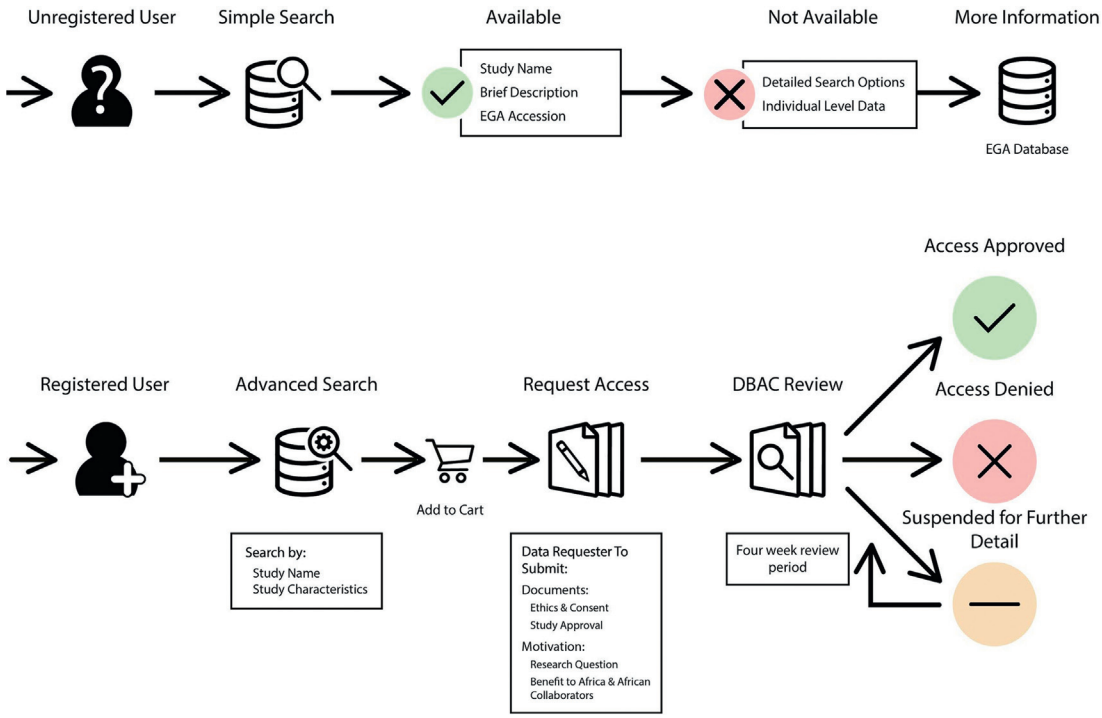

Figure 7.2
Data request procedure through the H3Affrica data and biospecimens catalog.

country, gender, and disease. The result is provided as a list of studies with general summary information about their participants and biospecimens. For ethical and security reasons, despite being anonymized, the participant and biospecimen IDs are hidden to protect the study individuals; only authorized users, such as the DBAC Secretariat, can view these when necessary. From search results, registered users can subsequently request access to available biospecimens and/or data through an online application form. Requests can be approved or rejected within the system to enable tracking of requests by the DBAC Secretariat. Fig. 7.2 illustrates the process of data access requests through the H3Africa catalog. More info on the catalog development can be found online: https://h3abionet.github.io/catalogue/.

7.4 Beacons

Deposition of data into public repositories enables access to the complete dataset, the access of which usually requires application and approval by a DAC. However, with regards to genomics data, researchers may simply want information about a single or small collection of variants from aggregate data. In this case, application for and downloading of a whole dataset is excessive. Solutions to this have been provided by the GA4GH in the form of data

streaming (htsget) and Beacons.[52] As a data standards organization serving the larger genomics and clinical communities, the GA4GH, is driving responsible genomics data sharing by developing standards and associated tools.[34]

Beacons enable limited querying of datasets for the presence of data to obviate the need for first applying for full access. At their most public level, a Beacon can alert a user to the presence of a particular allele at a particular chromosome position within a dataset.[52] The datasets simply respond with "yes/no," depending on whether they have seen the allele in that position in their dataset(s). Querying of small insertions and deletions can also be supported. Once deeper access has been negotiated depending on security and privacy concerns, different levels of information can be revealed to registered users, including allele frequencies, pathogenicity scores, and other associated phenotypes.[53] The Beacon Network (beacon-network.org), a distributed search engine across the world's public Beacons, facilitates single-level queries against multiple datasets. It provides a technically and conceptually simple approach to learning where data of interest resides and querying of the data stored in different formats (such as VCF), thereby enabling rapid and widespread adoption.

The network has grown into an internationally relevant resource for faster-anonymized data discoverability. At the time of writing over 100 Beacons serving more than 200 datasets have been created. Notably, Beacons can also implement match-making to facilitate the creation of sample sets with unrelated individuals. Beacons can be implemented with different levels of access, including public, registered, or controlled access. In addition to tiered access, the security of Beacons can be enhanced by a number of methods to prevent reidentification. These include limiting the number of queries (quotas), consent-codes, aggregation across many beacons, and information budgeting/metering.[52] H3Africa is a GA4GH driver project and is currently exploring sharing aggregate data through the EGA Beacon for those datasets which have been submitted to EGA. Discussions within the consortium are ongoing on whether Beacon data sharing should be endeavored.

7.5 Data commons model

A data commons is an alternative method to a data archive for sharing biological data. Since it can often take long periods to download large datasets, and it can be costlier to analyze data and then generate it, it is often better to provide the data analysis tools and infrastructure on-site. Data commons are an attempt to co-locate these tools with data storage in an interoperable manner.[54]

A number of bioinformatics-specific data commons are now available for public use. For example, NIH Data Commons (https://commonfund.nih.gov/commons) coupled with NIH Big Data to Knowledge (BD2K)[55] provides a cloud-based service where users can

find, access, and interact with digital resources generated and made available by peers. To support precision medicine, NCI Genomic Data Commons (https://gdc.cancer.gov/),[56] BloodPAC Data Commons,[56] and DNAdigest and Repositive[57] allow users to share genomic and clinical data (such as blood profiling) from patients with cancer. Similarly, brain commons (https://www.braincommons.org/) is designed to support neurology studies. To support proteomics and related studies, a number of data-sharing platforms exist[58] including, ProteomeXchange,[59] and the Clinical Proteomic Tumor Analysis Consortium Data Portal.[60] The Web Data Commons project (http://webdatacommons.org/) exploits the wealth of information on the Web by extracting structured data from web crawls and providing these data for public download. Additionally, nonspecific data commons include the Data Commons Cooperative (https://datacommons.coop/) where members can contribute, maintain, govern, and fund member-directed development. The Data Commons Framework aims at easing the task of developing data commons for cancer data from the National Cancer Institute (NCI) Cancer Research Data Commons while supporting FAIR data. Notably, similar platforms exist such as the Terra platform, which is a collaborative genome analysis platform built on a cloud computing infrastructure.

Operating data commons are associated with several ethical, cultural, legal, financial, and technical challenges.[61,62] For example, the aforementioned resources store anonymized information about patients. Some patients may want their data to be removed to avoid risks of remote reidentification, which could result in health discrimination. Furthermore, as argued by Escribano et al. (2018), the ease of access to valuable data sets has led many to use data sets without proper acknowledgment, leading some data publishers to place embargoes on metadata.[63] Indeed, the lack of metadata has been one of the main issues limiting the reuse of biological data.[38] There have been efforts to establish policies to formalize the use of public data and work, including, Open Data Commons (https://opendatacommons.org/) and the Creative Commons (https://creativecommons.org/), initiatives that allow users to share their data and work within a legal framework. That is, they allow users to choose licenses that best fit their objectives. Table 7.2 summarizes the general features and differences between the data archive data common models.

7.5.1 Data commons in Africa

A data commons such as the NCI Genomics Data Commons is the result of well-established international collaborations among cancer genome consortia like The Cancer Genome Atlas and International Cancer Genome Consortium. GDC was established in the United States as part of then-President Obama's Precision Medicine Initiative.[64] Indeed, the potential of precision medicine has fueled progress in international data sharing. To our knowledge, there is no published information on data commons specifically for genomics data sharing in Africa. There are several possible reasons; firstly, precision medicine is not yet a reality in Africa

Table 7.2: Comparison between a data archive model and a data commons model.

	Data archive model	Data commons model
Findability	Data submitters are responsible for metadata curation. In some cases, data submitters form working groups that map data to existing ontologies.	All data uploaded are harmonized by adding metadata to it through well established policies. This ensures that the data can be easily queried.
Types of service	Functions as a data sharing service, where the data are stored for users to download without any support tools to analyze or visualize data.	Their primary purpose is to provide storage facilities but they also provide large scale computation services. Researchers can access tools provided on the portal to analyze or visualize data such as ICGC genome browser.
Data category	Data stored in archives can be domain-specific (e.g., cancer genomics) or multidomain (e.g., genome sequence archive). An example of a multidomain data archive is the H3Africa data archive which collects genomic and clinical data pertaining to different diseases.	Often domain-specific such as genomics data commons (GDC) which stores cancer-related data such as whole genome sequence, whole exome sequence, methylation, RNA expression, proteomic, and clinical datasets.
Data infrastructure	The architecture normally used for deploying a data archive is the client- server. For example, H3Africa, data are stored in a secure data archive solution located at the University of Cape Town.	Depending on the amount of data being stored the architecture might be a client-server for small to medium data storage. But for large scale data storage cloud platforms are preferred.
Data security	Data access varies and can be controlled or open depending on the archive employed.	Normally they provide different levels of access to datasets depending on the type of data being accessed or whether the data is associated with a consortium, for example, open access for phenotype data and controlled access to genomics data.
Data ownership	The data stored in the archives are the property of the researchers who generated and shared the dataset.	Multiple owners: multiple data generators for shared custodianship, controlled through shared licensing or agreement with domain experts.
Data availability	Some data archives are open to the public while others are controlled by Data Access Committees.	Data are not always open to the public but to the domain experts for submission, request, and sharing.
Accountability/ responsibility	Generally, accountability and responsibility lies with the data repository.	Generally, accountability and responsibility lies with the data generator or domain expert.

because there is a lack of population-specific knowledge, skills, and resources.[29,65] However, initiatives to drive the field of precision medicine in Africa are currently underway.[29,65] Secondly, sharing data amongst multiple African countries remains a technical barrier. Ethical and legal issues are the main bottleneck,[66] and often each country ends up implementing their own national laws for data sharing, as is the case in Europe.[67] Thirdly, cloud-based platforms without advanced security technologies may not convince public health organizations to collaborate because of the risk of identifying individuals from their genetic data, among others.[68]

Notably, the landscape in Africa is slowly changing, for example, a data commons on Agriculture for Africa was recently introduced.[69] Data Commons for UK Tech aims to lead the world in providing open data and holds information on startups, investors, corporates, and accelerators as well as data on workspaces, service providers, and universities. Among the recorded companies, 54GENE (https://www.54gene.com/) is an African startup company (2019) aiming to provide services for genomics research and development. Their pan-African DNA biobank is located in Nigeria and it provides access to aggregated data, deidentified samples, and biospecimens for use by researchers (source). Similarly, The Registry of Stroke Care Quality (RES-Q; https://qualityregistry.eu/ is an initiative of the European Stroke Organisation-Enhancing and Accelerating Stroke Treatment (ESO EAST) Project set up to help countries improve their stroke care system.[70] African RES-Q members include Ghana and South Africa. RES-Q provides tools for continuous monitoring, for example, it facilitates the collection of indicators of stroke quality care. The data collected can then be analyzed and compared with global best practices in acute stroke care in order to implement improvements across countries. To promote data sharing and data reusability, RES-Q aims to integrate existing quality registries from multiple participating countries to allow countries to collect their own local data while also contributing to building a larger picture of stroke care. Policies and procedures related to how data is requested, accessed, and shared are available from the RES-Q web site (https://qualityregistry.eu/policies/). In addition, the African Commons (https://commons.africa/) is a platform showcasing tools built across Africa. However, as yet, none of the featured projects are related to genomics or genetics. This demonstrates the need to provide visibility of genomics data and the bioinformatics tools developed in Africa.

7.6 Common challenges in genomic data sharing and managing risks

Researchers often face many barriers to conducting data sharing despite the plethora of benefits that could be accrued from both sharing data and analyzing shared data. These barriers have been highlighted throughout the chapter and include unwarranted litigation, the desire to protect confidential and sensitive information, and a variety of concerns related to data integrity as it relates to data quality, data mining, and erroneous secondary analyses of data.[71] Given their efforts, once data are generated, researchers are reluctant to immediately share

it because they want an opportunity to optimally exploit their data. In addition, the phenotypic characterization, and collection of biological samples require an enormous effort from multiple stakeholders.[13] Some researchers face significant cultural barriers to sharing data and participating in longer-term collaborative efforts that stem from a desire to protect intellectual autonomy and a career advancement system built on the priority of publication and citation requirements.

The following sections discuss some of the primary challenges to data sharing related to technical and infrastructure concerns and ethical challenges, as well as additional challenges related to economic, political, and legal considerations.

7.6.1 ELSI

A primary challenge related to data sharing is concerns related to ethics. These have been highlighted throughout the chapter and inform many of the associated data sharing policies and considerations, for example, participant consent, prevention of exploitation, etc. Founded in 1990, ethical, legal, and social implications (ELSI) formed an important part of the Human Genome Project. Its founding mission was to determine and solve problems coming associated with genomics research that would affect society, individuals, and families.[72] The ELSI was funded by the Human Genome Project at the US Department of Energy and the National Institutes of Health. This program looks at the possible consequences of genomic research in four main areas[72]:

1. The potential for genetic discrimination in insurance and employment, and the privacy and fairness in the use of genetics information.
2. The integration of genetic technologies, including genetic testing, into the practice of clinical medicine.
3. Ethical problems around the conduct and design of genetic research with people, such as the process of informed consent.
4. The education of policymakers, public, students, and healthcare professionals on genetics and the complex problems that arise from genomic research.

These areas should be carefully considered as they relate to genomics data sharing. Ethical considerations and concerns of genomic data sharing differ between countries around the world. So there is a need to find a way to perform federated data analysis without data movement, in order to ethically and securely access anonymized data.[24,25] There are two major points of the ethical debate on data sharing: lack of reciprocity and lack of proportionality.[25,73,74] Ethical considerations are required for the sharing of human genome data, while protecting data privacy. The genome itself is personally identifiable information and standard anonymization techniques may be insufficient, since even after anonymization, the remaining information may still be subject to reidentification of the individual.[74] A qualitative study in

Nigeria showed that when study participants were informed of the aim of biobanking, they believed it to be beneficial and were willing to share their samples with other researchers. Moreover, when participants were requested to choose between broad, tiered, or restricted consent, half of the respondents chose broad consent, while others chose restricted or tiered consent for biobanking. The main reason for choosing tiered consent was a desire to maintain control over the types of research conducted with donated samples.[75]

Legal challenges to data sharing are related to the willingness (or lack thereof) to share data. The main challenges are associated with strict personal data protection laws preventing data sharing, or ambiguous legal frameworks which prevent the identifications and tracking of data sharing exceptions.[76] Additionally, the lack of formal data-sharing, copyright, and data ownership agreements, also hinder data sharing, particularly across borders and in LMICs.[76,77] Often, data ownership is not well documented either, leading to inconsistent or incomplete data sharing guidelines and policies.[76,77]

7.6.2 Motivational challenges

Motivation challenges are related to ELSI and often result from a lack of incentives to share data. Data sharing typically requires significant time and effort with no reward or credit in return for the work. Personal and institutional incentives are often needed to prioritize data sharing.[78,79] Often a key barrier to sharing data comes from the potential to lose the opportunity for publication if data recipients with advanced analysis capacity gain data access and publish first, receiving the majority of the credit, despite the data collection efforts that enabled publication.[80–82] This is particularly challenging in LMICs; where data providers are subject to heavy criticism when errors are found in secondary use of their data. Consequently, individuals collecting data restrict access to it until all the analyses are published to get the exclusivity on results published in scientific journals.[81,83] With respect to policy, the restriction of data that goes out of the country is often led by the lack of perceived benefits among policy-makers. This tendency toward restricting data access is reinforced by a lack of knowledge about how the information is going to be used. This is a particular concern when the data are sensitive and can lead to damaged relationships between institutes or countries.

7.6.3 Technical challenges

A number of technical challenges will need to be faced when analyzing potentially large and heterogeneous datasets.[4,62] The most common technical data-sharing challenges include:

- **Data harmonization**. Data harmonization typically requires a significant amount of work to make a consistent environment available to sites within which data could be extracted and documented. This is often due to different levels of available information technology, different databases and different data structures used across sites.

- **Research data management skills**. Many researchers possess limited data management skills which makes sharing data difficult to implement even when there is willingness to share data. This is often due to the difficulties in hiring and retaining staff with the necessary skills. There is a dire need to develop capacity in the skills required for adequate data management.
- **Data quality**. Maintaining data quality over many years of longitudinal individual-level surveillance is a challenge, particularly when dealing with highly mobile populations where there are no unique individual identifiers.
- **Language**. Clinical data are often collected in local languages, which limit the use and integration of data at the global level.
- **Data not collected**. There is a significant gap in data health systems including disease surveillance systems and lack of civil registration in many countries, particularly LMICs.
- **Data not preserved or, cannot be found**. With regards to public health, data archiving is not prioritized because the primary use of data is short-term. This is often associated with the lack of local infrastructure or capacity to maintain and store data.
- **Maintaining anonymity in the era of Machine Learning (ML)**. The struggle to find a balance between sharing participants' data to protect privacy and confidentiality and to benefit ML practices is continually debated. The fear of losing participant privacy or confidentiality through data sources such as social media, governments, and consumer giants is aggregated and potential of sensitive information being deciphered should be handled with caution.
- **Updating annotation**. The continuous addition of new variants and updating of variant annotation can be challenging. Such classification is frequently revised, and guidelines or practices to overcome this challenge are lacking. This also applies to updating annotation of other data types as knowledge increases and changes.

7.6.4 Infrastructure challenges

A major impediment to the implementation of genomics research and data sharing in Africa is the dearth of resources in general. These include a lack of expertise in specific fields, poorly equipped facilities, inadequate infrastructure (including internet and connectivity capabilities), and defaulting power supply.[4,84] The many challenges facing advanced technologies in developing countries make them underutilized in the majority of research laboratories and educational institutions. This is primarily due to economic factors, such as the cost to develop and maintain such capacity.

Data sharing infrastructure needs to provide services for researchers to submit their data, as well as facilitate data cleaning and curation for more widespread sharing. Storage infrastructure needs to be maintained in order to store cleaned data in knowledge bases such as commons platforms or data archives. Then additional tools should be provided for researchers to

integrate, analyze, and interpret the data to uncover new insights. The H3Africa initiative has supported the development of infrastructure for data and resource sharing in several locations across Africa.[4,84] This infrastructure could be used to contribute to the knowledge gap of genomics in African populations, provided there is proper maintenance, guidelines development, and constant funding. The burden of maintaining these newly developed infrastructures will, however, lie with the research stakeholders (particularly African scientists) and an important consideration in the regard is that funding is often finite.[4,84]

7.6.5 Economic and political challenges

Economic barriers currently restricting data sharing in LMICs. These barriers involve the cost of data sharing that need to be addressed. Sharing implies a process that includes curating, hosting, and preparing the data in a required format. This process requires skilled human resources and dedicated time. In East and Southern Africa, excluding South Africa, it has been previously highlighted how difficult it can be to find high-quality IT staff that can maintain data management standards.[62,76,85]

Political barriers to data sharing include restrictive data access policies; bureaucratic hurdles; hiding data for political reasons; lack of political will and commitment to push data sharing; lack of guidelines; and lack of trust. These barriers hinder effective sharing at the global level but are apparent at the local or national level as well. Difficulties in the public health sector related to obtaining data from regional centers or from other teams in the same institute have previously been described.[76] In LMICs, data sharing is usually more restrictive in terms of open data and data sharing policies. Countries are attentive to the data generated within them, and may heavily regulate the data sharing processes related to these data. This trend was substantiated by an epidemiologist working in a geographic area plagued during the Ebola outbreak and was, as a matter of policy, denied access by the geographical institute of the country he was working in, to the geo-coordinates of the localities where he was operating. Restrictive data access policies are often justified by the fear of information misuse. An expert in bioethics suggested that data generated for public health should only be used for public health purposes.[71,76]

7.6.6 Intellectual property rights

Innovation often results in the development or improvement of novel and existing products, processes, or services and can be protected through IP rights. Furthermore, a patent that provides exclusive rights for the development of a product or process that offers a new technical solution to a problem or provides a new way of doing something, may be granted.[86] In order to obtain a patent, an application in which technical information about the invention must be disclosed to the public, must be lodged to a relevant body. This means that published patent

documents become available as a potentially valuable source of technical and business information for inventors, enterprises, and researchers. For more information on how to apply for a patent see the specialized materials available on the World Intellectual Property Organization (WIPO) website (https://www.wipo.int/).

According to the WIPO, "IP refers to creations of the mind, such as inventions, literary and artistic works, designs, and symbols, names, and images used in commerce. IP rights aim to reward such creative human endeavor, thereby promoting innovation, economic growth and a higher quality of life." Nevertheless, not all creations of the mind can be subject to IP rights, and different types of IP rights have different criteria for protection, rights, and limitations. WIPO drafted a checklist for deciding whether or not to acquire IP rights. The most important issues related to data genomic sharing are[87]:

- Will the products or processes arising from the utilization of genetic resources have sufficient potential commercial value to justify the expense of seeking IP protection?
- Are these products and processes prone to rapid change and development? For example, synthetic biology, new genome engineering, and next-generation sequencing are providing new insights on the potential use of genetic resources which could quickly make prior discoveries obsolete or commercially nonviable.

A research consortium that encompasses large-scale epidemiology and state-of-the-art genomic medicine technology, with both developing and developed country partners, poses complex issues for data sharing and IP.[88] One of the most important considerations in creating policies that address these issues is to involve all consortium members in both the initial formulation of guidelines and subsequent evaluation of the policies over the period of the consortium. The H3Africa data sharing, access, and release policy instructs that all conclusions derived directly from the shared data should remain freely available, without any licensing requirements, for uses such as but not necessarily limited to, markers for developing assays and guides for identifying new potential targets for drugs, therapeutics, and diagnostics. H3Africa discourages any premature claims on precompetitive information that may impede research; however, it does encourage patenting of products suitable for private investment that address the healthcare needs of Africa. Patent applications should aim to prevent the restriction of H3Africa data access. The H3Africa IP policy closely follows the guidelines recommended by the NIH and Wellcome Trust.

7.7 Executive summary

Data sharing practices remain in development in LMICs. At worst, the value of data sharing is often forgotten or neglected, at best, data sharing is often secondary to answering the primary research questions for which data was generated. More widespread adoption can only be achieved if a culture of data sharing is encouraged, incentivised, and monitored. Though

funders request data-sharing strategies during funding applications, the implementation of such strategies often remains unmonitored. Therefore, data sharing is often staggered and only implemented when required by journals for publication purposes. Beyond the H3Africa consortium and data sharing community's preparedness, a culture of international data sharing can be achieved by the federation and use of metadata, respecting national and regional restrictions. To overcome the political and motivational challenges discussed in the chapter, realization of the benefits of higher-level and widespread data sharing need to be achieved. Overall, there is a lack of demonstration of the short- and long-term benefits of data sharing to policymakers and key stakeholders. A paradigm shift should be encouraged that values pride in data sharing rather than the current attitude toward the preservation of sovereignty.

Organizations and funders are key to monitoring, encouraging, and ensuring the adoption of best practices, standards, and policies and support and/or reward of responsible data sharing. This will accelerate a federated ecosystem for sharing of anonymized genomic and clinical data that will enable next-generation of FAIR, that is, Federated AI Ready, and create leaders and policymakers in the field, taking pride in sharing over preserving. A number of challenges still need to be addressed to achieve such an ecosystem. Protecting the privacy of individuals is a classical concern. Data anonymization and informed consent are considered traditional safeguards to protect the privacy of participants in research and clinical settings. Concerns that law enforcement, authorities, regulators, or policymakers might have access to or control over citizens' genetic information need to be addressed. There are also potential issues related to health insurance systems and policies, and thus, new guidelines need to be developed to address these threats. Importantly, tools and platforms will continuously evolve and change as the technology and demand change. However, the essence of protecting the participants and making FAIRer data available to the researchers beyond the initial investigators will grow exponentially in the future, demanding that we increasingly educate and inform the public, as well as the growing human resources involved in producing and analyzing the data.

References

1. Ristevski B, Chen M. Big data analytics in medicine and healthcare. *J Integr Bioinforma [Internet]*. 2018;15. (cited April 23, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6340124/.
2. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics – re-shaping scientific practice. *Nat Rev Genet*. 2009;10:331–335.
3. Kumuthini J, Chimenti M, Nahnsen S, et al. Ten simple rules for providing effective bioinformatics research support. *PLOS Comput Biol*. 2020;16:e1007531.
4. Mulder N, Adebamowo CA, Adebamowo SN, et al. Genomic research data generation, analysis and sharing – challenges in the african setting. *Data Sci J*. 2017;16:49.
5. Bakken S. The journey to transparency, reproducibility, and replicability. *J Am Med Inform Assoc*. 2019;26:185–187.
6. Dev SB. Unsolved problems in biology—The state of current thinking. *Prog Biophys Mol Biol*. 2015;117:232–239.

7. Gold ER, Ali-Khan SE, Allen L, et al. An open toolkit for tracking open science partnership implementation and impact. *Gates Open Res*. 2019;3:1442.
8. Medicine (US) i of. the benefits of data sharing [Internet]. *Sharing Clinical Research Data: Workshop Summary.* National Academies Press (US); 2013 (cited 2020, April 23, 2020), https://www.ncbi.nlm.nih.gov/books/NBK137823/.
9. Raza S, Hall A. Genomic medicine and data sharing. *Br Med Bull*. 2017;123:35–45.
10. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol*. 2019;20:128.
11. Siu LL, Lawler M, Haussler D, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. *Nat Med*. 2016;22:464–471.
12. Borry P, Bentzen HB, Budin-Ljøsne I, et al. The challenges of the expanded availability of genomic information: an agenda-setting paper. *J Community Genet*. 2018;9:103–116.
13. Modjarrad K, Moorthy VS, Millett P, Gsell P-S, Roth C, Kieny M-P. Developing global norms for sharing data and results during public health emergencies. *PLOS Med*. 2016;13:e1001935.
14. The PLOS Medicine.Can data sharing become the path of least resistance? *PLOS Med*. 2016;13:e1001949.
15. Villanueva AG, Cook-Deegan R, Robinson JO, McGuire AL, Majumder MA. Genomic data-sharing practices. *J Law Med Ethics*. 2019;47:31–40.
16. Dyke SOM, Philippakis AA, Argila JRD, et al. Consent codes: upholding standard data use conditions. *PLOS Genet*. 2016;12:e1005772.
17. Kaye J, Hawkins N. Data sharing policy design for consortia: challenges for sustainability. *Genome Med*. 2014;6:4.
18. Edemekong PF, Annamaraju P, Haydel MJ. Health insurance portability and accountability act (HIPAA). In: StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; 2020 (cited April 23, 2020): http://www.ncbi.nlm.nih.gov/books/NBK500019/.
19. La monaca G, Schiralli I. Data protection, privacy. *Clin Ter*. 2010;161:189–191.
20. Spencer A, Patel S. Applying the data protection act 2018 and general data protection regulation principles in healthcare settings. *Nursing Management (Harrow London England: 1994)*. 2019;10(7784). doi:10.7748/nm.2019.e1806. In this issue.
21. Gamulin S. The forthcoming era of precision medicine. *Acta Medica Acad*. 2016;45:152–157.
22. Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? *JAMA*. 2018;319:1981–1982.
23. Krzyszczyk P, Acevedo A, Davidoff EJ, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology*. 2018;6:79–100.
24. Takashima K, Maru Y, Mori S, Mano H, Noda T, Muto K. Ethical concerns on sharing genomic data including patients' family members. *BMC Med Ethics*. 2018;19:61.
25. Shabani M, Dove ES, Murtagh M, Knoppers BM, Borry P. Oversight of genomic data sharing: what roles for ethics and data access committees? *Biopreservation Biobanking*. 2017;15:469–474.
26. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The fair guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
27. Dandara C, Huzair F, Borda-Rodriguez A, et al. H3Africa and the African life sciences ecosystem: building sustainable innovation. *OMICS J Integr Biol*. 2014;18:733–739.
28. Adoga MP, Fatumo SA, Agwale SM. H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa. *Source Code Biol Med*. 2014;9:10.
29. Mulder N, Abimiku A, Adebamowo SN, et al. H3Africa: current perspectives. *Pharmacogenomics Pers Med*. 2018;11:59–66.
30. Piwowar HA, Vision TJ, Whitlock MC. Data archiving is a good investment. *Nature*. 2011;473:285.
31. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol [Internet]*. 2015;13. (cited April 23, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4640582/.

32. Kush R, Goldman M. Fostering responsible data sharing through standards. *N Engl J Med*. 2014;370:2163–2165.
33. Kumuthini J, Zass L, Chimusa ER, Chaouch M, Masimiremwa C. Chapter 9 – Minimum information required for pharmacogenomics experiments. In: Lambert CG, Baker DJ, Patrinos GP, eds. *Human Genome Informatics [Internet]* Academic Press; 2018:(cited January 13, 2020):179–193..
34. Knoppers BM. Framework for responsible sharing of genomic and health-related data. *HUGO J [Internet]*. 2014;8. (cited April 23, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4685158/.
35. Cheng HG, Phillips MR. Secondary analysis of existing data: opportunities and implementation. *Shanghai Arch Psychiatry*. 2014;26:371–375.
36. Tripathy JP. Secondary data analysis: ethical issues and challenges. *Iran J Public Health*. 2013;42:1478–1479.
37. Ulrich H, Kock A-K, Duhm-Harbeck P, Habermann JK, Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 fhir. *Stud Health Technol Inform*. 2016;228:162–166.
38. Wang F, Vergara-Niedermayr C, Liu P. Metadata based management and sharing of distributed biomedical data. *Int J Metadata Semant Ontol*. 2014;9:42–57.
39. Parker Z, Maslamoney S, Meintjes A, et al. Building infrastructure for african human genomic data management. *Data Sci J*. 2019;18:47.
40. Lappalainen I, Almeida-King J, Kumanduri V, et al. The european genome-phenome archive of human data consented for biomedical research. *Nat Genet*. 2015;47:692–695.
41. Leinonen R, Akhtar R, Birney E, et al. The European nucleotide archive. *Nucleic Acids Res*. 2011;39(Database issue):D28–D31.
42. Waithira N, Mutinda B, Cheah PY. Data management and sharing policy: the first step towards promoting data sharing. *BMC Med*. 2019;17:80.
43. Blasimme A, Fadda M, Schneider M, Vayena E. Data sharing for precision medicine: policy lessons and future directions. *Health Aff Proj Hope*. 2018;37:702–709.
44. Guttmacher AE, Nabel EG, Collins FS. Why data-sharing policies matter. *Proc Natl Acad Sci USA*. 2009;106:16894.
45. Michener WK. Ten simple rules for creating a good data management plan. *PLoS Comput Biol*. 2015;11:e1004525.
46. Grady C, Eckstein L, Berkman B, et al. Broad consent for research with biological samples: workshop conclusions. *Am J Bioeth*. 2015;15:34–42.
47. Armstrong S, Langlois A, Laparidou D, et al. Assessment of consent models as an ethical consideration in the conduct of prehospital ambulance randomised controlled clinical trials: a systematic review. *BMC Med Res Methodol*. 2017;17:142.
48. Shabani M, Knoppers BM, Borry P. From the principles of genomic data sharing to the practices of data access committees. *EMBO Mol Med*. 2015;7:507–509.
49. Cheah PY, Piasecki J. Data access committees. *BMC Med Ethics*. 2020;21:12.
50. de Vries J, Tindana P, Littler K, et al. The H3Africa policy framework: negotiating fairness in genomics. *Trends Genet TIG*. 2015;31:117–119.
51. Beiswanger CM, Abimiku A, Carstens N, et al. Accessing biospecimens from the H3Africa consortium. *Biopreservation Biobanking*. 2017;15:95–98.
52. Fiume M, Cupak M, Keenan S, et al. Federated discovery and sharing of genomic data using beacons. *Nat Biotechnol*. 2019;37:220–224.
53. Wan Z, Vorobeychik Y, Kantarcioglu M, Malin B. Controlling the signal: practical privacy protection of genomic data sharing through beacon services. *BMC Med Genomics*. 2017;10:39.
54. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A case for data commons: toward data science as a service. *Comput Sci Eng*. 2016;18:10–20.
55. Jagodnik KM, Koplev S, Jenkins SL, et al. Developing a framework for digital objects in the big data to knowledge (BD2K) commons: report from the commons framework pilots workshop. *J Biomed Inform*. 2017;71:49–57.

56. Grossman RL, Abel B, Angiuoli S, et al. Collaborating to compete: blood profiling atlas in cancer (BloodPAC) consortium. *Clin Pharmacol Ther*. 2017;101:589–592.
57. Kovalevskaya NV, Whicher C, Richardson TD, et al. DNAdigest and repositive: connecting the world of genomic data. *PLoS Biol [Internet]*. 2016;14. (cited 2020 April 24, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4807091/.
58. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics*. 2015;15:930–949.
59. Deutsch EW, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res*. 2020;48(D1):D1145–D1152.
60. Edwards NJ, Oberti M, Thangudu RR, et al. The cptac data portal: a resource for cancer proteomics research. *J Proteome Res*. 2015;14:2707–2713.
61. Bambauer JR. Tragedy of the data commons [Internet]. Rochester, NY: Social Science Research Network; 2011 Mar [cited April 24, 2020]. Report No.: ID 1789749. Available from: https://papers.ssrn.com/abstract=1789749.
62. Figueiredo AS. Data sharing: convert challenges into opportunities. *Front Public Health [Internet]*. 2017;5. (cited December 20, 2019). https://www.frontiersin.org/articles/10.3389/fpubh.2017.00327/full.
63. Escribano N, Galicia D, Ariño AH. The tragedy of the biodiversity data commons: a data impediment creeping nigher? *Database J Biol Databases Curation [Internet]*. 2018;2018. (cited April 24, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5892138/.
64. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI genomic data commons as an engine for precision medicine. *Blood*. 2017;130:453–459.
65. Radouani F, Zass L, Hamdi Y, et al. A review of clinical pharmacogenetics studies in African populations. *Pers Med*. 2020;17:155–170.
66. Dove ES, Joly Y, Tassé A-M. Public population project in genomics and society (P3G) international steering committee, international cancer genome consortium (ICGC) ethics and policy committee, knoppers BM. Genomic cloud computing: legal and ethical points to consider. *Eur J Hum Genet*. 2015;23:1271–1278.
67. Molnár-Gábor F, Korbel JO. Genomic data sharing in europe is stumbling—Could a code of conduct prevent its fall? *EMBO Mol Med [Internet]*. 2020;12. (cited April 24, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7059003/.
68. Carter AB. Considerations for genomic data privacy and security when working in the cloud. *J Mol Diagn*. 2019;21:542–552.
69. Baarbé J, Blom M, de Beer J. A proposed "Agricultural data commons" in support of food security. *Afr J Inf Commun*. 2019;23:1–33.
70. Mikulik R, Bar M, Grecu A, et al. The registry of stroke care quality (RES-Q): the first nation-wide data on stroke care quality. *J Neurol Sci*. 2017;381:91.
71. Barnes KI, Canario JA, Vernekar SS, et al. Equitable data sharing: challenges and suggestions for ways forward. *Wellcome Open Res*. 2019;4:172.
72. Oliver JM, McGuire AL. Exploring the ELSI universe: critical issues in the evolution of human genomic research. *Genome Med*. 2011;3:38.
73. Population C on, education D of B and SS and, the national academies of sciences E. Exploring the ethical imperative for data sharing [Internet]. Sharing Research Data to Improve Public Health in Africa: A Workshop Summary. National Academies Press (US); 2015 (cited 2020 April 24, 2020). https://www.ncbi.nlm.nih.gov/books/NBK321546/.
74. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet*. 2014;15:409–421.
75. Igbe MA, Adebamowo CA. Qualitative study of knowledge and attitudes to biobanking among lay persons in Nigeria. *BMC Med Ethics*. 2012;13:27.
76. Edelstein DM, Sane DJ, Epidemiologist S, Welfare NI for H and, Helsinki, Finland. Overcoming barriers to data sharing in public health: a global perspective. 2015 (cited 2020 April 24, 2020), https://www.chathamhouse.org/publication/overcoming-barriers-data-sharing-public-health-global-perspective.

77. Medicine (US) i of. barriers to data sharing [Internet]. Sharing Clinical Research Data: Workshop Summary. National Academies Press (US); 2013 (cited April 24, 2020) https://www.ncbi.nlm.nih.gov/books/NBK137824/.
78. Fan X, Yu P. A discussion about the importance of laws and policies for data sharing for public health in the people's republic of china. *Stud Health Technol Inform*. 2007;129(Pt 1):316–319.
79. Lopez AD. Sharing data for public health: where is the vision? *Bull World Health Organ*. 2010;88 467–467.
80. Chandramohan D, Shibuya K, Setel P, et al. Should data from demographic surveillance systems be made more widely available to researchers? *PLoS Med [Internet]*. 2008;5. (cited April 24, 2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2253613/.
81. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull World Health Organ*. 2010;88:462–466.
82. Tangcharoensathien V, Boonperm J, Jongudomsuk P. Sharing health data: developing country perspectives. *Bull World Health Organ*. 2010;88:468–469.
83. Lang T. Advancing global health research through digital technology and sharing data. *Science*. 2011;331:714–717.
84. Adebamowo SN, Francis V, Tambo E, et al. Implementation of genomics research in Africa: challenges and recommendations. *Glob Health Action [Internet]*. 2018;11. (cited December 20, 2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5769805/.
85. Richter H, Slowinski PR. The data sharing economy: on the emergence of new intermediaries. *IIC - Int Rev Intellect Prop Compet Law*. 2019;50:4–29.
86. Saha CN, Bhattacharya S. Intellectual property rights: an overview and implications in pharmaceutical industry. *J Adv Pharm Technol Res*. 2011;2:88–93.
87. WIPO, Initiative ACD. *A Guide to Intellectual Property Issues in Access and Benefit-sharing Agreements.* Geneva: World Intellectual Property Organization; 2018:90.
88. Chokshi DA, Parker M, Kwiatkowski DP. Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration. *Bull World Health Organ*. 2006;84:382–387.

CHAPTER 8

Data standardization in the omics field

Judit Kumuthini[a], Lyndon Zass[b], Melek Chaouch[c], Zoe Gill[b], Verena Ras[b], Zahra Mungloo-Dilmohamud[d], Dassen Sathan[d], Anisah Ghoorah[d], Faisal Fadlelmola[e], Christopher Fields[f], John Van Horn[g], Fouzia Radouani[h], Melissa Konopko[i], Emile R. Chimusa[j] and Shakuntala Baichoo[d]

[a]South African National Bioinformatics Institute, University of the Western Cape, Bellville, Cape Town, Republic of South Africa [b]H3ABioNet, UCT Computational Biology Division, Institute of Infectious Disease and Molecular Medicine, University of Cape Town Health Sciences Campus, Cape Town, South Africa [c]Laboratory of Bioinformatics, Biomathematics and Biostatistics LR16IPT09, Institut Pasteur de Tunis, Tunis, Tunisia [d]Faculty of Information, Communication and Digital Technologies, University of Mauritius, Réduit, Mauritius [e]Centre for Bioinformatics & Systems Biology, Faculty of Science, University of Khartoum, Al Khartoum, Sudan [f]High-Performance Computing in Biology (HPCBio), University of Illinois, Champaign, IL, United States [g]Department of Psychology and School of Data Science, University of Virginia, Charlottesville, VA, United States [h]Chlamydiae and Mycoplasma Laboratory, Research Department, Institut Pasteur du Maroc, Casablanca, Morocco [i]GA4GH, Sanger Institute, Hinxton, CB10 1SD, United Kingdom [j]Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle, Tyne and Wear, NE1 8ST, United Kingdom

8.1 Introduction

8.1.1 Defining standardization

Standardization is the process of developing and implementing a set of "standards" that are **specifications**, guidelines, or requirements regarding a particular subject matter.[1] These standards are typically set through consensus within groups such as organizations, government, end users, or interest groups with the intention of increasing quality and efficiency.[1] Standards are implemented in various settings depending on user and aim. The most commonly implemented standards are **controlled vocabularies** or **reporting guidelines**, and **standard practices** (see Table 8.1). These tend to dictate how field-specific information is described and processed.[2] These specifications all serve a similar purpose namely; **harmonizing data, promoting data quality** and **compatibility**.[1,2]

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00008-0
Copyright © 2023 Elsevier Inc. All rights reserved.

Table 8.1: The different types of standards.

Standard Type	Description
Data; Data-exchange	Standards that allow a consistent flow of data between systems and organizations. Specify data formats, elements, and structure.
Terminology; Ontologies	Standards that promote the use of consistent and agreed-upon vocabulary with regards to specific fields of application.
Document; Reporting	Stipulations as to the minimum information to be included when reporting data into a manuscript or system.
Conceptual; Application	Standards for methodology coherence. include storage, transportation, transformation, and distribution of information.

Standards are applicable across multiple branches of sciences and have been implemented in physical, life, social, applied and interdisciplinary sciences.[3] There are also several different types of standards, and depending on the application, standards can exist as a single type or a combination of different types.[2,4] These broad standard types are listed in Table 8.1.

Who creates a standard and how it is created is critical with regard to the uptake and utility of a standard. Standard creation needs to be both an objective and thorough process. If not, it can face significant governance and user adoption challenges related to credibility, suitability, and specificity. Consequently, standard-setting bodies must negotiate between the stakes of the community of interest, key stakeholders, and the public when engaging in standard creation. A standard must gain widespread acceptance within the community of interest as being comprehensive and relevant in order to be widely used.[5] Therefore, many standards are constructed by way of consensus-building. These consensus-building efforts are generally driven by initiatives, organizations, and knowledge leaders with a key interest in the construction of the standard.[5] These efforts are generally conducted through discussions at international meetings, correspondence with journals of interest, and using user-friendly online platforms or other asynchronous teamwork environments.[4,5] Once a set of standards has been created, it is typically submitted to a leading standards-housing or -sharing body within the community of interest for assessment. Afterwhich its importance and implementation need to be advocated. This can be achieved through scientific presentation, training events, and gaining organizational buy-in. Thereafter, the community of interest is responsible for feedback on these standards, and contribution to its maintenance and evolution.[6]

A typical standard's life cycle consists of three phases:

Formulation—Identification of a need for a standard. Collection of use cases that describe the breadth and depth of the requirements. Defining the gaps in the field and the scope of the standard.

Development—Drafts are developed by a core group. Feedback and evaluation by experts are requested. Standard formats are identified and developed.

Maintenance—Creation of implementations for the standard. These include documentation and education materials. Maintenance also includes sustainability of the standards ensuring accessibility, as well as the evolution of the standards, including backward compatibility of each version.

8.2 Omics data standardization

Since the advent of high throughput genome sequencing in the 1990s, health and life sciences technology has undergone rapid development, changing the scale and scope of biomedical research.[7] This evolution has led life and health sciences into an era known as the omics or "big data" era. The omics era is associated with the generation of immense and exponentially increasing amounts of very diverse biohealth data. With this rapid change in focus to big data in biology and life sciences, bioinformatics capabilities have become an integral part of the research.[8] Biohealth data can be both **qualitative**, such as information about qualities that can't be measured, and **quantitative**, values that can be measured, in nature. Data that have been collected or generated often require clean-up and processing before it is suitable for subsequent determination or resolution.[8] If the collection or generation of data changes without the knowledge to those analyzing the data, it can be misinterpreted as a change in the variables being measured. Thus, the quality of data collection and generation is invaluable; maintaining and assessing the quality and consistency of data requires data collection standards.

Therefore, data integration is key to tackling big data analyses in the omics era. **Data integration** is the process of combining and unifying similar or related data from diverse sources, for example, data collected across national borders; data collected in both clinical and research settings. The integration of big data requires both capacity and computational resources, but if done comprehensively, can add tremendous value to biomedical research.[9] Data integration or management is not an aim in itself, but rather the key conduit leading to knowledge discovery and innovation. Data integration can lead to data reuse and subsequent knowledge generation following data sharing.[9] The exponential increase of research outputs within biomedical sciences in the last two decades means that there is more publically available data than ever before.[10] Published biohealth data, in all its forms, provide a rich resource of crucial scientific information awaiting interpretation. However, several significant challenges prevent the extraction, interpretation and use, these include data sharing and access considerations, as well as data diversity and integration challenges.[11] If implemented comprehensively and effectively, data standardization can provide the vital stepping stone required to decipher these issues within the omics sphere.[11]

As illustrated in Fig. 8.1, data standardization can streamline the transformation of data to reduce unnecessary diversity and volume by utilizing consistent and uniform reporting of

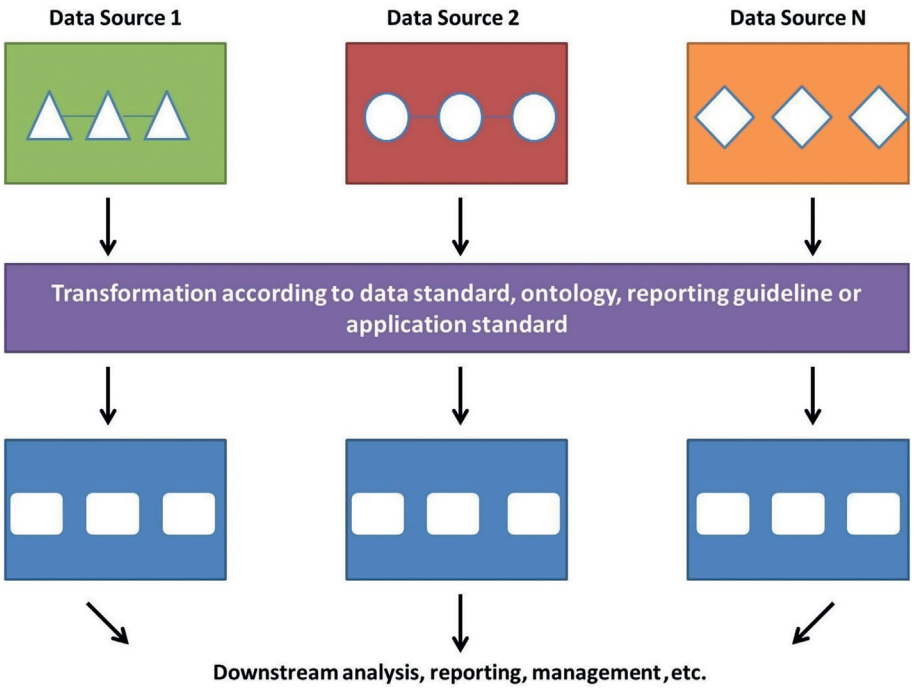

Figure 8.1
Graphical representation of data standardization.

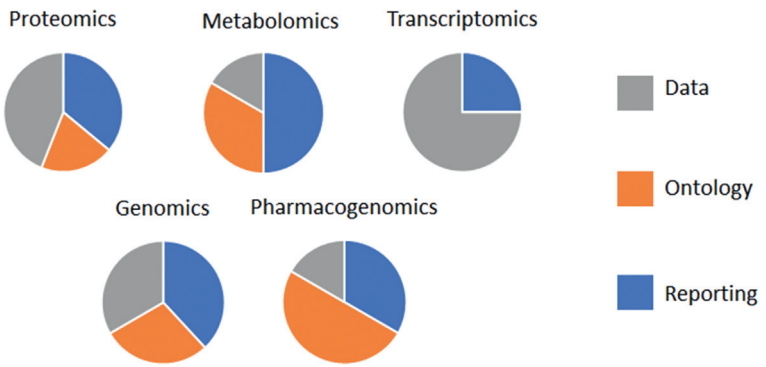

Figure 8.2
Proportion of omics standards found on FAIRsharing.org (accessed November 23, 2017).

data collected or generated by different researchers, technicians, managers, organizations, technologies, etc.[4] Consequently, this allows effective data management and integration, and produces data, information and research that adhere to **FAIR** (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable) **principles**.[12]

In an era of ever-evolving research and ever-increasing data generation, omics data standardization becomes, or rather should become, as essential to life and health sciences as bioinformatics. Data standardization holds great promise for interpreting previously generated data and information as well as increasing the power and diversity of datasets and bioinformatics studies, resulting in the generation of valuable and novel knowledge for future use.[4] The benefits of standardization are broadly captured within its aim to improve data reusability through maintaining quality, compatibility, and interoperability. However, these benefits can be further expanded upon within the context of the fields of application. Standardization supports quality knowledge accumulation by facilitating meta-analysis and data harmonization, producing data that is consistent and can be merged or aggregated, and it supports the use of emerging data sources due to increased comparability and understandability.[9] Omics data standardization can go a long way in promoting data management that is both effective and coherent. This simplifies data analysis to a point at which only computing/technological resources remain the limiting factor in large-scale analytics, allowing for the development and identification of novel data analysis strategies, innovative research studies, and therapeutic avenues for complex diseases.[2,4,5]

Omics data standardization also permits the ability to exchange data without losing meaning or values and thus provides a sound basis for both the sharing and reuse of diverse omics data types. It does so by using shared formats, structures, syntax, and terminology, producing research and data that are interpretable and intuitive while simultaneously reducing research redundancy.[4,5] This allows researchers to analyze factors that deviate between similarly conducted and aimed research studies. Moreover, omics data standardization also sparks research collaboration and interoperability within the life and health sciences, thereby lending itself to the production of improved research quality and data analysis strategies and activities.[4,5] The benefits of omics data standardization are perhaps best reflected in the potential it holds for the tackling of significant data analysis challenges.[4,11] Big data analytics endeavors are generally the result of large-scale collaborations between specialists from various disciplines and these collaborations tend to be unstructured.[11] This contributes to significant variance in data capturing, reporting, and altering.[4,5] Heterogeneous data should no longer be the limiting factor in large-scale investigations.

8.2.1 Existing standards and resources

Standards relevant to omics sciences are housed in databases that curate and supervise the maintenance of these standards. Some of these resources, as summarized in Table 8.2, have evolved over time, combining or curating their structure to cater to specific end users. For example, the Minimum Information for Biological and Biological Investigations (**MIBBI**) foundry culminated in the inception of the **BioSharing** Portal, and exists as the **FAIRsharing** today, facilitation standardization in several fields, including omics technologies. Various

Table 8.2: Standardization resources relevant to biomedical research and omics standardization.

Resource Description		
Open Biomedical Ontologies (OBO) Foundry[12]	Aim	Functions as primary biomedical ontology database, developing families of interoperable ontologies, based on the principles of open use, collaborative development, non-overlapping and strictly scoped content, common syntax and relation.
	Collection	Large collection of biomedical ontologies, including the gene ontology (GO), human disease ontology, protein ontology, plant ontology, and ontology of biomedical investigations.
	URL	http://www.obofoundry.org/.
Ontology for Biomedical Investigations (OBI)[23]	Aim	Functions as a biomedical ontology database, reusing biomedical knowledge ontologies from the OBO project while adding the ability to describe how this knowledge was derived.
	Collection	Large collection of biomedical ontologies, including the gene ontology (GO), chemical entities of biological interest (ChEBI), and phenotype attribute and trait ontology (PATO).
	URL	http://obi-ontology.org/.
BioPortal[24]	Aim	Functions as open-source, domain-independent comprehensive ontology repository, supporting community-based access, peer-review, mapping and annotation of ontology content.
	Collection	Large collections of biomedical ontologies developed in various formats and originating from various ontological databases, including The OBO Foundry, the proteomics standards initiative, biodiversity information standards, etc.
	URL	https://bioportal.bioontology.org/.
Enhancing the Quality and Transparency Of health Research (EQUATOR) Initiative[25]	Aim	Improve the reliability and value of published health research literature by promoting transparent and accurate reporting and use of robust reporting guidelines.
	Collection	Minimum reporting guidelines specific to experiment/study types, including systematic review guidelines, case report guidelines, qualitative research guidelines, randomized trial guidelines, etc.
	URL	http://www.equator-network.org/.
BioSchemas	Aim	Improve data interoperability in life sciences by encouraging schema.org markup use, so that websites and services contain consistently structured information.
	Collection	Collection of bioschemas of generic types (elements and datasets used in disciplines beyond life sciences) as well as biological types (elements and datasets used in life sciences specifically).
	URL	https://bioschemas.org/.
Health Level 7 (HL7) International[16]	Aim	Provide comprehensive frameworks and standards for exchange, integration, sharing, and retrieval of electronic health information, supporting clinical practice and management, delivery and evaluation of health services.
	Collection	Houses health-related standards across seven sections; primary standards, foundational standards, clinical and administrative domains, her profiles, implementation guides, rules and references, education, and awareness.
	URL	http://www.hl7.org/.

(*continued on next page*)

Table 8.2: Standardization resources relevant to biomedical research and omics standardization—cont'd

Resource Description		
Minimum Information for Biological and Biological Investigations (MIBBI) foundry[26]	Aim	Promote the development and use of coherent minimum reporting guidelines for biological and biomedical investigations.
	Collection	large collection of minimum reporting guidelines related to diverse biomedical methodologies including minimum information about a cellular assay (MIACA), minimum information about a microarray experiment (MIAME), and minimum information about a phylogenetic analysis (MIAPA).
	Since its inception in 2008, MIBBI has undergone significant evolution over time. Starting in 2009, MIBBI was reconstructed into BioSharing, culminating into the launch of the BioSharing portal in 2011.	
BioSharing Portal[27]	Aim	Promote development, implementation and use of curated and crowd-sourced metadata standards, databases, and data policies in the life sciences
	Collection	Expanded upon MIBBI coverage to include a large collection of terminology, model and data standards as well as standardization databases and policies.
	The BioSharing portal experienced similar evolution and growth over time. In 2016, BioSharing became the ELIXIR Registry of Standards, functioning as part of the ELIXIR Interoperability Platform. It launched BioSharing Educational, aiming to increase knowledge of standardization in life sciences. In response to user feedback, and to reflect the broadened scope of FAIR principles, it was redesigned into FAIRsharing in 2017.	
FAIRsharing[28]	Aim	Functions as curated, informative and educational resource on data and metadata standards, interrelated to databases and data policies, and aims to serve users across all disciplines in life sciences, by supporting the production of FAIR data.
	Collection	Includes reporting, terminology and data standards, interrelated databases, data policies, and education material. Houses a large collection of standards relevant to post-genomics investigations.
	URL	https://fairsharing.org/.
Genomic Standards Consortium[29]	Aim	Improving descriptions of genomes and metagenomes through consensus-driven solutions, open-membership organization, community-based.
	Collection	Has created a Minimum Information about any sequence which has three minimum information checklists for describing genomes, metagenomes, and environmental marker sequences upon submission to public databases and publication.
	URL	https://press3.mcs.anl.gov/gensc/.

standards tagged as omics standards in FAIRsharing are listed in Table 8.3 and these proportions are illustrated in Fig. 8.2, separated by omics discipline, as well as standard and application type.

To illustrate data standardization, three case studies are presented, pertaining to different data types and associated standardization resources. Fig. 8.3 illustrates the three different data, DNA, RNA, and protein.

Table 8.3: List of omics standards found on FAIRsharing.org (accessed November 23, 2017).

Classification		Name (Abbreviation)	FAIR ID	PTS
	D	Biological Dynamics ML (BDML)	bsg-s000674	*
	D	CHADO XML	bsg-s000220	*
	E	Big Browser Extensible Data Format (BigBed)	bsg-s000212	*
	E	Affymetrix Raw Intensity Format	bsg-s000219	
	E	Pre-Clustering File Format	bsg-s000245	
	E	Binary sequence information Format (2bit)	bsg-s000205	
	E	bigWig Track Format	bsg-s000213	
	E	microarray track data Browser Extensible Data Format	bsg-s000243	
	E	Wiggle Track Format (WIG)	bsg-s000271	
	E	Big Gene Prediction (bigGenePred)	bsg-s000695	X
	E	Resource Description Framework Schema (RDFS)	bsg-s000283	
	E	Common Workflow Language (CWL)	bsg-s000606	*
	E	Biological Pathway eXchange (BioPAX)	bsg-s000038	*
	E	EMBRACE Data and Methods Ontology (EDAM)	bsg-s000275	*
	E	Comparative Data Analysis Ontology (CDAO)	bsg-s002618	
	P	Medical Subject Headings (MESH)	bsg-s000294	*
	P	Robert Hoehndorf's Version of MeSH (RH-MeSH)	bsg-s002792	
	E	MI About a Bioinformatics investigation (MIABi)	bsg-s000650	*
	E	MI about a Stem Cell Experiment (MISCE)	bsg-s000659	
	E	MI About Particle Tracking Experiments (MIAPTE)	bsg-s000671	*x&
	D	Genomic Contextual Data ML (GCDML)	bsg-s000079	*
	D	Global Alliance for Genomics and Health Metadata Model (GA4GH)	bsg-s000599	&
	D	OmicsDI XML format	bsg-s000722	*x
	D	Gene Product Information Format (GPI)	bsg-s000046	
	E	Functional Genomics Experiment ML (FuGE-ML)	bsg-s000075	*&
	E	Browser Extensible Data Format (BED)	bsg-s000211	
	E	Multiple Alignment Format (MFA)	bsg-s000242	
	E	Variant Call Format (VCF)	bsg-s000270	
	D	SNP Ontology (SNPO)	bsg-s002573	
	D	Environment Ontology (ENVO)	bsg-s000060	*
	D	Metagenome/Microbes Environmental Ontology (MEO)	bsg-s002785	
	E	Sequence Ontology (SO)	bsg-s000046	*x
	P	Neomark Oral Cancer Ontology version 4 (NeoMark4)	bsg-s002766	*
	P	Solanaceae P Ontology (SPTO)	bsg-s002799	*
	E	Marine Microbial Biodiversity, Bioinformatics and Biotechnology data reporting and service standards (M2B3)	bsg-s000592	*

(*continued on next page*)

Table 8.3: List of omics standards found on FAIRsharing.org (accessed November 23, 2017)—cont'd

Classification		Name (Abbreviation)	FAIR ID	PTS
	E	MI about a (Meta)Genome Sequence (MIxS – MIGS/MIMS)	bsg-s000172	*
	E	MI about any (x) Sequence (MIxS)	bsg-s000518	*
	E	Minimal Metagenome Sequence Analysis Standard (MINIMESS)	bsg-s000176	*
	E	MI Required for A Glycomics Experiment - Sample Preparation (MIRAGE – Sample Preparation)	bsg-s000682	*
	E	MI Required for A Glycomics Experiment - Glycan Microarray Analysis (MIRAGE Glycan Microarray Analysis)	bsg-s000683	*
	E	MI Required for A Glycomics Experiment - Mass Spectrometric Analysis (MIRAGE MS)	bsg-s000523	*
	E	Nuclear Magnetic Resonance ML (NMR-ML)	bsg-s000069	x&
	E	Quality Control ML (QCML)	bsg-s000570	*&
	E	eXtensible Experiment ML (XEML)	bsg-s000686	*x
	E	Nuclear Magnetic Resonance Controlled Vocabulary (nmrCV)	bsg-s000563	*x&
	P	Ontology of Glucose Metabolism Disorder (OGMD)	bsg-s002632	*
	D	Core Information for Metabolomics Reporting (CIMR)	bsg-s000175	*
	E	CDISC Laboratory Data Model (CDISCLAB)	bsg-s000165	
	P	Toxicology Data ML (ToxML)	bsg-s000539	*&
	D	Pharmacogenomics Ontology (PharmGKB-owl)	bsg-s002745	*
	D	Suggested Ontology for PHARMacogenomics (SOPHARM)	bsg-s000099	*
	P	Pharmacogenomic Relationships Ontology (PHARE)	bsg-s002697	*
	E	MI required for a DMET Experiment (MIDE)	bsg-s000628	*&
	E	Gel electrophoresis ML (PSI GelML)	bsg-s000087	*&
	E	Transition ML (HUPO-PSI TraML)	bsg-s000113	*x&
	E	Molecular Interaction Tabular (MITAB)	bsg-s000120	*
	E	Molecular Interaction eXtensible ML (PSI-MI XML)	bsg-s000121	*&
	E	PRIDE XML Format	bsg-s000561	*x&
	E	mzML	bsg-s000112	*&
	E	MzTab	bsg-s000693	*x
	E	Probed	bsg-s000694	X
	E	International HLA and Immunogenetics Workshop XML (IHIW XML)	bsg-s000700	&
	D	Proteomics Pipeline Infrastructure for CPTAC (CPTAC)	bsg-s002591	
	D	Proteomics data and process provenance (ProPreO)	bsg-s002622	

(*continued on next page*)

Table 8.3: List of omics standards found on FAIRsharing.org (accessed November 23, 2017)—cont'd

Classification		Name (Abbreviation)	FAIR ID	PTS
	D	ImMunoGeneTics Ontology (IMGT)	bsg-s002665	*
	E	HUPO PSI Mass Spectrometry Controlled Vocabulary (PSI-MS CV)	bsg-s000068	*
	E	Sample processing and separations controlled vocabulary (sepCV)	bsg-s000073	
	E	MI about a Molecular Interaction Experiment (MIMIx)	bsg-s000179	*
	E	MI About Sample Preparation for a Phosphoproteomics Experiment (MIASPPE)	bsg-s000180	*
	E	MI About a Proteomics Experiment (MIAPE)	bsg-s000184	*
	E	MI About a Peptide Array Experiment (MIAPepAE)	bsg-s000280	*
	E	MIAPE: Mass Spectrometry Quantification (MIAPE-Quant)	bsg-s000516	*
	E	MIAPE: Mass Spectrometry (MIAPE-MS)	bsg-s000607	*
	E	MIAPE: Mass Spectrometry Informatics (MIAPE-MSI)	bsg-s000608	*
	E	MIAPE: Gel Electrophoresis (MIAPE-GE)	bsg-s000609	*
	E	MIAPE: Gel Informatics (MIAPE-GI)	bsg-s000610	*
	E	MIAPE: Column Chromatography (MIAPE-CC)	bsg-s000611	*
	E	MIAPE: Capillary Electrophoresis (MIAPE-CE)	bsg-s000612	*
	E	Short Read Archive eXtensible ML (SRA-XML)	bsg-s000084	&
	E	MicroArray Gene Expression Tabular Format (MAGE-TAB)	bsg-s000080	*x
	E	Minimal Information about a high throughput SEQuencing Experiment (MINSEQE)	bsg-s000174	
	E	MI About a Microarray Experiment (MIAME)	bsg-s000177	*x
	E	Minimum Information about an ENVironmental transcriptomic experiment (MIAME/Env)	bsg-s000168	*
	E	MI About a Microarray Experiment involving Plants (MIAME/Plant)	bsg-s000182	*
	E	MI about a Nutrigenomics experiment (MIAME/Nutr)	bsg-s000190	*
	E	MI about an array-based toxicogenomics experiment (MIAME/Tox)	bsg-s000191	*
Green	Bioinfomatics	Yellow Glycomics	Red	Pharmacogenomics
Blue	Genomics	Orange Metabolomics	Purple	Proteomics
Pink	Transcriptomics	Cyan		Ontology
Brown	Data	Majenta		Reporting

PTS, publication, tool, schema; *, publication; x, tool; &, schema; D, domain; E, experimental; P, phenotype; MI, minimum information; ML, markup language.

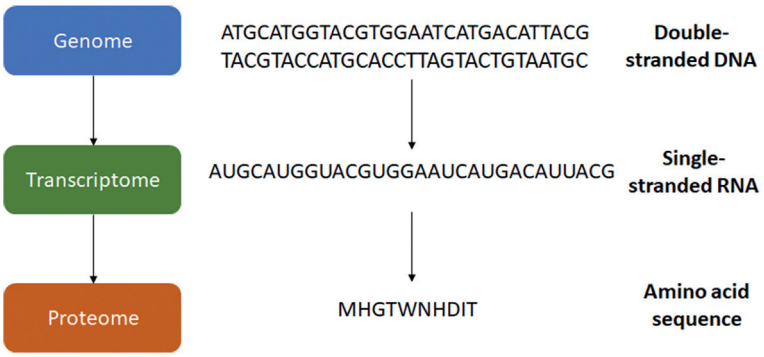

Figure 8.3
Data types associated with the Central Dogma.

8.2.1.1 Genomics

Genomics is the assembly of an organism's full complement of DNA.[13] A human genome is approximately 3Gb in size and contains DNA that will be encoded into RNA as well as DNA that will not be coded. The Global Alliance for Genomics and Health (**GA4GH**) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.[14] GA4GH has three focus areas: (1) genomic data toolkit; (2) regulatory and ethics toolkit; and (3) data security toolkit. Standards have been developed for both the structure of data in this field, such as the standard variant call format (**VCF**) format employed in next-generation sequencing technologies,[15] as well as the reporting of data in this field, such as the pharmacogenomics standard, the minimum information required for a DMET Experiment (**MIDE)** employed for reporting of data generated from the DMET Array.[16]

8.2.1.2 Transcriptomics

The transcriptome is the set of all RNA molecules in one cell or a population of cells.[13] RNA-Seq is the technique that produces sequencing data of the full complement of RNA, as opposed to microarrays that produce information about a preselected set of genesRNA-Seq experiments generate a large volume of raw sequence reads which have to be processed to yield useful information. Several transcriptomics standards have been developed under the **MIAME** (Minimum Information About a Microarray Experiment) brand,[17] including standards for nutrigenomic experiments, toxicogenomic experiments, and experiments involving plants.

8.2.1.3 Proteomics

Proteomics is the large-scale study of proteins in a cell.[13] A proteome is a set of proteins produced in an organism, system, or biological context. Like MIAME in transcriptomics, the

MIAPE (Minimum Information About a Proteomics Experiment) has also been adapted for several experimental techniques in protein research, including mass spectrometry, peptide array experiments, gel electrophoresis, gel informatics, and column chromatography.[18]

8.2.2 Data standardization and FAIR data

Data standardization is closely related to FAIR data as both aim to produce high-quality data, interoperable data, with FAIR including the added benefits of findability and accessibility.[19] Several cooperative projects manage both data standardization and FAIR data. The following section discusses the implementation of FAIR, as well as the key players and resources in FAIR and data standardization.

8.2.2.1 Implementation in FAIR

Data-intensive sciences facilitate knowledge discovery by assisting humans and computers to discover, access, integrate, and analyze data. The FAIR principles urge researchers to prepare for the concept of subsequent sharing and reuse of data from the outset.[19] A number of FAIR implementations are currently available.[12] Some of the most popular used ones are listed in Table 8.4.

1. **Findability** specifies that data should be identified, described, and registered or indexed in a clear and unambiguous manner. This implies that datasets are uniquely identifiable using persistent identifiers; that the main data properties are ideally specified using standards; and indexed in a public resource.
2. **Accessibility** specifies that datasets should be accessible through a step-by-step procedure for data access, preferably by automated means. In cases where the data are no longer available, relevant metadata should always be accessible.
3. **Interoperability** specifies that data as well as metadata are formulated, expressed, and organized using common, published standards; thus employing data standardization. Notably, these standards should themselves be made FAIR.
4. **Reusability** specifies the crux of other principles: the characteristics of the data, and their provenance, should be well described, according to the established standards for the field.

The FAIR framework must be backed by a number of policies. These policies constitute **legal interoperability** which is important aspect of this scientific advancement (illustrated by Fig. 8.4).[19]

- **citation policy**—specifying that researchers should give proper attribution and credit for data used.
- **fairness policy**—specifying that research data made publicly available should be equitably available to other research members.

Table 8.4: FAIR implementations relevant to the field of bioinformatics.

Resource	Description
DataVerse (https://dataverse.org)	Developed by the Institute for Quantitative Social Sciences (IQSS), this open-source web application allows for the sharing, preserving, citing, exploring, and analysis of research data. DataVerse assists in making data available to others, functioning as a container for datasets (research data, code, documentation, and metadata), which can be setup for individual researchers, departments, journals, and organizations.
FAIRDOMHub (https://fairdomhub.org/)[30]	A repository for publishing FAIR data, operating procedures and models for the systems biology community, accessible via the web, for storing and sharing systems biology research datasets. The platform assists with the management of collection management of data, storage and publishing your data, models, and operating procedures.
ISA (http://isa-tools.org/format/specification.html)[28]	ISA stands for investigation study and assay. It is a metadata tracking framework that facilitates standards-compliant collection, distribution, management and reuse in an increasingly diverse set of life science domains. ISA helps researchers to provide rich descriptions of experimental metadata (i.e., sample characteristics, technology and measurement types, sample-to-data relationships) so that the resulting data and discoveries are reproducible and reusable.
Open PHACTS (https://www.openphacts.org/)[29]	A data integration platform for information pertaining to drug discovery, enables researchers access to an expanse of pharmacological data from a simple interface. The primary aim of this platform is to reduce barriers to drug discovery in industry, academia and for small businesses. It contains all the data sources already in use, integrated and linked together in order to visualize the relationships between compounds, targets, pathways, diseases, and tissues.

- **access policy**—specifying that access to and reuse of research data should be open and unrestricted, or with as few limitations possible.
- **data usage policy**—should inform end users with associated terms and conditions of use.
- **usage rights policy**—establishes who or what entity has the rights to any given collection of data before the data are disseminated to other parties.

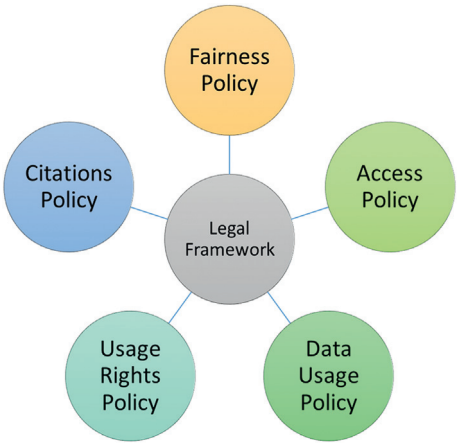

Figure 8.4
The subcomponents of a legal framework in support of FAIR.

8.2.2.2 Key players

Several key players exist in the research community which facilitates bioinformatics data standardization and the implementation of FAIR; many such key players are interrelated. These initiatives are summarized in Table 8.5.

8.3 Challenges to data standardization

Several challenges prevent the widespread implementation and adoption of omics standards in the biohealth research fields. These challenges can be broadly classified into two categories; (1) adoption and (2) policy.

8.3.1 Adoption challenges

The adoption of standards depends on several factors, including the availability and findability of the standards and associated materials and training events, education and training associated with standardization, community perception, the gaps, overlaps and maintenance of existing standards, and associated the technological requirements. These factors are further discussed hereafter.

Since data standardization is still in its infancy, many researchers have not adopted standardization because they are not aware of international standards relevant to their practice and how to implement those standards endorsed by a coordinating body.[20,21] **Education** and **training** are required to bridge this gap, therefore, emphasis needs to be placed on the teaching of standardization practices, either at late-undergraduate level or early postgraduate and scientific

Table 8.5: List of initiatives that drive FAIR principles.

Initiative	Description
ELIXIR (https://www.elixir-europe.org/)[31]	**ELIXIR** groups Europe's primary life science organizations in the supervision and maintenance of the increasing data volume produced by publicly funded research. It provides a stable infrastructure for bioinformatics resources for its member states, giving users in academia and industry access to services that are essential to conduct FAIR research.
CODATA (http://www.codata.org/)	The **Committee on Data of the International Science Council (CODATA)** is an international interdisciplinary scientific committee for the promotion of global collaboration to enhance the accessibility and usability of data in all areas of research.
RDA (https://www.rd-alliance.org/)	The **Research Data Alliance (RDA)** builds the social and technical infrastructure to enable open sharing of data. It provides a neutral space where its members can join dedicated global Working and Interest Groups to develop the infrastructure for data-sharing. It is associated with CODATA, ELIXIR, and RDA infrastructure and several community group initiatives.
Horizon 2020 (https://ec.europa.eu/programmes/horizon2020/en)	**Horizon 2020** is currently the biggest European research and innovation program. Standardization has been identified as one of the key measures to support this project.
CWA (https://www.nbic.nl/about-nbic/affiliated-organisations/cwa/)	The **Concept Web Alliance (CWA)** is a community collaboration that has been created to handle FAIR issues arising from "Big Data" generation, such as data storage, interoperability and performing analyses on big and disparate data.
BD2K (https://commonfund.nih.gov/bd2k)	The **Big Data to Knowledge (BD2K)** program supports the research and development of innovative approaches and tools to maximize and accelerate the utility of big data and data science in biomedical research. It facilitates broad use of biomedical big data, develops analysis methods and software for large-scale data analysis, and supports efforts toward making datasets FAIR.

career levels—particularly for those individuals pursuing or specializing in omics research. Trainings can facilitate the adoption of newly developed standards, and should be targeted at every person involved in the process of collecting, generating, analyzing, and managing data.

Perception that standardization provides little to no value to current research is a major barrier to adoption in the scientific community.[4] Some may perceive standardization or standard compliance as too complicated to implement. Similarly, scientific communities may also be resistant to change, particularly if a standard recommends practices far removed from their norm.[4] Comprehensive resources need to be consulted when creating a standard. Developing comprehensive standards which balance future value with immediate value is required, and this is where support by publishing and funding bodies can play a significant role.

Gaps may be present in existing standards that need to be addressed through the extension or merging of standards. If extending and merging do not create appropriate standards, new standards can be developed. This may also be achieved through comprehensive and appropriate referencing of aligning standards. In some cases, there also exists a high degree of **overlap** across existing and (or) developed standards. Operational criteria and evaluative studies are required in order to facilitate objective comparisons of competing and overlapping data standards, which may facilitate merging of standards, as mentioned above. During FAIRsharing's evolution, it has endeavored to address these concerns and reduce redundancy by manually curating from a variety of sources, including BioPortal, MIBBI, and the Equator Network.[20,21] It is crucial to develop innovative evaluation methods to implement these methods within newly developed standards and prevent redundancy and overlap prior to approval.

Maintaining existing standards can be a burden to curators responsible for housing standards. Regular and coordinated communication between standard-housing resources and standard-developing groups is required to maintain activity and relevancy, and reduce or eliminate variation and redundancy.[20,21] Standards need to be updated to reflect and account for the continuous development of research; otherwise they will not represent the latest knowledge. Currently, FAIRsharing carries the bulk of the maintenance load. Developers should aim to maintain standards with appropriate regularity and rigorous quality, with the procedures documented by the standards-housing resources.[22]

Finally, in order to realize widespread adoption of omics standards, a number of **technological requirements** need to be addressed, including the accessibility and usability of standards.[20,21] Tools and resources that illustrate competing standards and easing their use, in relation to well-defined work processes and tasks are needed. Particularly so, given the limited amount of education and training endeavors associated with data standardization.[4] Similarly, tools are needed to aid the evaluation of newly developed, as well as previously developed, standards, in order to reduce redundancy. Notably, an increasing amount of standard-developing groups are actively developing tools or schemas to accompany their standards, in order to bridge this gap and aid use.

8.3.2 Policy challenges

Although the scientific community is beginning to endorse the use of FAIR principles and data standardization, there is room for improvement.[20,21] To speed up the adoption of standardization, publishing and funding bodies should make compliance to standards obligatory. This will also highlight specializations or applications for which standards are yet to be created, reducing the development of redundant standards.

Access and **licensing** are important considerations when developing data standards. Open (or publically available) standards may be used by all without constraint, however, its origin must be acknowledged and it should not be altered and redistributed in the altered form under the original name or identifier. Many biomedical and omics standards are available openly via the OBO foundry, BioPortal, FAIRsharing, or directly from their hosting websites. Standards that have license restrictions can impede interoperability, limiting their widespread application and utilization.[22] Licenses are often recommended by the aforementioned resources, for example, OBO data must be released under a Creative Commons CC-BY license version 3.0 (or later) or into the public domain under CC0.

8.4 Executive summary

Emphasis on the practice of omics data occurred during the turn of the 21st century, once the scientific community appreciated the complexity associated with data diversity and realized the value of data comparison.[4] The first omics standard was created in 2001 (MIAME)[25], and standardization has since undergone a significant surge (as illustrated in Table 8.2). Compliance to standards is becoming an important requirement for funding bodies, scientific journals, and data repositories. Compliance with standards is required by several publishers (including *Nature, Cell*, and *Lancet*), funders, and repositories.[4] Similarly, publications are beginning to endorse the FAIRsharing initiative. Making data available and accessible in a comprehensive and consistent manner is a principal component to driving, effective data sharing, and open science.

Education and training on the available standards are required to facilitate the adoption of data standardization. There are too many overlaps between existing standards, for available technologies and applications; cohesive mechanisms of integration and community adoption are essential to prevent the redundancy.

Nonetheless, concerted efforts by international bodies have encouraged and endorsed the use of standards in the omics field. Standardization can impact all aspects of the research-data lifecycle, including data collection, storage, management, transformation, and analysis.[4] It can also facilitate data and research usability and enhance research collaboration. Most

importantly, such standardization can facilitate the elucidation of novel information across all aspects of omics and significantly contribute to our current understanding of living organisms.[4]

Acknowledgments

The authors would like to acknowledge the Human Heredity and Health of Africa (H3Africa) consortium and the H3Africa Bioinformatics Network (H3ABioNet). H3ABioNet is supported by the National Institutes of Health Common Fund under **Grant Number**: 2U24HG006941-06 and **FAIN**: U24HG006941. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors of this chapter have no conflicts of interest to declare.

References

1. Xie Z, Hall J, McCarthy IP, Skitmore M, Shen L. Standardization efforts: the relationship between knowledge dimensions, search processes and innovation outcomes. *Technovation*. 2016;48–49:69–78. doi:10.1016/j.technovation.2015.12.002.
2. Kim K. Clinical data standards in health care: five case studies [Internet]. California Health Care Foundation. 2005 (accessed August 8, 2017). http://www.chcf.org/.
3. Council NR. National science education standards [Internet]. 1996 (accessed December 10, 2017). https://www.nap.edu/catalog/4962/national-science-education-standards.
4. Chervitz SA, Deutsch EW, Field D, et al. Data standards for omics data: the basis of data sharing and reuse. *Methods Mol Biol Clifton NJ*. 2011;719:31–69. doi:10.1007/978-1-61779-027-0_2.
5. Holmes C, McDonald F, Jones M, Ozdemir V, Graham JE. Standardization and omics science: technical and social dimensions are inseparable and demand symmetrical study. *Omics J Integr Biol*. 2010;14:327–332. doi:10.1089/omi.2010.0022.
6. Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. *Nat Genet*. 2012;44:121. https://doi.org/10.1038/ng.1054.
7. Boja ES, Kinsinger CR, Rodriguez H, Srinivas P. Integration of omics sciences to advance biology and medicine. *Clin Proteomics*. 2014;11. doi:10.1186/1559-0275-11-45.
8. Kiechle FL, Zhang X, Holland-Staley CA. The -omics era and its impact. *Arch Pathol Lab Med*. 2004;128:1337–1345. doi:10.1043/1543-2165(2004)128<1337:TOEAII>2.0.CO;2.
9. Oberkampf H, Gojayev T, Zillner S, Zühlke D, Auer S, Hammon M. From symptoms to diseasescreating the missing link. In: *The Semantic Web Latest Advances and New Domains [Internet]*. Springer, Cham; 2015 (accessed December 20, 2017). p. 652–667. (Lecture Notes in Computer Science). https://link.springer.com/chapter/10.1007/978-3-319-18818-8_40.
10. Boissier M-C. Benchmarking biomedical publications worldwide. *Rheumatology (Oxford)*. 2013;52:1545–1546. doi:10.1093/rheumatology/ket181.
11. Fan J, Han F, Liu H. Challenges of big data analysis. *Nat Sci Rev*. 2014;1:293–314. doi:10.1093/nsr/nwt032.
12. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The fair guiding principles for scientific data management and stewardship. *Sci Data.* 2016;3:sdata201618. doi:10.1038/sdata.2016.18.
13. Field D, Amaral-Zettler L, Cochrane G, et al. The genomic standards consortium. *PLoS Biol*. 2011;9:e1001088. doi:10.1371/journal.pbio.1001088.

14. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;18:83. doi:10.1186/s13059-017-1215-1.
15. Global Alliance for Genomics and Health. GENOMICS.A federated ecosystem for sharing genomic, clinical data. *Science*. 2016;352:1278–1280. PubMed PMID: 27284183. doi:10.1126/science.aaf6162.
16. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–2158. doi:10.1093/bioinformatics/btr330.
17. Kumuthini J, Mbiyavanga M, Chimusa ER, et al. Minimum information required for a DMET experiment reporting. *Pharmacogenomics*. 2016;17:1533–1545. doi:10.2217/pgs-2016-0015.
18. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*. 2001;29:365–371. doi:10.1038/ng1201-365.
19. Martínez-Bartolomé S, Binz PA, Albar JP. The minimal information about a proteomics experiment (MIAPE) from the proteomics standards initiative. *Methods Mol Biol*. 2014;1072:765–780. PubMed PMID: 24136562. doi:10.1007/978-1-62703-631-3_53.
20. Durinx C, McEntyre J, Appel R, et al. Identifying elixir core data resources. *F1000Res*. 2017;5(ELIXIR):2422. doi:10.12688/f1000research.9656.2.
21. Sansone S-A, Rocca-Serra P. On the evolving portfolio of community-standards and data sharing policies: turning challenges into new opportunities. *GigaScience*. 2012;1:10. doi:10.1186/2047-217X-1-10.
22. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc*. 2007;14:687–696. doi:10.1197/jamia.M2470.
23. Smith B, Ashburner M, Rosse C, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007;25:1251. doi:10.1038/nbt1346.
24. Bandrowski A, Brinkman R, Brochhausen M, et al. The ontology for biomedical investigations. *PLoS One*. 2016;11(4:e0154556). https://doi.org/10.1371/journal.pone.0154556.
25. Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*. 2009;37(Web Server issue):W170–W173. doi:10.1093/nar/gkp440.
26. Dolin RH, Alschuler L, Beebe C, et al. The HL7 clinical document architecture. *J Am Med Inform Assoc*. 2001;8:552–569.
27. Taylor CF, Field D, Sansone S-A, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol*. 2008;26:889–896. doi:10.1038/nbt.1411.
28. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database J Biol Databases Curation*. 2016;2016. doi:10.1093/database/baw075.
29. Sansone SA, McQuilton P, Rocca-Serra P, et al. FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol*. 2019;37:358–367. doi:10.1038/s41587-019-0080-8.
30. Boeckhout M, Zielhuis GA, Bredenoord AL. The fair guiding principles for data stewardship: fair enough? *Eur J Hum Genet*. 2018;26:931–936. doi:10.1038/s41431-018-0160-0.
31. Digles D, Zdrazil B, Neefs J-M, et al. Open phacts computational protocols for in silico target validation of cellular phenotypic screens: knowing the knowns. *Med Chem Commun*. 2016;7:1237–1244. https://doi.org/10.1039/C6MD00065G.
32. Pandis N, Fedorowicz Z. The international equator network: enhancing the quality and transparency of health care research. *J Appl Oral Sci*. 2011;19. doi:10.1590/S1678-77572011000500001.
33. Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, et al. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Res*. 2017;45:D404–D407. https://doi.org/10.1093/nar/gkw1032.
34. Malone J, Stevens R, Jupp S, Hancocks T, Parkinson H, Brooksbank C. Ten simple rules for selecting a bio-ontology. *PLOS Comput Biol*. 2016;12:e1004743. doi:10.1371/journal.pcbi.1004743.

CHAPTER 9

Data sharing: The public's perspective

James C. O'Leary
Independent Scholar

Since the completion of the Human Genome Project, it has become increasingly clear that genotypic data loses much of its value for discovery when stripped of its associated phenotypic data. When the scope of inquiry is artificially limited, it can become difficult to understand the underlying mechanisms at work. The same can be said about the public's perspective on genomic data sharing. A hyper-focus on public opinions regarding genomics in the absence of an understanding of people's experiences with data in general is both limiting and short-sighted. A review of the existing literature on the topic immediately reveals how quickly the world has changed since many of the related studies were conducted. While the participants in many focus groups and studies share their own perspectives on how genomic data relates to other data types, the studies themselves display a level of genetic-exceptionalism that is disappearing as precision medicine becomes not only a goal, but a reality.

While precision medicine holds great promise to improve outcomes, it also presents a fundamental shift in how medicine is practiced. It requires larger datasets, the successful interface between different data types, and the application of that information to clinically impactful interventions. In this new world, genomic data are not used in isolation but are powerful as part of a rich mix of data. All of that data working harmoniously to create better care will require investment in healthcare infrastructure, institutional commitment, support from the public, and patient/participant buy-in. Like all innovations, this transition could mean more clarity for participants or it could mean more confusion.

The public's perspective is important for a number of reasons. Lack of support can lead to burdensome regulatory hurdles, legal battles, refusal to share data, loss to follow-up, and a chilling effect on participation. And, most importantly, if the public does not see the value in a practice, it may point to the fact that the practice itself is not aligned with the outcomes that matter most to patients.

How can the public's perspective be understood when the population is made up of a myriad of different communities, cultures, and individual experiences? The goal of this chapter is not to serve as a reference point for how to collect and share genomic data, but rather to give

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00004-3
Copyright © 2023 Elsevier Inc. All rights reserved.

context to relevant case studies and a framework for how to approach the issue in the age of precision medicine, including:

- What are the publics concerns and motivations regarding the sharing of genomic and health data?
- How does the public feel about those aspects that make genomic data unique?
- Are there differences in perspective based on traits such a community and disease state?
- What is the context in which participants approach data and data sharing?
- What gaps still exist in the research community's understanding of this issue?

9.1 Public willing to participate?

While overall participation rates in medical research are low on a national level, successful attempts at recruitment by community-based organizations and researchers, health systems, and direct-to-consumer companies have shown that many more people are willing to contribute their data than currently do. Even as pharmaceutical companies spend millions of dollars in attempts to fill existing clinical trials, thousands of patient communities have self-organized to accelerate research on their conditions. When individuals are properly engaged, they are motivated to participate in research. And this willingness does include genomic data. But while participants are shown to be willing to contribute data for disease-specific registries and biobanks, many aspects of precision medicine require research on large databases of genomic information collected across multiple sources. This fact is reflected in the National Institutes of Health policy that requires recipients of funding to obtain consent for broad sharing of genomic data in the majority of cases.[1] While such policies help accelerate scientific inquiry, they also create barriers for individuals with privacy concerns or lack of trust in the research enterprise. Given this reality, the public's perspective must be looked at both in terms of willingness to contribute data for research and comfort regarding broad sharing of their data.

One of the most comprehensive surveys on public attitudes toward genomic data sharing in biobank research was conducted with 11 institutions in the Electronic Medical Records and Genomics (eMERGE) Network in 2017. Among the 13,000 respondents, trust of institutions and researchers was generally high (greater than 60%), with a majority of respondents reporting willingness to contribute their data to a biobank (66%).[2] In fact, across numerous studies, most participants expressed high levels of trust in their own health care institutions when asked, but with varying levels of confidence that their data could realistically be kept secure.[3] When it comes to participation in research, concerns tend to fall into a few, consistent areas. A 2014 review of literature on attitudes of research participants and the general public toward genomic data sharing broke attitudes down into four areas: perceptions of sensitivity and controllability of genomic data, governance level considerations, personal perceptions of potential risks, and personal perceptions of potential benefits.[4] Those who are unwilling to

participate cite risks such as insurance and employment discrimination, marketing companies targeting participants for the sale of products, and improper access and use of that data by the government or law enforcement.[5]

While most studies tend to focus on reasons why people won't share their data, it is equally important to analyze why they do. The few formal studies that ask this question for genomic data frequently point to motivations for participation that are consistent with those for research in general: a mix of personal hope for a treatment where none exists, a desire to help others with a similar condition, or a wish to provide a benefit to society.[2,6] And, when clearly presented with the goals of a research study, participants are shown to more strongly identify with benefits over the perceived risks.[6] But more in-depth analysis of why people participate and how they are connected to research opportunities is severely lacking. It is likely that motivations such as the desire for privacy, tolerance for risk, and trust will change over time based on life experiences. Similarly, a diagnosis with a serious medical condition can have a major impact on privacy concerns. For example, the individuals with Amyotrophic Lateral Sclerosis share their data with registries and biobanks and participate in clinical trials at very high rates.[7,8] With only 25% of individuals with ALS living more than 5 years after diagnosis, the promise of open data leading to a treatment vastly outweighs the privacy concerns for most involved.

9.2 Concerns unique to genomic data?

Public awareness of genomics has grown.[9] Americans are increasingly exposed to genomics through representations in popular media, advertising for ancestry testing, and the rise of widely used genetic services such as prenatal screening. Unfortunately, rising awareness does not necessarily translate to improved knowledge of how genomic information can be used.[10] One 2017 study of the US population using a national dataset showed that only 57% of respondents were aware of genetic testing and a minority were knowledgeable about the use of genetics to determine treatment or drug efficacy.[11] In other words, participants were not aware that genetics could be used to fuel precision medicine. Without a proper understanding of the potential of genomics, it is also difficult to understand the risks. Given the genetic literacy of the population, should there be separate rules applied to the sharing of genomic data? The answer to that question is largely dependent on whether the aspects that make genomic data unique are of particular concern to the public. In most studies, concerns fell into two related areas: privacy, confidentiality, control, and security of data and trust relationships with researchers or health institutions.

9.2.1 Data concerns

Compared to opinions on data sharing, the privacy of genetic/genomic data and how it affects willingness to participate has been more thoroughly researched, albeit with significant gaps.

A 2017 literature review on the individuals' perspectives on genetic information found 53 studies covering 47,974 participants. One of the key components of risk cited by all participants was privacy, though respondents to studies on the topic frequently conflated privacy, confidentiality, control, and security. Individuals reported greater concerns about both medical and genomic information being shared with employers, insurers, and the government than with other researchers and commercial entities.[3] Many of the concerns found throughout the literature are comparable to those for health data use in general, with the primary focus being on the discrimination and stigma that could result from identification.

In addition, these concerns have remained remarkably consistent over time. Fears of genetic discrimination led to the passage of the Genetic Information Nondiscrimination Act (GINA) in 2008, which prohibits the use of predictive genetic information in health insurance or employment. And perhaps more impactful to US citizens are the provisions in the Affordable Care Act and Americans With Disabilities Act, which prohibit discrimination for existing conditions in health insurance and employment, respectively. Unfortunately, awareness of GINA remains low and the public is well aware of the precarious nature of these protections given the current political climate. In addition, though research participants have expressed concern over access to life and disability insurance, there are no prohibitions against the use of genetic information for that purpose.[12]

Perhaps the most unique aspect of genomic data is its ability to provide information about relatives, including paternity and maternity. This connection creates downstream opportunities for care but also privacy concerns that have not been sufficiently addressed in common data use practices. This fact extends many of the privacy and confidentiality concerns of genomic data to family members without their consent. This fact provides both an opportunity and a concern for the use of such data by law enforcement. Because genetic profiles can be used to identify perpetrators of violent crimes where DNA evidence is available, large curated genealogical databases are a valuable tool. In a particularly well-publicized case, the Golden State Killer was identified and arrested by the FBI in 2018 by first identifying his relatives using a free, online genetic database populated by individuals researching their family trees. Research following the case has shown a mix of results. While some surveys have shown broad support for the use of such searches in the apprehension of violent criminals,[13] in an online survey of 1046 customers of three DTC genetic testing companies, nearly all respondents favored a policy to ensure that insurers and law enforcement officials could not access their information.[14] In addition, the use of genetic information by law enforcement may cause particular concern to individuals in communities that have a history of experiencing police violence or have faced threats of deportation.

9.2.2 Matters of trust

As numerous genomic studies have pointed out, willingness to participate is lower among individuals with lower levels of trust. While not surprising, trust is important because it varies

significantly based upon a number of key demographics, such as race, religiosity, and education.[2,15,16]

Given the lack of diversity in reference databases and genome association studies, barriers to the recruitment of minority populations have been given particular attention.[17–19] There are stark public examples of unethical research methods, such as early eugenics research and the Tuskegee Syphilis Study, that reinforce distrust of researchers and the healthcare system, particularly among people of color. Research has documented that some racial and ethnic groups have less trust in medical researchers,[20,21] with some studies confirming significantly lower willingness to participate in genomics studies.[2,22] However, the relationship between distrust and participation in trials has been shown to be much more complicated, with recent authors pointing out that distrust by minority populations does not fully correlate to unwillingness to participate. Societal causes such as segregation within the healthcare system, lack of diversity of investigators, and socioeconomic barriers to accessing trials may play a significant role.[23]

Genomic medicine, like all precision medicine, requires access to data for dual purposes: the generation of the large datasets used to power effective care and the application of those datasets to an individual case to create personalized interventions. While participants cite benefit to others as a key reason to participate in medical research, the community to which a participant belongs may not always receive the benefits derived from that research. For example, studies have shown that people of color receive lower referral rates for genetic testing and cancer genetic services.[24–26] This, in turn, has the downstream result of potentially excluding them from relevant clinical studies and registries, further aggravating the problem.

Finally, certain types of research, such as studies on ancestry, lineage, and mental health, create unique concerns for some ethnic or religious communities. While these concerns range significantly, serious harm can be done to trust when customs, beliefs, and attitudes of communities are not properly understood or respected. This creates particular challenges when broad consents are required for research, leading to nonparticipation which potentially further disenfranchises vulnerable populations. The most high-profile case involving genomic data was the 2004 lawsuit filed by Havasupai Tribe against the Arizona Board of Regents and Arizona State University. The suit alleged that research was inappropriately conducted using samples that were originally collected and consented for studies on diabetes. The case reached a settlement in 2010 in the tribe's favor and DNA samples were returned to be tribal members to be disposed of in a culturally appropriate way.[27] This example highlights not only the importance of proper informed consent, but ethical issues present in the deidentification and even anonymization of data. In the case of the Havasupai Tribe, not only did the samples themselves carry a deeply religious significance, but the identification of the samples as coming from an ethnic community had imported, even if the specific individual could not be identified. Furthermore, this and other examples significantly impeded future research collaborations with the Havasupai and other indigenous communities.

9.3 Support for broad data sharing

Widespread data sharing and broad consents are particularly complicated issues to understand because the long-term implications are poorly understood by both participants and researchers. Numerous focus groups and studies highlight concerns around blanket consent.[16,28–30] Typically though, biorepositories that use deidentified data receive widespread support as long as consent is granted by participants, especially when those biorepositories are managed by an organization or institution that participants trust.[2,15] Trust in organizations varied across studies, but was generally higher for nonprofits or the institution from which the individual received care and was lower for government or pharmaceutical companies.[3] In a 2016 systematic literature review of individuals' perspectives on broad consent and data sharing, it was found that broad consent was often selected over tiered or study-specific consent across 48 studies.[1] Similarly, participants in randomized studies have shown willingness to sign a blanket consent in equal numbers to other models if only one option is given.[2]

Even those who share their data openly still hold privacy concerns. A 2017 study looked at people who obtained their genomic data from a direct-to-consumer company and then voluntarily decided to share it on the publicly accessible web platform openSNP. They found that respondents were concerned about privacy, but found the risk to be reasonably acceptable.[31] And yet, it has been clearly shown that participants tend to favor models of where they are offered choice, even if that choice is simply to opt-out of participation.[16,28,32,33] Notably, depending on the data collected or the hosting institution, many biorepositories may be exempt from or fall outside of federal regulations that provide for the protection of human subjects. Little data exists on what proportion of these biorepositories voluntarily follow all applicable guidelines.

As with any consent, there is significant concern that people do not fully read or understand the implications of broad data consents. After 23andMe announced a research partnership with GlaxoSmithKline in 2018, a University of Michigan research team conducted a nationally representative survey of the US adult population. The results of that survey highlighted participants discomfort with commercialization of data, with 67% agreeing that clear notification of potential commercialization of biospecimens is warranted but only 23% comfortable with such use.[34] Individuals who contribute data for the public good have particular concerns about the commercialization of data from biobanks.

Similarly, the limits of public support for data sharing were exhibited in 2010 by the destruction of more than five million newborn screening bloodspots as part of a settlement reached between the state of Texas and a civil rights group representing a group of parents. In that case, bloodspots were collected legally by the state as part of its public health program. IRB review was conducted for any research using deidentified samples and prior consent was obtained for the small amount of research conducted with identifiable samples. Interestingly,

the plaintiffs and their attorney were open to discussion about introducing a model of consent for research.[35] Despite that fact, the fear of public backlash against the state's public health program was enough to lead to the loss of an important resource. While research uses of such databases are naturally limited by fear of such public response to using samples collected under a public health mandate, their value for truly representative population data cannot be ignored.

In recent years, a number of high-profile programs have attempted to create large representative biorepositories, including genomic data, with broad consent already in place. And, in many cases, they have exceeded recruitment from similar efforts. These programs offer return of ancestry information, wellness reports, and even clinically useful information to participants, raising questions of coercion or undue influence. While these questions are important, in many ways they ignore the reality of precision medicine, which draws its data as much from clinical sources as it does from independent research studies. The modern model of the learning health system more closely matches the public's assumptions about care—that it is evidence-based—and the existence of large, consolidated health systems upend many of the previous realities that kept patient information close to its original source.

These recent programs have used mechanisms that maximize participant benefit and/or convenience and give individuals some degree of control; for example, how and if their data are shared and the option to remove their data at any time. Rightly so, there has been a significant focus on what constitutes informed consent in these instances, but it is equally clear that people are participating in record numbers. Thousands of patient communities have self-organized to host registries using variable methods of governance and recruitment. Geisinger's MyCode Community Health Initiative has consented to over 250,000 individuals out of the ~2 million patients in the health system,[36] and 23andMe, Ancestry.com, and other direct-to-consumer companies have seen rapid growth. In addition, more advanced models of consent and shared benefit have emerged in recent years in the form of distribution of ownership shares to data contributors, downstream compensation of contributors for data sales, and dynamic, granular consent.

9.4 A question of context

In order to make responsible and proactive decisions regarding genomic data sharing, it is important to understand trends in the public's relationship with data in general. When an individual is asked to give any consent to share their data, let alone broad consent, they are doing so within the context of their lives and experiences. And for most of the public, their experiences with sharing, privacy, and security of data have changed dramatically during the course of their lives and will continue to do so. While the adage "privacy is dead" is commonly heard, the reality is much more complicated. Even as the public has come to accept that their data

are constantly being collected, support for the practice is still low. In 2014, the Pew Charitable Trust reported that 91% of adults in their survey "agree" or "strongly agree" that consumers have lost control over how personal information is collected and used by companies.[37] This feeling is the main impetus behind an increasing array of privacy regulations, including the EU-based, but globally enforced, General Data Protection Regulation and the California Consumer Privacy Act. And the debate around privacy, already almost omnipresent in Europe, is accelerating in the United States, with national coverage of the Cambridge Analytica scandal and congressional testimony by familiar technology titans like Mark Zuckerberg.

And yet, people are still willing to share their data with those they trust. The good news is that the public's trust in medical researchers has grown in recent years. In a 2019 Pew Research Poll of US adults, 87% reported either a great deal (35%) or a fair amount (52%) of confidence in medical scientists to act in the best interests of the public.[38] Additionally, consumers are frequently willing to trade access to their data for services or convenience, and data as a currency are a concept with which most people are familiar. It is, in fact, control that the public is after. A follow-up 2015 survey found that more than 90% of respondents reported that being able to control who can get information about them and what information is collected is important. And health information is ranked second only to social security numbers as the most sensitive data type.[39]

Unfortunately, who has control over health data and where data are being shared remains a mystery to most people. While health information is typically viewed as more tightly regulated and protected, few studies have looked at the public's knowledge of existing regulations and human subjects' protections outside of specific use cases. Furthermore, there are few regulations and little public information available about the massive secondary market for healthcare data, despite the fact that deidentified datasets are purchased from HIPAA-covered entities including health systems, pharmacies, and in some cases EHR vendors.[40,41] Even those individuals who read HIPAA disclosures and informed consent documents are left with little true understanding beyond the primary uses of their data. Given the range and extent of data sharing and the increasingly cloudy distinction between clinical and research data, that is not surprising.

Despite this fact, the public is becoming increasingly aware that the security of their health data is at risk. A 2019 Politico/Harvard T.H. Chan School of Public Health poll found that a majority of Americans polled were somewhat or very concerned that unauthorized people may be able to access the state of their health and the medications they take. Similarly, public concern over the security of their data when it is being shared with third parties has remained constant. In a survey conducted by Office of the National Coordinator for Health Information Technology each year between 2013 and 2017, over 65% of individuals consistently reported having concerns regarding unauthorized viewing of their information when it is electronically exchanged.[42] And the public is right to be concerned. *HIPAA Journal*, which

compiles healthcare data breach statistics, reports that between 2009 and 2018 there have been 2546 healthcare data breaches involving more than 500 records, with resulting exposure of 189,945,874 healthcare records covering an estimated 59% of the population of the United States.[36] And those reports are limited to instances where the institution discovered the data breach. While similar statistics are not available for research data and biobanks, larger datasets will inevitably be targeted by hackers.

Even these estimates of data breaches are largely centered on data that includes common identifiers. According to the HIPAA Breach Notification Rule, "the impermissible use or disclosure of PHI is presumed to be a breach unless you demonstrate there is a low probability the PHI has been compromised."[43] But much attention has been given to the ability of deidentified health data, including genomic data, to be linked back to an individual's identity using public and summary data.[44,45] Recent studies have suggested that deidentified open health datasets, genomic datasets, and even single datasets from wearable devices can be reidentified even after aggregation and removal of protected health information.[46,47] Furthermore, brokers who buy and sell medical data have long been in the business of aggregating multiple datasets and compiling them under unique identifiers, making reidentification much easier. Given the massive amount of financial and personal health information that has already been inappropriately accessed, consumers have very real cause for concern. Given this reality, sharing data across multiple entities can only heighten the risk of data breach and reidentification.

Alternately, while data are exiting the healthcare systems for profit, there is little consistency for how it is being used internally. Though the term "learning healthcare system" has recently come into vogue, the concept is not a new one. In fact, the novelty of the term in the medical world may seem confusing to most patients. Trust in the medical establishment is based upon a perception that the doctor is the expert. And for most people, that includes an assumption that the doctor is uniquely qualified to deliver care that is both evidence-based and tailored to the individual. Unfortunately, the reality is far more complicated. The evidence base for many interventions is lacking and cost-effectiveness data are even more rare. And while patients are familiar with the concept that drugs undergo a rigorous evaluation through clinical trials before being approved, their knowledge does not extend to other interventions. In fact, though individuals overwhelmingly support using their data to improve care for future patients, those same individuals report a significant gap between their interest in discussing medical evidence with their doctor and what actually occurs.[48] Despite the fact that healthcare analytics are a billion-dollar industry, it is unclear how much of that processing power is being used to improve care versus lower costs.

Within this context, the low level of participation in research is not surprising. In fact, it is a testament to the desire of the public to improve the lives of others that so many people are willing to contribute their information within the current environment. If the goal is to truly create a culture of participation, the healthcare and research community must address these issues and delve more deeply into the motivations and interests of the public.

9.5 Policy for the people

It is critically important that the healthcare and research communities continue to work to earn the public's trust. Unfortunately, current data-sharing policy and infrastructure priorities are not always aligned with the needs of individuals and families. Patients are frequently in a position where lack of data sharing creates gaps in information or redundancy of care. A 2018 poll by the Office of the National Coordinator for Health Information Technology showed that nearly one-third of individuals who went to a doctor in the past 12 months reported experiencing a gap in information exchange.[42] Patients who face challenges getting the most basic of data sharing requests met—data shared between their care team—may be surprised to know that there are mechanisms already in place to sell their data to third parties and share it with national registries. Data portability and interoperability requirements that would address these issues, once a hallmark of the meaningful use standards that drove the expansion of health IT in the United States, were eventually stripped under pressure from vendors and hospitals.

Similarly, many of the current institutional and regulatory policies regarding data use ignore both public confusion around deidentification of data and how clinical and research data use has changed over time. Simultaneously, these policies seem to hamper research by discouraging the collection of the longitudinal data needed for precision medicine. As evidence of this fact, condition-specific, aggregated datasets collected by data brokers are incredibly valuable precisely because they provide a source of longitudinal data for identifying unmet needs, targeting therapeutic development, and adjusting coverage decisions. And yet researchers tend to favor broad consents, not just out of a desire to contribute to national databases, but also because it gives them more latitude to use the data without re-engaging research participants. Ironically, the research culture that has been built is one where the richness and humanity of datasets are stripped away under the guise of privacy only to be reconstituted through machine learning and sold. These ethical and logical inconsistencies make it difficult to argue that the interests of the public are being prioritized.

New policies are needed that better reflect the current reality of clinical and research data and the interests of the public. And these policies must include the wide range of players who currently interact with consumer health data. Technology companies are arguably one of the biggest holders of health data, from search history to purchasing habits to contributed data and biospecimens, and yet many of the human subjects protections they practice are voluntary. In fact, there is no comprehensive privacy legislation in the United States and significant gaps are exposed when compared to regulations in other jurisdictions.[49]

9.6 Further research

As data sharing continues to accelerate and datasets become more robust, it is increasingly important that people are viewed, not just as participants, but also as stakeholders in the

research process. To do so, it is important that decisions are based on a thorough understanding of the people's awareness, interest, and discomfort around data sharing and the broader research enterprise. Most national studies about public perceptions of research solely focus on enrollment and the factors that influence it. In order to effectively make this transition, it is important to close research gaps in the understanding of participants' motivations and concerns.

In particular, four areas of study have received little attention:

- How does the public think their data is currently used?
- How do individuals and communities want their data to be used?
- How do perceptions of benefits impact the participant' willingness to contribute data and provide consent for broad data sharing?
- What level of risk are individuals willing to accept when they are properly informed and how does that change over time?

Answering these questions can and should inform the development of data sharing infrastructure, including education, consent, and security. In addition, a new national conversation involving a diverse group of stakeholders is needed around the appropriate balance of risk and benefit. The public is supportive of research to improve human health and is already making decisions that balance their desire to contribute data for the public good against concerns for the security of their data.

When combined with other clinical and personal health information, genomic data can serve as the lynchpin that connects discovery and care. Therefore, it is imperative that the trust of the public is earned through responsible use and handling of this data. Because individual experience with and perceptions of data sharing transcend medical research, decisions about genomic data cannot be made in a vacuum. If precision medicine is to be successful, it is important that investment, institutional policy, legislation, and regulation align to create a trustworthy and secure system of data sharing that benefits all.

References

1. Garrison NA, Sathe NA, Antommaria AH, et al. A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. *Genet Med*. 2016;18:663–671. Epub 2015/11/19; PubMed PMID: 26583683; PubMed Central PMCID: PMCPMC4873460. doi:10.1038/gim.2015.138.
2. Sanderson SC, Brothers KB, Mercaldo ND, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. *Am J Hum Genet*. 2017;100:414–427. Epub 2017/02/09; PubMed PMID: 28190457; PubMed Central PMCID: PMCPMC5339111. doi:10.1016/j.ajhg.2017.01.021.
3. Clayton EW, Halverson CM, Sathe NA, Malin BA. A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. *PLoS One*. 2018;13:e0204417.

Epub 2018/10/31; PubMed PMID: 30379944; PubMed Central PMCID: PMCPMC6209148. doi:10.1371/journal.pone.0204417.

4. Shabani M, Bezuidenhout L, Borry P. Attitudes of research participants and the general public towards genomic data sharing: a systematic literature review. *Expert Rev Mol Diagn*. 2014;14:1053–1065. Epub 2014/09/26; PubMed PMID: 25260013. doi:10.1586/14737159.2014.961917.
5. Middleton A, Niemiec E, Prainsack B, et al. Your DNA, your say': global survey gathering attitudes toward genomics: design, delivery and methods. *Per Med*. 2018;15:311–318. Epub 2018/06/01; PubMed PMID: 29856292. doi:10.2217/pme-2018-0032.
6. Oliver JM, Slashinski MJ, Wang T, Kelly PA, Hilsenbeck SG, McGuire AL. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics*. 2012;15:106–114. Epub 2011/12/30; PubMed PMID: 22213783; PubMed Central PMCID: PMCPMC3318928. doi:10.1159/000334718.
7. Atassi N, Berry J, Shui A, et al. The pro-act database: design, initial analyses, and predictive features. *Neurology*. 2014;83:1719–1725. Epub 2014/10/08; PubMed PMID: 25298304; PubMed Central PMCID: PMCPMC4239834. doi:10.1212/WNL.0000000000000951.
8. DasMahapatra P, Raja P, Gilbert J, Wicks P. Clinical trials from the patient perspective: survey in an online patient community. *BMC Health Serv Res*. 2017;17:166. Epub 2017/02/27; PubMed PMID: 28241758; PubMed Central PMCID: PMCPMC5327530. doi:10.1186/s12913-017-2090-x.
9. Haga SB, Barry WT, Mills R, et al. Public knowledge of and attitudes toward genetics and genetic testing. *Genet Test Mol Biomarkers*. 2013;17:327–335. Epub 2013/02/13; PubMed PMID: 23406207; PubMed Central PMCID: PMCPMC3609633. doi:10.1089/gtmb.2012.0350.
10. Hann KEJ, Freeman M, Fraser L, et al. Awareness, knowledge, perceptions, and attitudes towards genetic testing for cancer risk among ethnic minority groups: a systematic review. *BMC Public Health*. 2017;17:503. Epub 2017/05/25; PubMed PMID: 28545429; PubMed Central PMCID: PMCPMC5445407. doi:10.1186/s12889-017-4375-8.
11. Krakow M, Ratcliff CL, Hesse BW, Greenberg-Worisek AJ. Assessing genetic literacy awareness and knowledge gaps in the US population: results from the health information national trends survey. *Public Health Genomics*. 2017;20:343–348. Epub 2018/05/31; PubMed PMID: 29852491; PubMed Central PMCID: PMCPMC6095736. doi:10.1159/000489117.
12. Parkman AA, Foland J, Anderson B, et al. Public awareness of genetic nondiscrimination laws in four states and perceived importance of life insurance protections. *J Genet Couns*. 2015;24:512–521. Epub 2014/09/23; PubMed PMID: 25242499; PubMed Central PMCID: PMCPMC4702480. doi:10.1007/s10897-014-9771-y.
13. Guerrini CJ, Robinson JO, Petersen D, McGuire AL. Should police have access to genetic genealogy databases? capturing the golden state killer and other criminals using a controversial new forensic technique. *PLoS Biol*. 2018;16:e2006906. Epub 2018/10/02; PubMed PMID: 30278047; PubMed Central PMCID: PMCPMC6168121. doi:10.1371/journal.pbio.2006906.
14. Bollinger JM, Green RC, Kaufman D. Attitudes about regulation among direct-to-consumer genetic testing customers. *Genet Test Mol Biomarkers*. 2013;17:424–428. Epub 2013/04/06; PubMed PMID: 23560882; PubMed Central PMCID: PMCPMC3634146. doi:10.1089/gtmb.2012.0453.
15. Brothers KB, Morrison DR, Clayton EW. Two large-scale surveys on community attitudes toward an opt-out biobank. *Am J Med Genet A*. 2011;155A:2982–2990. Epub 2011/11/07; PubMed PMID: 22065592; PubMed Central PMCID: PMCPMC3222722. doi:10.1002/ajmg.a.34304.
16. Kaufman DJ, Murphy-Bollinger J, Scott J, Hudson KL. Public opinion about the importance of privacy in biobank research. *Am J Hum Genet*. 2009;85:643–654. Epub 2009/10/29; PubMed PMID: 19878915; PubMed Central PMCID: PMCPMC2775831. doi:10.1016/j.ajhg.2009.10.002.
17. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. *Nature*. 2007;445:259. PubMed PMID: 17230172. doi:10.1038/445259a.
18. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends Genet*. 2009;25:489–494. PubMed PMID: 19836853. doi:10.1016/j.tig.2009.09.012.

19. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. *Nat Rev Genet*. 2010;11:356–366. PubMed PMID: 20395969; PubMed Central PMCID: PMCPMC3079573. doi:10.1038/nrg2760.
20. Shavers VL, Lynch CF, Burmeister LF. Knowledge of the Tuskegee study and its impact on the willingness to participate in medical research studies. *J Natl Med Assoc*. 2000;92:563–572. PubMed PMID: 11202759; PubMed Central PMCID: PMCPMC2568333.
21. Bates BR, Harris TM. The Tuskegee study of untreated syphilis and public perceptions of biomedical research: a focus group study. *J Natl Med Assoc*. 2004;96:1051–1064. PubMed PMID: 15303410; PubMed Central PMCID: PMCPMC2568492.
22. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am J Public Health*. 2014;104:e16–e31. Epub 2013/12/12; PubMed PMID: 24328648; PubMed Central PMCID: PMCPMC3935672. doi:10.2105/AJPH.2013.301706.
23. Fisher JA, Kalbaugh CA. Challenging assumptions about minority participation in US clinical research. *Am J Public Health*. 2011;101:2217–2222. Epub 2011/10/20; PubMed PMID: 22021285; PubMed Central PMCID: PMCPMC3222419. doi:10.2105/AJPH.2011.300279.
24. Cragun D, Weidner A, Lewis C, et al. Racial disparities in BRCA testing and cancer risk management across a population-based sample of young breast cancer survivors. *Cancer*. 2017;123:2497–2505. Epub 2017/02/09; PubMed PMID: 28182268; PubMed Central PMCID: PMCPMC5474124. doi:10.1002/cncr.30621.
25. Manrriquez E, Chapman JS, Mak J, Blanco AM, Chen LM. Disparities in genetics assessment for women with ovarian cancer: can we do better? *Gynecol Oncol*. 2018;149:84–88. PubMed PMID: 29605055. doi:10.1016/j.ygyno.2017.10.034.
26. Underhill ML, Jones T, Habin K. Disparities in cancer genetic risk assessment and testing. *Oncol Nurs Forum*. 2016;43:519–523. PubMed PMID: 27314195. doi:10.1188/16.ONF.519-523.
27. Garrison NA. Genomic justice for native Americans: impact of the havasupai case on genetic research. *Sci Technol Human Values*. 2013;38:201–223. Epub 2012/12/21; PubMed PMID: 28216801; PubMed Central PMCID: PMCPMC5310710. doi:10.1177/0162243912470009.
28. McGuire AL, Hamilton JA, Lunstroth R, McCullough LB, Goldman A. DNA data sharing: research participants' perspectives. *Genet Med*. 2008;10:46–53. PubMed PMID: 18197056; PubMed Central PMCID: PMCPMC2767246. doi:10.1097/GIM.0b013e31815f1e00.
29. Trinidad SB, Fullerton SM, Ludman EJ, Jarvik GP, Larson EB, Burke W. Research ethics. Research practice and participant preferences: the growing gulf. *Science*. 2011;331:287–288. PubMed PMID: 21252333; PubMed Central PMCID: PMCPMC3044500. doi:10.1126/science.1199000.
30. Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K. Public perspectives on informed consent for biobanking. *Am J Public Health*. 2009;99:2128–2134. Epub 2009/10/15; PubMed PMID: 19833988; PubMed Central PMCID: PMCPMC2775766. doi:10.2105/AJPH.2008.157099.
31. Haeusermann T, Greshake B, Blasimme A, Irdam D, Richards M, Vayena E. Open sharing of genomic data: who does it and why? *PLoS One*. 2017;12:e0177158. Epub 2017/05/09; PubMed PMID: 28486511; PubMed Central PMCID: PMCPMC5423632. doi:10.1371/journal.pone.0177158.
32. Simon CM, L'heureux J, Murray JC, et al. Active choice but not too active: public perspectives on biobank consent models. *Genet Med*. 2011;13:821–831. PubMed PMID: 21555942; PubMed Central PMCID: PMCPMC3658114. doi:10.1097/GIM.0b013e31821d2f88.
33. Platt J, Bollinger J, Dvoskin R, Kardia SL, Kaufman D. Public preferences regarding informed consent models for participation in population-based genomic research. *Genet Med*. 2014;16:11–18. Epub 2013/05/09; PubMed PMID: 23660530; PubMed Central PMCID: PMCPMC3904287. doi:10.1038/gim.2013.59.
34. Spector-Bagdady K, De Vries RG, Gornick MG, Shuman AG, Kardia S, Platt J. Encouraging participation and transparency in biobank research. *Health Aff (Millwood)*. 2018;37:1313–1320. PubMed PMID: 30080467; PubMed Central PMCID: PMCPMC6143362. doi:10.1377/hlthaff.2018.0159.

35. Waldo A. https://theprivacyreport.com/2010/03/16/the-texas-newborn-bloodspot-saga-has-reached-a-sad-and-preventable-conclusion/: The Privacy Report. 2010 10/10/2019. [Accessed 12 July 2019].
36. HIPAA Data Breach Statistics. https://www.hipaajournal.com/healthcare-data-breach-statistics/: HIPAA J.; 2019 [cited October 10, 2019].
37. Madden M. Public perceptions of privacy and security in the post-snowden era. https://www.pewinternet.org/2014/11/12/public-privacy-perceptions/: Pew Research Center, 2014. [Accessed 20 July 2019].
38. Funk C, Johnson C, Hefferon M. 5 key findings about public trust in scientists in the U.S. https://www.pewresearch.org/fact-tank/2019/08/05/5-key-findings-about-public-trust-in-scientists-in-the-u-s/: Pew Research Center, 2019. [Accessed 10 October 2019].
39. Madden M, Rainie L. Americans' views about data collection and security. https://www.pewinternet.org/2015/05/20/americans-views-about-data-collection-and-security/: Pew Research Center, 2015. [Accessed 9 October 2019].
40. Tanner A. How data brokers make money off your medical records. https://www.scientificamerican.com/article/how-data-brokers-make-money-off-your-medical-records/: *Sci Am.* 2016. [Accessed 20 September 2019].
41. Leetaru K. How data brokers and pharmacies commercialize our medical data. *Forbes*. 2018. https://www.forbes.com/sites/kalevleetaru/2018/04/02/how-data-brokers-and-pharmacies-commercialize-our-medical-data/?sh=3af1dbac11a6. [Accessed 10 September 2019].
42. NIH Genomic Data Sharing [Internet]. https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/; 2015 [cited July 10, 2019].
43. Services CfMaM. HIPAA basics for providers: privacy, security, and breach notification rules. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/HIPAAPrivacyandSecurityTextOnly.pdf2018. [Accessed 10 October 2019].
44. Rothstein MA. Is deidentification sufficient to protect health privacy in research? *Am J Bioeth*. 2010;10:3–11. PubMed PMID: 20818545; PubMed Central PMCID: PMCPMC3032399. doi:10.1080/15265161.2010.494215.
45. McGuire AL. Identifiability of DNA data: the need for consistent federal policy. *Am J Bioeth*. 2008;8:75–76. PubMed PMID: 19003718; PubMed Central PMCID: PMCPMC2771195. doi:10.1080/15265160802478511.
46. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*. 2008;4:e1000167 Epub 2008/08/29; PubMed PMID: 18769715; PubMed Central PMCID: PMCPMC2516199. doi:10.1371/journal.pgen.1000167.
47. Na L, Yang C, Lo CC, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA Netw Open*. 2018;1 e186040. Epub 2018/12/07; PubMed PMID: 30646312; PubMed Central PMCID: PMCPMC6324329. doi:10.1001/jamanetworkopen.2018.6040.
48. . Roundtable on value & science-driven health care. *Best Practices Innovation Collaborative. Communicating with Patients On Health Care Evidence*. Washington, DC: Institute of Medicine of the National Academies; 2012:14.
49. Kloss LL, Brodnik MS, Rinehart-Thompson LA. Access and disclosure of personal health information: a challenging privacy landscape in 2016–2018. *Yearb Med Inform*. 2018;27:60–66. Epub 2018/08/29; PubMed PMID: 30157506; PubMed Central PMCID: PMCPMC6115206. doi:10.1055/s-0038-1667071.

CHAPTER 10

Genetic data sharing in the view of the EU general data protection regulation (GDPR)

Pieter De Smet and Mahsa Shabani

Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium

10.1 Introduction

Access to large-scale genetic data is essential for biomedical research. This access will allow researchers to analyze datasets from across the world and improve analyses of data. However, sharing genetic data implies processing sensitive personal data from patients and research participants. This may lead to harmful use of data, for example, to discriminate against individuals in employment and insurance settings.[1] Historically, the rights to privacy and data protection of individuals have been protected under European Union ("EU") and international regulations.[2] In the EU, the General Data Protection Regulation ("GDPR" or "Regulation") regulates processing personal data, including health and genetic data.[3]

To assure a homogeneous interpretation, the GDPR explicitly defines the scope of genetic data in article 4(13) and recital 34. The GDPR recognizes genetic and health data as "special categories of personal data" or sensitive data, providing higher legal requirements for processing such data (article 9). Genetic data sharing is a form of personal data processing protected under the GDPR, which comes with responsibilities for those processing the data ("controllers").[4] The EU legislator explicitly opted for a stringent regulatory framework for sensitive data, as indicated in recital 51. However, this framework also clearly provides derogations for legitimate data sharing such as *bona fide* research, which is specified in recitals 52–54. Nonetheless, concerns about concrete rules and obligations for data controllers have emerged.[5] Even though the European legislator sought to foster the use of data for scientific research purposes with the GDPR, some have pointed out that the Regulation has had a reverse effect on scientific research.

Genetic data sharing is favored by biomedical researchers for a variety of reasons, notably because of the advances in molecular biology that helped identify and categorize genetic elements that cause diseases.[6] The importance of genetic data sharing has been underlined in

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00012-2
Copyright © 2023 Elsevier Inc. All rights reserved.

the framework of guidelines and principles set out by the scientific community, for example, the Fort Lauderdale Agreement.[7] In response, online platforms have been established to facilitate genetic data sharing (e.g., the European Genome-Phenome Archive[8]).[9] Such agreements and platforms encourage and support cross-border data sharing in accordance with ethical principles and legal rules.[10]

Several ethical standards relevant to the processing of health and genetic data (for scientific research) were agreed upon in international documents. Significant examples include the World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and the Declaration of Taipei on Ethical Considerations Regarding Health Databases and Biobanks.[11] The United Nations ("UN") and its subsidiaries (notably UNESCO), as well as the Organization for Economic Co-operation and Development have also issued important declarations, guidelines, and recommendations related to biomedical research and genetic data.[12] On a pan-European level, the Council of Europe ("CoE") has been active in setting up a framework for fundamental principles. An important example of a CoE document includes the 2019 recommendation on the protection of health-related data, which states: "*the law may provide for the processing of health-related data for scientific research without the data subject's consent.*"[13] These international legal and ethical standards should be considered when assessing EU legislation such as the GDPR. They are important to understand the derogations formulated in the GDPR.

In this chapter, we will try to render a clear picture of the challenges and uncertainties the GDPR brings for genetic data sharing for research purposes. In the first part, we will elaborate on the particular status of genetic data under the GDPR, and the special requirements that processing of sensitive data implies. Subsequently, we will elaborate on the GDPR provisions that are of importance for scientific researchers aiming to share genetic data. In the third part, the link between the concept of "informed consent" in a research context and that of "consent" under the GDPR will be discussed.

Finally, recognizing the current legal uncertainty surrounding consent for secondary processing, the fourth part focuses on a legal alternative to consent and the necessary safeguards that are required.

10.2 The special status of genetic/genomic data

Under the EU data protection law, certain types of personal data are marked as special categories of personal data or sensitive data.[14] Those categories have been attributed specific protection for various reasons, such as the prevention of unfair discrimination.[15] This special categorization must be seen in the context of the fundamental rights and freedoms of EU citizens and residents.[16] Article 21 of the EU Charter of Fundamental Rights ("EU Charter") states discrimination based on genetic characteristics is explicitly prohibited. In the GDPR,

genetic data is included in the list of sensitive data (article 9), which was not the case with the GDPR predecessor, the Data Protection Directive ("DPD").[17]

Interestingly, the GDPR offers a clarification of the legal definition of genetic data (article 4(13) GDPR) that allows an assessment of its scope. This can be seen as a source-based definition. Recital 34 states:

"*Genetic data should be defined as personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained.*"

The introduction of genetic data in the list of sensitive data clearly underlines the augmenting importance of such data. The so-called exceptionalistic or particularistic approach toward genetic data (treating genetic information separately from other sorts of health data) has its reasons.[18] To begin with, one could ask who else is considered as a "data subject" or impacted individual under the GDPR once the genetic data of a genetically related person is processed, as genetic data of one person could also have implications for others such as family members.[19] This would raise questions about the biological relatives' privacy and their data protection rights in this context. In addition, there's a question on how far scientific and clinical utility—in favor of "the community"—may overrule or undo individual rights and control over personal data, if at all. As Panagiotopoulos pointed out, processing genetic data by specific controllers "*is often considered a priori legally and morally justifiable*" and this "*communitarian*" approach also bears risks.[20]

These questions and issues may not always directly be answered by data protection law. However, they should be considered in the design of data protection measures (article 25 GDPR). This, at times, implies the need for additional technical and organizational safeguards, when processing genetic data.

10.3 The GDPR framework for scientific research

Within the scientific research community, the adoption of the GDPR raised concerns about ongoing and forthcoming projects requiring the sharing and further processing of genetic data. There are several reasons why the GDPR has had an undesired chilling and even impeding effect on scientific research.[21] The lack of clarity and fragmentations between the laws of EU Member States have led to problems for certain research projects and cooperation schemes.[22] This has also led to criticisms for its effects on the response to the Covid-19 pandemic, notably with regard to international data transfers.[23] Consequently, the lack of clarity endangers "*the right to enjoy the benefits of scientific process and its applications*," a right described

by Slokenberga.[24] After all, the freedom of sciences and right to health care also constitute fundamental rights. The EU Charter itself states that "*scientific research shall be free of constraint*" and that everyone has the "*right to benefit from medical treatment*" (articles 13 and 35). The importance for data sharing in a health context can also be derived from other EU initiatives, such as the "European Health Data Space" initiative by the European Commission, which seeks to "*promote safe exchange of patients' data*".[25] These fundamental rights, therefore, need to be balanced consistently.

Under the GDPR, specific derogations for scientific research under a "research regime" or via the so-called "research exemption rules" have been foreseen. The preamble of the regulation indicates scientific research must be interpreted "*in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research*" (recital 159). On the other hand, the European Data Protection Board ("EDPB") has stressed that "*the notion may not be stretched beyond its common meaning and understands that 'scientific research' in this context means a research project set up in accordance with relevant sector-related methodological and ethical standards, in conformity with good practice*."[26] This interpretation is highly relevant as the GDPR explicitly tasks the EDPB with the drafting of guidelines with regard to the Regulation in order to ensure a consistent application (article 70). While according to the GDPR preamble, scientific research of a commercial nature may be generally allowed, the EDPB asserts that a *bona fide* and cautious methodological approach within an established ethical framework is always necessary.

Within the research regime of the GDPR, article 89 plays a central role, as it enables derogations from individual rights, including the right to object to certain processing, in some cases of scientific research. However, other relevant provisions are key to maintain the full overview of the research regime. Notably, article 5(1)(b) GDPR states that "*further processing* [of personal data] *for . . . scientific . . . research purposes . . . shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes.*" This is important, as the processing of personal data under the GDPR is normally limited to the initial purpose communicated to the data subject. The latter constitutes the purpose limitation principle.[27] Article 89 states that processing for scientific research purposes should be subject to "appropriate safeguards". This means that if one shares genetic data for scientific research purposes and makes use of the aforementioned exemption under article 5(1)(b), that researcher still needs to foresee "appropriate safeguards." Such safeguards could include additional information for data subjects or pseudonymization of the data. Despite the presence of a general beneficial regime for scientific research under the GDPR, many concrete and practical questions still need to be clarified. Further guidance from the EDPB on scientific research and personal data sharing is pending.[28]

The following part will show that practical issues arise for researchers working with the GDPR.

10.4 Consent for genetic data sharing under EU law

10.4.1 (Informed) consent for genetic data sharing: two distinct requirements arising from regulatory and ethics frameworks

In the context of genetic and health data sharing, the requirement of informed consent can arise from two sets of regulations. First, as a legal basis under data protection laws. Second, as a requirement by the laws on human subjects research and relevant ethics guidelines related to research with human subjects.[29] First, we will focus on the latter form of "informed consent," which applies in the context of (bio)medical and genetic research on human subjects.

Founded on the principle of integrity, all individuals have a right to decide what happens to their bodies. Informed consent implies that patients and research participants are informed about the—where applicable—risks, benefits, research goals, as well as withdrawal options[9]. As Vansweevelt and Tack described it, the protection of the personality rights of natural persons *vis-à-vis* third parties in a medical and scientific context is twofold.[30] On the one hand, personality rights offer negative protection through refusal rights protecting one's physical integrity, while on the other hand personality rights offer positive protection through granting self-determination rights.[31] The latter protection implies the need for legal provisions governing information duties. This allows natural persons to make an informed decision and consent to *inter alia* medical treatment or scientific research.[32]

From a historical perspective, "informed consent" has been a recurring concept throughout (international) legal documents governing healthcare, as well as biomedical and genetic research.[33] Article 5 of the Convention on Human Rights and Biomedicine of 1997 of the CoE ("Oviedo Convention") provided that *"an intervention in the health field may only be carried out after the person concerned has given free and informed consent to it."* Additionally, the CoE provided more specific guidance on how accompanying information should look like, as well as other necessary elements for informed consent in the context of health and genetic data sharing.[34] For example, one recommendation on biomedical research of the CoE lists the categories of information that should be given to research participants prior to consent, including the nature of any envisaged research and the conditions applicable to the storage of the materials.[35] Similar guidance for informed consent in the context of genetic testing for health purposes was also issued by the CoE in an additional protocol to the Oviedo Convention.[36] Last but not least, national regulations govern patients' rights within the national legal order, leading to *de facto* and *de iure* fragmentation of the rules related to "informed consent" within the EU[34].

In a preliminary opinion on data protection and scientific research by the European Data Protection Supervisor ("EDPS Opinion"), it was reaffirmed that the concept of consent in EU data protection law differs "*conceptually and operationally from informed consent of*

human participants in research."[37] Indeed, "consent" as a legal basis under data protection law must not be confused with the concept of "informed consent" for medical treatment or health and genetic research. The EDPB has stated that "*the requirement for informed consent for participation in a scientific research project can and must be distinguished from explicit consent as a possibility to legitimize the processing of personal data for scientific research purposes.*"[38] The EDPB refers to "*other options*" under articles 6 and 9 GDPR that can assure lawful data processing, apart from consent.[39] The mere possibility of other legal bases (e.g. public interest) for data sharing under the GDPR, does not mean an "informed consent" (as an ethical requirement) should *not* be obtained.

The distinct origin of the two concepts (informed consent for human subjects research vs consent under EU data protection law) does not imply the concepts do not share common elements. For example, informational self-determination is one of the elements fundamental to both concepts.[40] In practice, a real distinction between two types of consent is not always evident for the individuals. Often, they are merely asked to consent once for their treatment or research participation *and* the processing of their personal data. This leads to the misperception that consent is the only option for personal data processing under the law, whereas it is usually "informed consent" that is necessary and made use of to comply with data protection regulations in a single stretch. That said, various international legal documents often require controllers to obtain informed consent, making consent the most *a priori* attractive legal basis under the GDPR. It is, therefore, crucial to scrutinize consent as a legal basis under the GDPR for genetic data sharing. In the following part, we will first focus on the elements, advantages, and challenges of obtaining consent as a legal basis under the GDPR, before discussing other legal possibilities for genetic data sharing.

10.4.2 What type of consent is considered valid under the GDPR?

Genetic data sharing requires two legal bases under the Regulation: one "general" for the processing of personal data (article 6(1) GDPR), and one "additional," due to the sensitive nature of genetic data (article 9(2) GDPR).

First, genetic data sharing is only possible under the GDPR if the controller has a lawful basis for the processing under article 6(1). Processing under the GDPR has an elaborate meaning and entails the collection, storage, or use of personal data such as genetic data, but also the mere transfer of it (article 4.2. GDPR). According to article 4, paragraph 7, the controller is "*a natural or legal person, public authority agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.*" Consent as a legal basis under article 6(1)(a) implies that the data subject should consent in a free, specific, informed, and unambiguous way.[42] These are four cumulative conditions. Due to its nature, consent under the GDPR may constitute a heavy (administrative) burden in some circumstances. More flexible approaches by controllers toward consent have not been considered

lawful by the EU's highest court, the Court of Justice of the EU ("CJEU").[43] Moreover, article 7 imposes conditions for valid consent, such as the obligation for the controller to be able to demonstrate that consent has been given, and the possibility for the data subject to withdraw consent at any time. The EDPB has stressed that the latter withdrawal option is a right for every data subject, even when the purpose of the personal data processing is scientific research, and even when this could undermine certain scientific research projects.[44]

Second, the genetic data sharing should have a legal basis under article 9(2), allowing a derogation from the principled prohibition to the processing of sensitive data in article 9. Article 9(2)(a) allows genetic data sharing when "*the data subject has given explicit consent to the processing of those personal data for one or more specified purposes.*" Formally, article 9(2)(a) is a distinct legal basis from the general legal basis of consent in article 6(1)(a). In practice, however, once all the—strict—criteria of consent under articles 6(1)(a) and 7 are met, there will also be a sufficient legal basis for genetic data sharing under article 9(2)(a), as long as the consent is also made in an explicit way.

As has become clear since the implementation of the GDPR, a one-fits-all solution in approaching consent requirements is not existing. This is definitely the case for scientific research and genetic data sharing. Often, the purpose of all future genetic data sharing is not fully clear from the outset, making obtaining specific consent hard, if not impossible, in that context. In the following parts, we will elaborate on the discussion surrounding "broad" or "open" consent, as a potential solution that can allow multiple downstream uses of data.

10.4.2.1 The case for broad consent

It remains unclear whether the use of consent as a legal basis for genetic data sharing is possible, when the purposes are not yet all clear at the moment of the initial data collection.[45] The condition of specificity implies that consent needs to include specific purposes.[46] If there is more than one purpose, each individual purpose should be identified and consented to by the data subject. The possibility to diversify consent based on various specified purposes is called granular consent.[47] Data sharing within the scientific research community intensifies this issue, as it implies various purposes, which makes it difficult—if not impossible—to ask for specific consent. Another question raised is the possibility for researchers to further share data, if the initial collection of the personal data had consent as a legal basis and did not mention the further processing.[48]

In this regard, Recital 33 states that:

"*it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognized ethical standards for scientific research. Data subjects should have the opportunity to give*

their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose."

Recital 33 thus seems to allow researchers to derogate from the specificity requirement for consent. The EDPB guidelines on consent however clarify that recital 33 does not "*disapply*" the requirement for *specific* consent and that "*the GDPR cannot be interpreted to allow for a controller to navigate around the key principle of specifying purposes for which consent of the data subject is asked*."[49] The EDPB also stated that recital 33 "*allows as an exception that the purpose may be described at a more general level*," contrary to the condition of "specific purpose" in article 6(1)(a).[50] Furthermore, the EDPB also stated that when sensitive data—such as genetic data—are being processed, "*applying the flexible approach of recital 33 will be subject to a stricter interpretation and requires a high degree of scrutiny.*"[51] However, the extent of this "stricter interpretation" is not clear. The guidance of the EDPB suggests its members (the national data protection authorities) would not allow a very broad interpretation of recital 33 and thus broad consent.[52] It is interesting to note that restricting the use of "broad consent" may not lead to the most ethical approach. As Woolley has pointed out, some consider that not allowing broad consent "*violates the wishes of many willing participants who want to share their data widely*."[53]

Despite the legal uncertainty considering the EDPB's advice, some legal scholars have argued in favor of the use of broad consent or called for further clarifications. For example, Dove called the EDPB's predecessor's (i.e., the Working Party 29 or WP29) consent guidelines—guidelines that are very similar to those of the EDPB cited above—confusing.[54] He argues that the EDPB "*should clarify in future guidance that it is legally acceptable under the GDPR for data subjects' consent to be obtained on a 'broad basis', i.e. on a description of the 'certain areas of scientific research' for which their personal data will be processed, including special categories of personal data, in the current time.*"[54] Similarly, Hallinan has discussed broad consent from a principled, legal technical, and practical perspective. Hallinan specifically referred to the WP29 guidelines on consent, and concluded that from all perspectives, broad consent ought to be the path forward for genetic data processing, and the EDPB should issue guidelines in favor of it.[55] He considers the consent guidelines of the WP29 guidelines vague, and argues that the guidance "*might legitimately be set aside if this proves unsuitable and obstructive in relation to the unique practices of genomic research . . . even if problematic aspects of the guidance cannot be interpreted away, it may still be argued that the powers of the* [WP29] *and the EDPB do not extend to offering interpretations of the GDPR which contradict the express wishes of the legislator.*"[56]

Furthermore, Chassang has argued that the GDPR does not exclude a broad approach to consent, but that this possibility depends on the member states' national laws. Chassang claims that "*broad consent in the respect of applicable national law* [is allowed], *provided that the individual received sufficiently clear information and that the given consent represents the unambiguous indication of the data subject's wishes*."[57] Chassang mentions countries such as

France and Italy privilege "*a specific consent at the basis of the involvement of an individual in research.*"[57]

The variety of academic opinions on broad consent and recital 33 reveal its precarious foundations. It is clear the EU legislator's desire was the facilitation of a research-friendly environment, stimulating the free flow of personal data for scientific research purposes. As recital 33 was only introduced in a late stage of the legislative process,[58] initial reactions from the scientific field were indeed enthusiastic and assumed broad consent was generally allowed.[59] Allowing broad consent intrinsically implies the eradication of specificity, which constitutes a fundamental element of consent. Indeed, this would benefit genetic data sharing, and offset the burden of asking reconsent for every (further) research purpose after the initial data collection. On the other hand, the problem is that "broad consent" foundations merely appear in the GDPR's preamble, *not* as a provision of law. The CJEU confirmed that recitals—forming part of the act's preamble—cannot be relied on as a ground for derogating from the actual provisions.[60]

To conclude, many legal scholars are looking at the EDPB and national authorities to bring legal certainty for use of broad consent. There are various theoretical and practical reasons to argue in favor of using broad consent. However, in our opinion, the unclarity is intrinsic to the GDPR, making it impossible to force the EDPB and national authorities to issue a *contra legem* interpretation of the law. There are two possible solutions to implement broad consent and bring "absolute" legal certainty. First, the GDPR itself could be amended. However, this first option seems unlikely as members of the European parliament ("MEPs") are of the opinion that there is no need to vote on amendments to the GDPR, as a political choice.[61] However, the European Commission's evaluation of two years of application of the GDPR, raised the possibility to adopt such GDPR amendments.[62] As a second option, *lex specialis* could be adopted. This would mean specific EU legislation could be adopted in the area of scientific research and/or data sharing, in order to provide for a derogatory scheme to consent under the GDPR. This might be a more achievable goal, as this could be part of legislation that is focused on scientific research, not privacy or data protection exclusively. In any case, not clarifying the current legislation will unavoidably lead to further legal uncertainty, and possibly adversely impacting genetic data sharing.

10.4.2.2 Alternative approaches to consent requirements

There are various alternative approaches for broad consent that pose fewer formal legal problems, specifically with regard to the specificity element of consent. The most discussed examples include tiered-layered and dynamic consent.[63] Interestingly, the EU's legislative bodies have recently adpoted the EU Data Governance Act ("DGA").[64] The DGA seeks to facilitate the reuse of data held by the public sector while upholding high protection for relevant rights (e.g., intellectual property rights, rights regarding trade secrets, and personal data protection rights).[65] The proposed Act also introduces the concept of "data altruism." This concept will—or at least should—bring additional legal certainty while contributing

to additional transparency.[66] As the Act only applies 15 months after its entering into force (Art. 38), more specifically from 24 September 2023, we refrain from elaborating on the concrete consequences of the legislation.

For research and genetic data sharing, the concept of dynamic consent has gained considerable attention, even though it might not be suitable for every research context.[67] In short, dynamic consent allows research participants to consent to participate in each specific stage of (a) research project(s), if it is not possible to identify all specific purposes at the time of the data collection. In the case of genetic data sharing, this would imply research participants consent to each transfer of the data, once it is clear to whom the data could be transferred. This "*two-way communication*" between the researcher and research participant, say Kaye et al., "*enables participants to consent to new projects or to alter their consent choices in real time as their circumstances change . . .* "[68]

The EDPB stated in its consent guidelines that "*when research purposes cannot be fully specified, a controller must seek other ways to ensure the essence of the consent requirements are served best, for example . . . for specific stages of a research project that are already known to take place at the outset* [and] *as the research advances, consent for subsequent steps in the project.*"

Furthermore, the EDPS, in its preliminary opinion on data processing for scientific research, has specifically mentioned the added value of dynamic consent in the scenario "*where participants are asked to consent to different activities over time via an IT interface* [as] *trialled in the field of biobanks.*"[69]

Adopting dynamic consent for genetic data sharing requires building adequate IT infrastructure including online platforms to allow for interactions to take place.[70] Furthermore, dynamic consent requires a constant follow-up of communications between the researcher and research participant.[71] Moreover, as Budin-Ljøsne et al. stressed, collaboration with research ethics committees and assuring accessibility of the platform (i.e., for those who don't have access to the internet) is of salient importance to the success of dynamic consent[71]. Even though it might seem that the alternative would be very costly and time-consuming, the process might deepen the social relationship between researchers and research participants, and enhance trust in that research. Studies have shown that two-way communication is considered as important by research participants, in the case of data sharing.[72]

10.5 Alternative legal bases for genetic data sharing: shifting attention away from consent

With several provisions in the GDPR, the EU legislator explicitly put forward a more lenient framework for researchers, notably in articles 5(1)(b), 6(4), 9(2)(j), and 89. The set of legal

provisions governing this research-friendly framework has been called the "research regime" or "research exemption rules" in recent academic literature.[72] First and foremost, it must be stressed that none of the GDPR's provisions establishing the research regime create a stand-alone legal basis for personal data processing.[70] To build a complete overview of what the research regime actually entails, one must read all of the aforementioned provisions, along with the relevant recitals in the preamble. It is important to have this complete overview, as the EDPB has stressed the research regime may only be used "*for exceptions to specific requirements in specific situations "and that it is"[…] dependent on 'additional safeguards'*."[71] Only if all conditions are fulfilled, an alternative legal basis to consent could be found in the research regime.

As genetic data are considered as sensitive data, there will be a need for cumulative legal bases in articles 6 *and* 9 GDPR.[72] The "public interest" presents an interesting alternative to consent for genetic data sharing and can be found in article 6(1)(e) in conjunction with article 9(2)(g), (i) or (j). Article 6(1)(e) states that personal data processing is lawful if "*processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller*." In any case, sensitive data processing—including genetic data sharing—based on the public interest under article 9(2) will require a separate legal basis in EU or member state law and thus requires a legal provision to that effect.[73] Most EU Member States have adopted specific provisions allowing for personal data processing in the context of scientific research.[74]

Apart from the aforementioned legal grounds, it should be considered that article 9(4) provides for a possible complication, as this provision allows member states to "*maintain or introduce further conditions, including limitations, with regard to the processing of genetic data* …." In fact, the European Commission assessment on the member states' rules on health data already showed that this led to complex circumstances, specifically for cross-border genetic data sharing. For example, countries such as France uphold more strict approaches with a focus on individual protection, compared to approaches focused on general societal benefits such as in the Netherlands.[75]

Regardless of the legal basis on which the genetic data sharing for scientific research purposes supports, the principles of the research regime, in particular those mentioned under Article 89, must be respected. According to article 89 the controller must take "*appropriate safeguards,*" including "*technical and organizational measures … in particular in order to ensure respect for the principle of data minimization.*"

Currently, there is an ongoing discussion on the scope of "appropriate safeguards." In fact, the EDPB acknowledged in a letter to the European Commission "*that the present lack of specification on what could or should be considered adequate safeguards . . . can be considered a serious impediment for the proper use of the exceptions foreseen in the GDPR for processing personal data for scientific research purposes.*"[76] In principle, these safeguards

should include, among others, transparency and information requirements, anonymization and pseudonymization techniques, data access governance mechanisms, and risks assessment measures including Data Protection Impact Assessment (DPIA). It however remains to be seen whether or not future guidance can streamline the current divergent approaches in this regard.

10.6 Concluding remarks

The EU's aim to foster scientific research offers some tools to facilitate genetic research and genetic data sharing. However, the unclarities surrounding the relevant provisions including consent requirements had an impeding effect on data sharing and data-intensive scientific research in general. The key to compliance lies with the careful choice of the legal basis, as well as the appropriate safeguards put in place in preparation for genetic data sharing. Consent is a way to assure the former, but other possible legal bases—such as the public interest—should not be forgotten and could even be practically preferable.

Apart from the alternative approaches to consent that were presented in this chapter, the future may bring more specific guidance from the authorities, including verified sectoral agreements or so-called "codes of conduct." These codes could constitute a helpful tool, even though they cannot fix all the uncertainties.[77] Codes of conduct are instruments recognized by the GDPR, which the competent national data protection authorities can formally develop in accordance with article 40. Moreover, the European Commission can approve the general validity of codes of conduct across the EU (art. 40(9) GDPR). The recent approval of a transnational code of conduct for cloud computing ("EU Cloud CoC") has shown the possibilities of such codes.

References

1. Van Gyseghem JM, Les catégories particulières de données à caractère personnel. In: C De Terwagne and K Rosier, *Le Règlement Général Sur La Protection Des Données (RGPD/GDPR) – Analyse Approfondie*, Larcier, Brussels, 2018, 255–6.
2. European Union, Charter of fundamental rights of the European Union, official journal ("O.J.") C. of 26 October 2012, 326-391; Council of Europe, convention 108 and protocols, https://www.coe.int/en/web/data-protection/convention108-and-protocol (accessed August 12, 2021).
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection regulation), O.J.L L. 119 of 5 May 2016, https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed August 30, 2022), 1-88.
4. de Terwangne C, "Article 5: principles relating to processing of personal data" In: C Kuner, LA Bygrave, C Docksey and L Drechsler, eds. *The EU General Data Protection Regulation (GDPR): A commentary*, Oxford University Press, Oxford, 2020, (309)318–9.
5. Shabani M, Borry P. "Rules for processing genetic data for research purposes in view of the new EU general data protection regulation". *Eur J Hum Genet*. 2018;26:149–156. doi:10.1038/s41431-017-0045-7.

6. Sprumont D, Borry P and Shabani M, Chapter 16: genetic testing in Europe. An overview of the legal framework. In: A. DEN EXTER, eds. *Eur Health Law.*, Maklu, Antwerpen, 2017, 365.
7. Wellcome Trust, *Sharing data from large-scale biological research projects: a system of tripartite responsibility*, 2003, http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf (accessed August 12, 2021).
8. European Bioinformatics Institute and the Center for Genomic Regulation, European genome-phenome archive, https://ega-archive.org/ (accessed August 12, 2021).
9. Shabani M, Knoppers BM, Borry P. Genomic databases, access review, and data access committees. In: Kumar D, Antonarakis S, eds. *Med Health Genom*. Elsevier; 2015:29–35. https://doi.org/10.1016/B978-0-12-420196-5.00003-4 (accessed August 30, 2022).
10. Kalkman IS, Mostert M, Gerlinger C, Van Delden JJM, Van Thiel GJMW. Responsible data sharing in international health research: a systematic review of principles and norms. *BMC Medical Ethics*. 2019;20:21. doi:10.1186/s12910-019-0359-9.
11. World medical association declaration of Helsinki, ethical principles for medical research involving human subjects, 64th General Assembly (Fortazela, brazil), October 2013; World Medical Association Declaration of Taipei, Ethical Considerations Regarding Health Databases and Biobanks, 67th General Assembly (Taipei, Taiwan), 2016.
12. UNESCO, Universal declaration on bioethics and human rights, 19 October 2005; UNESCO, International Declaration on Human Genetic Data, 16 October 2003; UNESCO, Universal Declaration on the Human Genome and Human Rights, 11 November 1997; OECD, Guidelines for Human Biobanks and Genetic Research Databases, 2009.
13. Council of Europe, Recommendation CM/Rec(2019)2 of the Committee of Ministers to member states on the protection of health-related data (Adopted by the Committee of Ministers on 27 March 2019 at the 1342nd meeting of the Ministers' deputies), p. 4. https://search.coe.int/cm/pages/result_details.aspx?objectid=090000168093b26e (accessed August 30, 2022), par. 15.3.
14. Georgieva L and Kuner C, Article 9. Processing of special categories of personal data in C Kuner et al., The EU General Data Protection Regulation (GDPR): a commentary, o.c., (365)369 et seq; Judgement Court of Justice of the European Union ("CJEU") of 6 November 2003, Bodil Lindqvist, C-101/01; Judgment CJEU of 24 September 2019, GC and others, C-136/17.
15. Ernst S, "Art. 4 begriffsbestimmungen" in BP Paal and DA Pauly, Datenschutz-Grundverordenung Bundesdatenschutzgesetz, Beck, München, 3rd ed., (43)71.
16. Frenzel, EM, "Art. 9 verarbeitung besonderer kategorien personenbezogener daten" in BP Paal and DA Pauly, Datenschutz-Grundverordenung Bundesdatenschutzgesetz, B, München, 3rd ed., (153) 156, 159; United Nations Special Rapporteur on the Right to Privacy, Recommendation on the Protection and Use of Health-Related Data, 6 November 2019, https://www.ohchr.org/Documents/Issues/Privacy/SR_Privacy/DraftRecommendationProtectionUseHealthRelatedData.pdf (accessed August 12, 2021), 5.
17. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, O.J. L. 281 of 23 November 1995, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046 (accessed August 30, 2022).
18. Sprumont D, Borry P, Shabani M. Chapter 16: Genetic testing in Europe An overview of the legal framework. In: Exter A Den, ed. *Eur Health Law*, 365; 2017:366–367.
19. Heaton TJ, Chico V. Attitudes towards the sharing of genetic information with at-risk relatives: results of a quantitative survey. *Hum Genet*. 2016;135(1):109–120. https://doi.org/10.1007/s00439-015-1612-z. WP29 working document on genetic data, 17 March 2004, WP91, 8; W.G. URGESSA, "The Protective Capacity of the Criterion of 'Identifiability' under EU Data Protection Law", *Eur Data Prot Law Rev.* 2016, Vol. 2, Is. 4, https://doi.org/10.21552/EDPL/2016/4/10, 521-531.
20. Panagiotopoulos A. Genetic information and communities: a triumph of communitarianism over the right to data protection under the GDPR? *Eur Data Prot Law Rev.*. 2018;4(459):469. doi:10.21552/edpl/2018/4/8.
21. Rabesandratana T. Researchers sound alarm on European data law. *Science*. 2019;366:936. doi:10.1126/science.366.6468.936.

22. Peloquin D, Dimaio D, Bierer B, Barnes M. Disruptive and avoidable: GDPR challenges to secondary uses of data. *Eur J Hum Genet*. 2020;28:697–705. https://doi.org/10.1038/s41431-020-0596-x. Federation of European Academics of Medicine, International Sharing of Personal Health Data for Research report, April 2021, https://www.feam.eu/wp-content/uploads/International-Health-Data-Transfer_2021_web.pdf(accessed August 12, 2021).
23. Bentzen HB, et al. Remove obstacles to sharing health data with researchers outside of the European Union. *Nat Med*. 2021;27:1329–1333. doi:10.1038/s41591-021-01460-0.
24. Slokenberga S, Setting the foundations: individual rights, public interest, scientific research and biobanking in S Slokenberga, O Tzortzatou and J Reichel, eds. *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation Across Europe,*Springer Law, Governance and Technology Series, Cham, Vol. 43, 2021, https://doi.org/10.1007/978-3-030-49388-2, (11)12.
25. European Commission, Webpage the European health data space, 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12663-Digital-health-data-and-services-the-European-health-data-space_en (accessed August 12, 2021).
26. European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, version 1.1, Adopted on 4 May 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf (accessed August 30, 2022), 30.
27. de Terwagne C, Article 5. Principles relating to processing of personal data in C Kuner et al., The EU General Data Protection Regulation (GDPR): A Commentary, o.c., (309)315-7.
28. On 30 April 2021, an "EDPB stakeholder event on processing of personal data for scientific research purposes" took place online, in which the main author of this chapter participated as an academic stakeholder, https://edpb.europa.eu/news/news/2021/edpb-stakeholder-event-processing-personal-data-scientific-research-purposes_en (accessed August 12, 2021).
29. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU general data protection regulation. *Eur J Hum Genet*. 2018;26:149–156. https://doi.org/10.1038/s41431-017-0045-7.
30. Vansweevelt T, Tack S, Hoofdstuk V: het recht op gezondheidstoestandinformatie en geïnformeerde toestemming. In: T Vansweevelt and F Dewallens, eds. *Handboek Gezondheidsrecht,* Volume II: rechten van de patiënten – van embryo tot lijk, Intersentia, Antwerpen, 2014, (355)355.
31. Vansweevelt T, Tack S, Hoofdstuk V. het recht op gezondheidstoestandinformatie en geïnformeerde toestemming. In: Vansweevelt T, Delwallens F, eds.. *Handboek Gezondheidsrecht*, Volume II, rechten van de patiënten - van embryo tot lijk, Intersentia; 2014:355.
32. Paterick TJ, Carson GV, Allen MC, Paterick TE. Medical informed consent: general considerations for physicians. *Mayo Clin Proc*. 2008;83:313–319. https://doi.org/10.4065/83.3.313. J.R. BOTKIN, "Informed Consent for Genetic and Genomic Research", *Curr Protoc Hum Genet.* 2020, Vol. 108, Is. 104, https://doi.org/10.1002/cphg.104.
33. Veny L. Patient rights: the right to give informed consent to medical treatments from European and Belgian perspectives. In: Apan RD, Fodor EM, eds. *Health Law*. Pro Universitaria; 2018:348–367. https://lib.ugent.be/catalog/pug01:8562580 (accessed August 30, 2022).
34. Council of Europe, Recommendation CM/Rec(2016) of the committee of ministers to member states on research on biological materials of human origin, 2016, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=090000168064e8ff (accessed August 12, 2021), Chapter III; Council of Europe, Additional Protocol to the Oviedo Convention on Human Rights and Biomedicine, concerning Biomedical Research, no. 195, 25 January 2005, https://www.coe.int/en/web/conventions/full-list/-/conventions/rms/090000168008371a (accessed August 12, 2021); Council of Europe, Recommendation CM/REC(2016)8 of the Committee of Ministers to the member states on the processing of personal health-related data for insurance purposes, including data resulting from genetic tests, 26 October 2016, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016806b2c5f (accessed August 12, 2021).

35. Council of Europe, CM/Rec(2016) of the committee of ministers to member states on research on biological materials of human origin, 2016, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=090000168064e8ff (accessed August 19, 2021).
36. Council of Europe, Additional protocol to the Oviedo convention on human rights and biomedicine, concerning genetic testing for health purposes, Treaty No. 203, 27 November 2008, https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/203 (accessed August 12, 2021).
37. European data protection supervisor*, a preliminary opinion on data protection and scientific research,*2020, https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf (accessed August 12, 2021), 2.
38. European Data Protection Board, Document on response to the request of the European Commission for clarifications on the consent application of the GDPR, focusing on health researc h, Adopted on 2 February 2021, https://edpb.europa.eu/our-work-tools/our-documents/other-guidance/edpb-document-response-request-european-commission_en (accessed August 30, 2022), 4.
39. Verhenneman G. *The Patient, Data Protection and Changing Healtcare Models*. Cambridge: Intersentia; 2021:47.
40. EDPB, Guidelines 05/2020 on consent under regulation 2016/679, v. 1.1., 2020.
41. Judgment Court of Justice of the European Union of 11 November 2020, Orange Romania SA v Autoritatea Nationala de Supraveghere a Prelucrarii Datelor cu Caracter Personal (ANSPDCP), C-61/19, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62019CJ0061 (accessed August 30, 2022); Judgment Court of Justice of the European Union of 1 October 2019, Bundesverband der Verbraucherzentralen und Verbraucherverbände — Verbraucherzentrale Bundesverband eV v. Planet 49 GmbH, C-673/17, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62017CJ0673 (accessed August 30, 2022).
42. EDPB, Guidelines 05/2020 on consent under regulation 2016/679 v. 1.1., 2020, par. 163.
43. Peloquin D, Dimaio D, Bierer B, Barnes M. "Disruptive and avoidable: gdpr challenges to secondary uses of data". *Eur J Hum Genet*. 2020;28(697):700. https://doi.org/10.1038/s41431-020-0596-x.
44. Verhenneman G. *The Patient, Data Protection and Changing Healthcare Models*. Cambridge: Intersentia; 2021:194–198.
45. EDPB, Guidelines 05/2020 on consent under regulation 2016/679, version 1.1., 2020, 12.
46. Shabani M, "The data governance act and the EU's move forward towards facilitating data sharing", Mol Syst Biol., vol. 17, 10.15252/msb.202110229, 2.
47. EDPB, Guidelines 05/2020 on consent under regulation 2016/679, v. 1.1., 2020, par. 156.
48. EDPB, Document on response to the request from the european commission for clarifications on the consistent application of the GDPR, focusing on health research, 2 february 2021, par. 26; EDPB, guidelines 05/2020 on consent under regulation 2016/679, v. 1.1., 2020, par. 157.
49. Verhenneman G, Claes K, Derèze JJ, et al. How GDPR enhances transparency and fosters pseudonomysation in academic medical research. *Eur J Health Law*. 2019;27:40.
50. Woolley JP, "How data are transforming the landscape of biomedical ethics: the need for ELSI metadata on consent" in BD Mittelstadt and L Foridi (eds.), The Ethics of Biomedical Big Data, Springer International Publishing, Cham, 2016, https://doi.org/10.1007/s11673-017-9812-y, (171)175.
51. Dove ES. The EU general data protection regulation: implications for international scientific research in the digital era. *J Law Med Ethics*. 2021;46(1013):1022–1023. doi:10.1177/10731105188220003.
52. Hallinan D. Broad consent under the GDPR: an optimistic perspective on a bright future. *Life Sci Soc Policy*. 2020;16(Ed. 1). https://doi.org/10.1186/s40504-019-0096-3.
53. Chassang G. The impact of the EU general data protection regulatation on scientific research. *Ecancermedicalscience*. 2017;11(709):11. https://doi.org/10.3332/ecancer.2017.709.
54. Eur-Lex, Preparatory documents GDPR: https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=CELEX:32016R0679; Hallinan D and Friedewald M, Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation?, *Life Sci Soc Policy.* 2015, Vol. 11, 10.1186/s40504-014-0020-9.

55. Marelli L, Testa G. Scrutinizing the EU general data protection regulation. *Science*. 2018;360(496):497. doi:10.1126/science.aar5419.
56. Klimas T, Vaicuikaite JCJEU Judgment of November 19 1998, Nilsson, C-162/67, par. 54. The law of recitals in European community legislation. *Ilsa J Int Comp L.*. 2008;61:90.
57. Manancourt V. EU privacy law's chief architect calls for its overhaul, Politico EU (news organisation), 2021, https://www.politico.eu/article/eu-privacy-laws-chief-architect-calls-for-its-overhaul/ (accessed August 12, 2021).
58. European Commission, Communication from the commission to the European parliament and the Council, data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition – two years of application of the general data protection regulation, COM(2020)264 final, 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0264&from=EN (accessed August 12, 2021).
59. Bunnik EM, Janssens AC, Schermer MH. A tiered-layered-staged model for informed consent in personal genome testing. *Eur J Hum Genet*. 2013;21:596–601. https://doi.org/10.1038/ejhg.2012.237.
60. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), O.J. L. 152 of 3 June 2022, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868 (accesssed August 30, 2022).
61. Baloup J. The data governance act: new rules for international transfers of non-personal data held by the public sector. *European Law Blog*. 2021. https://europeanlawblog.eu/2021/06/10/the-data-governance-act-new-rules-for-international-transfers-of-non-personal-data-held-by-the-public-sector/. (accessed August 12, 2021).
62. Shabani M. The data governance act and the EU's move towards facilitating data sharing. *Mol Syst Biol*. 2021;17. doi:10.15252/msb.202110229.
63. Verhenneman G. *The Patient, Data Protection and Changing Healthcare Models*. Cambridge: Intersentia; 2021:197.
64. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melhalm K. Dynamic consent: a patient interface for twenty-first century research networks. *Eur J Hum Genet*. 2015;23:141. doi:10.1038/ejhg.2014.71.
65. European data protection supervisor, *a preliminary opinion on data protection and scientific research*, 2020, https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf (accessed August 12, 2021), 14.
66. Budin-Ljøsne I, Teare HJA, Kaye J, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. *BMC Med Ethics*. 2017;18:6. doi:10.1186/s12910-016-0162-9.
67. Ludman EJ, Fullerton SM, Sprangler L, et al. Glad you asked: participants' opinions of re-consent for dbGaP data submission. *J Empir Res Hum Res Ethics.*. 2010;5(3):9–16. https://doi.org/10.1525/jer.2010.5.3.9. A.L. McGuire, J.A. Hamilton, R. Lunstroth, L.B. McCullough and A. Goldman, "DNA data sharing: research participants' perspectives", *Genet Med: Official J Am College Med Genet.* 2008, Vol; 10, Is. 1, https://doi.org/10.1097/GIM.0b013e31815f1e00, 46–53.
68. Slokenberga S, Tzortzatou O and Reichel J, Introduction in in S Slokenberga et al. (eds.), *GDPR and Biobanking*, o.c., 1-6; EDPB, Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research, 2021, 3.
69. Pauly DA, Art. 89 garantien und ausnahmen in bezug auf die verarbeitung zu im öffentlichen interesse liegenden archivzwecken, zu wissenschaftlichen oder historischen forschungszwecken und zu statistischen zwecken, in BP Paal and DA Pauly, Datenschutz-Grundverordenung Bundesdatenschutzgesetz, Beck, München, 3ed ed., (1093)1094.
70. Hallinan D. Oxford: Oxford University Press; 2021:8–9.
71. Hallinan D. *Protecting Genetic Privacy in Biobanking through Data Protection Law*. Oxford: Oxford University Press; 2021:166–167.
72. European Commission, *Assessment of the EU member states' rules on health data in the light of the GDPR,*2021, https://ec.europa.eu/health/ehealth/key_documents_en#anchor1 (accessed August 12, 2021).

73. European Commission, *Assessment of the EU member states' rules on health data in the light of the GDPR, 2021*, https://health.ec.europa.eu/system/files/2021-02/ms_rules_health-data_en_0.pdf (accessed August 30, 2022), par. 26.
74. European Commission, *Assessment of the EU member states' rules on health data in the light of the GDPR, 2021*, https://health.ec.europa.eu/system/files/2021-02/ms_rules_health-data_en_0.pdf (accessed August 30, 2022), par. 53.
75. Molnár-Gábor F, Korbel JO. Genomic data sharing in europe is stumbling – could a code of conduct prevent its fall? *EMBO Mol Med*. 2020;12:11421. doi:10.15252/emmm.201911421.
76. Vander Maelen C. First of many? first GDPR transnational code of conduct officially approved after edpb opinions 16/2021 and 17/2021. *Eur Data Prot Law Rev.*. 2021;7:228–231. https://doi.org/10.21552/edpl/2021/2/12. Belgian Data Protection Authority press release, The BE DPA approves its first European code of conduct, 20 May 2021, https://www.dataprotectionauthority.be/the-be-dpa-approves-its-first-european-code-of-conduct(accessed August 12, 2021).
77. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), O.J. L. 152 of 3 June 2022, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868 (accesssed August 30, 2022).

CHAPTER 11

Genomic data sharing and intellectual property

Jorge L. Contreras
Department of Human Genetics, S.J. Quinney College of Law, University of Utah

Intellectual property protection has figured prominently in the ongoing debate over access to scientific data, particularly in the context of genomic data sharing and its implications for precision medicine. When speaking about genomic and other data, this chapter refers to collections of data that are accessible to the public—either on an unrestricted or restricted basis—as an information or knowledge commons[1] and, more specifically, as the "genomic commons."[2–4] Intellectual property can have two principal effects on such a knowledge commons: it can prevent the entry of data into the commons, and it can limit the ability of others to utilize data that are already in the commons. It is thus important to understand the different types of intellectual property that may apply to such data, and how they interact. This chapter discusses the principal forms of intellectual property protection that impact genomic data, as well as their implications for precision medicine. The perspective of this chapter is largely one of US law, and while international intellectual property treaties and agreements have harmonized national laws to some degree, there are still numerous points of divergence from country to country. A discussion of the laws of multiple countries is beyond the scope of this chapter.

11.1 Forms of intellectual property protection for genomic data

11.1.1 Copyright

Copyright law offers authors protection for their original works of authorship and gives such authors certain exclusive rights to exploit those works, including the rights to display, distribute, reproduce, and create derivatives of them. Copyright law is of significant importance in the area of scientific journal articles, and there have been extensive debates concerning the high cost of access to scientific publications.[5]

Unlike written expression, such as articles, scientific facts, and data themselves are not subject to copyright protection in the United States. As observed by Justice Louis Brandeis more than

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00005-5
Copyright © 2023 Elsevier Inc. All rights reserved.

a century ago, facts are "free as the air to common use."[6] Yet there is some indication that collections and compilations of facts are entitled to a degree of "thin" copyright protection.[7] This protection encompasses the organizational structure of such facts, headings, labels, indexes, pagination, and the like.

In the context of data, copyright principles are most relevant when discussing limitations on access and use of data that are the subject of scholarly publications. That is, even though facts and conclusions are not themselves copyrightable, the articles in which they are presented (including the text and any illustrations) are subject to copyright and thus controlled by the publishers of the journals carrying those articles. In some cases, even data that might otherwise be in the public domain (such as mapping and geographic data developed under a contract with the US government) may be stored in proprietary databases that are accessible only by paid subscribers.[8] In several areas, the privatization of governmental data is proceeding rapidly, leading to fears that increasing amounts of data will become "enclosed" (evoking the historical fencing-off of commonly held land) and thereby unavailable for public use.[9]

11.2 Databases, data protection, and terms of use

Databases that contain genomic data may be protected under various legal regimes. In the European Union, databases with commercial value have legally protected status under the so-called Database Directive of 1996 (Dir. 96/9/EC). While formal legal protection for databases is not recognized in the United States, access to data in electronic databases can be controlled by database operators through technological means, such as password and authentication restrictions. Thus, while data itself is not subject to legal protection, the circumvention of technical protection measures is prohibited under the "anti-circumvention" provisions of the 1998 Digital Millennium Copyright Act (17 U.S.C. Sec. 101 et seq.).

Likewise, many online databases require the user to agree to contractual "terms of use," which are often presented as electronic "click to accept" agreements. Despite the fact that few users actually read or understand these dense legal documents, courts have routinely held them to be enforceable,[10] (Ch. 17.B). Such terms of use are now customary with respect to databases of genetic, genomic, and even genealogical data.[11] Even governmental genomic databases such as dbGaP contain legally binding terms of use that limit a user's right to utilize data that is accessed, often in ways that are intended to protect the privacy and identity of data subjects.[12]

The violation of contractual terms of use can result in civil liability including monetary damages and the revocation of the right to use data. In these ways, scientific information that might otherwise be in the public domain can become encumbered when compiled in proprietary databases. Such restrictions were adopted by Celera Genomics when it announced its intention to sequence the human genome in competition with the

publicly funded Human Genome Project (HGP) and offer the resulting data to commercial users on a paid basis.[13] A similar approach was adopted by Myriad Genetics, which maintains a proprietary database that includes tens of thousands of *BRCA1/2* variants that may have a bearing on cancer susceptibility but which are not generally accessible to researchers or the public.[14]

11.3 Patents

The US Patent Act authorizes the granting of patents for "any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof" (35 U.S.C. § 101). A patent gives its owner a period of 20 years during which it has the exclusive right to make, use, sell and import such a patented invention in the United States. Data, information, and other facts about the world are not themselves included within the scope of patentable subject matter, but they can be included as important elements of other inventions that make use of them.

11.3.1 Early biotech patents

Patents on molecular biology discoveries were first issued in the United States beginning in the 1970s,[15] but it was not until 1980 that the Supreme Court ruled that a patent could cover a living organism (*Diamond v. Chakrabarty)*[16] The organism in question was a bacterium modified to accelerate the breakdown of crude oil. In a split 5–4 decision, the Court held that "anything under the sun made by man" was eligible for patent protection, opening the door to patents covering bioengineered organisms.

11.3.2 Genetic patents and utility

By the early 1990s, patents were being sought on human DNA. In 1991, J. Craig Venter, a researcher at the National Institutes of Health, filed a series of patent applications claiming hundreds of short DNA fragments known as expressed sequence tags (ESTs). These filings resulted in a public outcry, which eventually led the NIH to reverse its position on the patenting of human DNA. [15,17] The NIH's position, which now reflects that of the US Patent and Trademark Office, is that raw genomic sequence data, the function of which is not known, fails to meet the requirement that a patented invention be "useful."[18]

Despite these agency positions, patent applicants continued to argue that ESTs and other DNA sequences satisfied the legal utility requirement for patentability. Among other things, they argued that ESTs were useful as probes and markers along the massive genome. By the time the first EST patent was issued to Incyte Pharmaceuticals in 1998, that company alone had filed patent applications claiming more than 1.2 million DNA sequence fragments.[19] The

question was not settled by the Court of Appeals for the Federal Circuit until 2005.[20] In the case *In re. Fisher*, the court confirmed that ESTs, without an identified function, do not meet the statutory utility requirement.

11.3.3 Bermuda and official patent deterrence

By the mid-1990s, concerns over the patenting of human DNA were growing among policymakers. By the time the HGP was ready to begin sequencing the human genome in 1996, the international scientific leadership of the project met in Bermuda to develop a new policy to govern the sharing and release of HGP-funded data. From the meeting emerged a set of guidelines that became known as the "Bermuda Principles."[21] These principles require that all DNA sequence information from large-scale sequencing projects be made "freely available and in the public domain." Most significantly, the Bermuda Principles mandated that such sequencing data be released a mere 24 hours following generation.

The unprecedented data release requirements of the Bermuda Principles were justified on a number of grounds, including the need to coordinate data among geographically dispersed sequencing centers and to make the discoveries of the publicly funded HGP data as broadly available as possible. But another real, though less explicit, goal of the Bermuda Principles was to deter researchers from obtaining patents on sequence data generated by the HGP.[4] This goal was accomplished in two ways.

First, the 24-hour data release requirement ensured that HGP data would be made public before laboratories performing sequencing work could file patent applications claiming such data. In jurisdictions such as the European Union and Japan, which have so-called "absolute novelty" requirements, an invention may not be patented if it has been publicly disclosed prior to the filing of a patent application (i.e., if it is deposited in a public database such as GenBank). In the United States, under the then-prevailing rule, a patent application could be filed up to one year after its description in a "printed publication" or its first "public use." Thus, if an inventor wished to seek patent protection for an invention in the United States, it was required to file its application no more than one year after disclosing the invention in a publication or, with respect to DNA sequence information, depositing it in a public database.

These measures acted as deterrents to patenting by the HGP sequencing centers, but patenting by the sequencing centers was not the primary concern of the HGP planners. Rather, they were concerned that DNA sequence data could be patented independently by commercial enterprises (a fear that was realized a few years later with the emergence of Craig Venter's new company Celera Genomics—see below). The rapid data release requirements of the Bermuda Principles were intended to deter this form of patenting as well.

Under then-prevailing US law, if an applicant's invention was described in a printed publication or otherwise "known" to the public before it was invented by the applicant, it would be

considered "prior art," causing the patent to be denied on grounds of "anticipation." Thus, a third party that sequenced ("invented") a relevant DNA segment after its release by the HGP would be prevented from patenting that sequence. In short, every DNA sequence released by the public HGP would act as prior art against later patent filings by private laboratories.

Moreover, the publicly released sequence would also serve as prior art for purposes of determining whether a different but related third-party sequence overcame the "nonobviousness" requirement for patentability. For example, the full sequence of a length of human DNA along a particular chromosome might include several genes. Separating these genes are noncoding segments of DNA, and as well as noncoding DNA segments within genes themselves (introns). The DNA sequence disclosed by the HGP would include all of this DNA, including genes, noncoding DNA and introns. Yet the most biologically significant DNA is the coding DNA found within the genes, and an enterprise seeking a patent might choose to claim only such coding DNA. If the applicant claimed an isolated gene sequence that had previously been disclosed by the HGP with its attendant introns and noncoding DNA, it is likely that the disclosed sequence could be used as prior art against the claimed gene even though the sequences would not be identical, strictly speaking.

The rapid data release requirements of the Bermuda Principles thus functioned as NIH's first official policy of patent deterrence. This approach, though praised by many, was also criticized by those who believed that the NIH's adoption of an antipatenting stance contravened the requirements of federal statutes, including the Bayh-Dole Act of 1980 (35 U.S.C. § 200–212), which expressly encourages federally funded researchers to patent their inventions for the benefit of the US economy.

In response to this criticism, NIH's 1996 policy implementing the Bermuda Principles acknowledges the Bayh-Dole Act, stating that recipients of NIH funding have the right to patent inventions that "reveal convincing evidence for utility."[18] But at the same time, NIH cautions that it "will monitor grantee activity in this area to learn whether or not attempts are being made to patent large blocks of primary human genomic DNA sequence."[18] The policy is silent, however, regarding the consequences of seeking such patents. The NIH's approach in this respect has thus been one of norm-setting—the formation of institutional expectations regarding behavior—rather than legally enforceable regulation, an approach that it has used with varying degrees of success over the years.[22]

11.3.4 The Ft. Lauderdale principles

By requiring that data generated by the HGP be placed into the public domain within 24 hours after being generated, the Bermuda Principles effectively eliminated the head start that data generating researchers traditionally had to analyze and publish conclusions derived from such data. Not surprisingly, researchers engaged primarily in the production of large datasets began

to balk at rapid data release requirements. This issue was central to a 2003 summit held in Ft. Lauderdale, which led to an agreement on another set of data release principles.[23] While the Ft. Lauderdale participants "enthusiastically reaffirmed" the 1996 Bermuda Principles, they also drew a distinction between large "community resource projects" (CRPs), as to which rapid data release was considered appropriate, and more directed "hypothesis-driven" research, as to which researchers might be given time to analyze the data prior to releasing it publicly. Thus, by 2003 the genomics research community has already begun to accept theoretical limitations on the rapid release of data into the public domain.

11.3.5 NIH's evolving policy toward patenting

As the 21st century advanced, the NIH's aversion to patents seems to have softened. In its 2007 policy on genome-wide association studies (GWAS), NIH merely expressed a "hope" that "genotype-phenotype associations identified through NIH-supported and NIH-maintained GWAS datasets and their obvious implications will remain available to all investigators, unencumbered by intellectual property claims" and stated that "[t]he filing of patent applications and/or the enforcement of resultant patents in a manner that might restrict the use of NIH-supported genotype-phenotype data could diminish the potential public benefit they could provide."[24] However, in an effort to support patent seekers, the policy also "encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs" (id.). And in its 2014 Genomic Data Sharing policy, the NIH appears largely to have discarded the patent-deterring early data release mandate that it championed in the Bermuda Principles, permitting data generators to keep human genomic data confidential for periods of up to six months and nonhuman data until the time of publication.[25,26] The agency, somewhat optimistically, cites the *Myriad* decision as providing sufficient protection against DNA sequences becoming encumbered by patents (see Part 11.3.9, below).

NIH's most recent data sharing policy, which applies to numerous forms of scientific data in addition to genetic and genomic data, has few specific requirements regarding public data release and does not mention patents or other intellectual property.[27]

11.3.6 Patent deterrence outside the United States

US funding agencies were not the only ones to discourage patenting activity with respect to genomic data. Genome Canada adopted its first formal data release policy in 2005.[28] While acknowledging the Bermuda and Ft. Lauderdale Principles, the Canadian policy does not adopt their 24-hour data release requirement. With respect to patents, Genome Canada "recognizes the need to protect patentable and other proprietary data" (p. 1) and thus requires that data generators release data following publication or the filing of a patent application, whichever occurs first.

Since the beginning of the HGP, the UK-based Wellcome Trust has funded genomic research, both through grants and through its Sanger Institute in Cambridge, UK, a leading sequencing center. In 2006, the Wellcome Trust funded a large-scale study of seven complex human diseases by more than 50 research groups across the United Kingdom (the Wellcome Trust Case Control Consortium). The study generated a large quantity of data, including aggregated and individual-level genotypic and phenotypic information. Most of these data were released to the public in accordance with the Bermuda and Ft. Lauderdale Principles, and the project self-designated itself as a CRP. The Consortium required each prospective data user to sign a data access agreement under which access is granted only to qualified investigators for "appropriate use" (Sec. 4). The data access agreement does not, however, contain any specific restriction on patenting activity.

11.3.7 Nongovernmental limitations on patenting genomic data

In 2002, near the end of the HGP, an international group of researchers and funders undertook the development of a haplotype map of the human genome through the International HapMap Project.[29] The data release policy eventually developed by the HapMap Project was based on the Ft. Lauderdale Principles, and the project designated itself as a CRP. The project also took several affirmative steps to ensure that patents would not be obtained on the results of its research either by data generators or data users. Specifically, each user of HapMap data was contractually prohibited from restricting access to the HapMap database and, in particular, from filing patent applications on the data generated by the project. The project also adopted the legal position that raw genomic data lacks specific utility and is therefore ineligible for patent protection. These policies were applauded by policymakers at NIH, which lauded the HapMap Project's success at deterring "parasitic patents."[30]

11.3.8 The SNP consortium and defensive patenting

An innovative approach to patent deterrence for genomic data was developed during the HGP by the SNP Consortium Ltd. This nonprofit entity was formed in 1999 by a group of pharmaceutical companies and the Wellcome Trust to identify and map genetic markers known as "single nucleotide polymorphisms" (SNPs) and to release the resulting data to the public.[4,31] As the project progressed, these data were published on the consortium's web site and also deposited in GenBank. By 2002, the SNP Consortium had mapped 1.4 million SNPs and created a genome-wide SNP-based human linkage map, all of which were made publicly available.

The consortium also wished to ensure that its data would remain unencumbered by patents. To achieve this goal, it first filed US statutory invention registrations (SIRs), documents officially published in the patent office database for the purpose of disclosing prior art that would

prevent the patenting of these discoveries by others. Then, following a statutory change in 1998, the consortium shifted to filing patent applications that claimed the SNPs that it identified. It then abandoned these applications prior to allowance, ensuring that the consortium's discoveries would act as prior art defeating subsequent third-party patent applications, with a priority date extending back to their initial filing dates. This innovative "protective" patenting strategy has been cited as a model of the private industry's potential to contribute to the public genomic commons,[32–34] and has served as the model for additional private sector data sharing projects such as the International Severe Adverse Events Consortium.[35] In these cases, it is important to understand that the private-sector research funders preferred to release data free from patent encumbrances not out of a public interest in the free availability of data, but in order to avoid capture of valuable research tools by other private-sector firms—in other words, they preferred that the results be free to all rather than owned by someone else.[4]

11.3.9 Genetic sequence patents under Myriad[1]

Throughout the 2000s, the practice of patenting human genes came under increasing pressure in the public arena. In 2005, a widely cited article estimated that 20% of all human genes were subject to patent protection.[36] Two years later, bestselling author and *Jurassic Park* creator Michael Crichton published an op-ed in the *New York Times* claiming that "You, or someone you love, may die because of a gene patent that should never have been granted in the first place."[37] Crichton and other public critics of gene patenting were reacting, among other things, to the business practices of Myriad Genetics, a biotech firm that obtained patents covering the *BRCA1* and *BRCA2* genes, certain variants of which were known to have a strong association with an elevated risk of breast and ovarian cancer.

Myriad's patents, which began to issue in 1997 (and some of which were co-owned by the University of Utah), claimed numerous aspects of newly discovered and sequenced genes. Most significantly, Myriad claimed the complete DNA sequence of the isolated and purified *BRCA* genes as new compositions of matter. It also claimed as compositions of matter complementary DNA (cDNA) constructs containing only the protein-coding regions of the genes (exons), as well as any consecutive sequence of at least 15 bases contained within the genes (15-mers). Myriad also obtained method claims for various diagnostic and drug discovery activities utilizing the *BRCA* genes, particularly the diagnosis of an elevated risk of breast and ovarian cancer. Myriad granted exclusive licenses of its patents to Eli Lilly with respect to therapeutic uses of the *BRCA* genes, and to Hybridon (a subsidiary of Eli Lilly) with respect to test kits. Myriad reserved to itself the use of the genes for laboratory diagnostic testing of the *BRCA1/2* genes.

[1] Detailed accounts of the gene patenting litigation involving Myriad Genetics can be found in Refs. [50] and [54].

In 1998, Myriad began to send cease-and-desist letters to clinics and academic laboratories that had been performing diagnostic tests for *BRCA* mutations. As a result, by 2000 Myriad was the only remaining US lab performing diagnostic *BRCA* testing. Given the lack of competition, Myriad charged more than $3000 to test for the three most common *BRCA* variants, with a $700 add-on for large-scale rearrangements that were harder to detect. Both private insurers and Medicare/Medicaid were initially reluctant to cover Myriad's tests, making many individuals at risk for deleterious *BRCA* mutations unable to afford testing. And because Myriad was the only provider of *BRCA* testing in the United States, individuals with positive test results from Myriad could not obtain second opinions before undergoing prophylactic surgery or chemotherapy. Finally, the existence of the patents, coupled with Myriad's aggressive enforcement campaign, was found to chill research on the *BRCA* genes, hindering scientific understanding of their function.[38]

In 2009, the American Civil Liberties Union and the Public Patent Foundation brought suit on behalf of a coalition of 20 medical associations, advocacy groups, researchers, healthcare providers, and cancer patients to challenge Myriad's *BRCA* patents.[39]

Though the district court rejected all of Myriad's patent claims, the Federal Circuit, in a divided decision, upheld Myriad's composition of matter claims on the theory that the *BRCA* genes, when purified and isolated from the larger chromosomes on which they reside, become new chemical entities that are not naturally occurring. The Federal Circuit did, however, reject all but one of Myriad's challenged method claims as abstract ideas and mental processes (see discussion of *Mayo*, below).

The Supreme Court granted *certiorari* to decide the question "Are human genes patentable?" In 2013, a unanimous Court held that the DNA sequences of genes found in living organisms—"products of nature"—could not be patented. The Court's opinion began by explaining that Myriad's discovery of the *BRCA* genes—even if it involved substantial time, expense, and skill—was not "an act of invention." The substance of the Court's opinion holds that Myriad's claims covering the precise DNA sequences that exist in the human body fail to meet the statutory requirements for patent eligibility, as these sequences previously existed in nature. Nevertheless, the Court also found that artificially synthesized cDNA sequences that contain only the coding regions of the *BRCA* genes do not occur naturally, and thus are eligible for patent protection.

In a subsequent case, Myriad Genetics asserted patents claiming artificially synthesized primers (short DNA segments used in the diagnostic process) against competing *BRCA* test vendors. The Federal Circuit held that even though Myriad's primers were synthetic molecules, their nucleotide sequences duplicated sequences found in nature and thus rendered them ineligible for patent protection.[40]

The prohibition on patenting naturally occurring genetic sequences yielded notable benefits during the COVID-19 pandemic. Where research and collaboration on earlier viral outbreaks such as SARS and the H1N1 and H5N1 influenza strains were stymied by races to patent their genetic sequences,[41–43] this did not occur with SARS-CoV-19, the virus responsible for COVID-19. When the SARS-CoV-2 viral RNA sequence was elucidated in early 2020, researchers immediately deposited it in GenBank without, it is believed, any attempt to seek patent protection. This patent-free zone may have enabled researchers around the world to study the virus and more rapidly develop vaccines and drugs to combat it.

11.3.10 Diagnostic patents under Mayo

The year before the *Myriad* decision, the Supreme Court ruled in a case that did not involve patents on genetic or genomic discoveries, but which has arguably had a far greater impact on personalized and precision medicine techniques.[44] In *Mayo*, the court considered a patent claiming a method of adjusting a patient's drug dosage based on the measured level of metabolites in the patient's blood after the drug was administered.

The interaction of the drug with the patient's blood, of course, is a natural phenomenon, as are many biomedical discoveries. In view of this, the Court reasoned that it must "determine whether the claimed processes have transformed ... unpatentable natural laws into patent-eligible applications of those laws." In fashioning a test to assist with this determination, the Court seemingly raised the bar on what counts as a "new application" of a natural law. That is, in order to be eligible for patenting, the Court required that there be some "inventive concept" above and beyond the natural law and its application through "well-understood, routine, conventional activity." Thus, the Court in *Mayo* found that the claimed method for adjusting a patient's drug dosage based on metabolite levels in the patient's bloodstream was merely a straightforward application of a natural correlation observed in the human body. It was thus ineligible for patent protection.

Thus, while *Myriad* eliminated patents seeking to claim human genes as compositions of matter, *Mayo* eliminated patents on many applications of knowledge arising from the observation of human genes. But it is not only genetic diagnostics that have been affected. Recent judicial rulings have severely limited patents on other biomedical innovations that depend on natural correlations. Critics of these decisions argue that the unavailability of patents on diagnostic methods could discourage firms from investing in the development of new personalized medicine techniques.[45–47]

11.3.11 Licensing of genomic inventions

The mere fact that patents cover particular genomic inventions does not mean that those inventions are inaccessible to users other than the patent holder. In fact, many patented inventions are licensed by their owners to others. This pattern is common in the biotechnology

sector, in which academic research institutions routinely license patents to biotechnology and pharmaceutical companies for commercial exploitation.[10] When the *CFTR* gene associated with cystic fibrosis was discovered in 1989 and patented by a team of researchers at the University of Michigan and Hospital for Sick Children, they licensed that patent on a nonexclusive, royalty-free basis to any lab that wanted to test the gene.[48] A similar situation existed with the *HEXA* gene associated with Tay-Sachs disease, which was discovered at NIH in 1984.[49] NIH obtained a patent on the gene but chose to make licenses available without charge.

The prevailing wisdom in the technology licensing industry holds that a single exclusive licensee will often pay more than a slew of nonexclusive licensees. And while the institutions that held the *CFTR* and *HEXA* patents chose to make them broadly available without charge, that altruistic approach was quickly supplanted in the 1990s by more profit-oriented strategies. Thus, most gene patents, and a broad range of other biomedical innovations, were licensed exclusively to companies seeking to earn both testing and licensing revenue from them.[50]

While exclusive licensing may yield the greatest profit to the patent holder and its licensees, it also restricts the number of users of a patented technology. This type of restriction might be appropriate for a particular molecular targeting a disease, but is perceived as problematic when it covers a research tool that could be broadly applicable across the field. Cognizant of this distinction, in 1999 NIH advised that "[w]here the subject invention is useful primarily as a research tool, inappropriate licensing practices are likely to thwart rather than promote utilization, commercialization and public availability."[51] The agency followed in 2005 with a set of *Best Practices for the Licensing of Genomic Inventions* in which it urged NIH-funded researchers to grant nonexclusive licenses with respect to broadly applicable genomic research tools and resources.[52]

Nevertheless, exclusive licenses of genomic technologies, particularly from university laboratories, remain common. In 2017, Jacob Sherkow and I coined the term "surrogate licensing" to describe the increasingly common arrangement in which a research institution effectively relinquishes all of its rights in a patented technology to a private company—often a university "spinout" in which university researchers hold significant equity shares—and abdicates its corresponding social responsibilities.[53] This phenomenon is particularly acute in the case of CRISPR-Cas9 gene editing technology, which has potential uses across multiple fields and human genes, yet it is licensed exclusively by the academic institutions that made the foundational discoveries to two university spinout companies (id.)

11.4 Conclusion

Several forms of intellectual property protection can be obtained with respect to genetic and genomic data. While the United States does not recognize copyright in data or databases to any appreciable degree, contractual restrictions imposed by online and click-wrap terms

of use can impose similar or greater burdens than default legal regimes. Likewise, though patents can no longer be obtained for naturally occurring genomic sequences or diagnostic methods involving the application of natural laws, many attributes of products based on these discoveries can still be patented. It is thus important to consider intellectual property issues in any large-scale precision medicine project.

References

1. Hess C, & Ostrom E (Eds.). *Understanding Knowledge as a Commons: From Theory to Practice*. MIT Press; 2007.
2. Contreras JL, Knoppers BM. The genomic commons. *Ann Rev Genomics & Human Genet*. 2018;19:429.
3. Contreras JL. Constructing the genome commons. In: Frischmann B, Madison M, Strandburg K, eds. *Governing Knowledge Commons* Oxford University Press; 2014:99.
4. Contreras JL. Bermuda's legacy: patents, policy and the design of the genome commons. *Minn J L Sci & Tech*. 2011;12:61.
5. Contreras JL. Confronting the crisis in scientific publishing: latency, licensing and access. *Santa Clara L Rev*. 2013;53:491.
6. Int'l News Serv. V. Associated Press, 248 U.S. 215 (U.S. 1918).
7. Feist Publ'ns, Inc. V. Rural Tel. Serv. Co., 499 U.S. 340 (U.S.) 2022.
8. Reichman JH, Uhlir PF. A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment. *Law & Contemp Probs*. 2003;66:314.
9. Boyle J. *The Public Domain: Enclosing the Commons of the Mind*. Yale University Press; 2008.
10. Contreras JL. *Intellectual Property Licensing and Transactions: Theory and Practice*. Cambridge University Press; 2022.
11. Contreras JL, Schultz K, Teerlink CC, Maness T, Meyer LJ, Cannon-Albright LA. Legal terms of use and public genealogy websites. *J L & Biosciences, LSAA*. 2020;063(1). doi:10.1093/jlb/lsaa063.
12. Paltoo DN. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nature Genet*. 2014;46:934.
13. Williams H. Intellectual property rights and innovation: evidence from the human genome. *J Political Econ*. 2013;121:1.
14. Guerrini CJ, McGuire AL, Majumder MA. Myriad take two: can genomic databases remain secret? *Science*. 2017;356:586.
15. Sherkow JS, Greely HT. The history of patenting genetic material. *Ann Rev Genet*. 2015;49:161.
16. Chakrabarty Dv, 447 U.S. 303 (U.S. 1980).
17. Cook-Deegan R (1994). The gene wars: science, politics, and the human genome. Norton.
18. National Human Genome Research Institute. NHGRI policy regarding intellectual property of human genomic sequence. 1996.
19. Murray J. Owning genes: disputes involving DNA sequence patents. *Chi-Kent L Rev*. 1999;75:231.
20. In re Fisher, 421 F.3d 1365 (Fed. Cir. 2005).
21. International Human Genome Sequence Organisation. Summary of principles agreed upon at the first international strategy meeting on human genome sequencing (Bermuda, 25-28 February 1996) as reported by HUGO). 1996.
22. Contreras JL. Leviathan in the commons—biomedical data and the state. In: Strandburg K, Frischmann B, Madison M, eds., *Governing Medical Knowledge Commons*, Cambridge University Press; 2017.
23. Wellcome Trust. Sharing data from large-scale biological research projects: a system of tripartite responsibility. 2003.
24. National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). 2007.

25. National Institutes of Health.Final NIH genomic data sharing policy. *Fed Reg*. 2014;79:51345.
26. Contreras JL. NIH's genomic data sharing policy: timing and tradeoffs. *Trends Genet*. 2015;31:55.
27. National Institutes of Health.Final NIH policy for data management and sharing and supplemental information. *Fed Reg*. 2020;85:68890.
28. Genome Canada. Data release and resource sharing. 2008. http://www.genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf. (Accessed date: 27-08-2022)
29. International HapMap Consortium.The International HapMap Project. *Nature*. 2003;426:789.
30. National Human Genome Research Institute. ENCODE Project data release policy (2003–2007). 2003.
31. Holden A. The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques*. 2002;32:22.
32. Merges RP. A new dynamism in the public domain. *Univ Chi L Rev*. 2004;71:183.
33. Cook-Deegan R, McCormack SJ. A brief summary of some policies to encourage open access to DNA sequence data. *Science*. 2001;293(217 supp).
34. Eisenberg RS. The promise and perils of strategic publication to create prior art: a response to Professor Parchomovsky. *Mich L Rev*. 2000;98:2358.
35. Contreras JL, Floratos A, Holden A. The international serious adverse events consortium's data sharing model. *Nat Biotechnol*. 2013;31:17.
36. Jensen K, Murray F. Intellectual property landscape of the human genome. *Science*. 2005;310:239.
37. Crichton M. Patenting life. NY Times. 2007.
38. Cho MK, Samantha I, Weaver MA, Leonard DGB, Merz JF, et al. Effects of patents and licenses on the provision of clinical genetic testing services. *J Molecular Diag*. 2003;5:3.
39. Assn. For Molecular Pathology v. Myriad Genetics, 569 U.S. 576 (U.S. 2013).
40. Univ. Of Utah Rsch. Fund v. Ambry Genetics Corp., 774 F.3d 755 (Fed. Cir. 2014).
41. Rimmer M. The race to patent the SARS Virus: the TRIPS agreement and access to essential medicines. *Melbourne J Intl L*. 2004;5:335.
42. Greene H. Patent pooling behind the veil of uncertainty: antitrust, competition policy, and the vaccine industry. *BUL Rev*. 2010;90:1397.
43. Beldiman D. Patent choke points in the influenza-related medicines industry: can patent pools provide balanced access? *Tulane J Tech & Intell Prop*. 2012;15:31.
44. Mayo Collaborative Services v. Prometheus Laboratories, Inc., 566 U.S. 66 (U.S. 2012).
45. Holman CM. Mayo, Myriad and the future of innovation in molecular diagnostics and personalized medicine. *NC J L Tech*. 2014;15:639.
46. Noonan K. Diagnostic patents at risk after Federal Circuit decisions. *Nature Rev Drug Discovery*. 2016;15:377.
47. Eisenberg RS. Diagnostics need not apply. *J Sci Tech L*. 2015;21:256.
48. Minear MA, Kapustij C, Boden K, Chandrasekharan S, Cook-Deegan R. Cystic fibrosis patents: a case study of successful licensing. *Les Nouvelles*. 2013;48:21.
49. Colaianni A, Chandrasekharan S, Cook-Deegan R. Impact of gene patents and licensing practices on access to genetic testing and carrier screening for Tay-Sachs and Canavan disease. *Genetics in Med*. 2010;12:S5.
50. Contreras JL. Association for Molecular Pathology v. Myriad Genetics: a critical reassessment. *Mich Tech L Rev*. 2020;27:1.
51. National Institutes of Health.Principles and guidelines for recipients of NIH research grants and contracts on obtaining and disseminating biomedical research resources: final notice. *Fed Reg*. 1999;64:72090.
52. National Human Genome Research Institute.Best practices for the licensing of genomic inventions: final notice. *Fed Reg*. 2005;70:18413.
53. Contreras JL, Sherkow JS. CRISPR, surrogate licensing, and scientific discovery. *Science*. 2017;355:698–700.
54. Contreras JL. *The Genome Defense: Inside the Epic Legal Battle to Determine Who Owns Your DNA*. Algonquin; 2021.

CHAPTER 12

Data governance

Dimitri Patrinos, Michael Lang and Ma'n H. Zawati

Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

12.1 Background: precision medicine genomics and governance

There has been significant progress in genomics over the last two decades. Recent advances in our understanding of human genetics and genomics have accelerated the development of highly targeted healthcare. Genomic data are now available in unprecedented volumes and are increasingly integrated with individual health records, permitting significant improvements in the kind of phenotype characterization that makes highly targeted medical care possible. But achieving the goal of precisely targeted healthcare will require a significant degree of coordination among clinicians, health researchers, and patient interest communities. It will require effective, ethical, and legally compliant data sharing that both facilitates the discovery of new treatments and maintains trust between health institutions and the public. Data governance, in short, is essential for genomics to have a significant impact on precision medicine.

Constructed broadly, data governance refers to the policies and frameworks, both internal and external to a biobank, consortium, or other genomics initiative, that modulate the conduct of researchers and institutions with respect to data under their custody or control.[1] Data governance performs a range of crucial functions, from facilitating external data access, to ensuring compliance with applicable law, supporting the ethical use of genomic resources, protecting against potential harm, and promoting public benefit in the conduct of genomics research.[1] Adequate data governance regimes are a necessary feature of efforts toward an expansion of precision medicine.[2] This is so for three principal reasons. First, the collection, storage, and curation of large genomic datasets are fundamentally important for the kind of ambitious genomics research that undergirds precision medicine. Biobanking at a populational level provides both a requisite volume and diversity of data on which to draw biomedical findings with potential precision applications. Collecting, storing, and annotating large assemblages of genomic data is a massive organizational and logistical undertaking, one that is unlikely to be executed successfully absent a robust and directed governance regime. Second, precision medicine depends on precisely the kinds of genomic data sharing facilitated by data governance structures. Accelerating the discovery of precision treatments requires efficient,

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine.
DOI: https://doi.org/10.1016/B978-0-12-819803-2.00002-X
Copyright © 2023 Elsevier Inc. All rights reserved.

widespread, and internationalized genomic resource sharing in order to promote diversity in data cohorts, cost control, and research equity. Third, data governance works to ensure participant and public trust in genomic research initiatives by ensuring that the management of sensitive individual data is carried out in compliance with applicable legal and ethical norms. Far from acting as a barrier to effective genomics research, data governance works to encourage the kind of scientific exchange that is vital to achieving precision medicine.

This chapter outlines some of the ways that genomic data governance works to facilitate precision medicine. We focus in particular on three issues: retrospective data integration, prospective data collection, and data access. Each of these functions is deeply implicated in the work of precision medicine, and each is dependent on effective and efficient data governance. We suggest that the acceleration of precision medicine depends on the development and maintenance of data governance regimes capable of facilitating data availability for research and of maintaining public support for large-scale genomics research.

12.2 How data governance shapes precision medicine

In 2020, the World Economic Forum (WEF) identified policy and governance gaps that hinder the widespread implementation of precision medicine. Governance gaps in data sharing and interoperability were identified as significant barriers to progress in precision medicine. Data sharing is critical to precision medicine, which depends on the generation of large volumes of data that can subsequently be processed, linked, and analyzed by researchers.[3] Genomic and other health-related data from a wide variety of sources need to be pooled to be able to develop personalized health care approaches for individuals. Moreover, datasets must be interoperable, meaning they must be able to be exchanged across different systems. However, the lack of interoperability greatly restricts the ability of datasets to be widely and efficiently utilized by researchers, thereby limiting developments in precision medicine.[2]

It is, therefore, important to consider how these challenges related to data governance can be addressed so that they do not unduly thwart advances in precision medicine. The implementation of efficient and effective data governance structures and frameworks is key. Indeed, data governance should be viewed as an important facilitating factor in the precision medicine ecosystem. As stated by the WEF, guidelines for data standardization and governance principles can serve to increase data sharing between systems and facilitate collaborative research for precision medicine.[2] Three key facets of data governance, in particular, can help address some of these specific issues: (1) retrospective data integration, (2) prospective data collection, and (3) data access. In the following sections, we discuss how these governance mechanisms can help ensure the efficient and scientifically meaningful use of datasets and allow them to be shared widely across systems to ensure the greatest possible advances in precision medicine.

12.2.1 Retrospective data integration

Precision medicine requires leveraging significant volumes of data to be able to generate research results that are statistically significant. Assembling large enough cohorts of research participants to generate these required volumes of data, however, can be a challenge. This is especially true in the rare disease context, for instance, where it is difficult to assemble research cohorts as not enough participants may be available at a given location to conduct meaningful research.[4] Advances in precision medicine, therefore, rely not only on the generation of new datasets but also making use of existing datasets that have been generated in the context of previously conducted research. We refer to these types of datasets as "retrospective data." Usage of retrospective data can help facilitate diagnoses for patients, discover new drugs, and validate clinical trials,[4–6] as well as save costs and time for researchers.[4,7] A well-defined set of governance mechanisms should be implemented to make use of these types of data and reap the benefits of their use. Rich datasets are very often generated from previous research projects and constitute a highly valuable resource that should be tapped into. This is especially important within the genomic context, where the integration of genomic and phenotypic data is imperative in the development of precision medicine tools.[2] The ability to harness retrospective datasets is therefore an important piece of the precision medicine puzzle.

Nevertheless, repurposing retrospective data for novel research use is not without its challenges. Various ethical and legal considerations could potentially limit the ability to benefit from the secondary use of retrospective datasets, namely issues surrounding consent and the legal framework within which the datasets were generated. Regarding consent specifically, one must consider the purpose for which the datasets were collected and the scope of the informed consent provided by participants for the potential research uses of their data. Retrospective consent materials may have provided for limited use of collected datasets (e.g., research on a specific disease or disorder) or may be silent or even ambiguous as to their potential uses.[8] This may best be explained by the fact that many retrospective datasets were generated before data sharing became commonplace in research.[9] Consequently, the specific data use to which participants consented may limit the possibility of broad data sharing, as this possibility may not have been foreseen by researchers at the time of data collection.[9]

A further challenge in retrospective data integration is the legal framework within which the datasets were generated.[10] Specific legal requirements may limit the ability to exchange datasets across jurisdictional lines.[10] Significant differences between legal systems across jurisdictions can be an important barrier for research collaboration. Given that precision medicine relies on collaborative research across international borders, legal interoperability limitations pose a significant challenge for data sharing. At the same time, the potential benefits retrospective datasets can bring for precision medicine are considerable and should militate in favor of their usage to the greatest and most efficient extent possible. Data governance tools play a key role in retrospective data integration and should be used by researchers to

assess the ethical and legal permissibility of integrating retrospective datasets, allowing their maximum scientific potential. The Global Alliance for Genomics & Health's (GA4GH) recent consent policy is an illustrative case in point of a such a governance "toolkit." The GA4GH's policy serves as a guide for researchers on the sharing of genomic and health-related data and contains special considerations regarding retrospective datasets.[11] It provides a list of factors that researchers can review and assess prior to retrospective data sharing.

Under this guide, to make use of retrospective datasets, researchers should first review relevant consent materials and policy documents to assess whether broad data sharing was addressed at the time of data collection.[11] As previously mentioned, data sharing has not always been standard practice in research and, in many cases, broad consent for data sharing will not be foreseen in retrospective data collections. If not foreseen, researchers should then attempt to reconsent participants for the novel proposed use of their data, if possible and practical, or notify participants of this proposed use and give them the opportunity to opt-out if they do not consent to this use.[11] If neither reconsent or notification with opt-out are possible or practical, researchers may then consider sharing anonymized versions of datasets or obtaining a consent waiver from an Institutional Review Board to allow for broader or different use of the datasets.[11] Authors Tassé et al. (2016) similarly provide a three-step assessment process for retrospective data integration involving reviewing relevant consent materials, reconsent, and the possibility of obtaining a consent waiver.[8]

This three-step approach, therefore, helps guide researchers in assessing the usability of retrospective datasets for precision medicine research. Rather than automatically ruling out their potential use for a novel research purpose, researchers can use these tools and maximize the research value of these retrospective datasets. In this manner, they can help advance precision medicine initiatives in scientifically meaningful and impactful ways. Nevertheless, it should still be noted that these governance tools do not completely eliminate all barriers relating to the integration of retrospective datasets. For one, the possibility of obtaining a consent waiver, where applicable, will necessarily depend on domestic legal norms, that is, those in place in the jurisdiction in which the data was collected.[8] This can limit the possibility of using the retrospective datasets for a novel research purpose.[8] Nonetheless, these tools can help researchers in targeting usable retrospective datasets for precision medicine research in a streamlined manner, allowing researchers to make use of valuable existing resources in an ethically and legally compliant way.

12.2.2 Prospective data collection

Contrary to retrospective datasets, prospective data refer to data that has yet to be collected and that are required for a specific research project. Within the precision medicine context specifically, this will refer to data collected for purposes relating to precision medicine research. Still, many of the governance-related considerations related to prospective data

collection are similar to those concerning retrospective data integration, though they differ in their nature and scope. When prospectively collecting data, researchers will have the opportunity to directly and pre-emptively address some of the ethical and legal challenges previously discussed. In this manner, the data collection process can be optimized for the specific needs of precision medicine research. As previously stated, precision medicine greatly depends on the ability to widely share different types of datasets for study and analysis. Researchers should be cognizant of this key consideration when prospectively collecting data and how data governance plays a key role therein.

In our overview of retrospective data integration, we discussed how consent for data sharing is often a major barrier, given that many retrospective datasets that could be useful for precision medicine research were generated before data sharing become a common scientific practice. Today, data sharing is not only common practice but widely viewed as both an ethical and scientific imperative.[12] Global data sharing and interjurisdictional collaboration between researchers greatly facilitate the creation of statistically significant sample sizes and richer, more valuable datasets.[13] Consent materials for precision medicine research should therefore encapsulate permission for broad and global data sharing to eliminate potential downstream ethical hurdles. As precision medicine research is rapidly expanding and evolving, ensuring that data can be shared widely and easily for long periods of time is critical. Obtaining the appropriate consent for data sharing achieves this objective, while respecting participants' personal autonomy and self-determination.[13]

Consent in the precision medicine context can also extend to the possibility of linking different types of datasets. For instance, genetic and clinical data can be linked to determine the best possible treatment option for a particular patient.[14] Linkage of genomic data with population-based cancer registry records has been explored to inform individualized treatment and testing options for breast cancer patients.[15] Data linkage can also help improve data completeness and interpretation, while also reducing participants' burden through minimizing the collection of additional data.[13] Ensuring this possibility early on during the data collection process constitutes an important factor in advancing precision medicine initiatives.

While prospective data collection allows researchers to directly address how datasets may be used, interoperability can still pose a challenge, much as it does for retrospective data integration. Again, this is complicated by the increasingly globalized nature of data-sharing initiatives, particularly within the precision medicine milieu. Inevitably, this will lead to situations in which a complex network of different laws and regulations will have to be navigated to enable interjurisdictional data sharing, linkages, and analyses.[16] Data governance mechanisms should, therefore, be implemented to optimize the usage of the data in compliance with relevant legal and normative requirements. While this does not eliminate all interoperability barriers—indeed, some jurisdictions may be highly restrictive and territorial in the export of datasets—it can still serve to facilitate sharing and collaboration in the future.

This feature may be most illustrative in the context of international consortiums that bring together researchers from different institutions across global jurisdictional lines. Initiatives of this kind make collected clinical and genomic data widely available to the research community with the goal of fostering genomics research. Effective data governance is crucial for ensuring data interoperability in these types of international collaborative research initiatives. While full legal harmonization across jurisdictions will be difficult to achieve, data governance tools can nevertheless maximize data interoperability, maximizing scientific output.

Standardized consent clauses and data sharing agreements, for example, should be viewed as important data interoperability governance tools. Implementing common confidentiality and privacy protection measures can promote compliance between legal norms across jurisdictions. As an illustration, GA4GH's Regulatory & Ethics Toolkit provides for various consent clauses, policies, and other tools that can aid researchers in facilitating the open sharing of data across jurisdictional lines. Making use of such governance tools can help promote the exchange of data across different systems while respecting legal norms as well as participants' rights and interests. When prospectively collecting data, especially within international research consortiums, researchers should coordinate, to the extent possible, to standardize consent materials, protocols, and agreements to promote data interoperability. In short, while prospective data collection presents certain governance-related challenges, especially related to interoperability, it does afford researchers the opportunity to directly address these challenges at the time of collection, so that datasets may be used in the most efficient and impactful manner for precision medicine research. Making use of available data governance toolsets should be viewed as a facilitating factor for precision medicine research and can help guide researchers in their downstream data sharing plans.

12.2.3 Data access

The promotion and facilitation of data access are one of the most important functions of data governance. Data access is likely to have a significant impact on the future of precision medicine. For our purposes, we mean data access to refer to practices on the part of biobanks, research consortiums, and other genomics initiatives related to the sharing of genomic resources with external stakeholders for the purposes of conducting biomedical research. Data access regimes are conventionally described in governance framework documents, which set out the structure of and rules surrounding external data access. There are numerous ways to facilitate data access, though mechanisms are often open, registered, or controlled. Each of these options aims to balance multiple objectives, some of which are in considerable tension. How data governance enables access and the conditions to which external researchers are subject will play an important role in realizing precision medicine's promise.

Data access models can be thought to exist on a spectrum organized according to the degree of ex-ante review that modulates an external researcher's capacity to obtain genomic

data. At one end of the spectrum are open access models. Under this kind of approach, data is freely distributed on request. Often, it is accessible through an open online portal or is distributed through a shared cloud computing network. In the center of the spectrum are registered access models, which aim to find a middle ground between the flexibility of open access and the stringency of controlled access. In a registered model, proposed data uses are not reviewed on a case-by-case basis. Rather, applicant researchers are reviewed to ensure, among other things, that they are a *bona fide* researcher holding an eligible research position at an academic or clinical institution.[17] Controlled access models, then, exist at the farthest end of the stringency spectrum. In this kind of approach, proposals for data access are reviewed and approved on a case-by-case basis.[18] Project approval in a controlled access regime is usually facilitated by an application and review process overseen by a "data access committee" (DAC).[19] Access to controlled resources is often managed through an online application process in which researchers convey project information, including a description of project objectives and methods, an accessible project summary, and applicable research ethics approval. Applicant researchers and an institutional signing authority sign a data access agreement binding the research team to certain data management and research use conditions.

Each of these models offers unique strengths. Open access models, for example, have the advantage of being highly administratively flexible. Data access does not usually require filing application paperwork, awaiting access decisions, or demonstrating ethics approval. Open access might also enable greater diversity of access to genomics data. All potential data users, whether professional researchers or citizen scientists, are on equal footing with respect to their ability to use genetic resources in their work. These factors might generate significant research value, but they come at a cost. These kinds of access regimes might, for example, produce risks to data subject privacy. Considering that open access models do not typically formally restrain data users from processing data in combination with that obtained from other sources, there is a risk that individual subjects, even when data were initially anonymized, could be reidentified. Biobanks, research initiatives, and data repositories might also find it difficult to recruit research participants willing to have their personal information shared for an indeterminate range of purposes and with an undefined group of people. Registered and controlled access models attempt to address these weaknesses without substantially encumbering access to genomic resources. In both models, researchers will typically agree to keep accessed data confidential, refrain from intentionally reidentifying data subjects, and use accessed data only for purposes outlined in the project consent documents. Registered access promises to facilitate data access with somewhat greater efficiency than controlled models. Researchers might view controlled access regimes as unnecessarily bureaucratic and time intensive. Preparing application documents, responding to DAC comments, and obtaining institutional signatures on data access agreements all requires time and considerable effort. But these procedures may, apart from working to safeguard participant privacy, also help to promote public trust in genomic data sharing. To the extent that research participants worry

that genomic data sharing entails potential risk to their personal privacy, a controlled access regime may work to palliate this concern.

In practice, governance decisions about data access significantly impact the rate at which data is shared, the administrative burdens on researchers, and the capacity of participants to forge relationships with research initiatives founded on trust and mutual respect. As a way of illustrating some of the challenges raised by the governance of data access, we might consider the development of federated access regimes within the context of large constitutive cohorts. Researchers are increasingly assembling "cohorts of cohorts," initiatives composed of several subsidiary cohorts. These efforts promise to promote increased research efficiency and harmonization. A notable example is a Canadian Partnership for Tomorrow's Health (CanPath), formerly the Canadian Partnership for Tomorrow Project, the largest populational health study in Canada.[20] Composed of seven regional cohorts spanning the country, CanPath facilitates harmonized access to genomic data collected from more than 300,000 Canadians. CanPath's access regime permits researchers "who want to obtain core questionnaire data or biosamples from 2 or more regions to benefit from a 'one-stop shop' process for access requests."[20] Rather than submit multiple access applications, researchers engage with a single controlled process enabling access to multiple cohorts. While the benefit of increased access efficiency for researchers often justifies these efforts, creating single points of access for data initially collected by multiple cohorts presents an array of governance challenges. In an earlier part of the chapter, we described governance issues associated with the retrospective collection of genomic data. By analogy, the retrospective design of harmonized access governance might be similarly challenging.

Consider a group of several independent cohorts, each initiated for similar kinds of genomic research and each with their own data access procedures. Harmonizing data access across these cohorts would, by reducing administrative burden on applicant researchers, improve data access efficiency and commensurately increase the statistical power of associated work. What governance considerations would need to be addressed in order to transform this sort of multiple-track access model into a one-stop shop? For one thing, cohorts in custody of genomic data are limited in their sharing capacities by participant consents. A model of harmonized access must account for limitations on data sharing to which participants have consented; harmonized access is not possible, for example, where participants have consented to data collection on the promise that their data would not be shared. To be sure, few modern genomics research cohorts would make such a promise. Much more likely is that participants will consent to data sharing only within a certain jurisdiction or with certain kinds of researchers. In any case, the bounds of participant consent will operate as a primary control over what is ethically and legally permissible in the creation of a harmonized access regime. Another challenge is that harmonized access models will need to be developed in consideration of the layers of contractual relationship between researchers, institutions, and cohorts. Certain cohorts, for example, might require that agreement for researcher access to genomic

data are entered between specific institutions or include specific provisions. Requirements of this kind will require careful consideration and, potentially, accommodation. One option is for individual cohorts to derogate from standard operating practices by entering into an agreement between themselves, setting out the terms according to which harmonized access will operate.

It is worth stressing that while decisions about data access governance within individual cohorts will have a significant impact on precision medicine, harmonized access may play an even more powerful role. As we described above, broad data sharing performs a crucial function in ensuring that the kind of research capable of producing targeted therapies is maximally productive. Harmonized access models generate precisely the efficiencies useful in achieving this goal. It may be, as a consequence, that challenges associated with creating governance instruments that adequately address the above concerns may be well worth the effort. Likely the most direct approach for designing harmonized access governance is to first understand how already existing cohort governance documents and policies overlap. By identifying areas of agreement and divergence, particularly with respect to limitations on sharing and requirements for data access agreements, a common denominator access model may be developed. To be sure, large research consortiums aggregating data from multiple projects will need to consider how regulatory frameworks and local law applicable in each of the initiative's constituting projects might limit how genomic data are shared with external researchers. Governance decisions for data access, whether within a single cohort or for harmonized access, will need to attempt to reconcile perspectives that are sometimes in tension or disagreement. Critically, the governance decisions made in the creation of a data access model will have a substantial effect on the conduct of genomic research and on the realization of precision medicine. As we outlined above, two related factors have uniquely considerable implications for the success of precision medicine: the sharing of genomic data and public trust. Data access governance models modulate both of these considerations by controlling the manner in which genomic data is shared and by protecting the interests of participants.

12.3 The road ahead: how data governance should shape the future of precision medicine

Significant strides have been made in the postgenomic era to advance precision medicine. These efforts have resulted in greatly improved and enhanced clinical care. It is now generally possible to approach disease treatment and prevention in a significantly more patient-centered manner, contrasting with the disease-focused approach that characterized earlier eras. But precision medicine has yet to realize its full potential. Governance gaps in data sharing and data interoperability have limited progress in precision medicine, which greatly depends on the collection, sharing, and use of large volumes of individual genomic data.

In this chapter, we have illustrated some of the ways that decisions made about data governance play an indispensable role in shaping the future of precision medicine and, specifically, how retrospective data integration, prospective data collection, and data access can be important facilitating factors in helping precision medicine realize its full potential. We have considered how data governance tools can help researchers overcome some of the challenges related to the integration of retrospective datasets, thereby allowing researchers to make use of these valuable resources. Indeed, retrospective datasets are critical in generating statistically significant research results in a time- and cost-efficient manner. However, retrospective datasets pose unique ethical and legal challenges, which could potentially limit researchers' ability to leverage this highly valuable resource. Data governance tools can help circumvent many of these challenges, making way for enhanced research opportunities in precision medicine. Similarly, data governance ensures that prospective datasets are able to be widely shared, linked, and analyzed to meet the specific needs of the precision medicine context. Prospective data collection affords researchers the opportunity to directly address some of the data governance challenges currently facing precision medicine. We demonstrated how standardizing consent and data processes, for instance, can help overcome ethical, legal, and interoperability barriers to widespread precision medicine implementation.

Perhaps most critically for the future progress of precision medicine, data access regimes should be detailed in governance frameworks for precision medicine initiatives to facilitate external access while respecting participant interests and applicable ethics rules. As precision medicine efforts move forward, the need for adequate data governance mechanisms will only become more important, especially in the context of large-scale and international precision medicine initiatives. Greater sets of technical, legal, and ethical norms will inevitably come into play in the precision medicine ecosystem, necessitating the need for appropriate and efficient data governance mechanisms. The role of data governance as an enabler toward the widespread implementation of precision medicine will therefore only continue to grow. These developments depend meaningfully on the joint effort of a diverse array of stakeholders, including researchers, research ethics boards, DACs, and research participants. Understanding and engaging with the perspectives of each of these groups of stakeholders will be essential for developing and implementing effective data governance regimes. In particular, data governance will require engaging with communities and the public, both at the moment a governance regime is being developed, as well as throughout its lifecycle. Periodic participant contact and public communication can help to engender trust in precision medicine and promote further participation in ongoing health research. Securing researcher confidence similarly requires a transparent process. DACs and research ethics boards can work to promote such confidence by ensuring open communication and cooperation. In the case of DACs, it is particularly critical that governance strategies incorporate clear and accessible terms of reference that promote efficiency and are responsive to changing research dynamics. DACs and research ethics boards, as intermediaries between researchers, data subjects, patients, and

the public, play an essential role structuring governance practices in ways that both assure due diligence in data access and the efficient distribution of valuable scientific resources.

In years to come, developments are likely to arise that can facilitate the creation of robust data governance for precision medicine initiatives. Automation technologies, such as advanced algorithms, for instance, can become a critical component of data governance. Indeed, this can be particularly efficacious in managing large datasets from different jurisdictions, which can greatly complicate data governance. Harnessing automation technologies can help precision medicine stakeholders overcome many of the challenges posed by data management, such as compliance with regulatory requirements and data sharing restrictions.[21] In this manner, data governance will be able to evolve alongside the changing and advancing precision medicine landscape, allowing precision medicine to realize its objective of improving and enhancing patient care.

References

1. O'Doherty KC, Shabani M, Dove ES, Bentzen HB, Borry P, Burgess MM, et al. Toward better governance of human genomic data. *Nat Genet*. 2021;53:2–8.
2. World Economic Forum. *Precision Medicine Vision Statement: A Product of the World Economic Forum Global Precision Medicine Council*. Geneva: World Economic Forum; 2020.
3. Blasimme A, Fadda M, Schneider M, Vayena E. Data sharing for precision medicine: policy lessons and future directions. *Health Aff*. 2018;37:702–709.
4. Bernier A. Rare disease data stewardship in Canada. *FACETS*. 2020;5:836–863.
5. Boycott KM, Ardigo D. Addressing challenges in the diagnosis and treatment of rare genetic diseases. *Nat Rev Drug Discovery*. 2018;17:151–152.
6. Thorogood A. International data sharing and rare disease: the importance of ethics and patient involvement. In: Wu ZH, ed. *Rare Diseases*. London: IntechOpen; 2020:221–236.
7. Li X, Song Y. Target population statistical inference with data integration across multiple sources—an approach to mitigate information shortage in rare disease clinical trials. *Stat Biopharm Res*. 2020;12: 322–333.
8. Tassé AM. A comparative analysis of the legal and bioethical frameworks governing the secondary use of data for research purposes. *Biopreserv Biobanking*. 2016;14:207–216.
9. Wallace SE, Kirby E, Knoppers BM. How can we not waste legacy genomic research data? *Front Genet*. 2020;11:446.
10. Tassé AM. From ICH to IBH in biobanking? A legal perspective on harmonization, standardization and unification. *Stud Ethics Law Technol*. 2013;7:1–15.
11. Global Alliance for Genomics and Health. Consent Policy; 2019. https://www.ga4gh.org/wp-content/uploads/GA4GH-Final-Revised-Consent-Policy_16Sept2019.pdf [accessed August 21, 2002].
12. Bredenoord AL, Mostert M, Isasi R, Knoppers BM. Data sharing in stem cell translational science: policy statement by the International Stem Cell Forum Ethics Working Party. *Regen Med*. 2015;10:857–861.
13. Nguyen MT, Goldblatt J, Isasi R, Jagut M, Jonker AH, Kaufmann P, et al. Model consent clauses for rare disease research. *BMC Med Ethics*. 2019;20:1–7.
14. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. *Nature*. 2015;526: 336–342.
15. Kurian AW, Friese CR. Precision medicine in breast cancer care: an early glimpse of impact. *JAMA Oncol*. 2015;1:1109–1110.

16. Parchomovsky G, Mattioli M. Quasi-patents and semi-patents in biobanking. In: Pascuzzi G, Izzo U, Macilotti M, eds. *Comparative Issues in the Governance of Research Biobanks*. Berlin and Heidelberg: Springer; 2013:251–266.
17. Dyke SOM, Linden M, Lappalainen I, et al. Registered access: authorizing data access. Eur J Hum Genet. 2018;26:1721–1731.
18. Thiebes S, Schlesner M, Brors B, Sunyaev A. Distributed ledger technology in genomics: a call for Europe. Eur J Hum Genet. 2020;28:139–140.
19. Shabani M, Dyke SOM, Joly Y, Borry P. Controlled access under review: improving the governance of genomic data access. PLoS Biol. 2015;13:1–6.
20. Dummer TJB, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. CMAJ. 2018;190:E710–E717.
21. Paul S, Gade A, Mallipeddi S. The state of cloud-based biospecimen and biobank data management tools. *Biopreserv Biobanking*. 2017;15:169–172.

Index

Page numbers followed by "*f*" and "*t*" indicate, figures and tables respectively.

Genomic Data Sharing

Case Studies, Challenges, and Opportunities for Precision Medicine

Edited by

Jennifer B. McCormick, PhD, MPP
Associate Professor, Department of Humanities, College of Medicine, Pennsylvania State University, Hershey, PA, United States

Jyotishman Pathak, PhD
Professor of Medical Informatics, Professor of Psychiatry, Chief of the Division of Health Informatics, Vice Chair of the Department of Population Health Sciences, Weill Cornell Medicine, New York-Presbyterian Hospital, New York, United States

Over the last decade, increasing emphasis has been placed on biomedical data sharing with the goal of making data sets publicly available to a wide range of researchers. This broad data sharing is critical to move the genomic science forward. However, it is equally important that the ethical, legal, and social issues (ELSI) including privacy and confidentiality, access, informed consent, and return of both individual and aggregate research results to participants are recognized and addressed.

Genomic Data Sharing: Case Studies, Challenges, and Opportunities for Precision Medicine provides a comprehensive overview of current and emerging issues, such as intellectual property, informed consent, and data governance in genomic data sharing as applied in new genomic research and precision medicine. Here, international leaders in genomic data sharing examine these issues in-depth, and offer practical case studies highlighting key successes, challenges, and opportunities. Cases of data sharing discussed come from the Marshfield Clinic Biobank, All of Us Research program, Global Alliance for Genomics and Health (GA4GH), and Human Heredity and Health in Africa (H3Africa). In addition to these perspectives from the front lines, this book provides succinct overviews of ethical, legal, social, and informatics related challenges relevant to research employing and sharing genomic data for the academic research and academic administrator. Clinician investigators, clinicians affiliated with academic medical centers, policymakers, and regulators will also gain insight allowing them to navigate the increasingly complex ethical, social, and technical landscape of genomic data sharing.

Key features

- Covers both technical and ELSI (ethical, legal, and social implications) perspectives on genomic data sharing
- Includes applied case studies of how data are being shared
- Features chapter contributions from international leaders in genomic data sharing

ELSEVIER

ACADEMIC PRESS
An imprint of Elsevier
elsevier.com/books-and-journals

ISBN 978-0-12-819803-2