Please cite this article in press as: Lewis et al., Navigating data sharing in research, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2026.04.004

PERSPECTIVE

Navigating data sharing in research

Anna C.F. Lewis,[1,2] Ellen W. Clayton,[3] Hana Bangash,[4] Harris T. Bland,[5] Katherine E. Bonini,[6] James J. Cimino,[7] John J. Connolly,[8] Robert R. Freimuth,[9] Stephanie M. Fullerton,[10] Hakon Hakonarson,[8] Ingrid A. Holm,[2,11] Gail P. Jarvik,[12] Atlas Khan,[13] Krzysztof Kiryluk,[13] Jodell E. Linder,[14] Yuan Luo,[15] John A. Lynch,[16] Kathleen F. Mittendorf,[17] Jennifer A. Pacheco,[18] Luke V. Rasmussen,[15] Susannah L. Rose,[3,5] Robb Rowley,[19] Richard R. Sharp,[20] Theresa L. Walunas,[21] Wei-Qi Wei,[5] Chunhua Weng,[22] and Maya Sabatello[13,23,*]

Summary

Sharing biomedical research data can accelerate scientific discovery, leading funders and journals to increasingly mandate sharing. However, data openness must be balanced with protecting research participants from harm in an evolving legal and social landscape. Drawing on experiences from the Electronic Medical Records and Genomics (eMERGE-IV) Network—a US-based, multi-site consortium gathering genomic and medical data focused on underrepresented groups to refine disease risk prediction—we examine challenges in implementing data sharing that are "as open as possible, as closed as necessary." Recent US legal developments, including the Dobbs decision and gender-affirming care bans, highlight the urgency of considering data-sharing risks and required the Network to rethink strategies to prevent individual- and group-level harms from genomic analyses. eMERGE-IV implemented several strategies to mitigate concerns, including cell suppression for race/ethnicity data and not extracting certain diagnostic codes from participants' electronic health records. These decisions balanced immediate protection and long-term scientific benefits for relevant populations. Participant agreement to broad data sharing in informed consent is often required for research participation to make data as open as possible. No consent form, however, can define the terms of "as closed as necessary"—a construct that is subject to sociolegal changes across the life cycle of research studies. Providing protection requires robust data governance, including engagement with prospective and actual participants. The research enterprise must reconsider its consenting approach and develop transparent, inclusive governance structures responsive to evolving vulnerabilities while maintaining scientific progress. Public trust depends on the research enterprise successfully navigating these competing demands.

Introduction

Medical research aims to understand the causes of disease and to identify approaches to improve health outcomes. Extensive individual-level data are essential for this work. Sharing research data is cost-effective, allows result verification, provides opportunities to study new questions, and holds out the promise of advancing health for all.[1] Such sharing is increasingly expected, even required, by scientific journals and funders like the National Institutes of Health (NIH). Yet these data can be and have been used in ways that harm both the individuals to whom they pertain and larger groups, through breaches of privacy and the promotion of stigma.[2–4] Thus, advancing these important scientific goals must be accompanied by the research enterprise's ethical responsibility to ensure participants' data are used in ways that promote benefit and prevent harm. This responsibility, shared by investigators, their institutions, and data holders, is heightened in studies enrolling groups that have been previously harmed or underrepresented in biomedical research. This tension requires deciding what data to share, who has access, which future research questions can be studied, and who gets to decide.[5]

We share perspectives from phase four of the Electronic Medical Records and Genomics (eMERGE-IV) Network, a large, 10-site consortium funded by the National Human Genomics Research Institute (NHGRI). The funding opportunity both encouraged recruiting "persons who come from racial or ethnic minority populations, underserved

[1]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; [2]Harvard Medical School, Boston, MA, USA; [3]Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, TN, USA; [4]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA; [5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [6]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [7]Biomedical Informatics and Data Science, Heersink School of Medicine at the University of Alabama-Birmingham, Birmingham, AL, USA; [8]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [9]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA; [10]Department of Bioethics & Humanities, University of Washington School of Medicine, Seattle, WA, USA; [11]Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA; [12]Department of Medicine, University of Washington Medical Center, Seattle, WA, USA; [13]Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA; [14]Vanderbilt Institute of Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, USA; [15]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [16]School of Communication, Film, and Media Studies, University of Cincinnati, Cincinnati, OH, USA; [17]Division of Hematology/Oncology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; [18]Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ, USA; [19]General Internal Medicine, Mayo Clinic, Rochester, MN, USA; [20]Biomedical Ethics Program, Mayo Clinic, Rochester, MN, USA; [21]Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [22]Department of Biomedical Informatics, Columbia University, New York, NY, USA; [23]Department of Medical Humanities and Ethics, Columbia University Irving Medical Center, New York, NY, USA

*Correspondence: ms4075@cumc.columbia.edu

https://doi.org/10.1016/j.ajhg.2026.04.004.

Please cite this article in press as: Lewis et al., Navigating data sharing in research, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2026.04.004

populations, or populations who experience poorer medical outcomes"[6] and required data sharing. Here, we discuss enduring and evolving scientific, legal, and ethical challenges in sharing large datasets, particularly those that combine genetic, electronic health record (EHR), and participant self-reported data, while avoiding immediate and downstream harms.

We first outline the growing pressure to share research data despite the potential associated risks to both individuals and groups. We then describe how eMERGE-IV navigated the tensions and sought to share data for research while mitigating harms and complying with federal and local data privacy laws. Next, we argue that meeting participants' and the public's likely expectations—rather than limiting protection of participants' interests to the strict language of consent forms—offers a useful frame for these decisions. We conclude by considering what the broader research ecosystem must do to keep shared data as "open as possible and as closed as necessary."[3]

A world of increasing data sharing

Biomedical data-sharing requirements have grown over time. Federal funders increasingly encourage, even require, researchers to maximize the sharing "with appropriate protections" of data generated with public funds, beginning with NIH's 2014 Genomic Data Sharing policy,[7] followed by a broader policy for data management and sharing (DMS).[8] The DMS policy requires researchers intending to generate scientific data to submit a DMS plan with their grant application, comply with the plan, and obtain NIH approval on any subsequent revisions. More recently, the NIH has increased funding to develop sustained repositories to enable sharing while instituting federal data access committees (DACs) to help ensure that investigators proposing to use the data have legitimate needs and credentials for their use.[9]

In the US, participant research data without identifiable private information are not considered "human subjects" under the Common Rule and therefore do not fall within its protections.[10] In particular, institutional review boards (IRBs), which implement the Common Rule, typically do not require ongoing review of projects using such data after the determination of non-human-subject research. Typically, data, once de-identified, are submitted to "controlled access" repositories. Access requests go to a DAC, which evaluates whether proposed studies align with data use limitations (generally based on participants' "informed," typically broad consent) and determines whether the requesting institution and researcher are *bona fide*.[11] This determination process may be insufficient to prevent harm, as evidenced by the allegedly inappropriate access granted to users who misused data from the federally funded Adolescent Brain Cognitive Development (ABCD) study for race research.[3]

A world of increasing vulnerabilities

Vulnerability to direct harm from research participation and data sharing—including loss of privacy with potential legal and social consequences—is not new. A systematic review revealed that the most common concerns of prospective participants in precision medicine research relate to privacy, security, and confidentiality.[12] These concerns are commonly addressed via security measures.[13] Certificates of confidentiality (CoCs) reduce the risk by prohibiting disclosure of research data to anyone not connected to the research, particularly in legal processes, and are automatically granted to NIH-funded research and data repositories. Yet, advances in reidentification technologies[14] and the rapidly evolving legal and policy landscape have altered levels of risks to individuals and groups.

For example, the Dobbs vs. Jackson Women's Health decision in 2022,[15] which eliminated the constitutional right to abortion, and laws that financially incentivize reporting an abortion (e.g., in Texas) might make women reconsider participating in studies involving sharing EHR data that may contain codes for pregnancy termination. This concern is not abstract: prior data support lower research enrollment in states with restrictive abortion policies.[16] Transgender individuals may similarly rethink taking part in research that involves data sharing of any kind, given recent bans on gender-affirming care, healthcare coverage, and other rights[17] as well as actions that risk patient privacy and confidentiality in clinical care, such as the Department of Justice (DOJ)'s sending subpoenas for identifying health information to clinics that provide gender-affirming care.[18] Over time, such concerns may compromise research outcomes, as groups who are most directly impacted by changes in politics and policy will more likely decline research participation and thus may receive fewer benefits from studies that address their health needs.

Researchers often consider de-identification, which mitigates but may not eliminate privacy risks since reidentification may nonetheless be possible in unique combinations of participant demographics and/or unusual data distributions. Studies must thus weigh how to guard against these risks—especially for genomic information collection, which can, under certain circumstances, be used to identify individuals—while maintaining the benefits of sharing comprehensive data.[19] Addressing the tension between risk and benefit further requires investigators to decide what participant data are made available for wider sharing in light of potential (mis)uses of data by secondary users. Researchers can choose not to collect, or to collect without sharing, data from individuals or communities where reidentification would expose them

to heightened risk of harm. Such approaches may, however, limit future research that could benefit such groups.

In addition to potential direct harm to individual participants, research can be stigmatizing to groups by the ways data are analyzed, interpreted, and represented. This concern has been raised regarding communities for whom there is a history of mistreatment in research studies, such as Black,[20] Indigenous,[21] and disability communities.[22] Such harms could be intentional but may also arise when investigators are unaware of possible harms or where there is disagreement over what could be harmful.

Harms to individuals and groups can significantly undermine participants' trust in research.[23] Evidence further suggests that the trustworthiness of researchers, research institutions and the broader research enterprise can suffer if data are accessed and used in ways that clash with participants' values and expectations, beyond the provisions in the informed consent form.[24] The risk for loss of research trustworthiness is exacerbated if individual or group harms manifest.[25,26] The resulting mistrust can compromise otherwise well-meaning research that promotes care for all people. Accordingly, eMERGE-IV adopted measures to protect participants from both individual and group harms. Below, we discuss these decisions, the process for their determination, and their advantages and shortcomings.

Protecting against individual harms

As is common in translational genomic research studies, eMERGE-IV participants were informed about both data de-identification and risk for reidentification during the consent process. The consent form stated circumstances whereby "we might need to give out your name or other information that identifies you," including situations that could cause harm to the participant or others, and when necessary to comply with US laws. Participants were also told that their privacy would be protected to the maximum extent possible. This combination reflects an inherent challenge in consent forms: they inform participants about potential risks and researchers' commitment to take steps to reduce them but cannot guarantee absolute protection. In addition, there is overwhelming data that consent forms are poorly understood.[27] We examine the balancing act required for this commitment to be fulfilled.

The risk of reidentification can be reduced via statistical disclosure control, a set of techniques that restrict the amount of data released or modify them—for example, by adding noise to the data.[28] One way in which eMERGE-IV chose to reduce reidentification risks is by using the technique of cell suppression, i.e., combining categories with small numbers of participants. For example, eMERGE-IV collected participants' self-reported race and/or ethnicity via a single question with multiple possible responses (per the updated Office of Management and Budget Categories[29]), including follow-up subcategory questions. Following discussions, eMERGE-IV opted for internally aggregating data with responses from fewer than 10 individuals—a threshold also used by the Centers for Medicare and Medicaid Services[30]—and externally sharing only the top-level categories. These decisions run counter to recommendations emphasizing the value of more granular data[31–33] and risk obscuring meaningful variation or rendering certain groups "invisible." In our cohort, the Network identified cell suppression as important to reduce risks for reidentification and to increase participant privacy protection, thus aligning with our responsibility to participants.

To further minimize individual harm, the eMERGE-IV Network collectively considered whether certain diagnosis or procedure codes in the EHRs should be excluded from participant data collection. Each site suggested codes that should not be included, a follow-up survey determined the level of agreement across sites, and an all-site discussion resulted in consensus on which codes to leave out of data pulls. Agreed-upon codes to exclude from EHR extraction were those associated with elective pregnancy termination, gender dysphoria/transition, and suspected/actual child abuse (see Table 1). Codes ultimately included in the data pulls, despite having been considered for exclusion, included those related to medical errors and adverse events, substance use, mental health, and sex-specific and/or hormone-related codes (a broad category including reproductive organ issues and sexually transmitted infections). Decisions about which codes to collect or not were ultimately driven by the study's scientific goals. For example, hormone-related categories are intertwined with breast and prostate cancer—diseases that are central to eMERGE's investigation—and were collected. More generally, while the data collected included codes that are loosely connected to potentially stigmatizing traits, we avoided codes that may be used more directly as the basis for legal action against participants.

In eMERGE-IV, reidentification risks are also heightened by the unique integration of longitudinal EHR data with other information. Repeated encounters over many years can create unique clinical and temporal fingerprints that can allow identity inference. For example, a distinctive sequence of diagnoses, medication changes, or procedure dates can link an otherwise de-identified record to a single individual when cross-referenced with public or commercial data. To mitigate these risks, eMERGE-IV employed temporal aggregation (e.g., collapsing dates to broader intervals such as month or year), event sampling (sharing representative rather than complete encounter histories), and localization control (removing or generalizing facility and provider identifiers).

One could argue that decisions about what to include in data pulls should be informed by the views and expectations of prospective participants since they are directly in harm's way. In eMERGE-IV, pre-implementation

Please cite this article in press as: Lewis et al., Navigating data sharing in research, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2026.04.004

Table 1. List of codes excluded from EHR data extraction

Issue	Concept class	Vocabulary	Category	Example (code and term)
Early termination of pregnancy	diagnosis	ICD10CM	direct codes for elective termination of pregnancy	Z33.2, encounter for elective termination of pregnancy
Early termination of pregnancy	diagnosis	ICD10CM	failed medical abortion	O07, failed attempted termination of pregnancy
Early termination of pregnancy	diagnosis	ICD9CM	codes indicating complications from a procedure	639, complications following abortion and ectopic and molar pregnancies
Early termination of pregnancy	diagnosis	ICD9CM	possible complications after procedure	996.32, mechanical complication due to intrauterine contraceptive device
Early termination of pregnancy	diagnosis	ICD9CM	induced abortion	635, legally induced abortion
Early termination of pregnancy	diagnosis	ICD9CM	codes indicating the involvement of legal authorities or social services	E978, legal execution
Early termination of pregnancy	diagnosis	ICD9CM	history of past medical procedures	V45.77, acquired absence of other genital organ(s)
Early termination of pregnancy	diagnosis	ICD9CM	termination of pregnancy due to medical necessity for the health of the fetus or newborn	779.6, termination of pregnancy (fetus)
Early termination of pregnancy	procedure	CPT	abortion	59840, induced abortion, by dilation and curettage
Early termination of pregnancy	procedure	CPT	treatment of incomplete abortion	59812, treatment of incomplete abortion, any trimester, completed surgically
Early termination of pregnancy	procedure	CPT	induced abortion	59855, induced abortion, by 1 or more vaginal suppositories (e.g., prostaglandin) with or without cervical dilation (e.g., laminaria), including hospital admission and visits, delivery of fetus and secundines
Early termination of pregnancy	medication	HCPCS	medication used in chemically induced abortion	S0190, mifepristone, oral, 200 mg (Mifeprex)
Gender-affirming care	diagnosis	ICD10CM	psychological	F64, gender identity disorders
Gender-affirming care	diagnosis	ICD10CM	encounters for treatment and care related to gender transition	Z87.890, personal history of sex reassignment
Gender-affirming care	diagnosis	ICD10CM	physical indication	Q52.X, other congenital malformations of female genitalia
Gender-affirming care	diagnosis	ICD9CM	psychological indication	302, sexual and gender identity disorders
Gender-affirming care	diagnosis	ICD9CM	hormone treatment	255.2, adrenogenital disorders
Gender-affirming care	procedure	CPT	vaginal surgery	construction of artificial vagina
Gender-affirming care	procedure	CPT	breast surgery	19325, mammaplasty, augmentation; with prosthetic implant[a]
Gender-affirming care	procedure	CPT	urethroplasty	53420, urethroplasty, 2-stage reconstruction or repair of prostatic or membranous urethra; first stage[b]
Gender-affirming care	procedure	CPT	surgery of male genitalia	54125, amputation of penis; complete[b]
Gender-affirming care	procedure	CPT	intersex surgery	55970, intersex surgery, male to female
Gender-affirming care	procedure	CPT	hysterectomy	58150, total abdominal hysterectomy (corpus and cervix), with or without removal of tube(s), with or without removal of ovary(s)[a]
Suspected child abuse	diagnosis	ICD9CM	nonaccidental injuries	Y07.1, parent (adoptive) (biological), perpetrator of maltreatment and neglect
Suspected child abuse	diagnosis	ICD9CM	fetus or newborn affected by maternal conditions that may be unrelated to present pregnancy	760.71, alcohol affecting fetus or newborn via placenta or breast milk
Suspected child abuse	diagnosis	ICD9CM	other conditions originating in the perinatal period	7795, drug withdrawal syndrome in newborn

(*Continued on next page*)

Please cite this article in press as: Lewis et al., Navigating data sharing in research, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2026.04.004

Table 1. ***Continued***

Issue	Concept class	Vocabulary	Category	Example (code and term)
Suspected child abuse	diagnosis	ICD10CM	drug-induced mental disorders	F19950, other psychoactive substance use, unspecified with withdrawal, unspecified
Suspected child abuse	diagnosis	ICD10CM	congenital malformations, deformations, and chromosomal abnormalities	Q86.0, fetal alcohol syndrome (dysmorphic)
Suspected child abuse	diagnosis	ICD10 CM	counseling for perpetrator of physical/ sexual abuse	Z69021, encounter for mental health services for perpetrator of non-parental child abuse

ICD9CM, International Classification of Diseases, 9th Edition with Clinical Modifications; ICD10CM, International Classification of Diseases, 10th Edition with Clinical Modifications; CPT, Current Procedural Terminology; HCPCS, Healthcare Common Procedure Coding System.
[a]Remove IF record contains ICD9CM or ICD10CM codes related to gender identity or other sexual disorders, personal history of sex reassignment, congenital malformations of genitalia, OR participants who have any of the following from any source that are not "female" or "woman": sex assigned at birth, sex, legal sex, gender identity; they can be included if all variables listed above are female/woman.
[b]Remove IF record contains ICD9CM or ICD10CM codes related to gender identity or other sexual disorders, personal history of sex reassignment, congenital malformations of genitalia, OR participants who have any of the following from any source that are not "male" or "man": sex assigned at birth, sex, legal sex, gender identity; they can be included if all variables listed above are male/man.

patient feedback focused primarily on the creation of understandable patient materials, while details of EHR outcome measures and related phenotyping variables were left to later phases of the study. The sociolegal developments that occurred during the study (the Dobbs decision and policies on gender-affirming care) highlighted the growing potential risks to participants and the need for the Network to center attention on our responsibility to study participants. Indeed, transgender or gender-diverse eMERGE-IV participants, when interviewed, expressed concerns about the safety of sharing gender identity data given the changing sociolegal landscape and fear that malicious actors may access these data (unpublished data). Future study designs should include co-development of data-sharing plans with community members to address policy landscape changes and revisit these definitions, as those at risk from harm are ever-changing.

Protecting against group harms

Efforts to reduce the risk of group harm raise further concerns, not least because of the requirement that data be placed in a centralized federally controlled repository and made broadly available to both Network and non-Network members. While such repositories promote broad scientific inquiry, their structure means researchers lose control over future uses of the collected data, despite their ongoing ethical responsibility to their participants. It thus raises the need for both internal and external measures before data release that can balance these duties and mitigate potential group harm.

During the course of eMERGE, several strategies were implemented to address these challenges. eMERGE researchers request data access through a manuscript concept sheet, which is then circulated to eMERGE-IV members for feedback, to invite collaboration, and to address issues prior to approval. Concept sheets contain a checkbox for researchers to indicate whether their work involves data that could be considered sensitive or stigmatizing. If checked, the concept sheet undergoes additional review by the steering committee and potential referral to relevant external experts, including community members. Individual sites may choose not to include data from their site in specific analyses or manuscripts, should they find the risks to participants outweigh the benefits. Concept sheets are also used by researchers external to eMERGE-IV and require that an eMERGE collaborator be concurrently involved in the proposed study to enhance continued implementation of eMERGE-IV decisions and internal guidelines during the life of the study. The concept sheet process helps to prospectively identify potential scientific or ethical concerns and provides a forum to address them through a dialogue and prior to data analyses. The vetting process, however, is limited to direct requests to access eMERGE-IV data during the course of its funding.

What might replace this process when the collected data are shared externally for secondary use? Here, eMERGE-IV faces unresolved challenges. As mentioned above, each study site was required to transfer research data to a centralized repository and make them available to consortium members for internal use. Most of these data will also be made available to non-consortium members via the Analysis Visualization and Informatics Lab-space (AnVIL), NHGRI's platform for sharing large-scale genomic and linked biomedical data.[34] Thus, unlike other studies, such as the UK Biobank[35] or the All of Us Research Program,[36] eMERGE-IV cannot protect the data once they are deposited. For example, none of the above strategies implemented internally to mitigate potential harms are available for secondary data uses after data are posted in a centralized repository.

Instead, NHGRI's DAC controls access to these data based "primarily on conformance" of the researcher's request with the data use limitations established by the depositing institution.[36] Under this governance system, data users, who must be affiliated with an institution, are required to sign a legally binding data use agreement allowing the NIH to hold both researchers and their

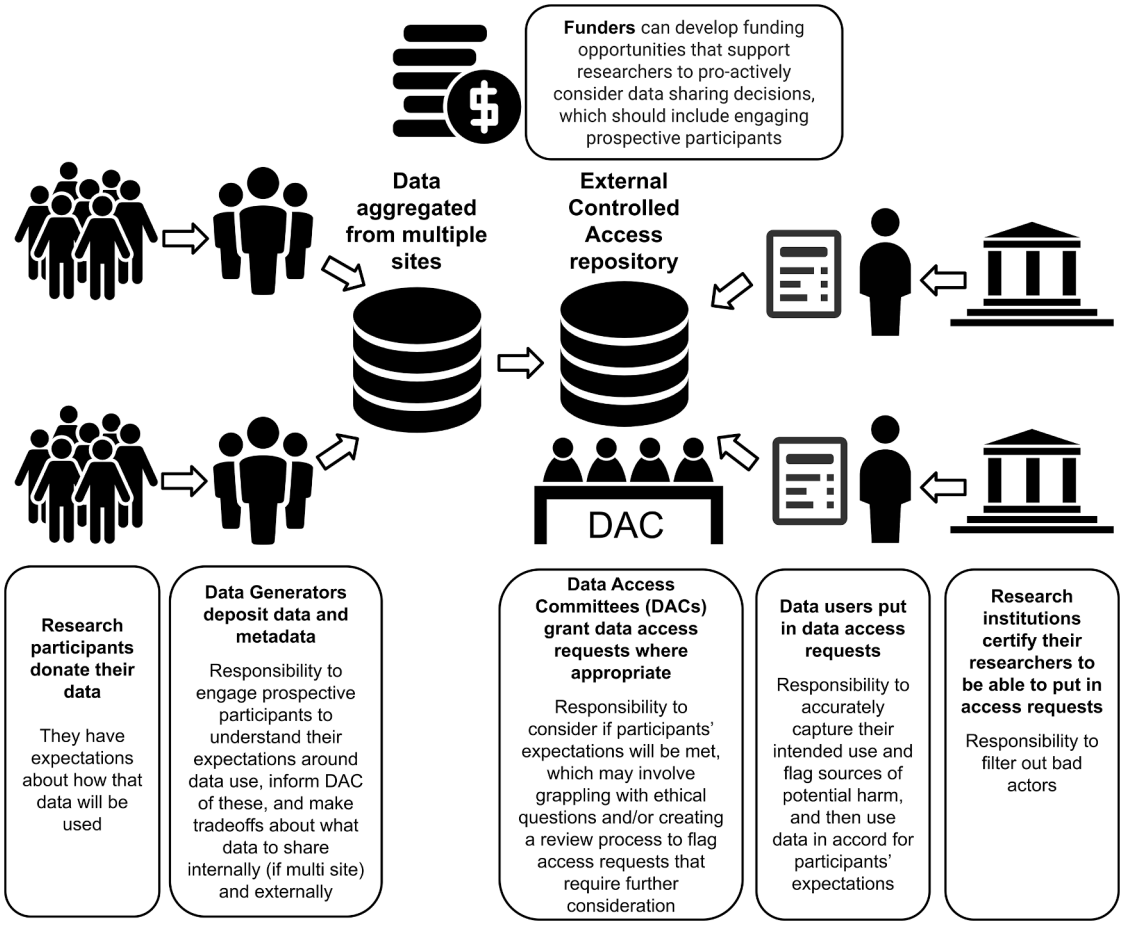

Figure 1. All the players within the research ecosystem have a responsibility to meet the expectations of the research participants who donated their data

institutions liable for misuse. Additionally, researchers are required to submit a project renewal request (or close-out), including a report of how the data were used, the publications generated, and whether any data management incidents occurred.[36,37] The DAC can also examine the presence of malicious or potentially concerning requests. Data access requests are publicly available, allowing anyone to flag issues (for example, the data access requests made for the eMERGE phase 3 data can be seen on this page: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000888.v1.p1). Taken together, these steps offer substantial protection.

Yet, like other DACs, the NHGRI DAC process has limitations. First, data access decisions are based on depositing institutional certifications stipulating restrictions regarding the study's informed consent provision on data sharing—yet this provision is often broad and not well understood by participants.[27] Moreover, consent is often obtained at the beginning of the study—a single time point that does not account for sociolegal developments during or after a study, as had occurred in eMERGE-IV. Finally, since the NHGRI is a federal body, all members involved in its DAC must be federal employees, precluding review by community members who may be directly affected. Data generators such as eMERGE-IV can provide non-binding recommendations but cannot further influence DAC decisions.

Meeting research participants' expectations

Our experiences navigating data-sharing decisions made us reflect on the limitations of relying on the informed consent process to define the scope of data sharing. Individuals agreeing to participate in research studies cannot be expected to think through all eventualities of how their data could be used. Nor can all possible eventualities be spelled out in consent forms, not least because they evolve with time. However, researcher-participant relations give rise to certain expectations that extend beyond the study cycle: just as research participants expect researchers to take necessary actions to protect them from harm from study procedures, they have trust-based expectations about how their data might be used in the future. This expectation stands regardless of how the data-sharing provision was stipulated in the informed consent form,[38] evidenced by episodes such as those experienced by the Havasupai,[39] Henrietta Lacks,[4] and the ABCD project.[3] The solution to

making data "as open as possible" is thus to design responsive processes—data governance—through which data stewardship responsibilities reflect the common expectations of data donors (see Figure 1).[40,41] This concept of meeting research participants' expectations is central to the ethical concepts of respect for research participants and trustworthiness of studies.

Accordingly, data generators should engage prospective participants to understand their expectations around data use before developing study material and consent forms and allow participants to provide guidance regarding their expectations around data sharing once recruitment has begun. Those generating data as part of a clinical trial (such as eMERGE-IV) can also probe prospective participants about whether external data sharing is a barrier to participation, and, if this is identified as a concern, they can consider providing an option to opt out of external data sharing.[42] Although this approach could be seen as reducing the datasets that are available for future uses, it adds important value to scientific endeavors by facilitating broader research participation, bringing in participants who may decline research participation if external data sharing is required, and promoting the generalizability of findings. This approach is also ethically sound, upholding respect for participants, rewarding trustworthy researchers, and bolstering public trust in research and researchers who fulfill their commitments in the long run. Of note, this approach does not aim to undermine efforts to encourage data sharing but to highlight the need for consciousness of the context of each specific study. Data generators may make trade-offs—such as those we have outlined above—favoring broader sharing if they are confident that the data governance approach adopted for secondary research will uphold the expectations of the original study's participants. Data generators can also use their institutional certificates to inform DACs and data users about community concerns and identify specific unauthorized categories of research use.

Conclusion

Decisions about data collection and sharing require trade-offs between advancing scientific discoveries in health and mitigating the risk of individual and group harms to our participants. Research participants expect that if their data are shared, they will be used in ways that do not cause harm to them, their families, and their communities—and that researchers will take steps to minimize the risks. This is the bedrock for trust in biomedical research. All players in the research ecosystem have a role to play in the task of creating data governance processes that are transparent, inclusive, and accountable, explicitly weighing foreseeable benefits and risks and meeting participants' expectations for how their data are used.

Acknowledgments

We would like to acknowledge our colleagues at the NHGRI for their support and advice related to this manuscript. This phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG011172 (Cincinnati Children's Hospital Medical Center), U01HG011175 (Children's Hospital of Philadelphia), U01HG008680 (Columbia University), U01HG011176 (Icahn School of Medicine at Mount Sinai), U01HG008685 (Mass General Brigham), U01HG006379 (Mayo Clinic), U01HG011169 (Northwestern University), U01HG011167 (University of Alabama at Birmingham), U01HG008657 (University of Washington), U01HG011181 (Vanderbilt University Medical Center), and U01HG011166 (Vanderbilt University Medical Center serving as the coordinating center). A.C.F.L. is funded by the NHGRI, 1K99HG012809.

Author contributions

Conceptualization, A.C.F.L., E.W.C., and M.S.: writing – original draft, A.C.F.L. and M.S.; writing – review & editing, all authors.

Declaration of interests

A.C.F.L. serves on CERA's external scientific panel and the Broad Institute's Bioethics Initiative. E.W.C. serves on the advisory boards of CERA and TRans/Forming. J.J. Connolly serves on the advisory board of Rhythm Pharma, IHCC. S.L.R. serves as an ethics consultant for the Blue Cross Blue Shield Association's pharmacy and therapeutics committee. I.A.H. serves as an associate editor and is on the editorial board of *AJHG* and serves as an IRB member of the All of Us Research Program. M.S. serves on the IRB of the All of Us Research Program, as a steering committee member of the Trisomy Collaborative, as a member of ASHG's Professional Practice and Social Implications Committee, and as a consultant to NIH's INCLUDE Project.

References

1. Knoppers, B.M. (2014). Framework for responsible sharing of genomic and health-related data. HUGO J. *8*, 3. https://doi.org/10.1186/s11568-014-0003-1.
2. Garrison, N.A. (2013). Genomic Justice for Native Americans: Impact of the Havasupai Case on Genetic Research. Sci. Technol. Human Values *38*, 201–223. https://doi.org/10.1177/0162243912470009.
3. McIntire, M. (2026). Genetic Data From Over 20,000 U.S. Children Misused for 'Race Science'. https://www.nytimes.com/2026/01/24/us/children-genetics-race-science.html.
4. Skloot, R. (2011). The Immortal Life of Henrietta Lacks (Crown).
5. Nebeker, C., Bélisle-Pipon, J.C., Collins, B.X., Cordes, A., Ferryman, K., McInnis, B.J., McWeeney, S.K., Novak, L.L., Rose, S., Yracheta, J.M., et al. (2025). Ethical sourcing in the context of health data supply chain management: a value sensitive design approach. JAMIA Open *8*, ooaf101. https://doi.org/10.1093/jamiaopen/ooaf101.
6. National Institutes of Health. (2019). Expired RFA-HG-19-014: The Electronic Medical Records and Genomics (eMERGE): Genomic Risk Assessment and Management Network - Enhanced Diversity Clinical Sites (U01 Clinical

Trial Required). https://grants.nih.gov/grants/guide/rfa-files/rfa-hg-19-014.html
7. NOT-OD-14-124: NIH Genomic Data Sharing Policy https://grants.nih.gov/grants/guide/notice-files/not-od-14-124.html.
8. NOT-OD-21-013: Final NIH Policy for Data Management and Sharing https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.
9. NOT-OD-26-023: Request for Information on Draft NIH Controlled-Access Data Policy and Proposed Revisions to NIH Genomic Data Sharing Policy https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-023.html.
10. 45 CFR Part 46 – Protection of Human Subjects. https://www.ecfr.gov/current/title-45/part-46.
11. Lawson, J., Rahimzadeh, V., Baek, J., and Dove, E.S. (2024). Achieving Procedural Parity in Managing Access to Genomic and Related Health Data: A Global Survey of Data Access Committee Members. Biopreserv. Biobank. *22*, 123–129. https://doi.org/10.1089/bio.2022.0205.
12. Ahmed, L., Constantinidou, A., and Chatzittofis, A. (2023). Patients' perspectives related to ethical issues and risks in precision medicine: a systematic review. Front. Med. *10*, 1215663. https://doi.org/10.3389/fmed.2023.1215663.
13. Trinidad, S.B., Fullerton, S.M., Bares, J.M., Jarvik, G.P., Larson, E.B., and Burke, W. (2010). Genomic research and wide data sharing: views of prospective participants. Genet. Med. *12*, 486–495. https://doi.org/10.1097/GIM.0b013e3181e38f9e.
14. Wan, Z., Hazel, J.W., Clayton, E.W., Vorobeychik, Y., Kantarcioglu, M., and Malin, B.A. (2022). Sociotechnical safeguards for genomic data privacy. Nat. Rev. Genet. *23*, 429–445. https://doi.org/10.1038/s41576-022-00455-y.
15. Dobbs v. (2022). Jackson Women's Health Organization, 597 U.S. 215.
16. Willis, M.D., Hoffman, M.N., Wang, T.R., Sabbath, E.L., Kuriyama, A.S., Wesselink, A.K., and Wise, L.A. (2024). Evaluating participant engagement in a preconception cohort study in relation to the Dobbs decision. Paediatr. Perinat. Epidemiol. *38*, 627–634. https://doi.org/10.1111/ppe.13080.
17. Kraft, S.A., and Mittendorf, K.F. (2024). Can Open Science Advance Health Justice? Genomic Research Dissemination in the Evolving Data-Sharing Landscape. Hastings Cent. Rep. *54*, S73–S83. https://doi.org/10.1002/hast.4932.
18. Mulvihill, G., and Press, A. "Justice Department demanded details on transgender patients from at least 1 hospital", https://www.pbs.org/newshour/nation/justice-department-demanded-details-on-transgender-patients-from-at-least-1-hospital (2025).
19. Wilkinson, K., Green, C., Nowicki, D., and Von Schindler, C. (2020). Less than five is less than ideal: replacing the "less than 5 cell size" rule with a risk-based data disclosure protocol in a public health setting. Can. J. Public Health Rev. Can. Santé Publique *111*, 761–765. https://doi.org/10.17269/s41997-020-00303-8.
20. Kahn, J.P., Mastroianni, A.C., and Sugarman, J. (1998). Beyond Consent: Seeking Justice in Research, 1st edition (Oxford University Press).
21. Claw, K.G., Anderson, M.Z., Begay, R.L., Tsosie, K.S., Fox, K., Garrison, N.A.; and Summer internship for INdigenous peoples in Genomics SING Consortium (2018). A framework for enhancing ethical genomic research with Indigenous communities. Nat. Commun. *9*, 2957. https://doi.org/10.1038/s41467-018-05188-3.
22. Chapman, C.R., Quinn, G.P., Natri, H.M., Berrios, C., Dwyer, P., Owens, K., Heraty, S., and Caplan, A.L. (2025). Consideration and Disclosure of Group Risks in Genomics and Other Data-Centric Research: Does the Common Rule Need Revision? Am. J. Bioeth. *25*, 47–60. https://doi.org/10.1080/15265161.2023.2276161.
23. Kraft, S.A., Cho, M.K., Gillespie, K., Halley, M., Varsava, N., Ormond, K.E., Luft, H.S., Wilfond, B.S., and Soo-Jin Lee, S. (2018). Beyond Consent: Building Trusting Relationships With Diverse Populations in Precision Medicine Research. Am. J. Bioeth. *18*, 3–20. https://doi.org/10.1080/15265161.2018.1431322.
24. Sabatello, M., Martschenko, D.O., Cho, M.K., and Brothers, K.B. (2022). Data sharing and community-engaged research. Science *378*, 141–143. https://doi.org/10.1126/science.abq6851.
25. Milne, R., Morley, K.I., Almarri, M.A., Anwer, S., Atutornu, J., Baranova, E.E., Bevan, P., Cerezo, M., Cong, Y., Costa, A., et al. (2021). Demonstrating trustworthiness when collecting and sharing genomic data: public views across 22 countries. Genome Med. *13*, 92. https://doi.org/10.1186/s13073-021-00903-0.
26. Aguirre, A., Lee, S.S.-J., Callier, S., Spicer, P., and Sabatello, M. (2025). Towards trustworthiness of precision medicine research for people with disabilities. Nat. Genet. *57*, 1321–1324. https://doi.org/10.1038/s41588-025-02195-1.
27. VandeVusse, A., Mueller, J., and Karcher, S. (2022). Qualitative Data Sharing: Participant Understanding, Motivation, and Consent. Qual. Health Res. *32*, 182–191. https://doi.org/10.1177/10497323211054058.
28. Duncan, G.T., Elliot, M., and Salazar, G.J.J. (2011). Statistical Confidentiality: Principles and Practice (Springer).
29. Federal Register, Management and Budget Office. "Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity" https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and (2024)
30. ResDAC. CMS Cell Size Suppression Policy. https://resdac.org/articles/cms-cell-size-suppression-policy (2024).
31. Wilson, M.R., Beachy, S.H., and Schumm, S.N. (2025). Rethinking Race and Ethnicity in Biomedical Research (National Academies Press). https://doi.org/10.17226/27913.
32. UNESCO. UNESCO Recommendation on Open Science. https://unesdoc.unesco.org/ark:/48223/pf0000379949 (2021).
33. UN Committee on the Rights of the Child "General comment no. 15 (2013) on the right of the child to the enjoyment of the highest attainable standard of health (art. 24)". https://www.ohchr.org/en/documents/general-comments-and-recommendations/crccgc15-general-comment-no-15-right-child-highest (2013).
34. Schatz, M.C., Philippakis, A.A., Afgan, E., Banks, E., Carey, V.J., Carroll, R.J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R.L., et al. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genom. *2*, 100085. https://doi.org/10.1016/j.xgen.2021.100085.
35. UK Biobank. "Protecting the data", https://www.ukbiobank.ac.uk/about-our-data/protecting-the-data/ (2025).

36. National Institutes of Health. "NIH Data Use Certification Agreement" https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/accessing-data/certification-agreement (2023).
37. National Institutes of Health, "Expiration Date, Renewal, Project Suspension, and Closeout", https://www.ncbi.nlm.nih.gov/books/NBK570252/ (2013).
38. Jamal, L., Sapp, J.C., Lewis, K., Yanes, T., Facio, F.M., Biesecker, L.G., and Biesecker, B.B. (2014). Research participants' attitudes towards the confidentiality of genomic sequence information. Eur. J. Hum. Genet. *22*, 964–968. https://doi.org/10.1038/ejhg.2013.276.
39. Garrison, N.A., Barton, K.S., Porter, K.M., Mai, T., Burke, W., and Carroll, S.R. (2019). Access and Management: Indigenous Perspectives on Genomic Data Sharing. Ethn. Dis. *29*, 659–668. https://doi.org/10.18865/ed.29.S3.659.
40. Boers, S.N., van Delden, J.J.M., and Bredenoord, A.L. (2015). Broad Consent Is Consent for Governance. Am. J. Bioeth. *15*, 53–55. https://doi.org/10.1080/15265161.2015.1062165.
41. Koenig, B.A. (2014). Have We Asked Too Much of Consent? Hastings Cent. Rep. *44*, 33–34. https://doi.org/10.1002/hast.329.
42. Committee on Strategies for Responsible Sharing of Clinical, Board on Health Science Policy, and Institute of Medicine (2015). Sharing of Clinical Trial Data: Maximizing Benefits, Minimizing Risk (National Academies Press (US)). https://pubmed.ncbi.nlm.nih.gov/25590113/.