Cell Genomics

CellPress
OPEN ACCESS

Perspective

Aligning NIH's existing data use restrictions to the GA4GH DUO standard

Jonathan Lawson,[1,*] Elena M. Ghanaim,[2,*] Jinyoung Baek,[1] Harin Lee,[1] and Heidi L. Rehm[1,3,4]
[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[4]Department of Pathology, Harvard Medical School, Boston, MA, USA
*Correspondence: jlawson@broadinstitute.org (J.L.), elena.ghanaim@nih.gov (E.M.G.)
https://doi.org/10.1016/j.xgen.2023.100381

SUMMARY

It is widely accepted that large-scale genomic data (e.g., whole-genome sequencing, whole-exome sequencing, and genome-wide association study data) be shared through a controlled-access mechanism. This protects the privacy of research participants and ensures downstream uses of data align with participants' informed consent regarding future sharing of their data. In 2019, GA4GH approved the Data Use Ontology (DUO) standard to define data use terms with machine-readable representations to represent how a dataset can be used. We endeavored to determine the parity of existing data use restrictions ("Data Use Limitations" [DULs]) for datasets registered in the National Institutes of Health database for Genotypes and Phenotypes (dbGaP) with the DUO standard. We found substantial (93%) parity between the dbGaP DULs (n = 3,575) and DUO. This study demonstrates the comprehensiveness of the DUO standard and encourages data stewards to standardize data use restrictions in machine-readable formats to facilitate data sharing.

INTRODUCTION

The significance of data use terms (data use limitations) for controlled-access datasets

The expectation to share data (not just upon request, but via a broadly accessible data repository) is an increasing priority within biomedical research. Public and private funders, journals, and institutions have enacted policies to promote, if not require, data sharing, to enhance scientific rigor and reproducibility, de-duplicate data generation efforts, enable cross-study research, and maximize the utility of any given dataset. Data that are about people, however, must be shared in a manner that promotes public trust, maintains privacy of research participants, and aligns with arrangements with individual research participants and the values of communities participating in research.

When it comes to publicly funded genomic data, the expectation to share data has been in place for many years. Broad data sharing has been an ethos for the field of genomics since the early days of sequencing technology development.[1] The National Institutes of Health (NIH) first introduced a policy for sharing genome-wide association study (GWAS) data in 2008, which was later superseded by a broader NIH Genomic Data Sharing (GDS) Policy (2014).[2,3] These policies established a controlled-access framework for the vast majority of large-scale, de-identified data from human participants to minimize the (small) risk to privacy inherent to large-scale genomic data. The GDS Policy also set forth an expectation for informed consent of yet-to-be-enrolled research participants regarding future research use and broad data sharing of genomic data.

To submit large-scale, human genomic data to an NIH-designated data repository under the NIH GDS Policy, one must provide official documentation outlining the restrictions on data use, per the informed consent of research participants (i.e., Institutional Certification).[4] In developing the NIH Institutional Certification(s), NIH crafted standard Data Use Limitations (DULs) for datasets that describe how data may be used for secondary research. Institutions, in collaboration with their institutional review boards (IRBs), can select from these standard DULs, or they may choose to provide customized language for how data may be shared to reflect the informed consent process used for the study. Once data have been released, requestors must apply to individual "consent groups" to access the data, with each representing a unique DUL combination, and they will only be granted access if a data access committee (DAC) adjudicates that the Research Use Statement for the particular request is within the bounds of the given DUL(s) and thus within the bounds of the informed consent of research participants.

The database of Genotypes and Phenotypes (dbGaP) was the first NIH-designated data repository for controlled-access genomic data and the centralized registration site for genomic studies funded by the NIH. Today, dbGaP contains over 2,400 released datasets, with over 3,500 consent groups (each dataset has one or more consent groups). Consent groups from different datasets often have the same or similar DULs to represent their permitted data use. Further, there are examples of narrative

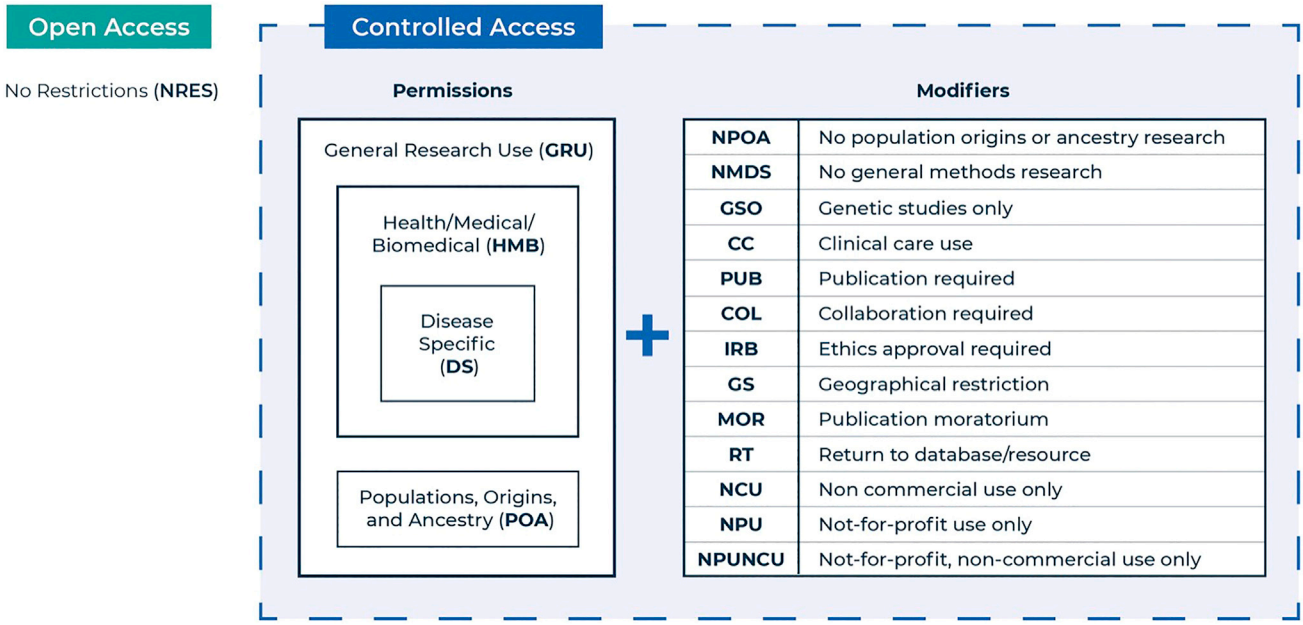

Figure 1. Illustrated representation of the GA4GH Data Use Ontology
Permission terms are displayed at center and left, and modifier terms are displayed at right.

DULs that are equivalent to one another but that use different phrases (e.g., health research versus Health/Medical/Biomedical). However, older studies tend to have more complicated DUL statements that are less aligned with standard NIH DULs, reflective of consent processes before broad data sharing was encouraged and the standard NIH DULs were in place. Though some DULs attributed to legacy datasets are functionally equivalent or amenable to alignment with standard NIH DULs, the standard NIH DULs are not organized ontologically, which leaves room for variation in interpretation. We contend that an ontological organization of the standard NIH DULs would facilitate assignment of DULs to datasets in a consistent manner across DACs, enabling more accurate dataset search results for researchers and streamlining the adjudication of data access requests (DARs).

AnVIL data access pilot

NHGRI's Analysis, Visualization, and Informatics Lab-space program has a specific goal to "develop and implement streamlined technical and administrative processes to review and authorize controlled-access data requests, while taking into account the Data Use Limitations of the studies hosted by the AnVIL."[5] To do this, the Broad Institute, one of the two collaborative grantees funded in 2018 to establish the AnVIL, is developing and testing the Data Use Oversight System (DUOS), an implementation of the GA4GH Data Use Ontology (DUO) standard, as a potential system for streamlining access to the controlled-access datasets stored in the resource.[6]

DUO is a hierarchical vocabulary of data use terms most often used to denote secondary usage conditions for controlled-access datasets. DUO does not aim to represent all possible data use terms, consent phrases, or complex logical permutations of permissions, limitations, or requirements. Currently, DUO contains 18 terms across two categories of data use terms, five permission terms and 13 modifier terms (Figure 1). Permission terms such as General Research Use (GRU), Health or Medical or Biomedical use (HMB), Disease Specific research (DS), Population Origins and Ancestry research (POA), and No Restrictions (NRES) standardize allowed usage of the datasets. Modifier terms are used to further qualify the permission terms of controlled access.[7] DUO's mapping of disease-specific permissions previously leveraged the Disease Ontology (DOID)[8] and have since been updated to leverage the Monarch Disease Ontology (Mondo), a "unified disease ontology, encompassing many disease terminologies, which aims to harmonize disease definitions across the world."[9]

The DUOS is an open-source software platform that enables research teams to register their datasets in a catalog for data sharing, allows researchers to submit requests for those datasets, and supports adjudication of those requests by DACs. DUOS leverages DUO to describe the datasets' data use restrictions through each of these processes to maintain consistent terms, definitions, and enable machine-readable representations as well as functions to facilitate search and access (https://github.com/DataBiosphere/duos-ui). DUOS also leverages an inference engine (referred to hereafter as "algorithm") for automated checking of the compatibility between DULs and DARs expressed via DUO.

For the last 4 years, DUOS has been piloted through a series of iterative phases.[10] Overseen by the AnVIL Data Access Working Group, which includes members of the NHGRI DAC, the evolution of the NIH DUOS pilot has involved close collaboration on policy and technology issues, with contributions from NHGRI program and policy officials, Broad Institute compliance and

Real NIH DULs (long form)	Real NIH DULs (short form)	Ontological code(s)
Use of the data is only limited by the terms of the model Data Use Certification	*General Research Use*	**DUO:0000042** (general research use, "GRU")
Research on health conditions	*All Health Conditions*	**DUO:0000006** (health or medical or biomedical research, "HMB")
Use of the data must be related to myelodysplasia	*Disease-Specific (Myelodysplasia)*	**DUO:0000007** (disease-specific research, "DS") **MONDO:0018881** (myelodysplastic syndrome)
These data may only be used for studies related to the human microbiome	*Human microbiome research*	**Cannot be mapped** *Areas of research not represented by DUO/MONDO*

Figure 2. Illustration of the mapping exercise

Italics on the left and in the middle are actual Data Use Limitations for studies registered in dbGaP. Bold codes on the right are the DUO and Mondo codes, with the associated definitions in parentheses.

technology representatives, large research consortium data coordinators, and ELSI experts.

Concurrent with the NIH pilot, the Broad Institute's DAC has used DUOS to oversee access requests for a number of Broad Institute datasets. Changes suggested by the NIH are often beneficial to the Broad DAC and would be available to other DACs using DUOS.

While the NIH DUOS pilot has many facets, this assessment focuses on the feasibility of structuring current NIH datasets' DULs with DUO. Given that dbGaP is the registration locus for most NIH-sponsored controlled-access genomic studies, we endeavored to determine whether the NIH DULs for datasets registered in dbGaP can be mapped to DUO with high fidelity.

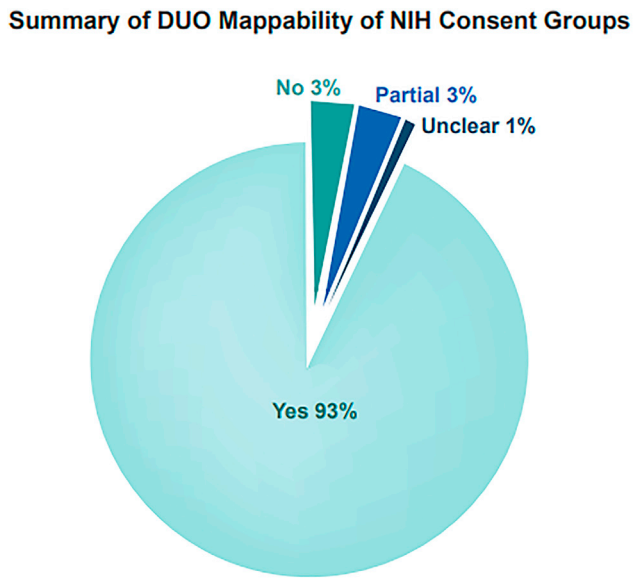

Figure 3. Percentage of NIH DULs that mapped to GA4GH DUO

Pie chart of the mappability of total NIH Data Use Limitations (DULs) to the GA4GH DUO standard for representing data use restrictions.

Design

To answer these questions, we obtained the public study details for 2,425 studies registered in dbGaP, comprising 3,598 consent groups in October 2021. We then took each NIH consent group's DUL ("consent_list" and "consent_title") and manually mapped it to one or more representative DUO terms and to corresponding Mondo or Disease Ontology terms when appropriate (Figure 2). This task was organized by a co-lead of the GA4GH DURI workstream that minted and oversees the DUO standard. The expertise of five data access committee experts (mostly NIH DAC chairs) was sought to adjudicate on the mappability for the inconclusive results.

RESULTS

Manual mapping of NIH DULs to GA4GH DUO

Initial results showed that 91% of current NIH datasets' DULs were mappable to DUO, yet roughly 5% of results were inconclusive and warranting further review for clear determinations, which was supplied by a cadre of NIH DAC experts. After these clarifications from NIH DACs, an additional 2% of DULs were identified as mappable to DUO, leading to an overall fidelity between NIH DULs and the GA4GH DUO of 93% (Figure 3). The vast majority of these consent groups are aligned to one or more of the standard NIH DULs (e.g., approximately 1,400 consent groups are GRU, 800 consent groups are HMB, and 1,300 consent groups are DS).

Examining DUO mappability across NIH DACs (Figure 4), we found that DACs with lower DUO fidelity tend to be lower-volume DACs (i.e., those managing fewer than 100 datasets, for instance). These DACs' consent groups reflect the more specific scientific use cases in their DULs, often denoting anatomy-specific or disease-specific limitations not found in the disease ontology leveraged by DUO. However, this is not always the case. Several NIH DACs managing a lower volume of datasets have high DUO fidelity, too. Excitingly, the NIH ICs stewarding the most consent groups and genomic data show high fidelity with the DUO (NCI: 1227, 98%; NHLBI: 532, 93%; and NHGRI: 467, 87%), which bodes well for the effort to allow for

DUO Mappability of NIH Consent Groups by NIH Data Access Committee (DAC)									
DAC	No		Unclear		Partial		Yes		Total
ADSP							6	100%	6
CDAC							3	100%	3
HeLa Genome							15	100%	15
NCATS							71	100%	71
NIDCD							12	100%	12
None							3	100%	3
RADx							76	100%	76
TCGA							5	100%	5
NIGMS	1	2%					64	98%	65
eNCI	1	2%					52	98%	53
NCI	6	<1%			21	2%	1199	98%	1227
NEI					1	2%	56	98%	57
NICHD	2	2%			1	1%	87	97%	90
NINDS	3	2%			4	3%	134	95%	141
ES					1	6%	17	94%	18
NHLBI	6	1%	22	4%	7	1%	497	93%	532
NIDDK	8	7%	1	1%	2	2%	106	91%	117
JAAMH	14	5%	1	0%	15	5%	249	89%	279
NIAID	7	6%			6	5%	105	89%	118
NHGRI	22	5%	3	1%	34	7%	408	87%	467
NIAMS	9	13%					59	87%	68
NINR	3	16%			1	5%	15	79%	19
NLM	2	22%					7	78%	9
Kids First	14	23%			1	2%	47	76%	62
NIDCR	14	23%			4	7%	43	70%	61
Grand Total	**112**	**3%**	**27**	**1%**	**99**	**3%**	**3337**	**93%**	**3575**

Figure 4. Percentage of NIH DULs mappability to GA4GH DUO by NIH DAC

Percent mappability broken down by the NIH DAC managing access to the study.

interoperability of data and research through their joint efforts in the NIH Cloud Platforms Interoperability (NCPI) efforts.[11]

Analyzing DUO mappability to NIH DULs over time, the lowest periods of DUO fidelity occur in the first 5 years of dbGaP's existence and NIH's use of DULs (Figure 5). Understandably, as data generators and stewards grew to understand common issues in DUL terms, consistency in DULs as well as alignment between standard NIH DULs and DUO grew. At initiation of the study, we would have presumed a more significant and consistent upward trend than what is shown in the data, which could be explained by large studies with numerous consent groups biasing annual results to higher or lower fidelity than presumed.

We observed 3% of DULs were only able to be mapped partially, often because the restriction described a disease but also the broader concept of "health" (e.g., brain health and disease). Where Mondo has a code for "brain disease," it does not capture the broader concept of brain health.

A thematic analysis of the 3% (112) of cases that did not map to DUO showed 41 datasets' DULs referred to an area of research rather than a specific disease available in the ontology (e.g., human microbiome, head and face, preterm birth, aging, smoking, childhood diseases, and aspirin-related research), and another 40 datasets' DULs were defined as a heterogeneous group of disorders rather than a single disease or ontological disease family (e.g., craniofacial, orofacial, eye movement, addictive, and aging-related disorders) (Figure 6). Further, a less common yet significant issue was the use of terms describing a clinical condition (e.g., preterm birth) or phenotypic feature (e.g., mammographic density, platelet function) rather than a disease, which was cause for another 16 datasets' DULs not be able to map to DUO.

DISCUSSION

DUO's high-fidelity alignment to NIH DULs, machine readability, hierarchical structure, and global adoption are clear benefits for genomic data stewards such as the NIH. Even in such cases where DUO's mappability to existing terms is partial (as with 3% of NIH DULs), it is foreseeable that partially mappable terms may be beneficial as helpful metadata.

Currently, there are no plans to change the GA4GH DUO standard to accommodate the small number of unmappable DULs. The reason is that a number of the non-mappable terms are derived from legacy consents that did not have broad data sharing in view and/or implicitly or explicitly contain logical fallacies that ought not to be ontologically replicated or for which demand has not surmounted among IRBs, participant representatives, or DACs in recent history.

The NIH Data Management and Sharing (DMS) policy, effective on January 25, 2023, requires applicants to submit a DMS Plan. Importantly, the policy expects researchers to "maximize appropriate data sharing when developing plans."[12] As a result, we predict a significant increase in human subjects data sharing that will require controlled-access (e.g., imaging data, clinical data, survey data, and more). NIH Supplemental Information on Protecting Privacy When Sharing Human Research Participant Data indicates that "[r]esearchers and institutions should develop robust consent processes that prioritize clarity regarding future sharing and use of scientific data, including limitations on future use, and general aspects regarding how data will be managed."[13] NIH has also released Points to Consider and Sample Language for Future Use and/or Sharing.[14] We see the high fidelity of DUO with existing datasets as a vote of confidence for NIH and other genomic data stewards to

DUO Mappability of NIH Consent Groups by Year									
Year	No		Unclear		Partial		Yes		Total
2008			1	4%			22	96%	**23**
2009	1	1%	6	9%	4	6%	59	84%	**70**
2010	2	2%	5	6%	6	7%	72	85%	**85**
2011	8	5%	7	4%	8	5%	152	87%	**175**
2012			3	3%	14	14%	82	83%	**99**
2013	2	2%			8	6%	114	92%	**124**
2014	4	1%	1	0%	5	1%	344	98%	**354**
2015	12	4%			1	0%	279	96%	**292**
2016	7	2%			1	0%	285	97%	**295**
2017	14	5%	1	0%	5	2%	235	93%	**255**
2018	9	3%			6	2%	260	95%	**275**
2019	17	5%			8	2%	324	93%	**349**
2020	16	4%	2	0%	6	1%	418	93%	**442**
2021	20	3%	1	0%	25	3%	691	94%	**737**
Grand Total	**112**	**3%**	**27**	**1%**	**99**	**3%**	**3337**	**93%**	**3575**

Figure 5. Percentage of NIH DULs mappability to GA4GH DUO over time

Percent mappability broken down by the dataset's release date (year).

standardize use restrictions even more than they are today, in ways that can be leveraged by systems to better facilitate the search and access of scientific data by implementing DUO.

What would a world look like where DUO is integrated in the data sharing endeavor from the start of a research project, particularly in light of the NIH DMS policy? Researchers and participants agree on one or more DUO term(s) that clearly define how data will be shared, and these concepts are incorporated into the participant consent form and informed consent process. Researchers and institutions communicate how this data can and will be shared to funders by indicating the DUO term(s) in their DMS Plan when writing their grant application and Institutional Certifications at data registration. When registered in data access systems, the DULs are expressed using the same DUO term(s). Finally, the DUO term(s) are used to facilitate search by prospective requestors and review of requests by DACs.[6]

Use of DUO throughout the process minimizes the possibility that data use restrictions will be misunderstood or miscommunicated and ensures participants' consent is respected consistently throughout the data sharing process. Given these benefits and in light of the high fidelity of existing DULs to the DUO standard, the AnVIL team is actively working to tag all NHGRI datasets in the AnVIL Catalog[15] with their corresponding DUO terms (when available) in order to facilitate search by researchers and to simultaneously enable the NHGRI DAC to manage DUO-backed DARs for those datasets in DUOS. This makes AnVIL the first NIH-supported data repository to adopt DUO. While the datasets stored in AnVIL are only a portion of the overall set of dbGaP-registered datasets that were mapped, it provides the opportunity to show that the DUO standard can be implemented, both retrospectively and prospectively. Simultaneously, the authors have provided the mapping of legacy NIH DULs produced in this analysis to dbGaP representatives for future use in dbGaP at their discretion.

Limitations of the study

Although dbGaP contains thousands of controlled-access datasets, one limitation of this study is that it only mapped the DULs for datasets registered in this one system. We have not compared the percent mappability to another controlled-access repository. It would be useful to assess DUO's mappability to the use restrictions used for non-genomic scientific datasets and those registered in a different resource.

We also did not assess whether mappability to DUO may have changed over time. For instance, legacy datasets are often deemed to have the most complicated data use restrictions. However, we were unable to assess whether there was any correlation between the age of a study and whether DUO could be used for describing the data use restrictions.

Lastly, as mentioned in the introduction, it is the submitting institution that determines the correct DUL(s) to apply to a controlled-access dataset registered in dbGaP by submitting an Institutional Certification to the NIH. Ideally, the submitting institution, in consultation with their IRB, would validate the assignment of DUO codes. In retrospective cases, NIH will have to decide, on a policy basis, who can determine the equivalency of a particular DUO code to the institutionally provided DUL. The NIH could consider a process similar to one used to seek institutional approval of the appropriate designation for Genomic Summary Results (GSRs) for existing dbGaP studies, where institutions were given an opportunity to designate GSRs as "sensitive" through the submission of a new institutional certification; datasets for which the institution did not reply in a certain amount of time were defaulted to open sharing of GSRs.

Future directions/areas to watch

While DUO was developed based upon scientific use cases from the genomics research community, DUO's terms and definitions are not specifically tied to or overly indexed for genomics. In fact, at initial review, researchers from various other scientific domains have inquired and suggested that DUO would be applicable for facilitating data sharing in their communities. Initial validation of such use cases is already underway via members of this research team along with colleagues from GA4GH. The broad applicability of DUO for all scientific data is especially notable given the expansion of NIH's aforementioned DMS Policy, which encourages data sharing plans for all scientific data generated using NIH funding.

Software systems such as DUOS are able to leverage DUO in order to fully automate the data access request process. While

Thematic Analysis of DULs Unable to Map to DUO		
Rationale Theme	**Counts**	**Frequency (%)**
Area of Research	41	34%
Clinical Condition/Phenotypic Feature	16	13%
Heterogeneous Disorders	40	33.5%
No Disease Term Exists	6	5%
Legal/Regulatory Term	2	1.5%
Non-standard Modifier	14	12%

Figure 6. Thematic categorization of NIH DULs not mappable to DUO

Percent of DULs not mappable to DUO by thematic rationale for their inability to be mapped.

the resulting benefits of expediting data sharing are desirable, caution has been expressed from select community members. Therefore, to pursue these benefits in an evidenced-based and minimum-risk manner, we plan to incrementally automate DARs from least-to-most restrictive use permissions (i.e., starting with GRU datasets).

Retrospectively mapping datasets to DUO is a significant manual undertaking. To avoid this work and risk, we contend that IRBs and primary study teams should strongly consider expressing consented data use terms using DUO, as resources such as the GA4GH Machine Readable Consent Guidance[16] describe. This avoids the ethical risk of reinterpreting consent form language post participant signature and removes any need for manual mapping work. Further, while the datasets addressed in this analysis and discussion are primarily research datasets, much opportunity exists for non-NIH funded entities such as clinical laboratories, hospitals, and pharmaceutical companies who may consent individuals for clinical trials or secondary data use, to align their datasets to the DUO, not only for more efficient and compliant use of data internally but also to elucidate the availability of datasets for inter-institutional collaborations to more expeditiously arrive at scientific discoveries.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100381.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of the NIH data access experts who helped to adjudicate the inconclusive cases: Barbara Thomas, Shu Hui Chen, Alicia Chou, and Lu Wang. We would also like to thank Valentina DiFrancesco, Carolyn Hutter, Stacey Donnelly, and Barbara Thomas for reviewing an early version of this manuscript.

J.L., J.B., and H.L. were funded by U24HG011025-01A1S1.

AUTHOR CONTRIBUTIONS

Corresponding author, J.L.; lead author, J.L.; conceptualization, J.L. and E.M.G.; data curation, J.L., J.B., and H.L.; formal analysis, J.L. and J.B.; funding acquisition, J.L. and H.L.R.; investigation, J.L. and E.M.G.; methodology, J.L. and E.M.G.; project administration, J.L.; visualization, J.L.; writing – original draft, E.M.G. and J.L.; writing – review & editing, E.M.G., J.L., and H.L.R.

DECLARATION OF INTERESTS

The authors have no competing interests to declare.

REFERENCES

1. Maxson Jones, K., Ankeny, R.A., and Cook-Deegan, R. (2018). The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. J. Hist. Biol. *51*, 693–805. https://doi.org/10.1007/s10739-018-9538-7.
2. National Institutes of Health (2007). Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-wide Association Studies (GWAS). https://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.
3. National Institutes of Health (2014). Final NIH Genomic Data Sharing Policy. https://grants.nih.gov/grants/guide/notice-files/not-od-14-124.html.
4. National Institutes of Health. About Institutional Certifications. https://sharing.nih.gov/genomic-data-sharing-policy/institutional-certifications/about-institutional-certifications.
5. National Human Genome Research Institute (2017). The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL) (U24). https://grants.nih.gov/grants/guide/rfa-files/rfa-hg-17-011.html.
6. Cabili, M.N., Lawson, J., Saltzman, A., Rushton, G., O'Rourke, P., Wilbanks, J., Rodriguez, L.L., Nyronen, T., Courtot, M., Donnelly, S., and Philippakis, A.A. (2021). Empirical validation of an automated approach to data use oversight. Cell Genom. *1*, 100031. https://doi.org/10.1016/j.xgen.2021.100031.
7. Lawson, J., Cabili, M.N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S.R., Boyles, R.R., Brookes, A.J., Brush, M., et al. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genom. *1*, 100028, None. https://doi.org/10.1016/j.xgen.2021.100028.
8. Disease Ontology. https://disease-ontology.org/.
9. Monarch Initiative (2019). New Release of Mondo Disease Ontology. https://monarchinit.medium.com/new-release-of-mondo-disease-ontology-9a48521353e3.

10. National Institutes of Health Office of Data Science Strategy. About the Data Use Oversight System (DUOS) Pilot. https://datascience.nih.gov/data-infrastructure/duos.

11. National Institutes of Health Office of Data Science Strategy. About the NIH Cloud Platform Interoperability (NCPI) Effort. https://datascience.nih.gov/nih-cloud-platform-interoperability-effort.

12. National Institutes of Health (2020). Final NIH Policy for Data Management and Sharing. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

13. National Institutes of Health (2022). Supplemental Information to the NIH Policy for Data Management and Sharing: Protecting Privacy when Sharing Human Research Participant Data. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-213.html.

14. National Institutes of Health (2022). Informed Consent for Secondary Research with Data and Biospecimens: Points to Consider and Sample Language for Future Use And/or Sharing. https://osp.od.nih.gov/wp-content/uploads/Informed-Consent-Resource-for-Secondary-Research-with-Data-and-Biospecimens.pdf.

15. NHGRI Analysis Visualization and Informatics Lab-Space. AnVIL Dataset Catalog. https://anvilproject.org/data/.

16. Global Alliance for Genomics and Health (2020). GA4GH Machine-Readable Consent Guidance: How to Map Data Sharing Consent Language to the GA4GH Data Use Ontology. https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
DUO GitHub repository	https://www.cell.com/cell-genomics/pdfExtended/S2666-979X(21)00035-5	http://purl.obolibrary.org/obo/duo
Data Use Oversight System GitHub repository	https://github.com/DataBiosphere/duos-ui	https://doi.org/10.5281/zenodo.8021267
Released DUO file	https://www.cell.com/cell-genomics/pdfExtended/S2666-979X(21)00035-5	http://purl.obolibrary.org/obo/duo.owl
dbGaP	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2031016/	https://www.ncbi.nlm.nih.gov/gap/
NIH DULs to DUO Mapping file	www.duos.org	DUOS ID: DUOS-000137

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jonathan Lawson (jlawson@broadinstitute.org).

Materials availability

This study did not generate new reagents or materials.

Data and code availability

The NIH Consent Groups' DULs aligned to the DUO data is available at duos.org under open access. The accession ID is available via the Key Resources Table.

METHOD DETAILS

1. We obtained a listing of all available NIH studies' registration info available via dbGaP
2. We then broke the study listing out by the individual consent groups within each study in order to map the DUO to each consent group individually
3. We then took each NIH consent group's DUL (which has a short form and a long form) and manually mapped it to one or more representative DUO terms based on a thorough understanding of the DUO and common terms used by NIH representatives in registering studies and consent groups.
4. For results that were not conclusively mappable or unmappable, we sought the expertise of multiple NIH Data Access Committee Chairs to adjudicate mappability.
5. Analysis of the results was visualized in a variety of formats including summary level, by DAC, and over time.