Cell Genomics, Volume *1*

Supplemental information

Empirical validation of an automated approach to data use oversight

Moran N. Cabili, Jonathan Lawson, Andrea Saltzman, Greg Rushton, Pearl O'Rourke, John Wilbanks, Laura Lyman Rodriguez, Tommi Nyronen, Mélanie Courtot, Stacey Donnelly, and Anthony A. Philippakis

Supplemental Information

Empirical Validation of an Automated Approach to Data Use Oversight

Authors

Moran N Cabili[†,1], Jonathan Lawson[†,1], Andrea Saltzman[1], Greg Rushton[1], Pearl O'Rourke[3], John Wilbanks[4], Laura Lyman Rodriguez[5], Tommi Nyronen[6], Mélanie Courtot[7], Stacey Donnelly[*,1], Anthony A. Philippakis[*,1]

[1] Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[3] Massachusetts General Brigham, Boston, Massachusetts, USA
[4] Biogen Inc., Cambridge, Massachusetts, USA
[5] Patient-Centered Outcomes Research Institute (PCORI), Washington, District of Columbia, USA
[6] ELIXIR Finland, CSC - IT Center for Science, FI
[7] European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

[†] These authors contributed equally to this work; * To whom all correspondence should be addressed.

Address for Correspondence

415 Main Street, Cambridge MA 02142

Supplemental Methods

Capturing Data Use limitations

DUOS uses the GA4GH Data Use Ontology[1], the Human Disease Ontology[2], to code the data use limitations for every dataset that is managed by the system. Data depositors use the DUOS user interface to answer a set of structured questions with regard to the restrictions on the dataset which are then captured in the backend using the ontologies. Depositors also log the free text describing the restrictions on the secondary use of the data as documented by the IRB overseeing the clinical study that collected the data. See Supplementary Figure S1.

Notes

A Human Subjects Protection expert from the Broad evaluated 123 DULs and attempted to structure them with DUO codes. Of 123 DULs 96% were successfully structured based on the DAC review. The following table includes the 5 DULs that we were unable to structure:

Table S3 describes the ontological representation of research purpose queries and the computation of which datasets with DU restrictions would they match when applying the DUOS matching using DUO and the Human Disease Ontology. To lower latency, the ontologies are indexed and pre-calculated for each ontology term.

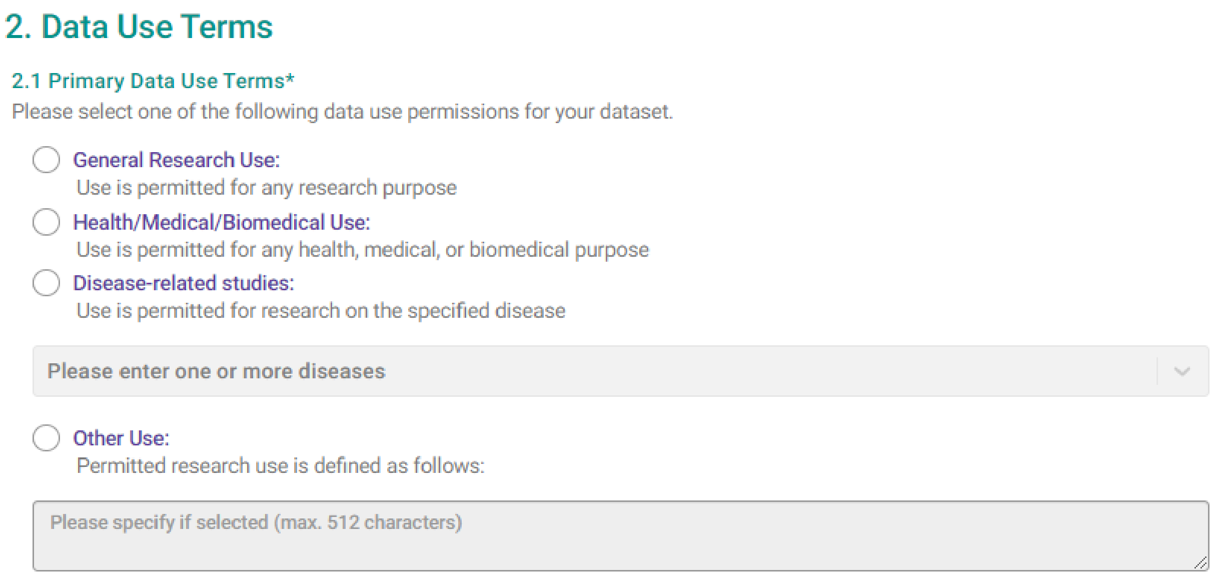
Capturing a Data Access Request

Data access requestors use the DUOS user interface to specify the datasets they would like to access and to answer a set of structured questions to describe their intended use of the data. DUOS then captures in the backend these answers using the previously mentioned ontologies. Requestors also log a free text description of their intended use of the data. See Supplementary Figure S2.

Automated matching examples

Once a data access request is submitted, the automated matching system computes if the intended use as represented by ontology terms is compatible with the restrictions of the specific dataset of interest. The companion manuscript by Lawson et al. graphically illustrates how an algorithm uses the ontology hierarchy to match only datasets labeled with ontology terms that are PARENT ontology terms relative to the ontology term capturing the requester's intended data use. See Supplementary Figure S3. For example, when a requestor would like to use a dataset to study Melanoma, the matching algorithm will approve datasets that are restricted for: (a) studying cancer (DS- cancer), (b) health biomedical research (HMB) or (c) general research uss (GRU) that are all PARENT nodes relative to the position of Melanoma in the Human Disease Ontology hierarchical tree/directed acyclic graph. The matching algorithm, however, will not match a dataset that is restricted for the study of Uveal Melanoma, since Uveal Melanoma is a specific subtype of Melanoma which is represented as a CHILD node relative to the hierarchical representation of Melanoma in the Human Disease Ontology tree.

Figures



2. Data Use Terms

2.1 Primary Data Use Terms*

Please select one of the following data use permissions for your dataset.

○ General Research Use:
Use is permitted for any research purpose

○ Health/Medical/Biomedical Use:
Use is permitted for any health, medical, or biomedical purpose

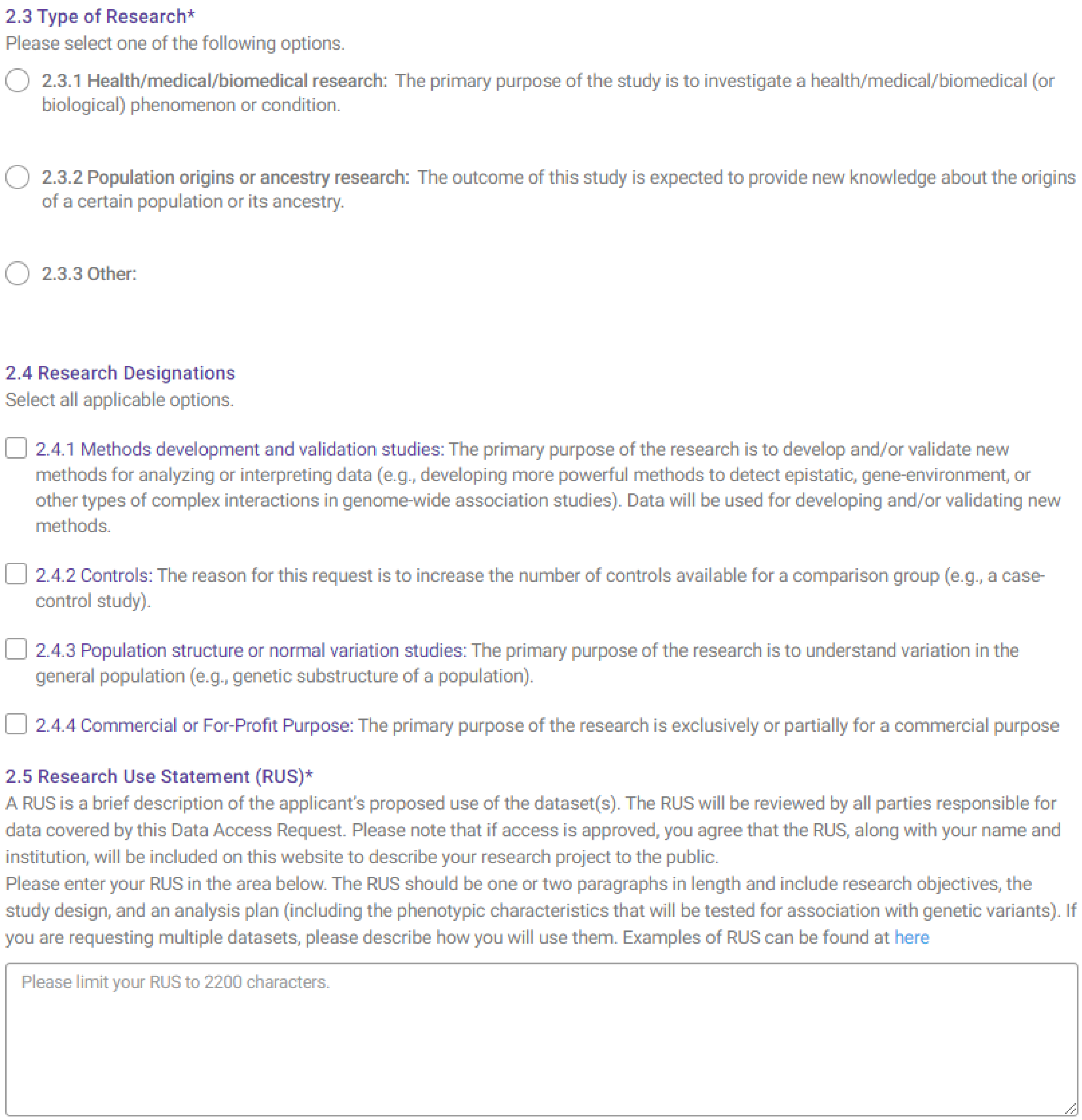
○ Disease-related studies:
Use is permitted for research on the specified disease

Please enter one or more diseases

○ Other Use:
Permitted research use is defined as follows:

Please specify if selected (max. 512 characters)

Figure S1. Data use limitation structuring interface, Related to STAR Methods.
Example of the DUOS interface capturing data use limitations on the dataset by the data depositor.



2.3 Type of Research*

Please select one of the following options.

○ 2.3.1 Health/medical/biomedical research: The primary purpose of the study is to investigate a health/medical/biomedical (or biological) phenomenon or condition.

○ 2.3.2 Population origins or ancestry research: The outcome of this study is expected to provide new knowledge about the origins of a certain population or its ancestry.

○ 2.3.3 Other:

2.4 Research Designations

Select all applicable options.

□ 2.4.1 Methods development and validation studies: The primary purpose of the research is to develop and/or validate new methods for analyzing or interpreting data (e.g., developing more powerful methods to detect epistatic, gene-environment, or other types of complex interactions in genome-wide association studies). Data will be used for developing and/or validating new methods.

□ 2.4.2 Controls: The reason for this request is to increase the number of controls available for a comparison group (e.g., a case-control study).

□ 2.4.3 Population structure or normal variation studies: The primary purpose of the research is to understand variation in the general population (e.g., genetic substructure of a population).

□ 2.4.4 Commercial or For-Profit Purpose: The primary purpose of the research is exclusively or partially for a commercial purpose

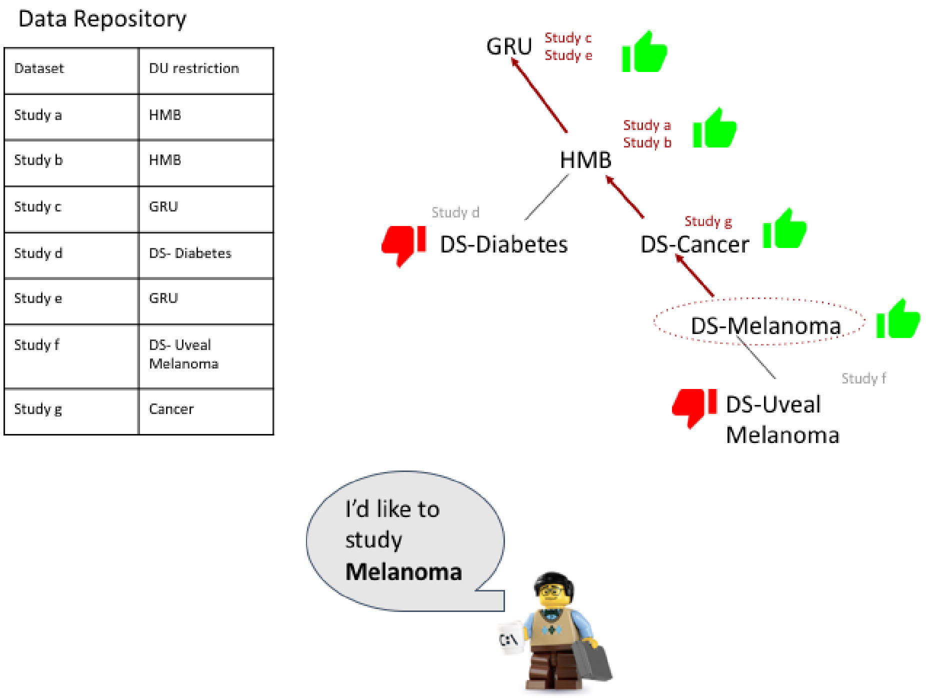
2.5 Research Use Statement (RUS)*

A RUS is a brief description of the applicant's proposed use of the dataset(s). The RUS will be reviewed by all parties responsible for data covered by this Data Access Request. Please note that if access is approved, you agree that the RUS, along with your name and institution, will be included on this website to describe your research project to the public.

Please enter your RUS in the area below. The RUS should be one or two paragraphs in length and include research objectives, the study design, and an analysis plan (including the phenotypic characteristics that will be tested for association with genetic variants). If you are requesting multiple datasets, please describe how you will use them. Examples of RUS can be found at here

Please limit your RUS to 2200 characters.

Figure S2. **Data access request structuring interface, Related to STAR Methods.**
Example of the DUOS interface capturing a data access request.



Data Repository

Dataset	DU restriction
Study a	HMB
Study b	HMB
Study c	GRU
Study d	DS- Diabetes
Study e	GRU
Study f	DS- Uveal Melanoma
Study g	Cancer

Figure S3. **Automated approval or denial based on the ontology terms hierarchy, Related to STAR Methods.**
Illustration of a data repository and the mapping of datasets to data use limitation in the form of ontology terms (left). Given a query to find all dataset that are approved for studying melanoma (bottom), a virtual mapping of datasets to an ontology tree hierarchy (right) illustrates how the DUOS algorithm approves ("green thumb up") or denies ("red thumb down") access to specific datasets based on their associated data use ontology term.

Tables

DUL number	Data use language in the DUL	Reason for inability to structure the DUL
#1	"The data is consented to be shared according to the terms included in the Consent Form".	We were unable to structure this DUL since no consent form was available.
#2	"General research use of aggregate level is prohibited".	Based on this language our Human Subject Protection experts were unable to determine what type of secondary use is allowed.
#3	"Data use is restricted to research in children	This DUL restricts the use of data

	under 18 years of age”	only for the purpose of pediatrics research. DUO does not have a way to encode such use and therefore this dataset was not available for automated data access requests.
#4	“No data may be used from participants who signed consent prior to” a given date”	There was no ability to encode this DUL since the date where each participant signed the consent form was not available for deposit in a repository. Therefore, there is no ability to deposit this data in a repository without collecting more information.
#5	“The data be held behind a firewall so that it is only available to qualified scientists and health care professionals.”	We were unable to structure this language with the existing DUO terms, and there is an ambiguity with regard to determining who is a qualified scientist.

Table S1. DULs that could not be structured.

Data use limitations clauses that could not be structured using the DUO ontology and an explanatory rationale.

DUL question in DUOS user interface	Data depositor’s answer	Ontology representation of DUL
1. Data is available for future general research use [GRU] (required)	Yes	GRU
	No	
2. Future use is limited for health/medical/biomedical research [HMB] (required)	Yes	HMB
	No	
3. Future use is limited to research involving the following disease area(s) [DS]	Ontology autocomplete	DS={node}

4. Future commercial use is prohibited [NCU] (required)	Yes	NCU
	No	
5. Future use by for-profit entities is prohibited [NPU] (required)	Yes	NPU
	No	
6. Future use for methods research (analytic/software/technology development) outside the bounds of the other specified restrictions is prohibited [NMDS] (required)	Yes	NMDS
	No	
7. Future use of aggregate-level data for general research purposes is prohibited [NAGR] (required)	Yes	NAGR
	No	
	Unspecified	
8. Future use as a control set for diseases other than those specified is prohibited [NCTRL] (required)	Yes	NCTRL
	No	
9. Future use is limited to research involving a particular gender [RS-G] (required)	Male	RS-M
	Female	RS-FM
	N/A	

10. Future use is limited to pediatric research [RS-PD] (required)	Yes	RS-PD
	No	
11. Future use is limited to research involving a specific population [RS-POP]	Free text input	RS-POP-XX
12. Future use is limited to data generated from samples collected after the following consent form date	Date input	

Table S2. Mapping logic used to structure DULs into ontology terms, Related to STAR Methods.

Mapping between structured questions and ontology codes in the backend. The following table lists the questions presented to a curator when cataloging a dataset in the DUOS repository. For each question, we illustrate how data use ontology codes are applied to the dataset according to the data depositor's choices.

If my Research Purpose has...	**The corresponding question in DUOS user interface**	**I should see ...**
Disease focused research	Future use is limited to research involving the following disease area(s) [DS]	Any dataset with GRU=true Any dataset with HMB=true. Any dataset tagged to this disease exactly. Any dataset tagged to a Human Disease Ontology/ Mondo Ontology Parent of disease X

Methods development/Validation study	Future use for methods research (analytic/software/technology development) outside the bounds of the other specified restrictions is prohibited [NMDS]	Any dataset where NMDS is false Any dataset where NMDS is true AND DS-X match
Control set	Future use as a control set for diseases other than those specified is prohibited [NCTRL]	Any dataset where NCTRL is false and is (GRU or HMB) Any DS-X match, if user specified a disease in the res purpose search
Study population origins or ancestry	Future use is limited to research involving a specific population [POA]	Any dataset tagged with GRU
Commercial purpose/by a commercial entity	Future commercial use is prohibited [NCU] Future use by for-profit entities is prohibited [NPU]	Any dataset where NPU and NCU are both false

Table S3. Mapping logic used to structure DARs into ontology terms, Related to STAR Methods.
Illustration of the DUOS algorithm mapping logic into ontology terms by specific data access request queries.

References

[1] 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 integrate and streamline access to ethically and legally diverse datasets. Cell Genomics.

[2] Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2018). Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic acids research 47(D1), D955–D962.