DATA ACCESS COMMITTEE REPORTING TOOL

Office of Data Science and Emerging Technology –
National Institute of Allergy and Infectious Diseases

Dr. Christopher S Marcum — Staff Scientist

CIVIC
DIGITAL
FELLOWSHIP

NIH

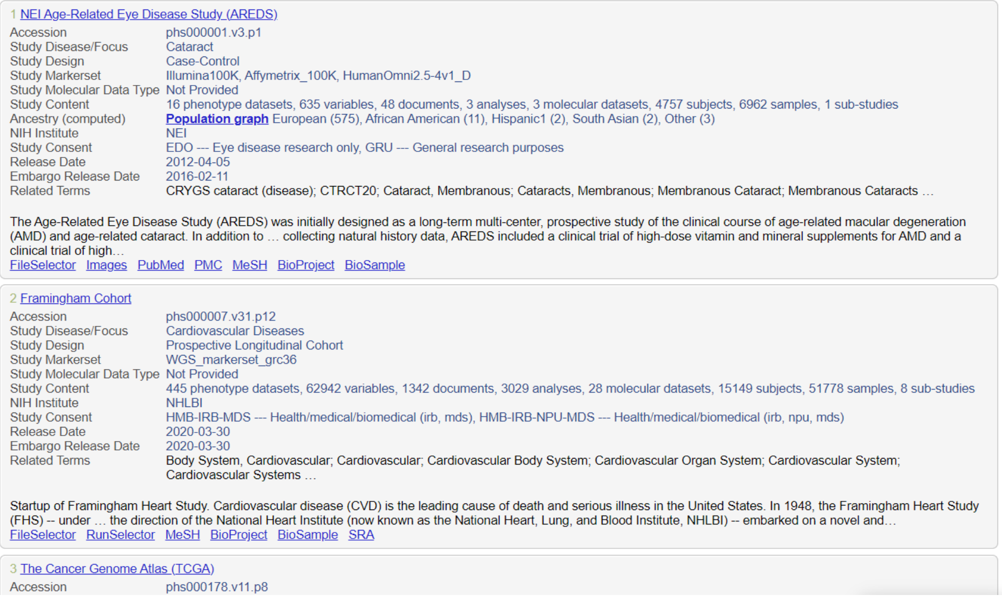
HOYIN CHU
Northeastern University
Computer Science and Mathematics

Background

- The database of Genotypes and Phenotypes (dbGaP) is a popular database hosting genomic data from institutes across the NIH

- Each institute has a Data Access Committee (DAC) which grants approval to data access requests (DAR)

dbGaP
GENOTYPES and PHENOTYPES



1 NEI Age-Related Eye Disease Study (AREDS)

Accession	phs000001.v3.p1
Study Disease/Focus	Cataract
Study Design	Case-Control
Study Markerset	Illumina100K, Affymetrix_100K, HumanOmni2.5-4v1_D
Study Molecular Data Type	Not Provided
Study Content	16 phenotype datasets, 635 variables, 48 documents, 3 analyses, 3 molecular datasets, 4757 subjects, 6962 samples, 1 sub-studies
Ancestry (computed)	**Population graph** European (575), African American (11), Hispanic1 (2), South Asian (2), Other (3)
NIH Institute	NEI
Study Consent	EDO --- Eye disease research only, GRU --- General research purposes
Release Date	2012-04-05
Embargo Release Date	2016-02-11
Related Terms	CRYGS cataract (disease); CTRCT20; Cataract, Membranous; Cataracts, Membranous; Membranous Cataract; Membranous Cataracts …

The Age-Related Eye Disease Study (AREDS) was initially designed as a long-term multi-center, prospective study of the clinical course of age-related macular degeneration (AMD) and age-related cataract. In addition to … collecting natural history data, AREDS included a clinical trial of high-dose vitamin and mineral supplements for AMD and a clinical trial of high…

FileSelector Images PubMed PMC MeSH BioProject BioSample

2 Framingham Cohort

Accession	phs000007.v31.p12
Study Disease/Focus	Cardiovascular Diseases
Study Design	Prospective Longitudinal Cohort
Study Markerset	WGS_markerset_grc36
Study Molecular Data Type	Not Provided
Study Content	445 phenotype datasets, 62942 variables, 1342 documents, 3029 analyses, 28 molecular datasets, 15149 subjects, 51778 samples, 8 sub-studies
NIH Institute	NHLBI
Study Consent	HMB-IRB-MDS --- Health/medical/biomedical (irb, mds), HMB-IRB-NPU-MDS --- Health/medical/biomedical (irb, npu, mds)
Release Date	2020-03-30
Embargo Release Date	2020-03-30
Related Terms	Body System, Cardiovascular; Cardiovascular; Cardiovascular Body System; Cardiovascular Organ System; Cardiovascular System; Cardiovascular Systems …

Startup of Framingham Heart Study. Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study (FHS) -- under … the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute, NHLBI) -- embarked on a novel and…

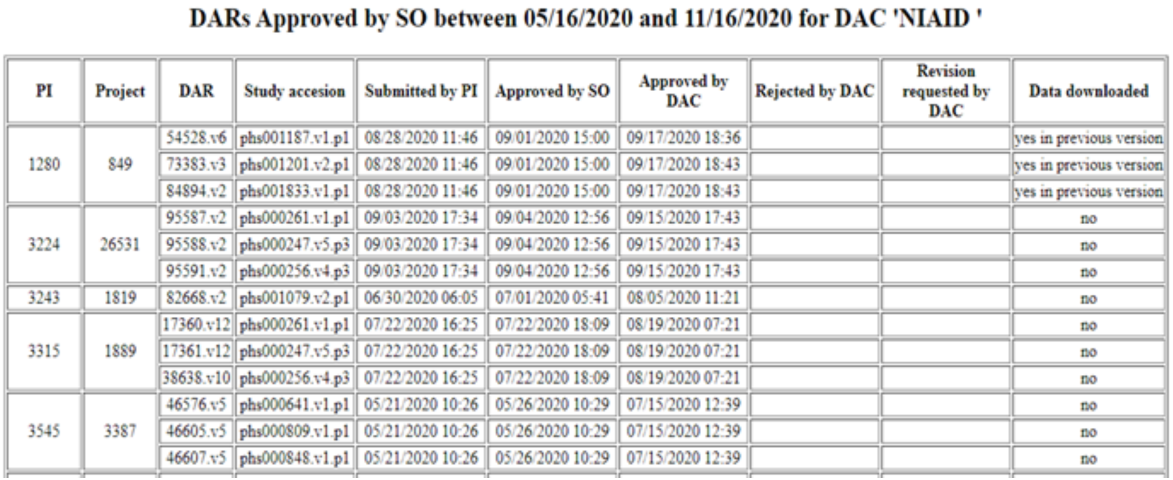
FileSelector RunSelector MeSH BioProject BioSample SRA

3 The Cancer Genome Atlas (TCGA)

Accession	phs000178.v11.p8

Current workflow

- Meta information about DARs are available as a table on a webpage
- They can help DACs understand research needs
- But it’s difficult to draw insights from just looking at this table

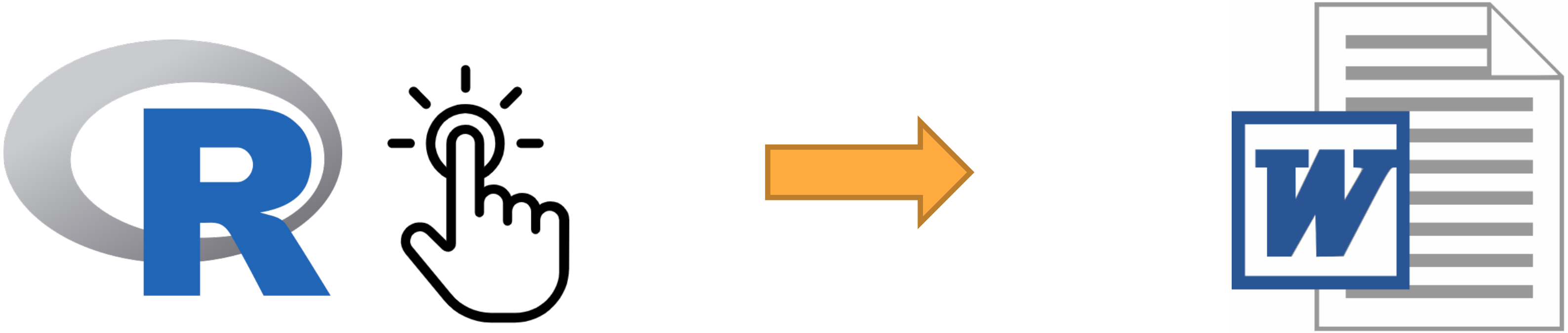


DARs Approved by SO between 05/16/2020 and 11/16/2020 for DAC 'NIAID '

PI	Project	DAR	Study accesion	Submitted by PI	Approved by SO	Approved by DAC	Rejected by DAC	Revision requested by DAC	Data downloaded
1280	849	54528.v6	phs001187.v1.p1	08/28/2020 11:46	09/01/2020 15:00	09/17/2020 18:36			yes in previous version
		73383.v3	phs001201.v2.p1	08/28/2020 11:46	09/01/2020 15:00	09/17/2020 18:43			yes in previous version
		84894.v2	phs001833.v1.p1	08/28/2020 11:46	09/01/2020 15:00	09/17/2020 18:43			yes in previous version
3224	26531	95587.v2	phs000261.v1.p1	09/03/2020 17:34	09/04/2020 12:56	09/15/2020 17:43			no
		95588.v2	phs000247.v5.p3	09/03/2020 17:34	09/04/2020 12:56	09/15/2020 17:43			no
		95591.v2	phs000256.v4.p3	09/03/2020 17:34	09/04/2020 12:56	09/15/2020 17:43			no
3243	1819	82668.v2	phs001079.v2.p1	06/30/2020 06:05	07/01/2020 05:41	08/05/2020 11:21			no
3315	1889	17360.v12	phs000261.v1.p1	07/22/2020 16:25	07/22/2020 18:09	08/19/2020 07:21			no
		17361.v12	phs000247.v5.p3	07/22/2020 16:25	07/22/2020 18:09	08/19/2020 07:21			no
		38638.v10	phs000256.v4.p3	07/22/2020 16:25	07/22/2020 18:09	08/19/2020 07:21			no
3545	3387	46576.v5	phs000641.v1.p1	05/21/2020 10:26	05/26/2020 10:29	07/15/2020 12:39			no
		46605.v5	phs000809.v1.p1	05/21/2020 10:26	05/26/2020 10:29	07/15/2020 12:39			no
		46607.v5	phs000848.v1.p1	05/21/2020 10:26	05/26/2020 10:29	07/15/2020 12:39			no

My Project

- A one-stop-shop R package that can automagically generate a data use report for any DAC using information scraped from the webpage



Report

CDAC Data Access Committee dbGaP Activity Report 2019-11-24-2020-12-14

Hoyin Chu, Christopher Steven Marcum

14 December, 2020

The CDAC Data Access Committee (DAC) currently manages 375 data access requests (DARs) for access to 341 projects in dbGaP.

1 Data Access Requests

Between 2019-11-24 and 2020-12-14 CDAC reviewed 223 DARs. Of these, 212 were accepted, 11 were downloaded, 57 had a previous version downloaded, and 4 were rejected. The average amount of time from when the Principle Investigator (PI) submited a DAR to the final decision by the DAC was 13.6 days. The average time to an accepted decision was 13.6 days, while the average time to a rejected decision was 10.2 days. Figure 1.1 is a barplot comparing the CDAC DAC to time to final decision to the average across all NIH DACs during the same time interval.

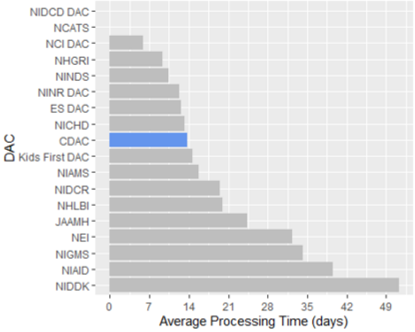


Figure 1.1: Comparison of DAR Processing Time among all DACs

CIVIC DIGITAL FELLOWSHIP

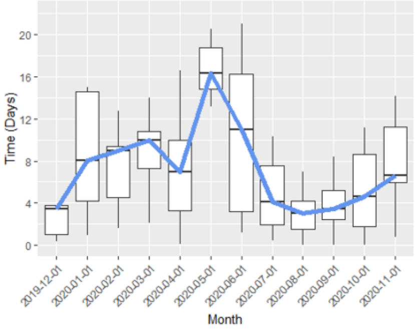


Figure 1.4: DAR Processing Time: From SO Approval to DAC Approval

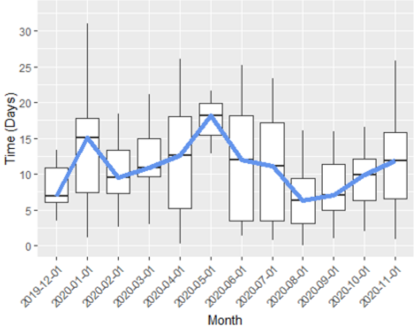
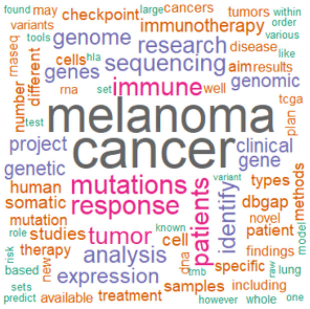
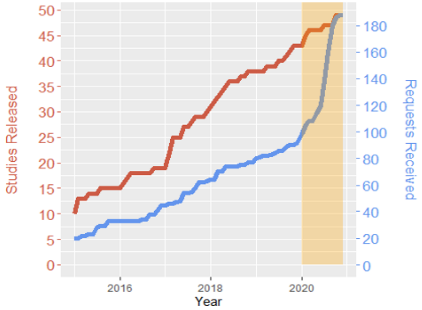
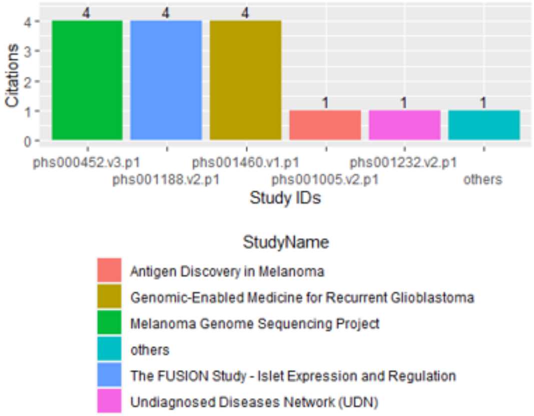


Figure 1.5: DAR Processing Time: From PI Submission to DAC Approval





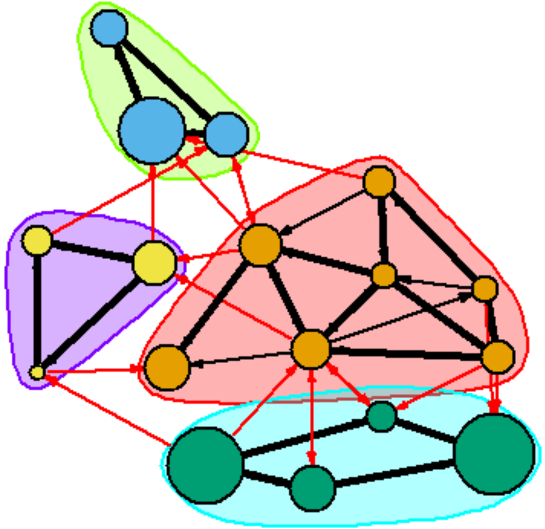


Future Work

- These data can help us answer more complicated research questions

- Examples: Study-to-Study Network

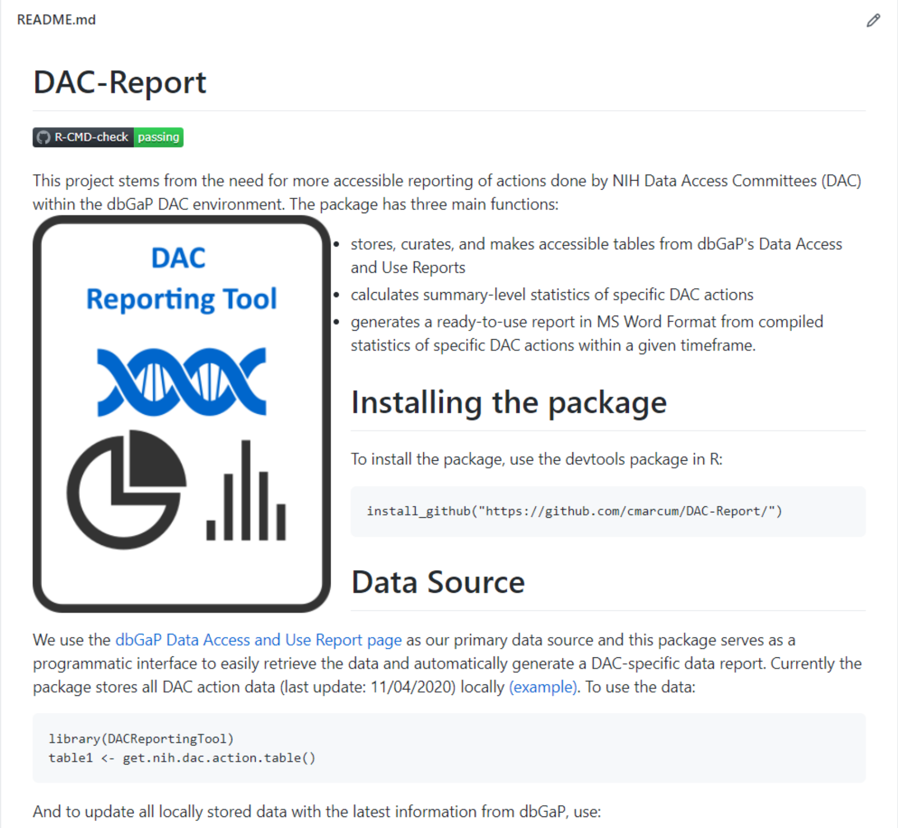
- What are some unexpected communities?
- How do these networks evolve over time?



A Starting Point

- Open-source practices (version control, continuous integration, documentation…) were followed during the development of the package

- Package will be available to the public and features can be added via pull requests



README.md

DAC-Report

R-CMD-check passing

This project stems from the need for more accessible reporting of actions done by NIH Data Access Committees (DAC) within the dbGaP DAC environment. The package has three main functions:

- stores, curates, and makes accessible tables from dbGaP's Data Access and Use Reports
- calculates summary-level statistics of specific DAC actions
- generates a ready-to-use report in MS Word Format from compiled statistics of specific DAC actions within a given timeframe.

Installing the package

To install the package, use the devtools package in R:

```
install_github("https://github.com/cmarcum/DAC-Report/")
```

Data Source

We use the dbGaP Data Access and Use Report page as our primary data source and this package serves as a programmatic interface to easily retrieve the data and automatically generate a DAC-specific data report. Currently the package stores all DAC action data (last update: 11/04/2020) locally (example). To use the data:

```
library(DACReportingTool)
table1 <- get.nih.dac.action.table()
```

And to update all locally stored data with the latest information from dbGaP, use:

Thank you!

- ODSET (Office of Data Science and Emerging Technology) Team
- Emerging Leaders in Data Science Fellows
- Everyone who made the fellowship possible

References

- Icons: thenounproject
- Network Image: https://kateto.net/network-visualization