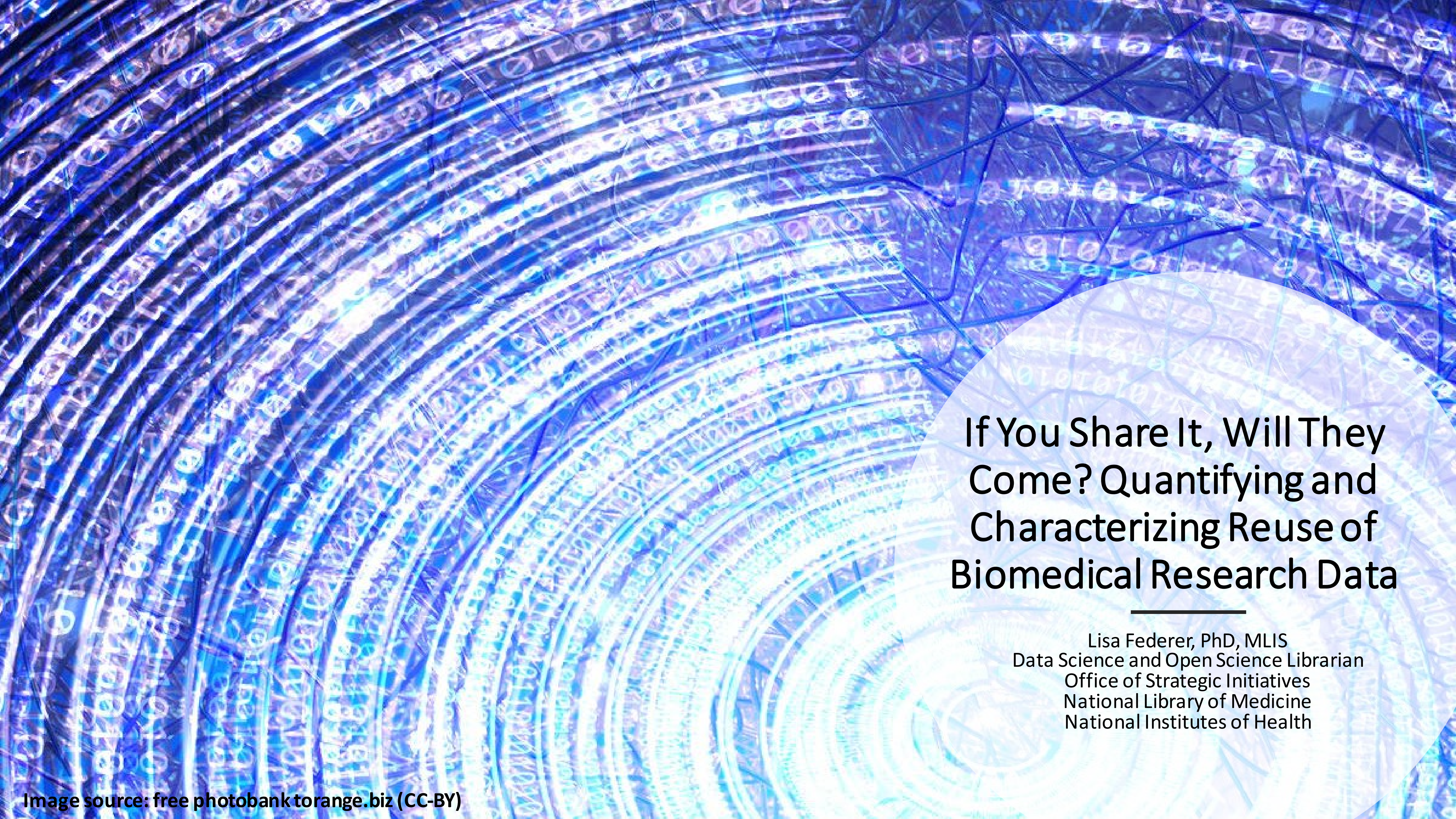


If You Share It, Will They Come? Quantifying and Characterizing Reuse of Biomedical Research Data

Lisa Federer, PhD, MLIS
Data Science and Open Science Librarian
Office of Strategic Initiatives
National Library of Medicine
National Institutes of Health

Image source: free photobank torange.biz (CC-BY)

Overview

- Background: where did all these datasets come from?
- Methods
- Findings: what happens with these datasets once they’re shared?
- Implications

Images from FlatIcon except where otherwise specified

Background

Cheaper and faster data generation

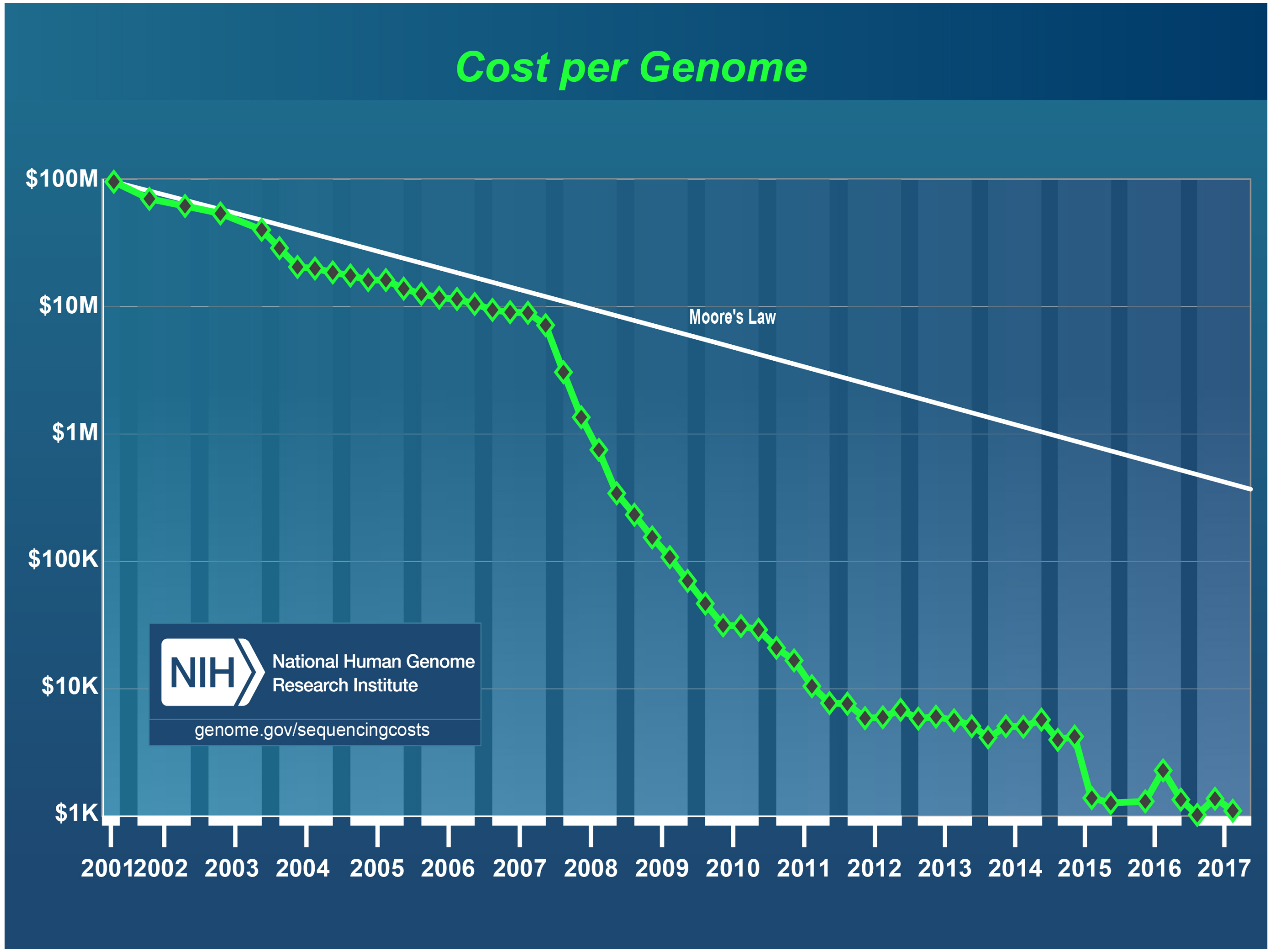


Image source: National Human Genome Research Institute

Greater availability of data

The Human Connectome Project
NIH The Common Fund
GTEx
ECHO
Environmental Influences on Child Health Outcomes
HMP
NIH HUMAN MICROBIOME PROJECT
CANCER MOONSHOT
Human Genome Project
BIOLOGY PHYSICS ETHICS INFORMATICS ENGINEERING CHEMISTRY
RESOURCES FOR THE BIOLOGY CENTURY
ADNI
ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE
TOPMed
FlyBase
MGI
ZFIN
The Zebrafish Model Organism Database
SGD Saccharomyces GENOME DATABASE
WormBase
Adolescent Brain Cognitive Development
Teen Brains. Today's Science. Brighter Future.
All of Us SM
THE FUTURE OF HEALTH BEGINS WITH YOU
The Precision Medicine Initiative®
THE BRAIN INITIATIVE
NATIONAL INSTITUTES OF HEALTH

Funder and journal sharing policies

NATIONAL INSTITUTES OF HEALTH

NSF

BMJ
British Medical Journal

npg
nature publishing group

INSTITUTE of Museum and Library SERVICES

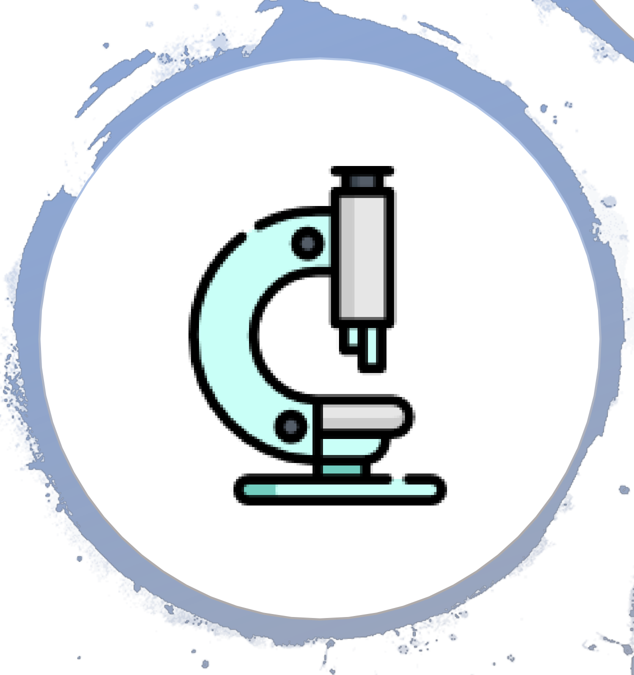
PLoS PUBLIC LIBRARY of SCIENCE

GORDON AND BETTY MOORE FOUNDATION

Science
AAAS

National Institutes of Health

- The primary biomedical and public health research agency of the United States
  - 27 Institutes and Centers focused on diseases, organ systems, and types of research
  - Invests nearly $37.3 billion annually in medical research
- Extramural research program: awards more than 50,000 competitive grants annually to research in every US state and around the world
- Intramural research program
  - World’s largest biomedical research institution
  - Nearly 6,000 scientists, primarily at the NIH campus in Bethesda, Maryland



National Library of Medicine

- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science
- The world's largest biomedical library (1836)
  - Create & host major resources, tools, & services for literature, data, standards, & more
    - Send > 115 terabytes of data to > 5 million users daily
    - Receive > 15 terabytes of data from > 3,000 users daily
  - Facilitate open science & scholarship by making digital research objects:
    - Findable, Accessible, Interoperable, & Reusable (FAIR)
    - As well as Attributable & Sustainable

But what’s happening with all the data?

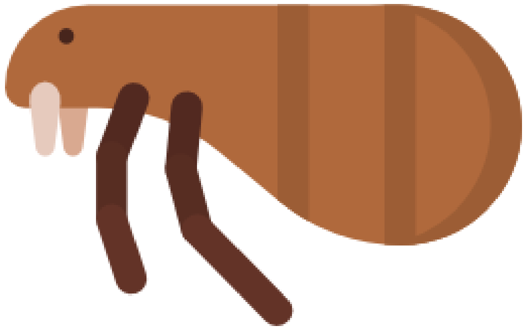
Existing research has explored:

- Researchers’ attitudes about data reuse
- Factors that influence researchers’ choice to use a particular dataset
- Subjective experiences of researchers in a few particular disciplines

Why does this matter?

Science as a credit economy

Bibliometrics as a means to quantify impact

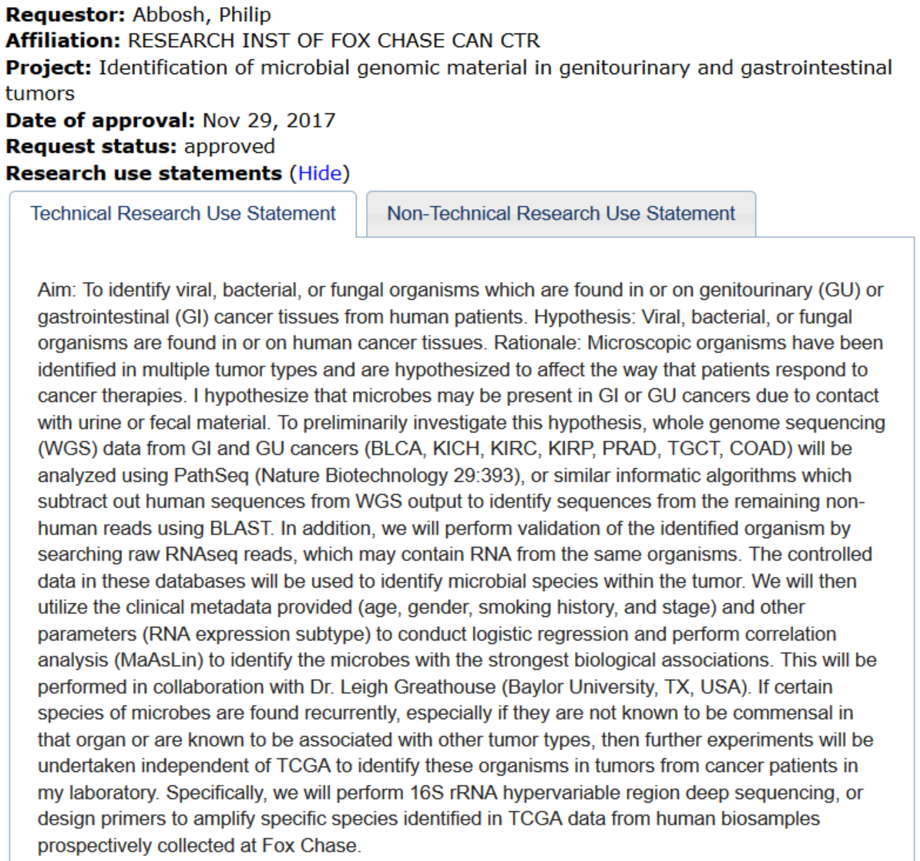


Quantifying impact of shared data enables reward to creators (no more “research parasites”)

Methods

Sampling and data collection

A proxy for reuse: use requests

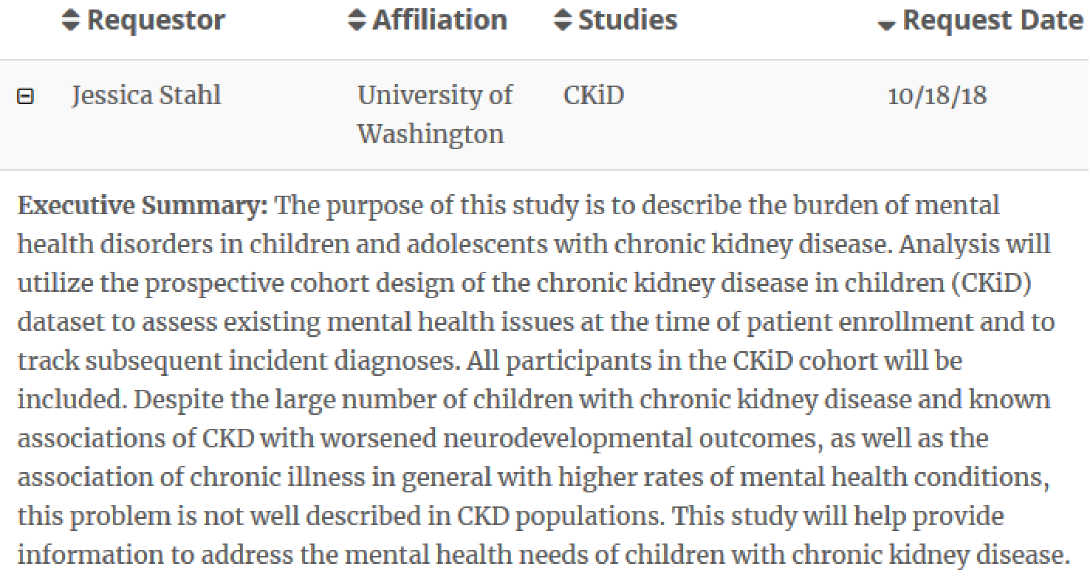


Requestor: Abbosh, Philip
Affiliation: RESEARCH INST OF FOX CHASE CAN CTR
Project: Identification of microbial genomic material in genitourinary and gastrointestinal tumors
Date of approval: Nov 29, 2017
Request status: approved
Research use statements (Hide)

Technical Research Use Statement
Non-Technical Research Use Statement

Aim: To identify viral, bacterial, or fungal organisms which are found in or on genitourinary (GU) or gastrointestinal (GI) cancer tissues from human patients. Hypothesis: Viral, bacterial, or fungal organisms are found in or on human cancer tissues. Rationale: Microscopic organisms have been identified in multiple tumor types and are hypothesized to affect the way that patients respond to cancer therapies. I hypothesize that microbes may be present in GI or GU cancers due to contact with urine or fecal material. To preliminarily investigate this hypothesis, whole genome sequencing (WGS) data from GI and GU cancers (BLCA, KICH, KIRC, KIRP, PRAD, TGCT, COAD) will be analyzed using PathSeq (Nature Biotechnology 29:393), or similar informatic algorithms which subtract out human sequences from WGS output to identify sequences from the remaining non-human reads using BLAST. In addition, we will perform validation of the identified organism by searching raw RNAseq reads, which may contain RNA from the same organisms. The controlled data in these databases will be used to identify microbial species within the tumor. We will then utilize the clinical metadata provided (age, gender, smoking history, and stage) and other parameters (RNA expression subtype) to conduct logistic regression and perform correlation analysis (MaAsLin) to identify the microbes with the strongest biological associations. This will be performed in collaboration with Dr. Leigh Greathouse (Baylor University, TX, USA). If certain species of microbes are found recurrently, especially if they are not known to be commensal in that organ or are known to be associated with other tumor types, then further experiments will be undertaken independent of TCGA to identify these organisms in tumors from cancer patients in my laboratory. Specifically, we will perform 16S rRNA hypervariable region deep sequencing, or design primers to amplify specific species identified in TCGA data from human biosamples prospectively collected at Fox Chase.

Sample dbGaP use request



	Requestor	Affiliation	Studies	Request Date
⊟	Jessica Stahl	University of Washington	CKiD	10/18/18

Executive Summary: The purpose of this study is to describe the burden of mental health disorders in children and adolescents with chronic kidney disease. Analysis will utilize the prospective cohort design of the chronic kidney disease in children (CKiD) dataset to assess existing mental health issues at the time of patient enrollment and to track subsequent incident diagnoses. All participants in the CKiD cohort will be included. Despite the large number of children with chronic kidney disease and known associations of CKD with worsened neurodevelopmental outcomes, as well as the association of chronic illness in general with higher rates of mental health conditions, this problem is not well described in CKD populations. This study will help provide information to address the mental health needs of children with chronic kidney disease.

Sample NIDDK use request

Repositories in the study

Genomic data

dbGaP

GENOTYPES and PHENOTYPES

Clinical data

NIH

National Institute of
Diabetes and Digestive
and Kidney Diseases

NIDDK Central Repository

BioLINCC

Biologic Specimen and Data Repository Information Coordinating Center

Web Access to NHLBI Biospecimens and Data

Data included in the study

	dbGaP	NHLBI	NIDDK	All combined
Datasets	1,014	146	77	1,237
Total requestors	5,260	N/A	253	5,513
Total institutional affiliations	1,230	1,001	195	2,426
Total requests	9,444	1,939	416	11,799
Total datasets requested	104,326	3,864	506	108,696

Findings

What’s happening with all these datasets?

Requests by reuse type

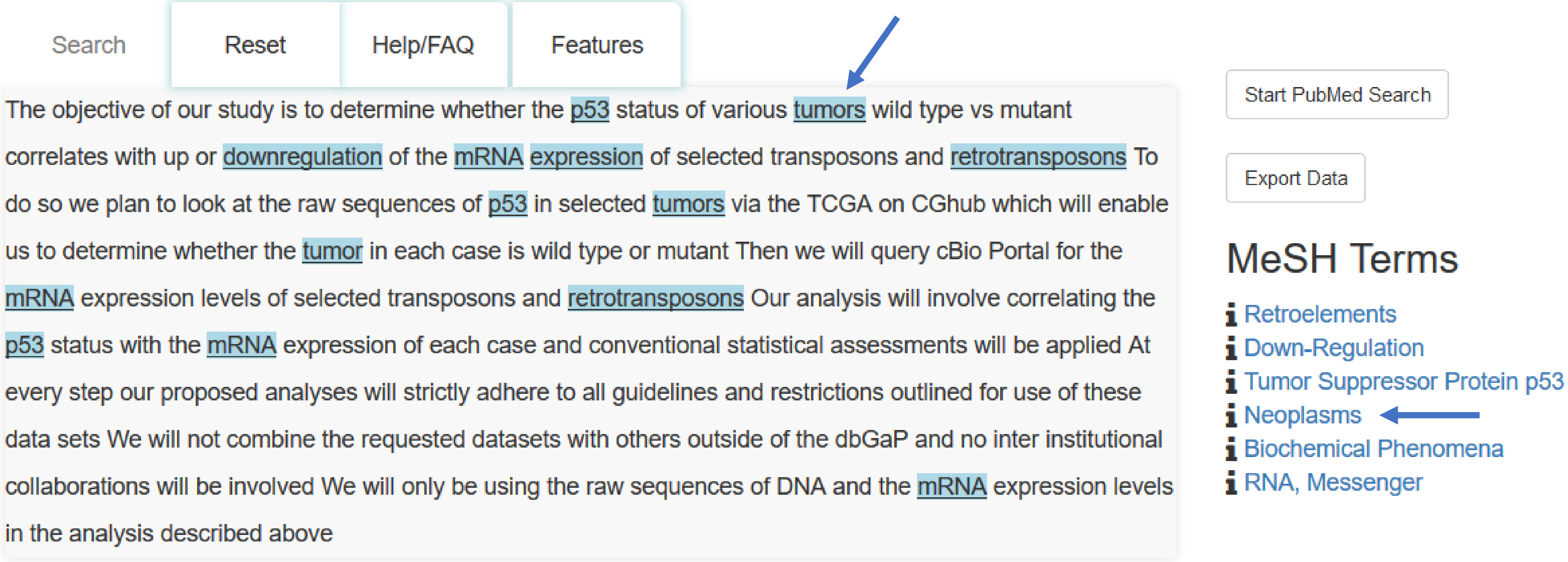
Category	Definition
Original research study	use of a single dataset to answer a new research question, distinct from the specific question for which the data were originally collected
Meta-analysis study	aggregation or integration of the dataset with other datasets to answer a research question or conduct a formal meta-analysis
Statistical methods study	use of one or more datasets to develop or verify new statistical methodology
Software or tool development study	use of one or more datasets to develop, test, or validate a new software product or analysis tool
Validation	use of one or more datasets to validate other findings, such as validating findings from an animal model in human subjects
Comparison or control	use of one or more datasets to validate the investigator's own data, provide comparison, or serve as a control group
Reproducibility or reanalysis study	reanalysis of one or more datasets to answer the same question for which the data were originally collected or to verify the original study's findings
Infrastructure	use of one or more datasets to populate a database or repository for internal or institutional use

Reuse types

Reuse type	dbGaP Requests		NIDDK requests	
	N	%	N	%
Original research	460	2.3%	282	50.27%
Meta-analysis	14,619	72.4%	139	24.78%
Comparison	858	4.3%	2	0.36%
Validation	221	1.2%	14	2.5%
Statistics	2,242	11.1%	84	15.0%
Software	1,097	5.4%	14	2.5%
Infrastructure	644	3.2%	0	0%
Re-analysis	11	0.05%	2	0.36%
Reuse type not specified	2	0.01%	24	4.28%

($\chi2 = 4547$, df = 8, $p < 0.01$)

Automated coding for reuse topic



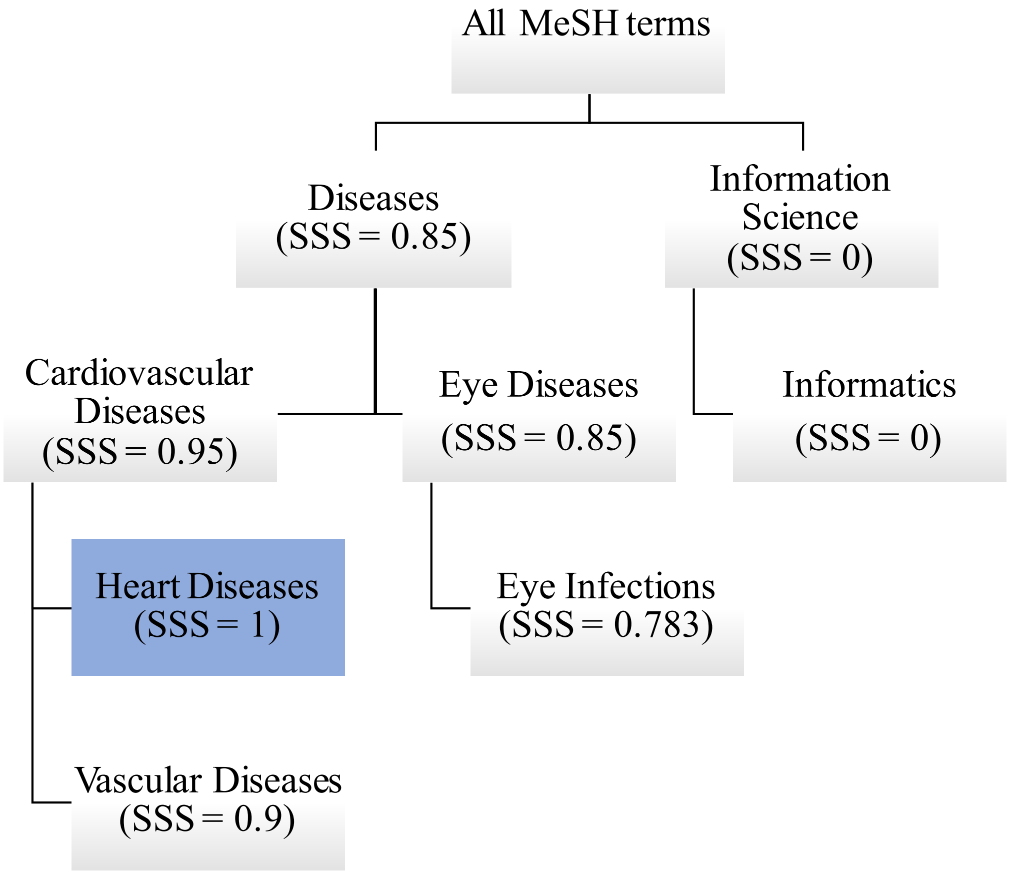
The objective of our study is to determine whether the p53 status of various tumors wild type vs mutant correlates with up or downregulation of the mRNA expression of selected transposons and retrotransposons To do so we plan to look at the raw sequences of p53 in selected tumors via the TCGA on CGhub which will enable us to determine whether the tumor in each case is wild type or mutant Then we will query cBio Portal for the mRNA expression levels of selected transposons and retrotransposons Our analysis will involve correlating the p53 status with the mRNA expression of each case and conventional statistical assessments will be applied At every step our proposed analyses will strictly adhere to all guidelines and restrictions outlined for use of these data sets We will not combine the requested datasets with others outside of the dbGaP and no inter institutional collaborations will be involved We will only be using the raw sequences of DNA and the mRNA expression levels in the analysis described above

MeSH Terms

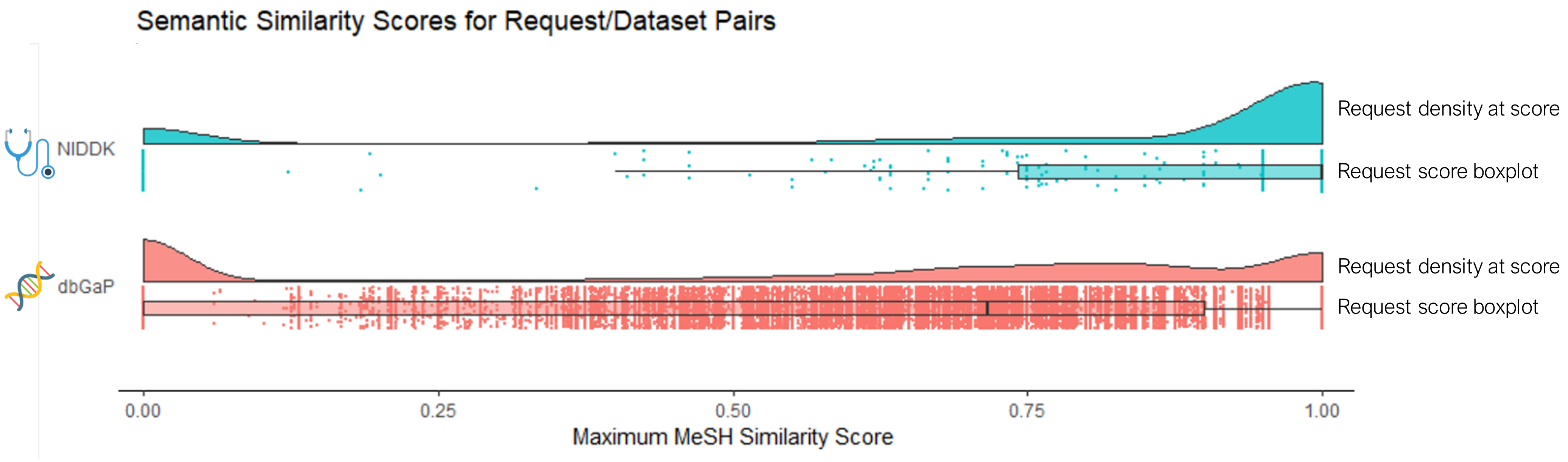
- Retroelements
- Down-Regulation
- Tumor Suppressor Protein p53
- Neoplasms
- Biochemical Phenomena
- RNA, Messenger

NLM Medical Text Indexer: https://ii.nlm.nih.gov/MTI/

MeSH terms and semantic similarity


All MeSH terms
Diseases (SSS = 0.85)
Information Science (SSS = 0)
Cardiovascular Diseases (SSS = 0.95)
Eye Diseases (SSS = 0.85)
Informatics (SSS = 0)
Heart Diseases (SSS = 1)
Eye Infections (SSS = 0.783)
Vascular Diseases (SSS = 0.9)

Request/dataset topic similarity

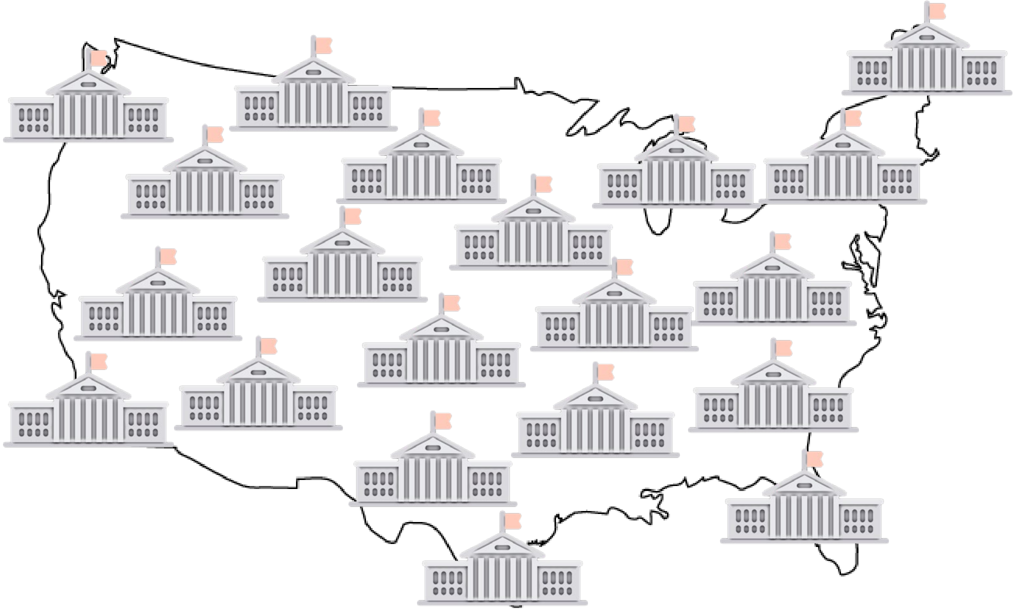


Coding for career stage and institution location

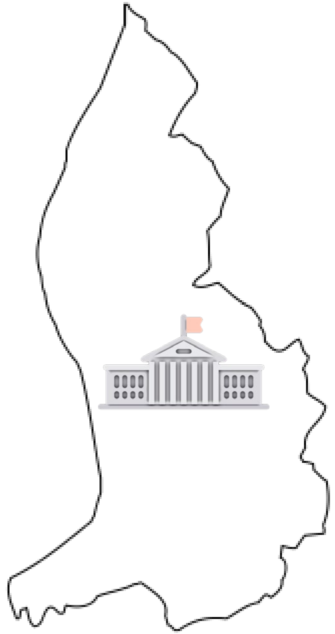
Name	Institution	Date	Status
Doe, John	Duke University	15-Jan-14	assoc_prof
Doe, John	Duke University	25-Jan-17	prof

Institution Name	# of requests	Latitude	Longitude	Country
University of Oulu	10	65.093	25.4663	Finland
deCODE Genetics, EHF	78	64.1265	-21.8174	Iceland

Calculating relative difference in composition

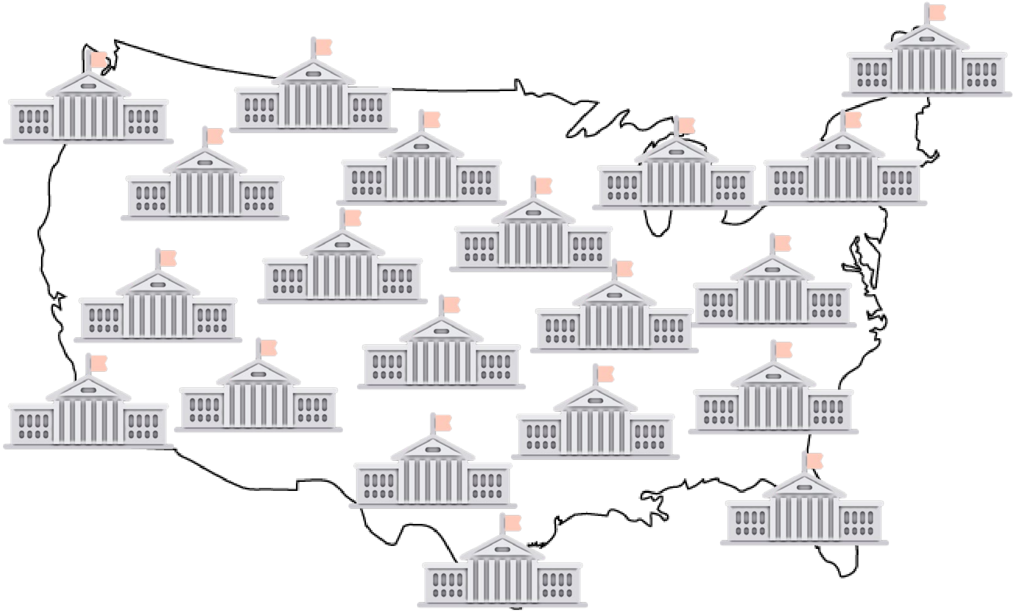


United States



Liechtenstein

Calculating relative difference in composition (RDC)



United States

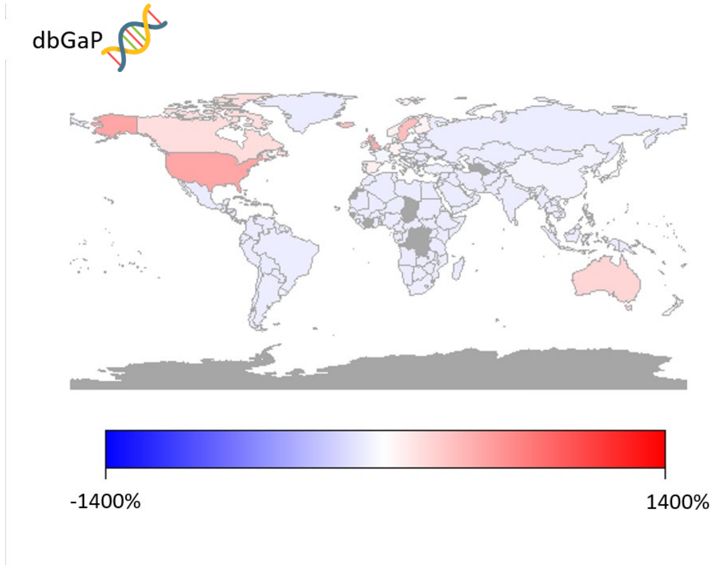
Difference in composition = % of world requests – % of world universities

$$\text{RDC} = \frac{\text{Difference in composition}}{\text{\% of world universities}} \times 100$$

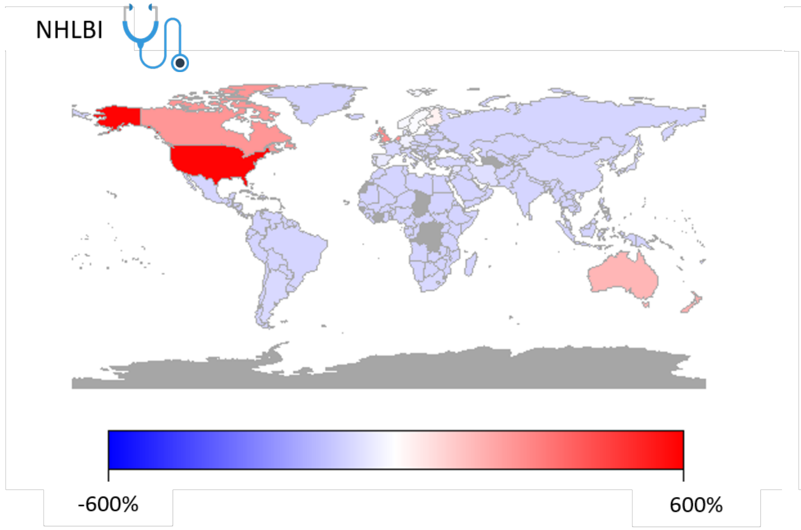
Difference in composition = 67.8% of requests came from US – 11.6% of all universities are in US = 56.2%

$$\text{RDC} = \frac{56.2\%}{11.6\%} \times 100 = 484.5\%$$

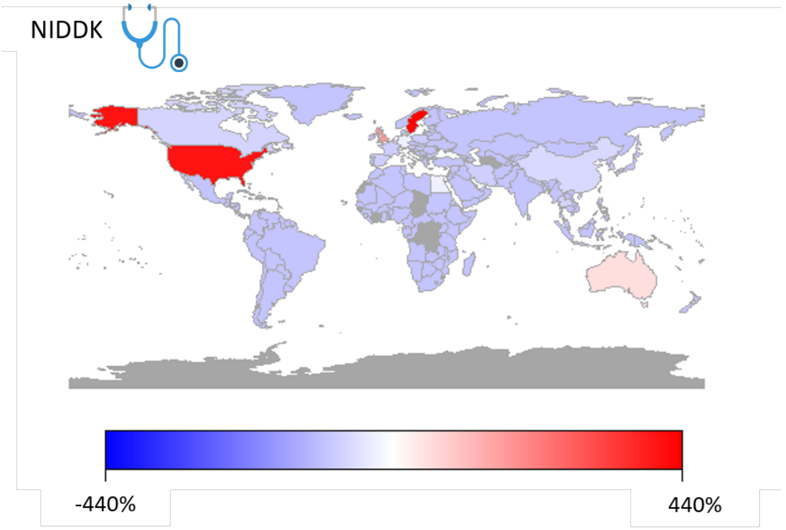
RDC of requests/research presence



dbGaP



NHLBI



NIDDK

Most overrepresented countries

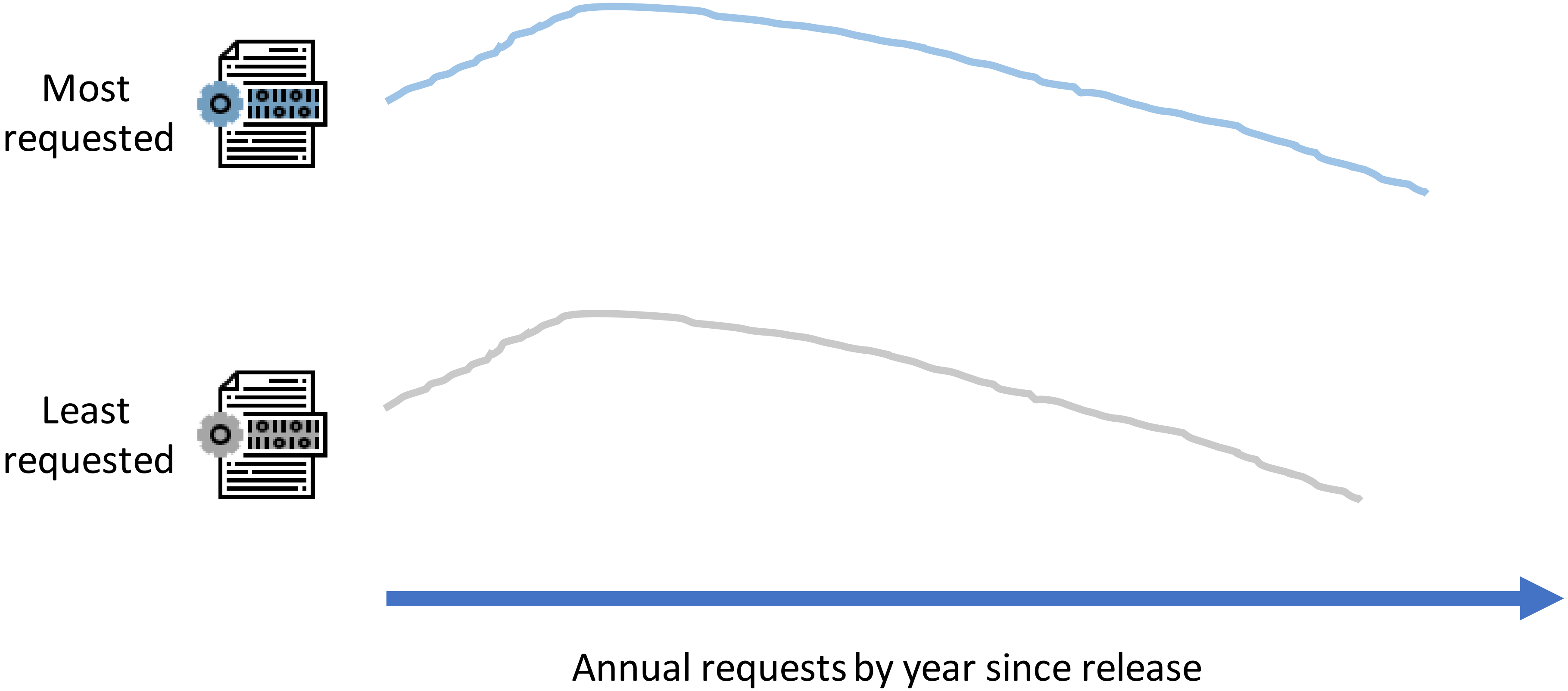
Country	University Count	dbGaP		NIDDK		NHLBI	
		N	RDC	N	RDC	N	RDC
Australia	188	183	221%	6	55%	35	170%
Canada	355	301	179%	2	-72%	85	246%
Cyprus	26	1	-89%	1	84%	0	-100%
Finland	46	23	65%	0	-100%	4	28%
Germany	465	223	58%	2	-26%	22	-32%
Iceland	9	12	337%	0	-100%	0	-100%
Israel	42	77	501%	0	-100%	10	248%
Italy	239	86	19%	5	2%	1	-94%
Luxembourg	3	14	1,397%	0	-100%	0	-100%
Netherlands	133	106	162%	2	-26%	32	248%
New Zealand	56	27	60%	0	-100%	11	186%
Qatar	9	0	-100%	0	-100%	1	56%
Singapore	45	44	224%	0	-100%	3	-6%
Sweden	46	63	352%	5	431%	3	-8%
Switzerland	102	59	90%	2	-4%	4	-42%
United Kingdom	280	471	484%	16	179%	71	267%
United States	3,257	5,773	484%	338	406%	1,556	592%

Career status of requestors

Career Stage	Title	Percent of dbGaP requests	Percent of NIDDK requests
Pre-professional	Student	0.7%	1.8%
	Fellow	0.7%	3.1%
	Total	1.4%	4.9%
Early career	Assistant Professor	19.1%	27.6%
	Resident Physician	0%	1.1%
	Lecturer	0.07%	0.4%
	Instructor	0.07%	0%
	Total	19.2%	29.1%
Mid-Career	Associate Professor	15.4%	13%
	Scientist	5.7%	3.9%
	Attending Physician	0%	0.2%
	Manager	0.7%	0.4%
	Total	21.8%	17.5%
Established	Professor	26.8%	24%
	Director	8.5%	5.5%
	Executive	3%	5.1%
	Senior Scientist	10.3%	6.7%
	Total	48.6%	41.3%
Unknown		9%	5.9%

(χ2 =81, df= 12, $p < 0.001$)

Tracking dataset requests over time

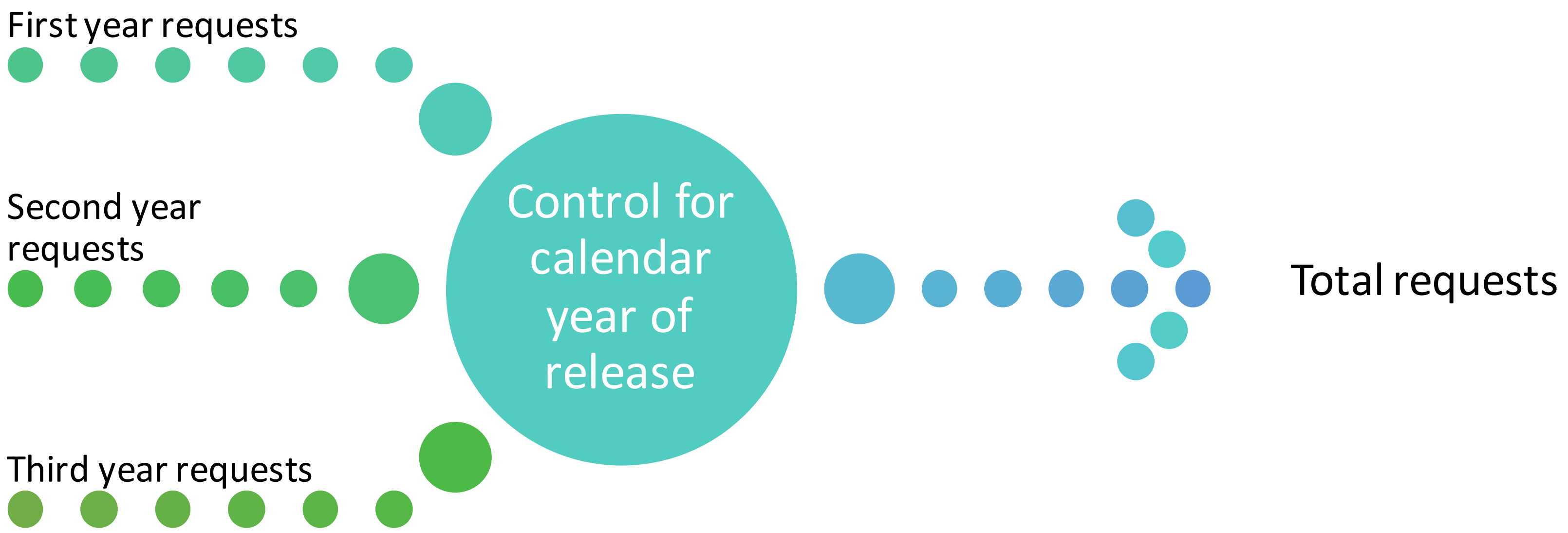


Most requested

Least requested

Annual requests by year since release

Predictive power of early requests



First year requests

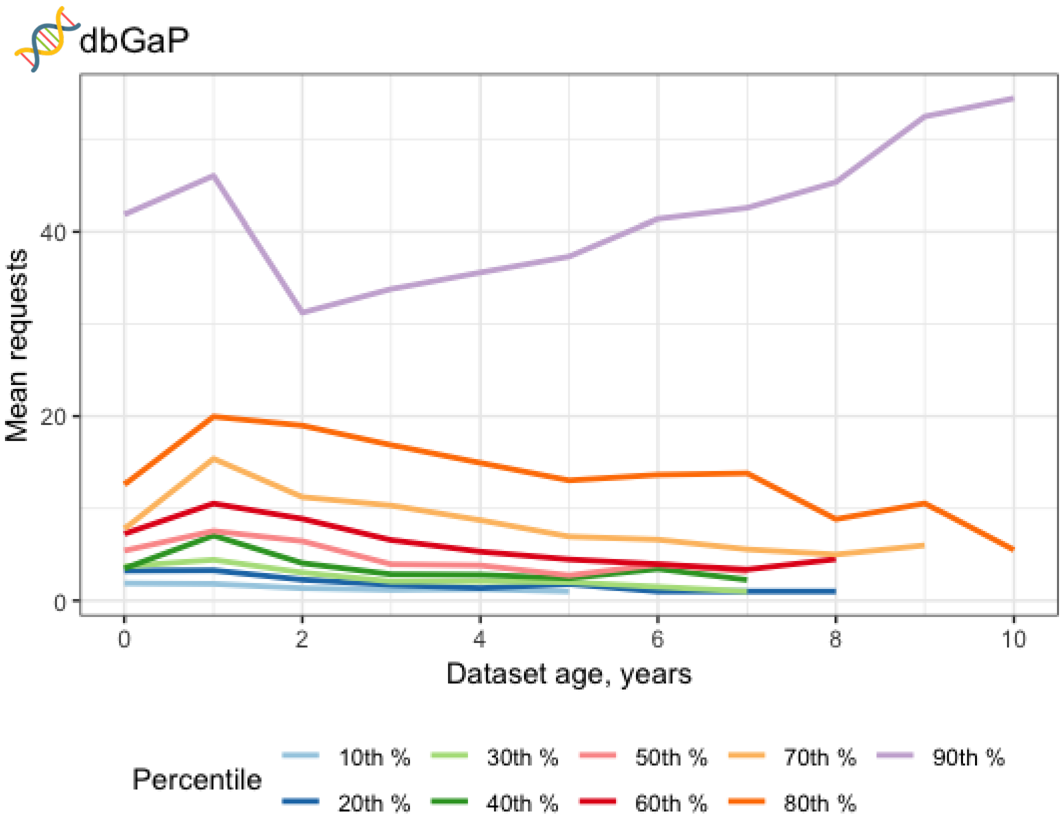
Second year requests

Third year requests

Control for calendar year of release

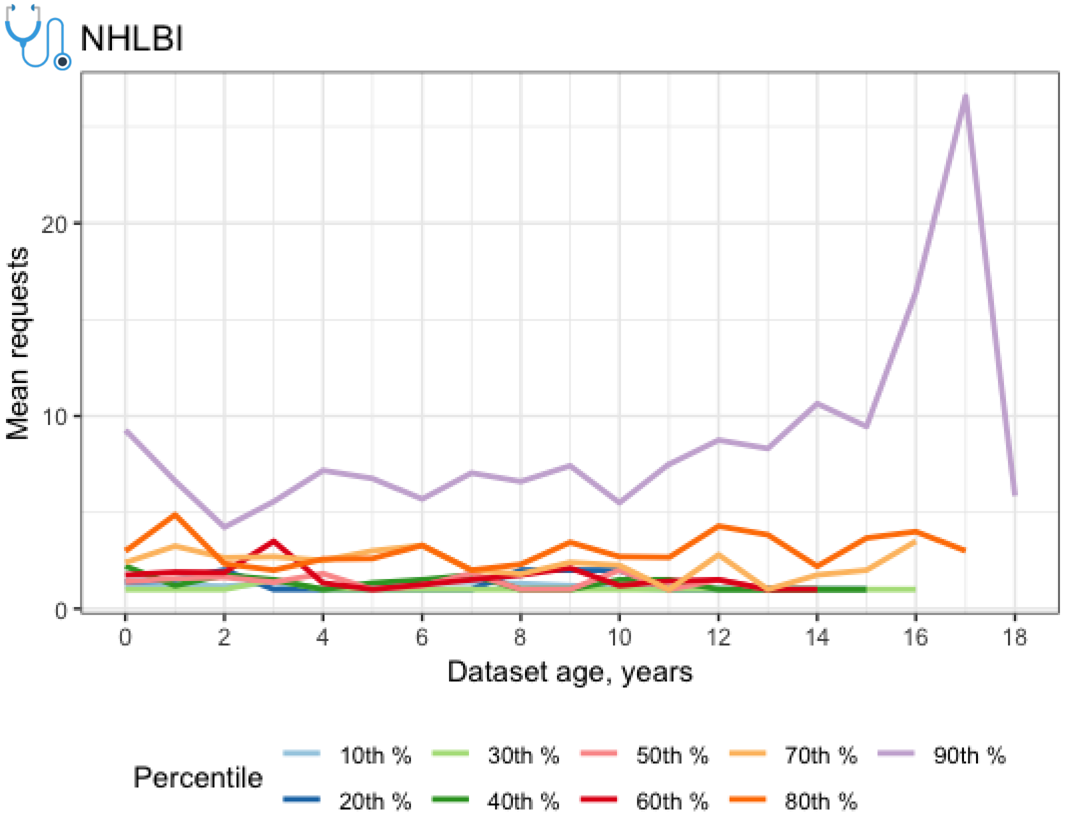
Total requests

Requests by year, dbGaP



Model	R-squared	p-value
One year	0.73	<0.001
Two years	0.8	<0.001
Three years	0.87	<0.001

Requests over time, NHLBI



Model	R-squared	p-value
One year	0.8	<0.001
Two years	0.89	<0.001
Three years	0.96	<0.001

Determining highly requested topics

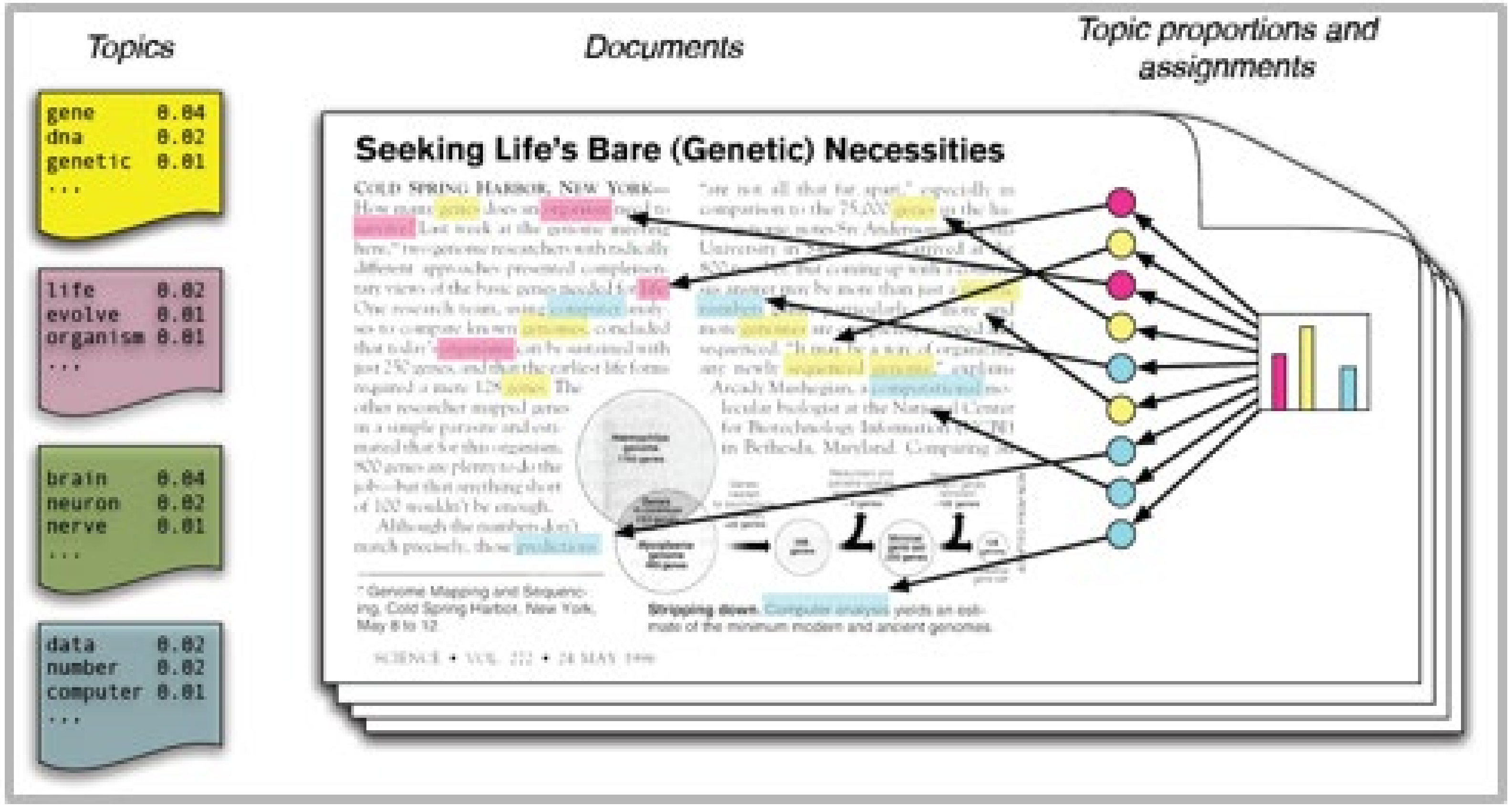
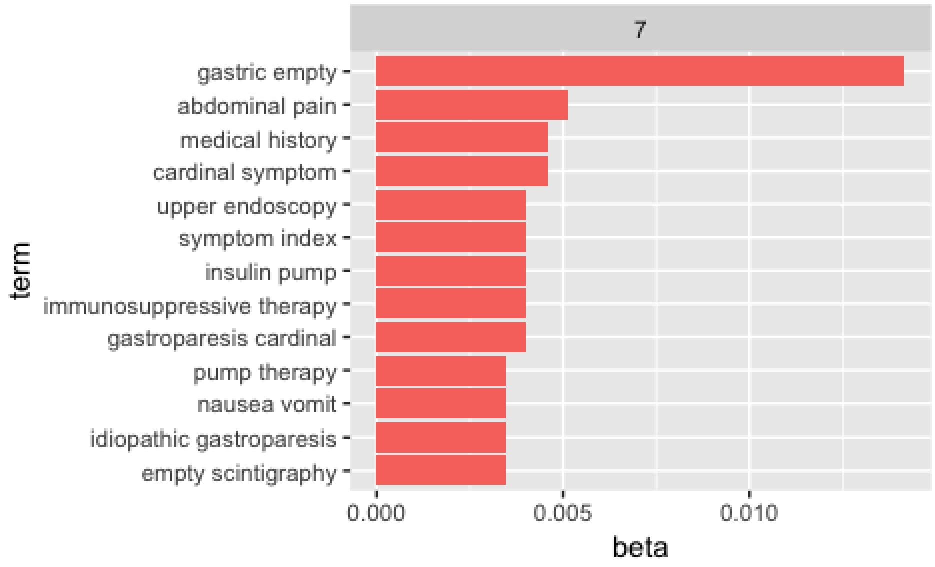
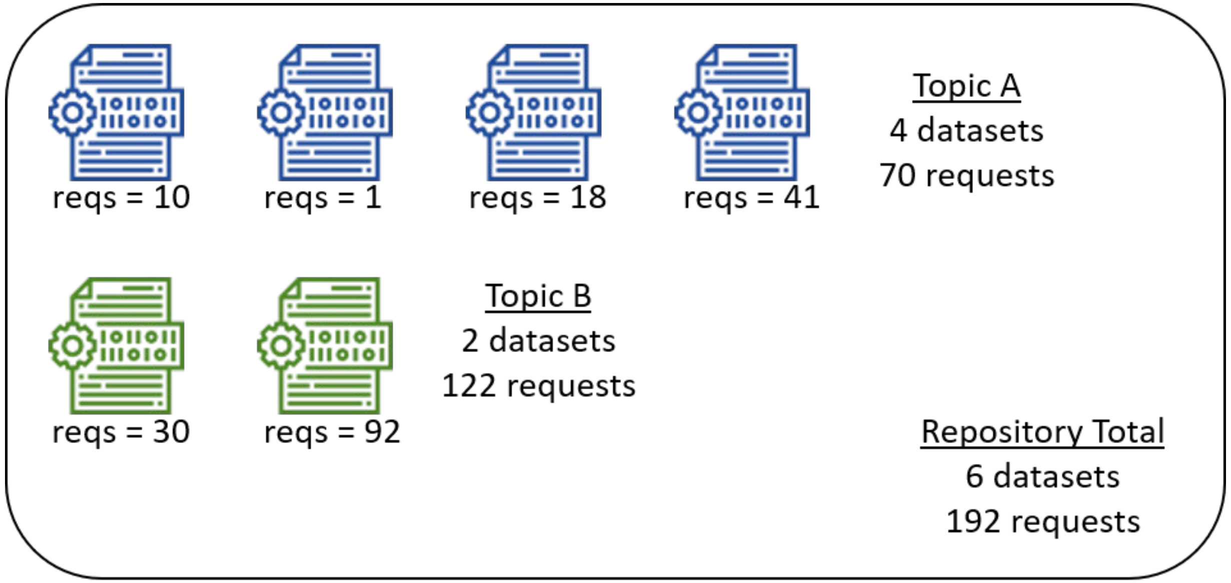

Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Image source: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

Sample topicmodels output



Request to Dataset (RTD) Ratio



reqs = 10

reqs = 1

reqs = 18

reqs = 41

Topic A
4 datasets
70 requests

reqs = 30

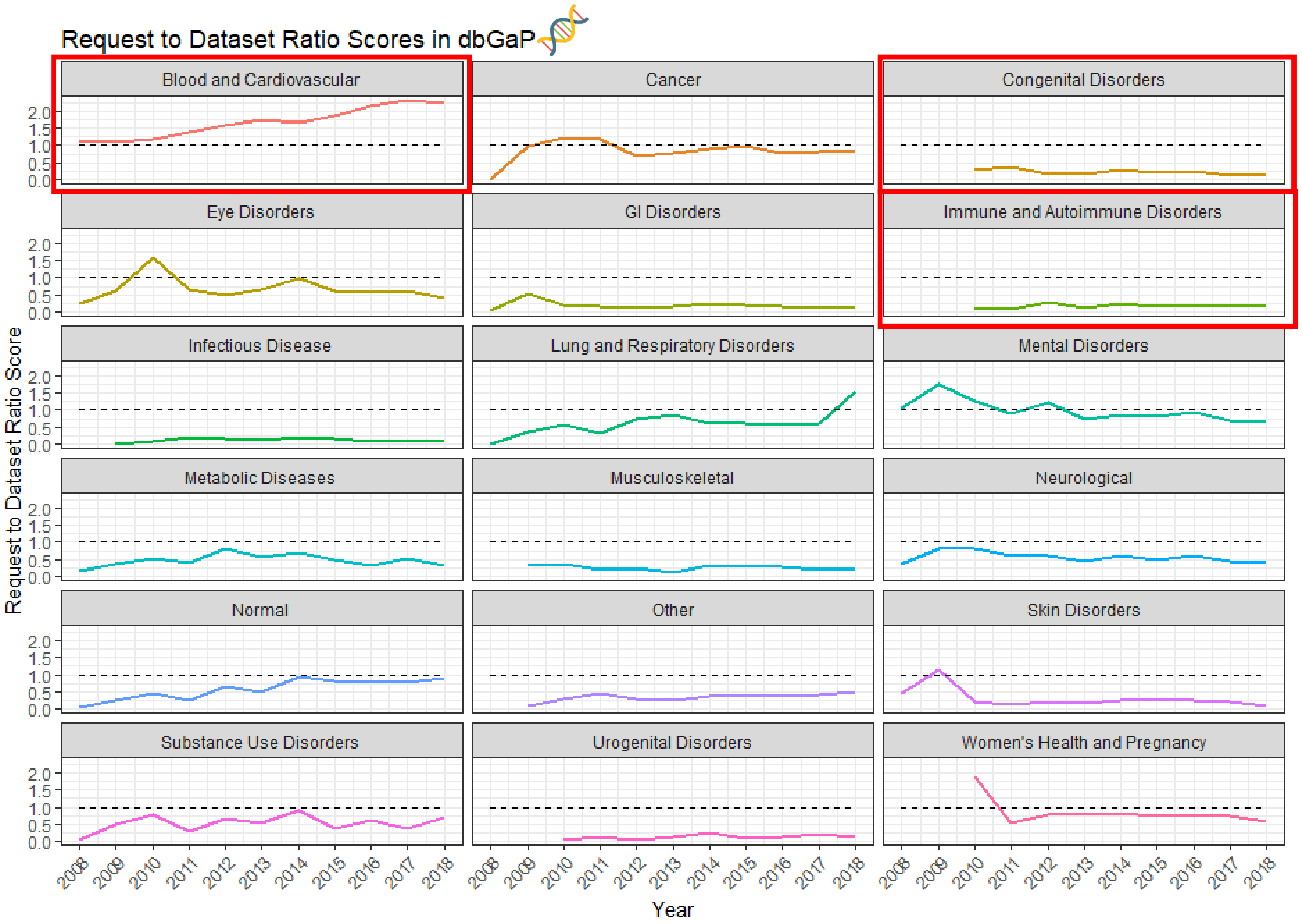
reqs = 92

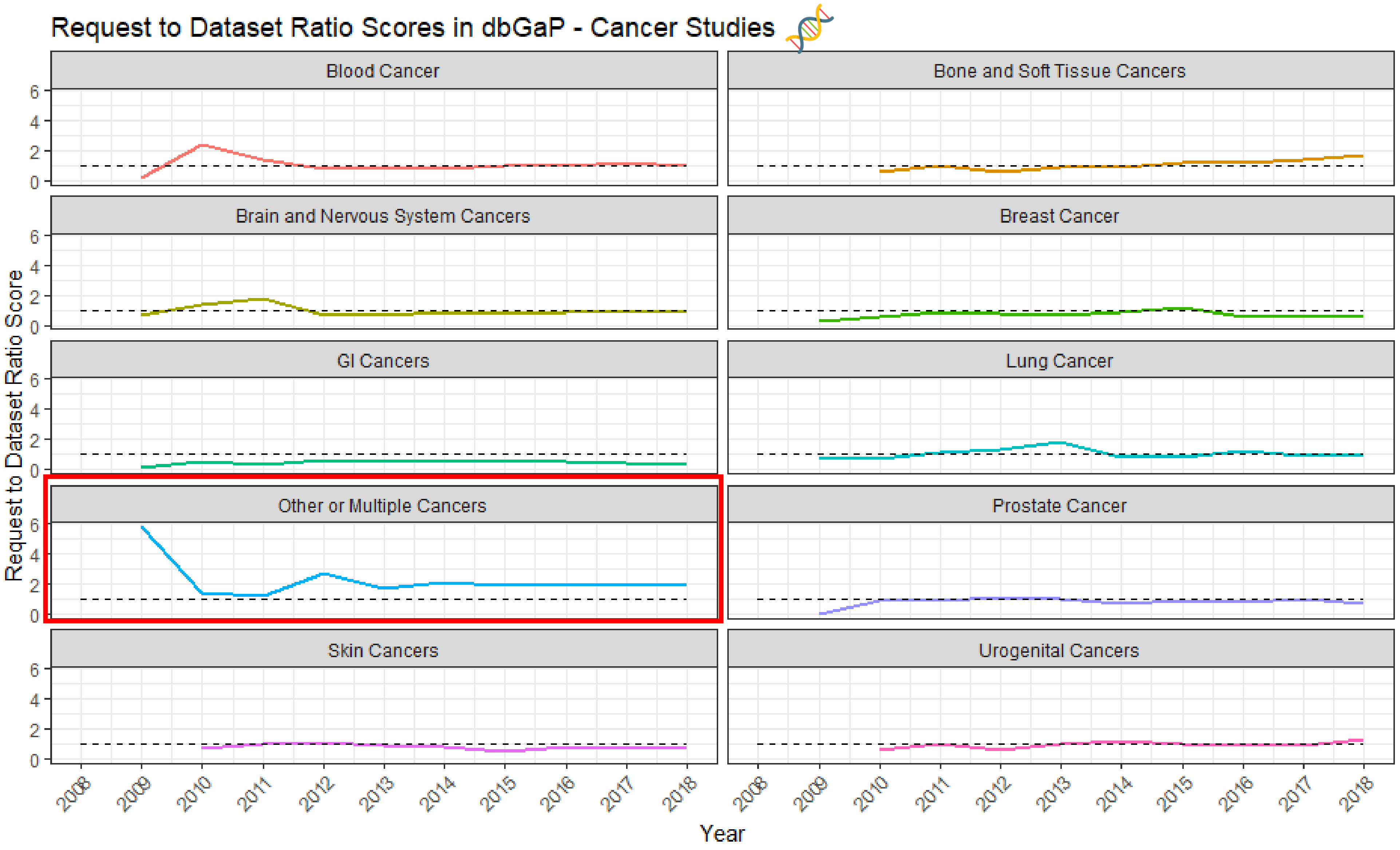
Topic B
2 datasets
122 requests

Repository Total
6 datasets
192 requests

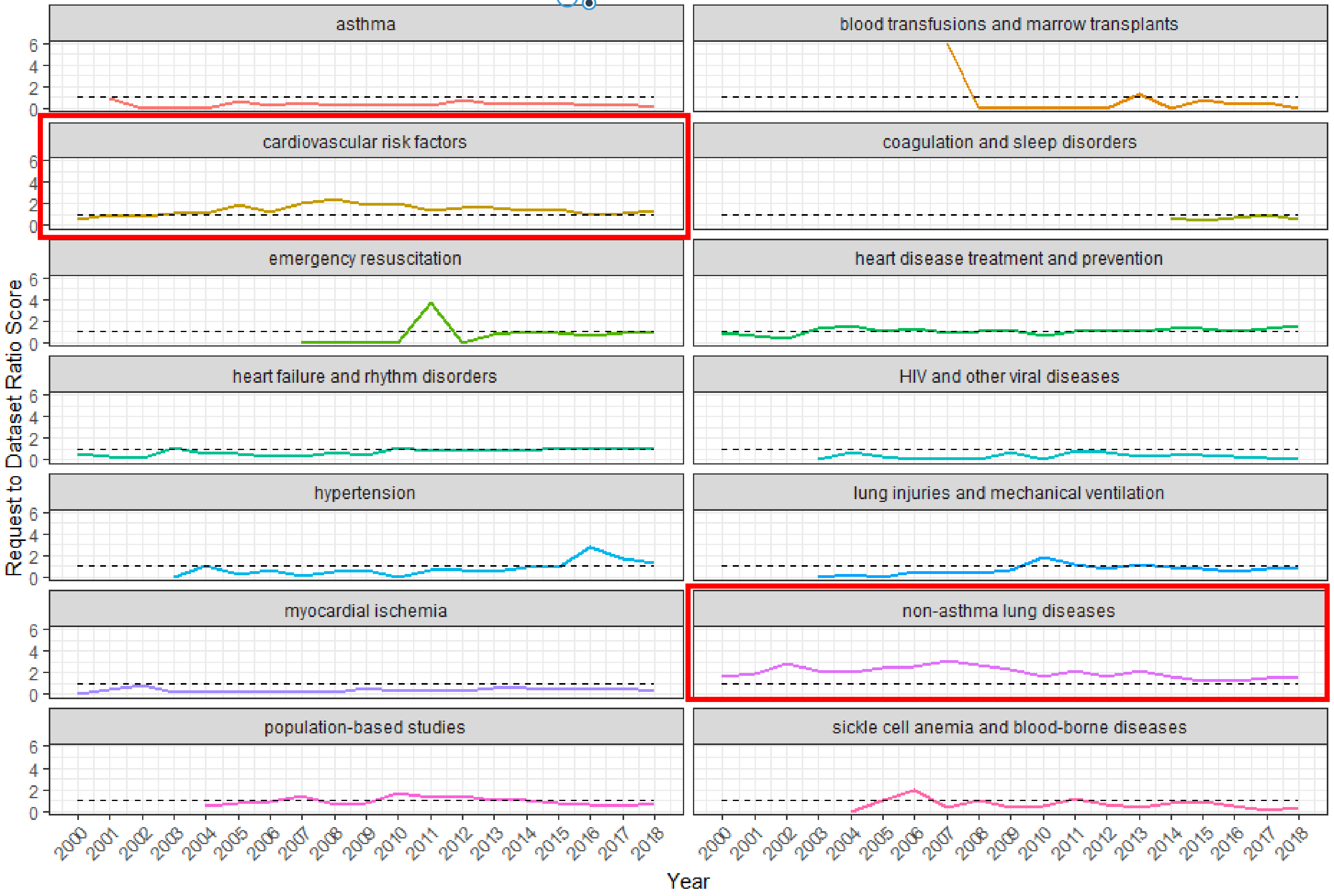
$$\text{RTD} = \frac{\text{proportion of requests in topic}}{\text{proportion of datasets in topic}}$$

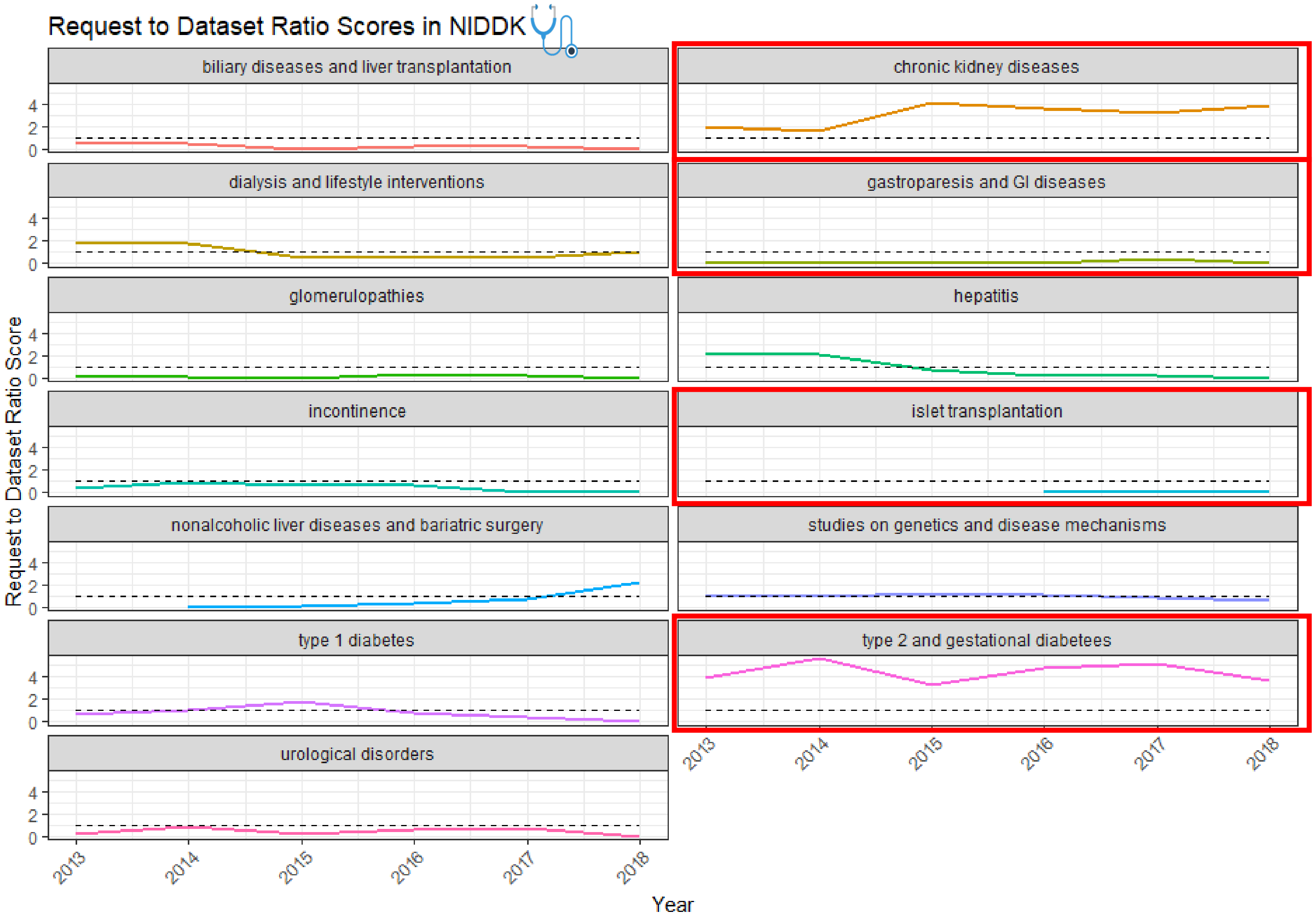
$$\text{RTD Topic A} = \frac{\dfrac{\text{70 requests in Topic A}}{\text{192 requests total}}}{\dfrac{\text{4 datasets in Topic A}}{\text{6 datasets total}}} = \frac{0.36}{0.67} = 0.54$$


Request to Dataset Ratio Scores in dbGaP
Blood and Cardiovascular
Cancer
Congenital Disorders
Eye Disorders
GI Disorders
Immune and Autoimmune Disorders
Infectious Disease
Lung and Respiratory Disorders
Mental Disorders
Metabolic Diseases
Musculoskeletal
Neurological
Normal
Other
Skin Disorders
Substance Use Disorders
Urogenital Disorders
Women's Health and Pregnancy
0.0
0.5
1.0
1.5
2.0
Request to Dataset Ratio Score
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year


Request to Dataset Ratio Scores in dbGaP - Cancer Studies
Blood Cancer
Bone and Soft Tissue Cancers
Brain and Nervous System Cancers
Breast Cancer
GI Cancers
Lung Cancer
Other or Multiple Cancers
Prostate Cancer
Skin Cancers
Urogenital Cancers
Request to Dataset Ratio Score
0
2
4
6
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

Request to Dataset Ratio Scores in NHLBI




Request to Dataset Ratio Scores in NIDDK
biliary diseases and liver transplantation
chronic kidney diseases
dialysis and lifestyle interventions
gastroparesis and GI diseases
glomerulopathies
hepatitis
incontinence
islet transplantation
nonalcoholic liver diseases and bariatric surgery
studies on genetics and disease mechanisms
type 1 diabetes
type 2 and gestational diabetees
urological disorders
Request to Dataset Ratio Score
0
2
4
2013
2014
2015
2016
2017
2018
Year

Implications

For researchers: sharing concerns may be unfounded

Getting “scooped” may not be a significant threat

Replication to refute results is not a major reuse of these datasets

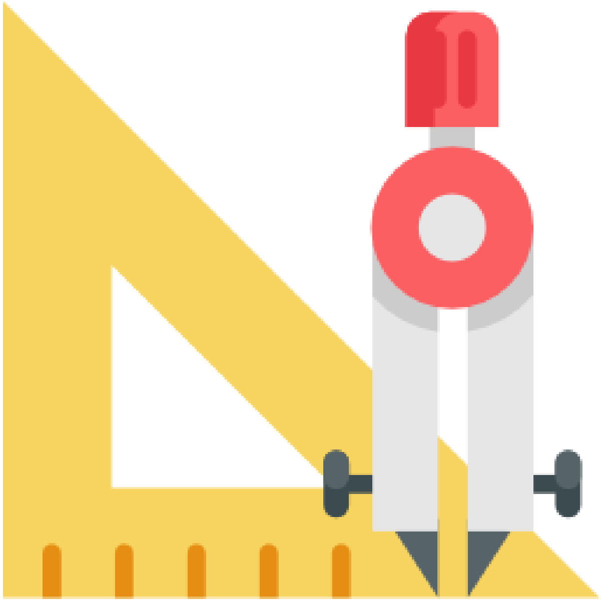
For repositories: evidence for preservation and curation decisions

Early requests for datasets are a predictor for long-term reuse

Certain topics may be expected to be more reused than others

For funders and institutions

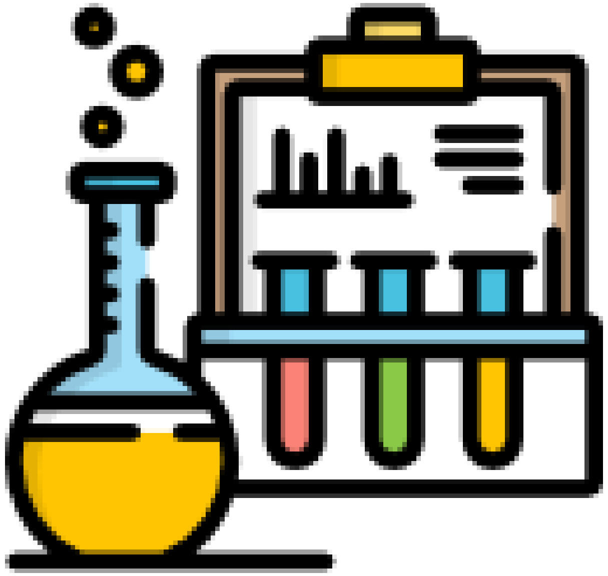
Datasets are reused in many ways – should creators be rewarded equally for all of them?



Need to carefully define metrics to avoid pitfalls such as those experienced in bibliometrics

Limitations

Unclear how closely requests track to actual reuse of datasets



Limited generalizability beyond biomedical repositories

NLM Office of Strategic Initiatives

Data Science & Open Science Team

Michael Huerta, PhD

Director

Lisa Federer, PhD, MLIS

Data Science & Open Science Librarian

Teresa Zayas-Caban, PhD

Coordinator, NIH FHIR Acceleration

Chief Scientist, ONC, DHHS

Rebecca Goodwin, JD

Policy Analyst & Open Science Specialist

Maryam Zaringhalam, PhD

Data Science & Open Science Specialist

Tony Chu, PhD, MLIS

Information Scientist

Questions?

Lisa Federer, PhD, MLIS
Lisa.Federer@nih.gov
@lisafederer

Image source: free photobank torange.biz (CC-BY)