End of fellowship presentation - May 19th, 2017

# Investigating the Coverage of Disease Concepts across SNOMED CT and the Disease Ontology… …and other projects

**Satyajeet Raje, PhD**

Lister Hill National Center for Biomedical Communications

U.S. National Library of Medicine




# Agenda

- Specific Project: Coverage of disease concepts across SNOMED CT and DO
  - Introduction
  - Methods and Results
  - Discussion
  - Conclusion – Practical Contribution
- Shout-outs to other projects
  - Leveraging lexical features of concept names using DL – Application to QA in SNOMED CT
  - Harmonizing user defined phenotypic variables across studies in dbGaP
  - MeSHgram: A tool for visual browsing of PubMed
  - A toolkit to extend TCGA queries to fetch analogous sequence data from NCBI datasets

# Investigating coverage of disease concepts in SNOMED CT and DO

# Introduction

## SNOMED CT

- Largest clinical terminology (over 300,000 concepts)
- About 100,000 concepts in “Clinical findings” hierarchy
- Integrated in UMLS
- *March 2016 US edition used in this study. Converted to OWL using script provided.*

## Disease Ontology (DO)

- Part of OBO. Widely used in the research domain.
- 6931 *active* disease concepts
- ***Not integrated in UMLS.*** Some concepts mapped to SNOMED CT via “obo:hasDbXref”.
- *August 2016 release used in this study. Available in OWL format.*

- ***Interoperability is critical for translational research***

# *Methods* - Overview

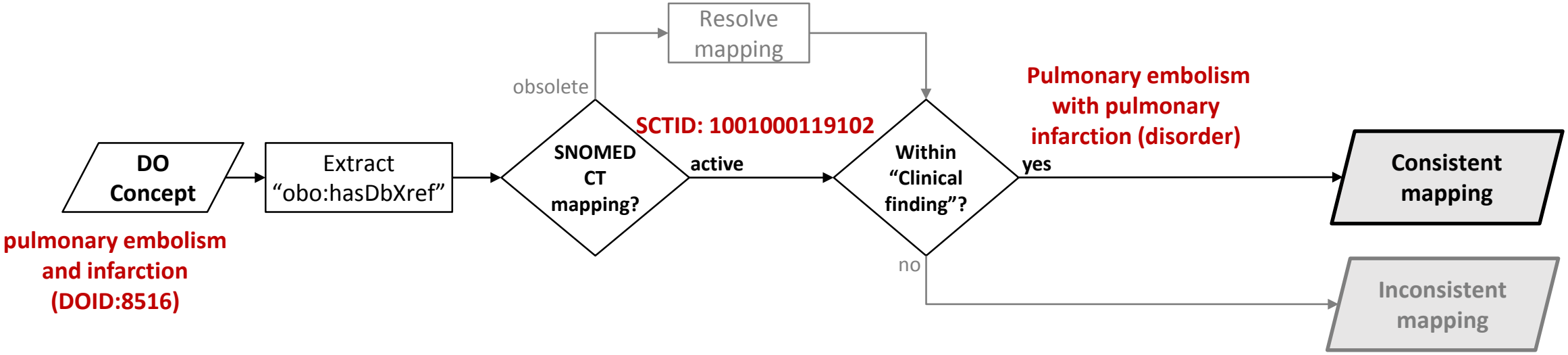
- Establishing a reference set of mappings
  - Apply semantic constraints on existing mappings from DO to SNOMED CT
  - Find additional mappings lexically
- Characterizing DO concepts not mapped to SNOMED CT
  - Distribution of mapped vs. unmapped concepts by top-level hierarchies in DO
  - Analysis of connected components of unmapped concepts
  - Manual review of semantic “differentia” for unmapped concepts
- Comparing the hierarchical organization based on mapped concepts

# *Methods* - Overview

- Establishing a reference set of mappings
  - Apply semantic constraints on existing mappings from DO to SNOMED CT
  - Find additional mappings lexically
- Characterizing DO concepts not mapped to SNOMED CT
  - Distribution of mapped vs. unmapped concepts by top-level hierarchies in DO
  - Analysis of connected components of unmapped concepts
  - Manual review of semantic “differentia” for unmapped concepts
- Comparing the hierarchical organization based on mapped concepts

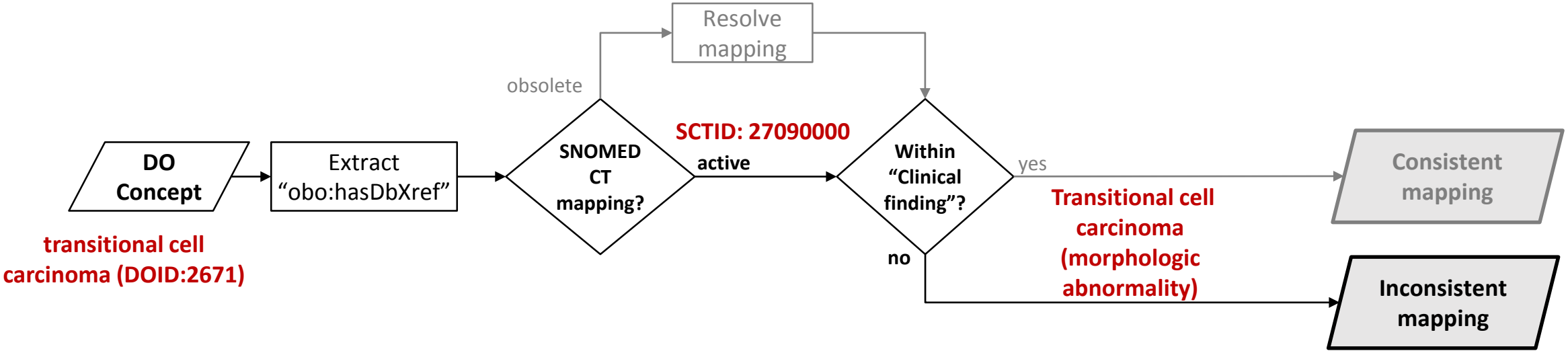
# *Methods* - Establishing a reference list of mappings (Semantic Constraint)

- Some DO concepts have “mappings” to SNOMED CT through “obo:hasDbXref”
- *Semantic constraint* - Mappings outside “Clinical Findings” hierarchy in SNOMED CT are considered semantically inconsistent



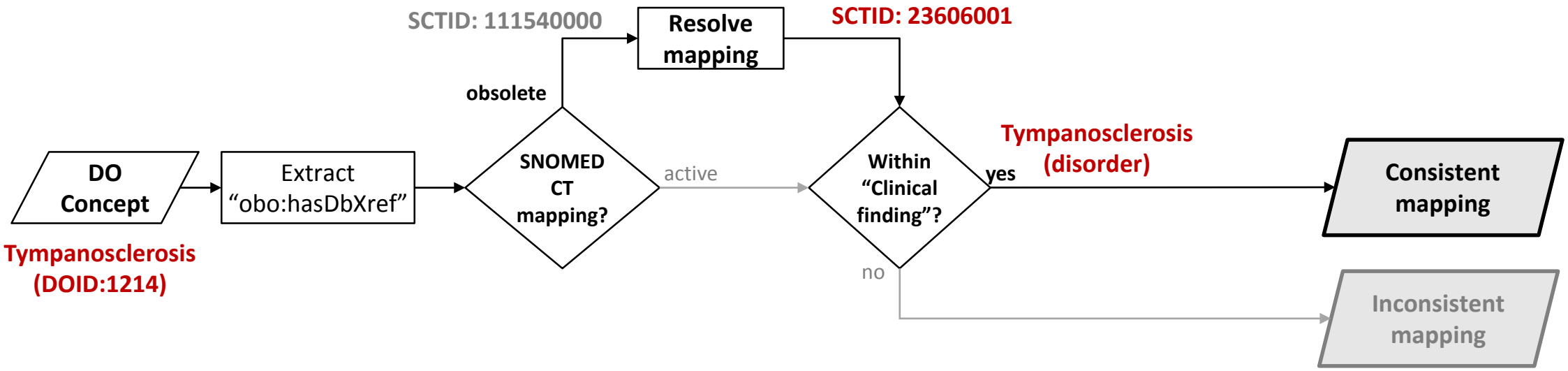
# *Methods* - Establishing a reference list of mappings (Semantic Constraint)

- Some DO concepts have “mappings” to SNOMED CT through “obo:hasDbXref”
- *Semantic constraint* - Mappings outside “Clinical Findings” hierarchy in SNOMED CT are considered semantically inconsistent



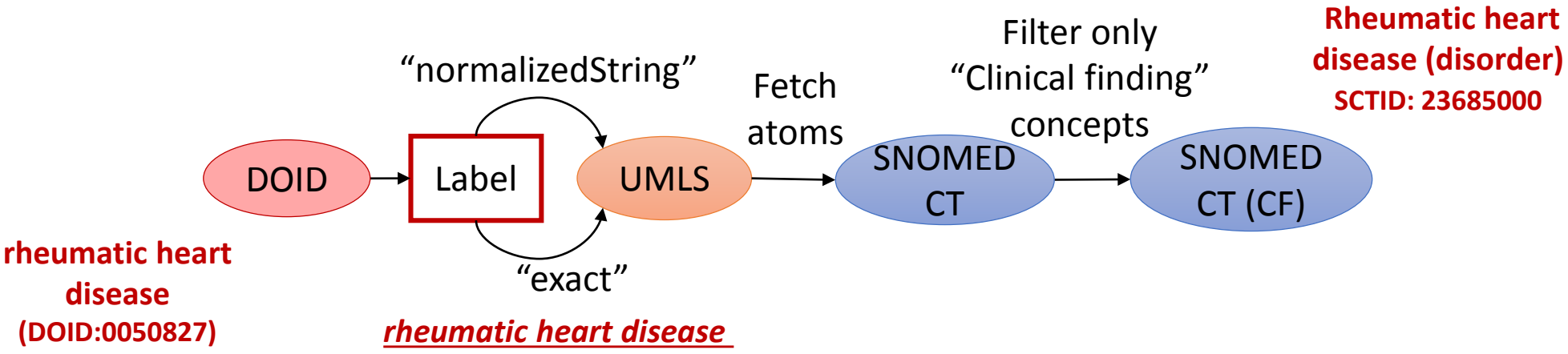
# *Methods* - Establishing a reference list of mappings (Semantic Constraint)

- Some DO concepts have “mappings” to SNOMED CT through “obo:hasDbXref”
- *Semantic constraint* - Mappings outside “Clinical Findings” hierarchy in SNOMED CT are considered semantically inconsistent
- Resolve mappings to obsolete SNOMED CT concepts

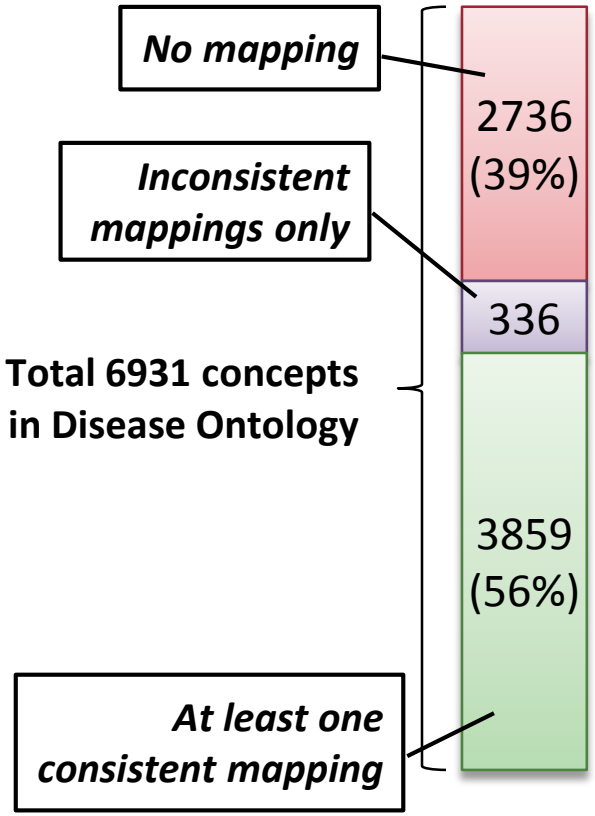


# *Methods* - Establishing a reference list of mappings (Lexical mappings)

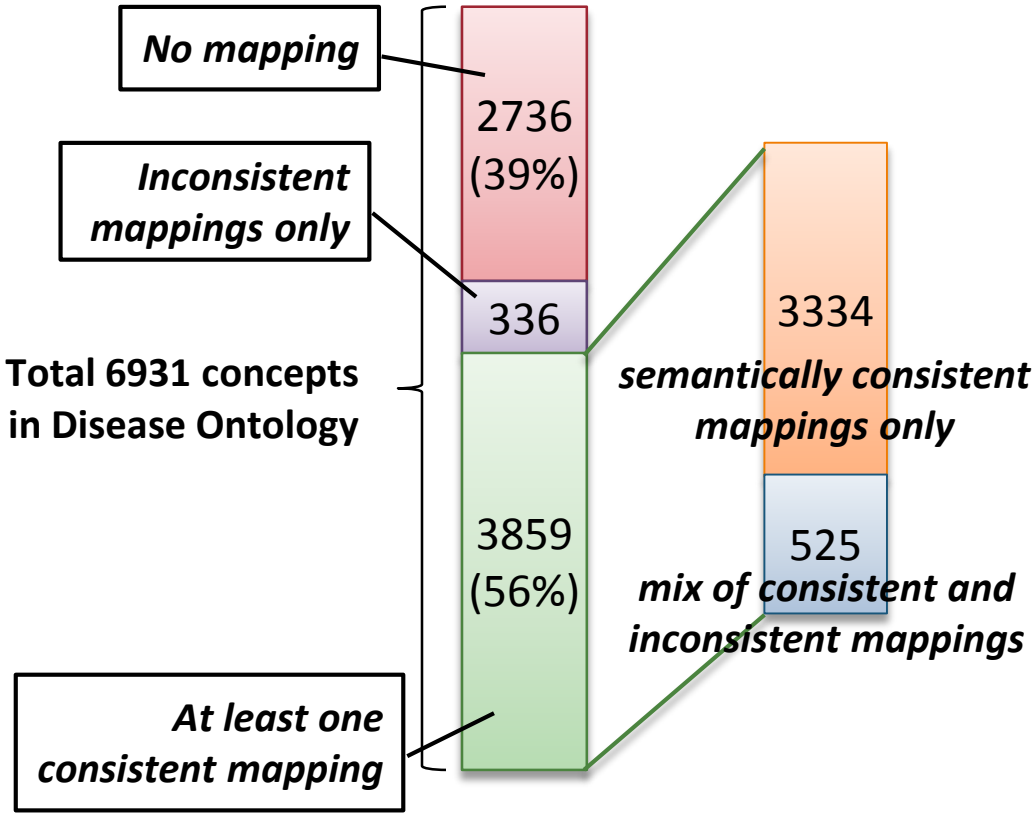
- Find additional mappings to SNOMED CT via UMLS concepts lexically
- Leveraging synonymy within UMLS
- UMLS Rest API for string match based on label of DO concept



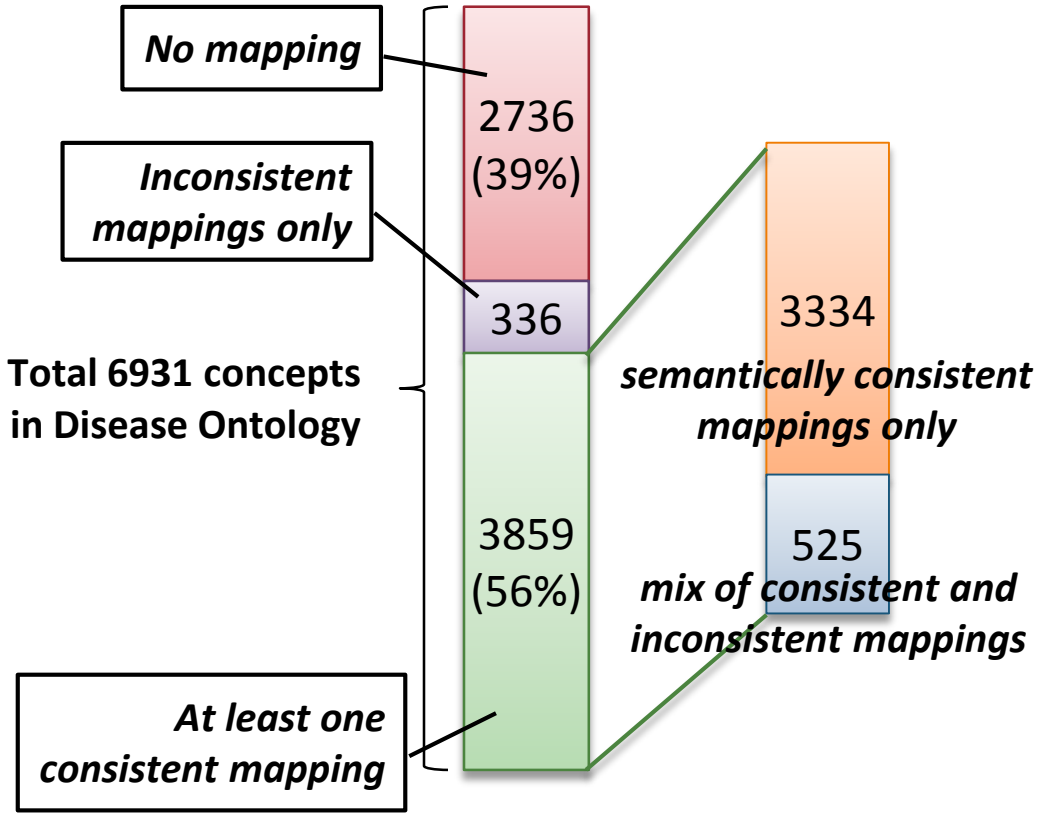
# *Results* - Establishing a reference list of mappings (Semantic Constraint)



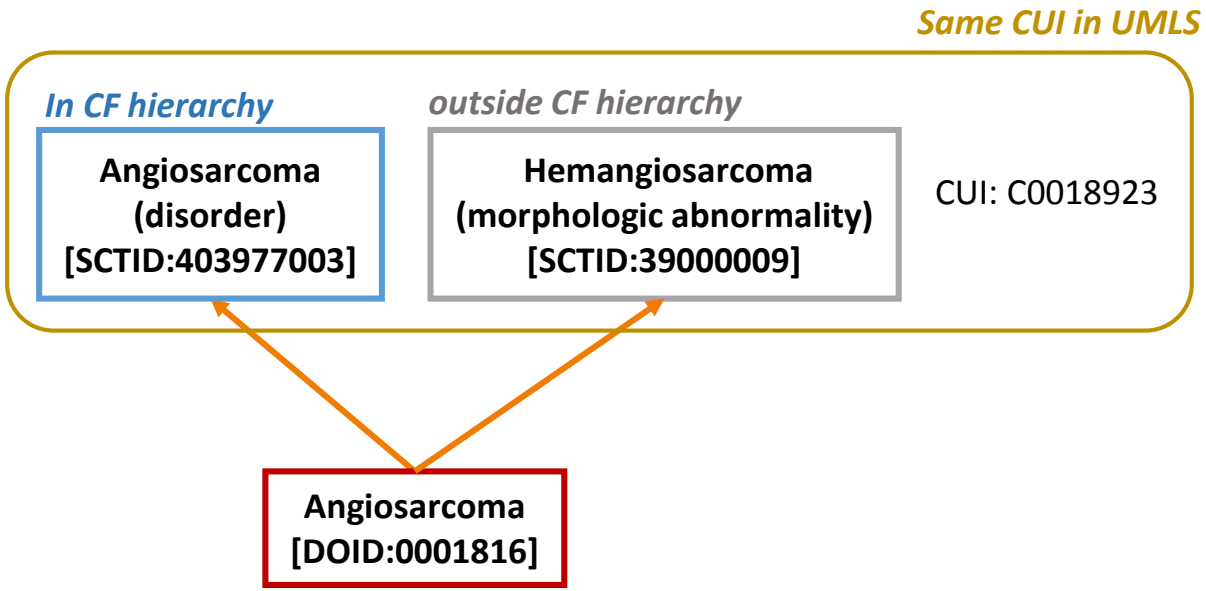
# *Results* - Establishing a reference list of mappings (Semantic Constraint)



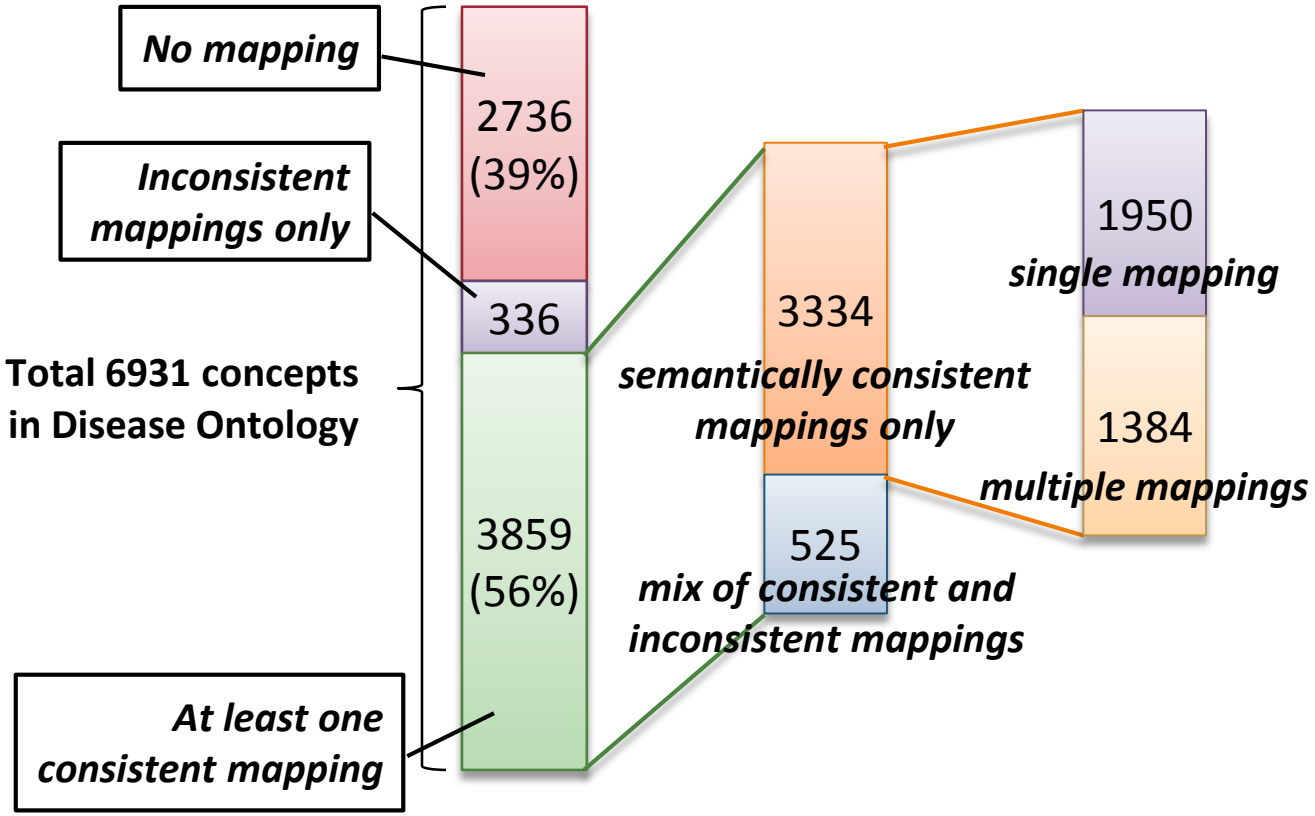
# *Results* - Establishing a reference list of mappings (Semantic Constraint)



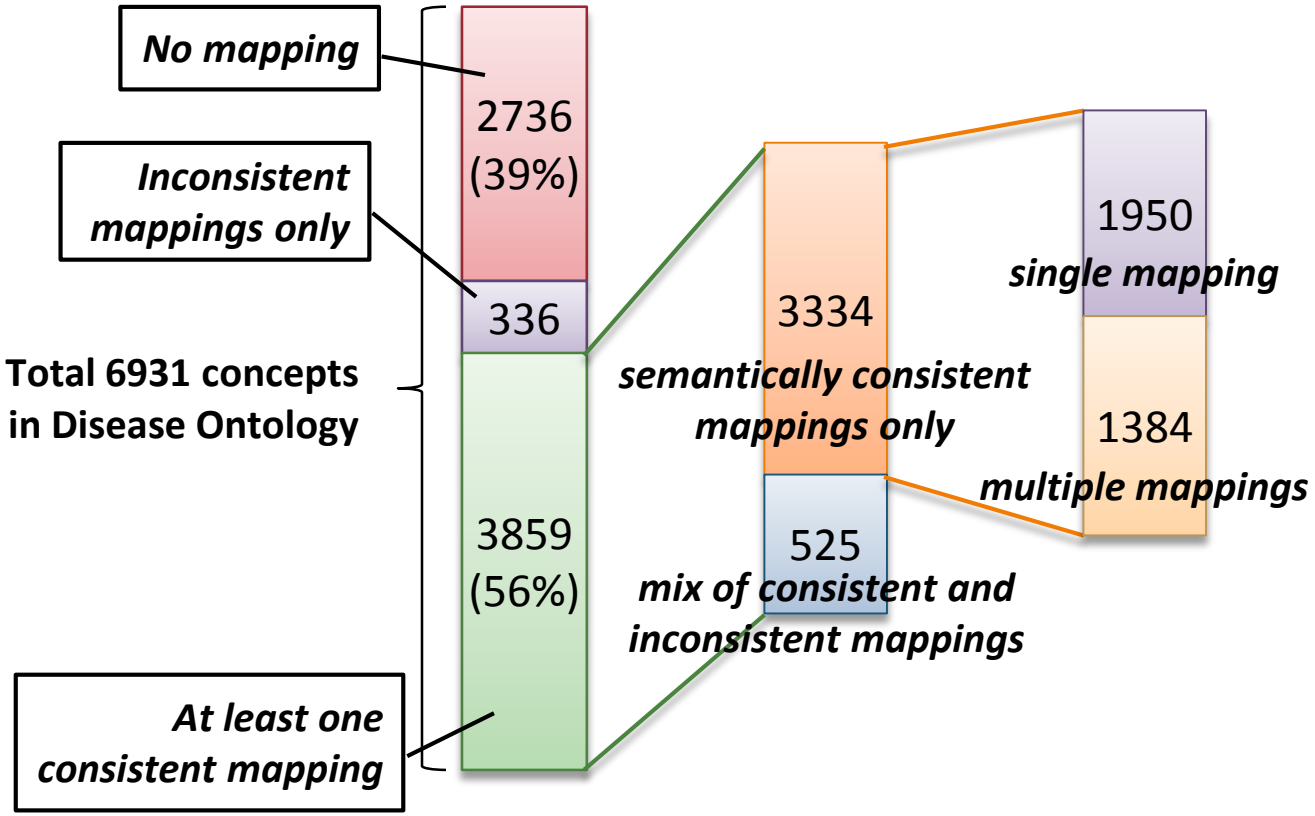

- 525 Concepts mapped with consistent and inconsistent mappings
- All of them were identified to be mapped to disorder and associated morphology
- Most had same UMLS CUI



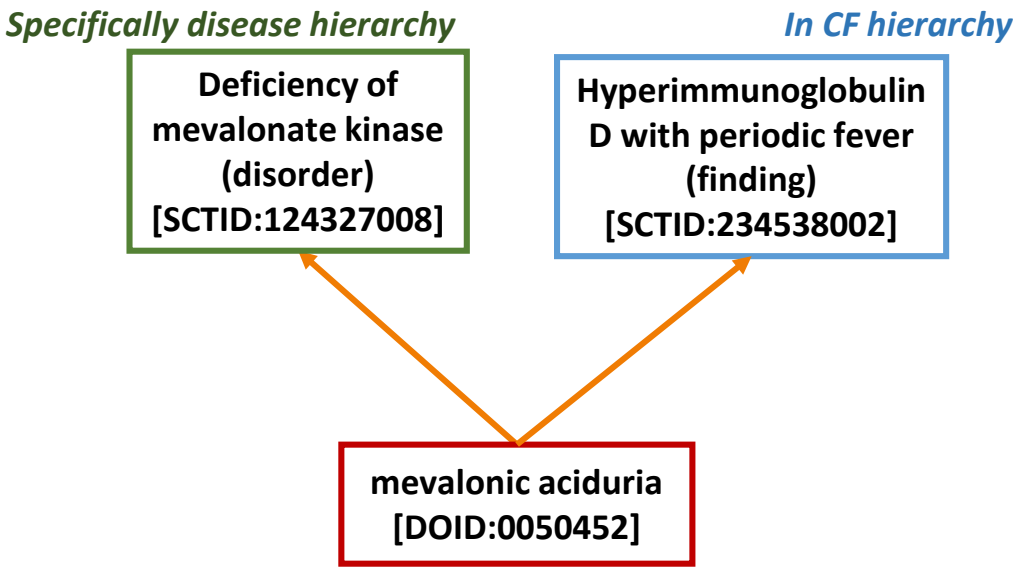
# *Results* - Establishing a reference list of mappings (Semantic Constraint)



# *Results* - Establishing a reference list of mappings (Semantic Constraint)




- 1384 concepts with multiple mappings that are all consistent
- 110 of these are mapped to disorder and associated finding
- 1274 concepts with multiple mappings within same hierarchy



# *Results* - Establishing a reference list of mappings (Lexical Mappings)

- 3072 DO concepts (44% of total)
  - 336 concepts with only inconsistent mappings
  - 2736 with no mapping
- Found mappings for **619 additional concepts**

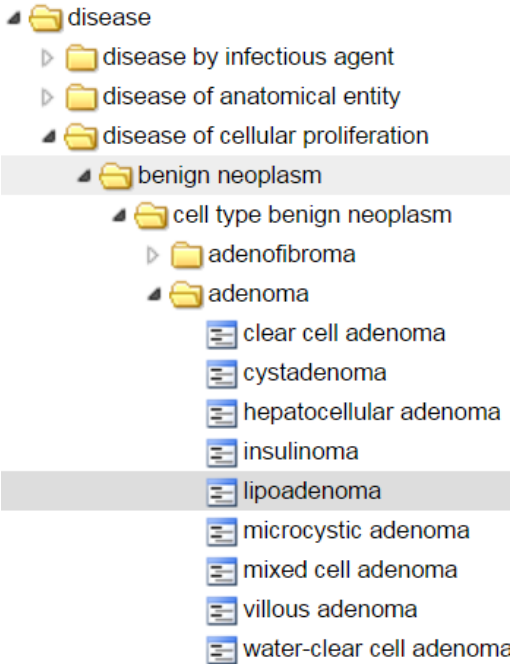
- ***2453 (35%) DO concepts remain unmapped***

# *Methods* - Overview

- Establishing a reference set of mappings
  - Apply semantic constraints on existing mappings from DO to SNOMED CT
  - Find additional mappings lexically
- Characterizing DO concepts not mapped to SNOMED CT
  - Distribution of mapped vs. unmapped concepts by top-level hierarchies in DO
  - Analysis of connected components of unmapped concepts
  - Manual review of semantic “differentia” for unmapped concepts
- Comparing the hierarchical organization based on mapped concepts

# *Methods* - Characterizing unmapped concepts (Distribution)

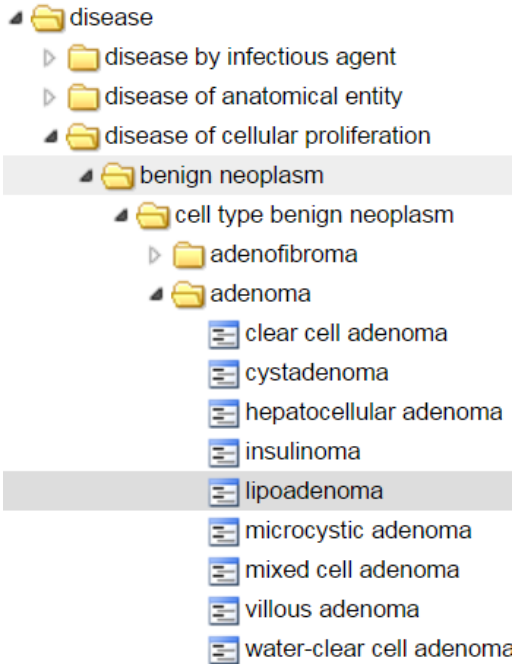
- Find the top level ancestor for each concept
- Plot the distribution of mapped versus unmapped concepts at the top level hierarchy in DO

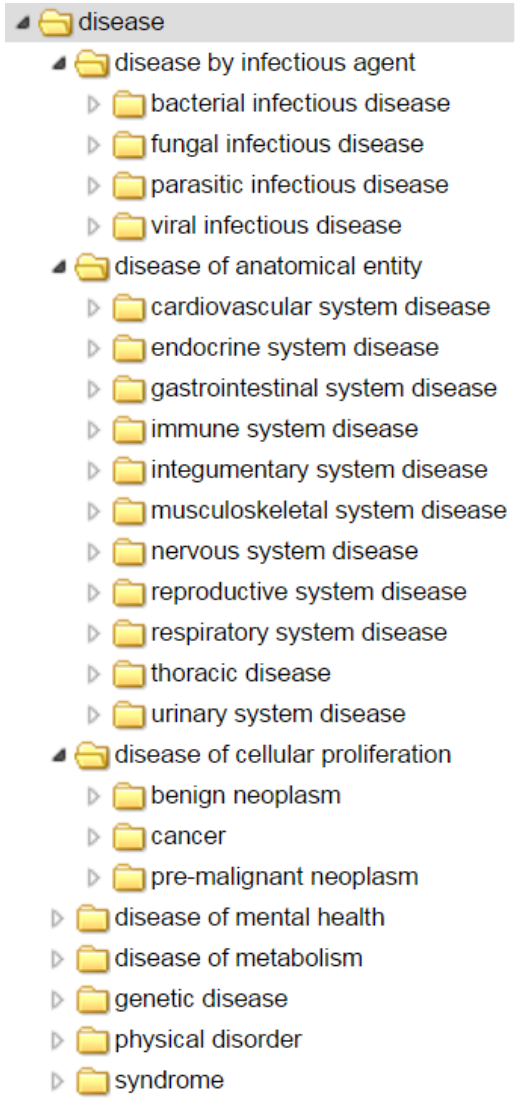



*Source:* disease-ontology.org

# *Methods* - Characterizing unmapped concepts (Distribution)

- Find the top level ancestor for each concept
- Plot the distribution of mapped versus unmapped concepts at the top level hierarchy in DO

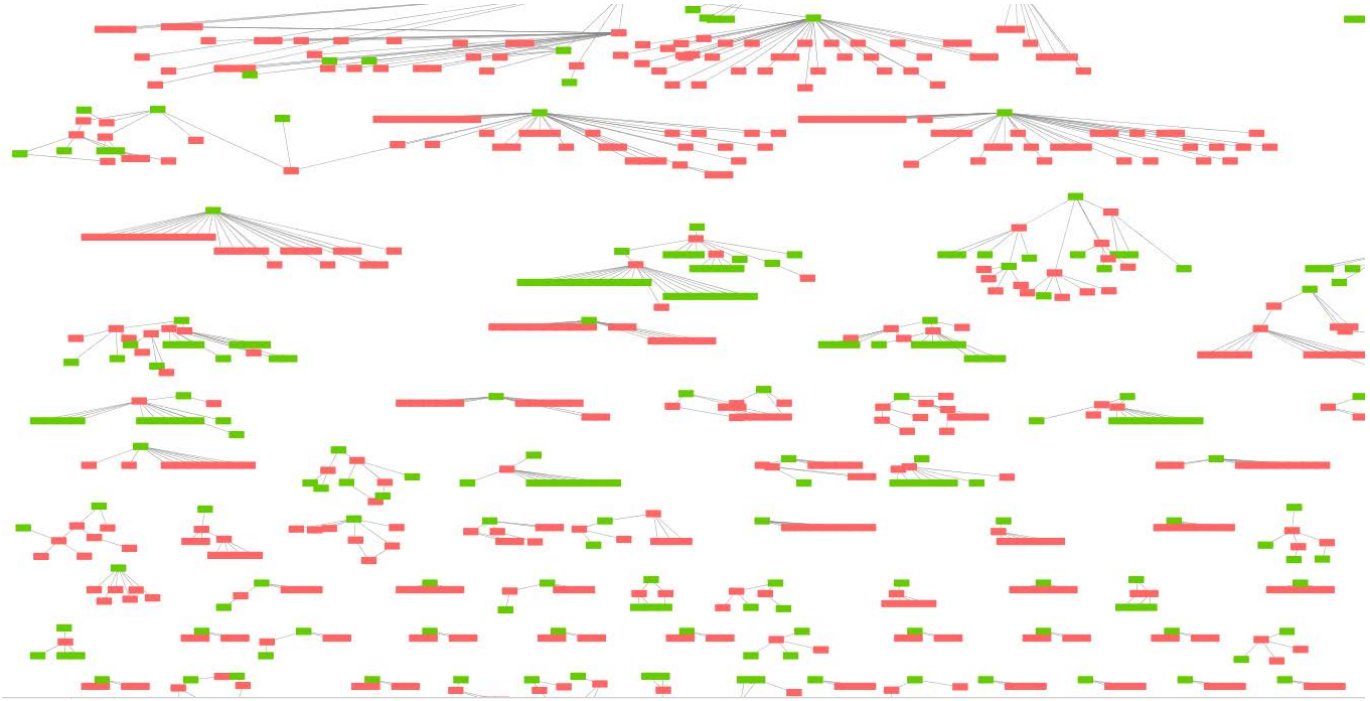



*Source*: disease-ontology.org

# *Methods* - Characterizing unmapped concepts (Structural)

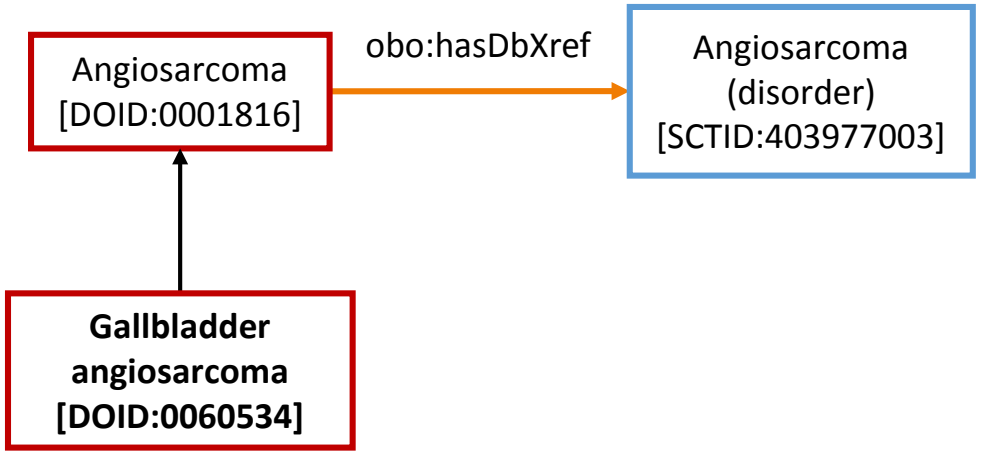
- Analyzing connected components on unmapped concepts
- Create graph of unmapped concepts with immediate (mapped or unmapped) parents and children
- Identify patterns among connected components generated



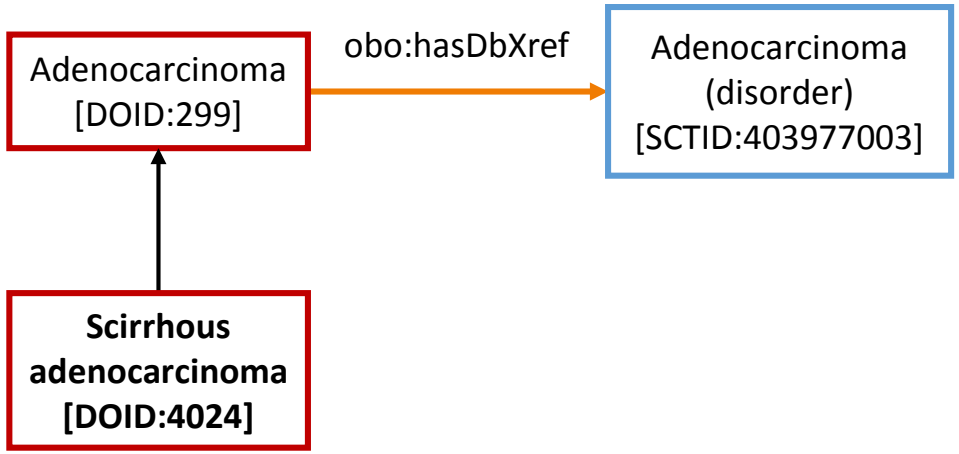
Visualization in Cytoscape

# *Methods* - Characterizing unmapped concepts (Semantic)

- Understand why some DO concepts are unmapped
- Applied to completely unmapped sub-hierarchies and leaves only
- Manually review the semantic "differentia" in comparison with parents

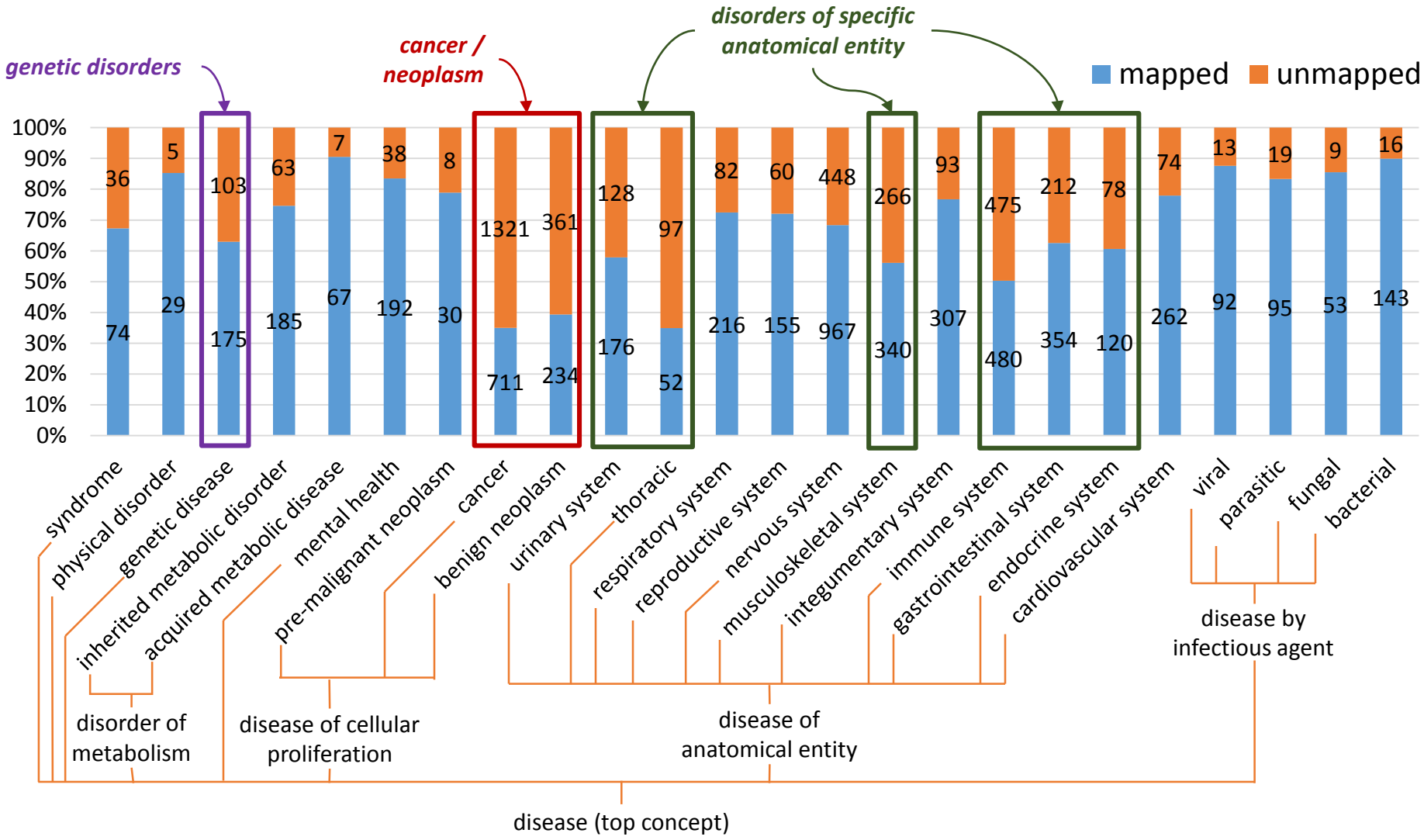



anatomic site



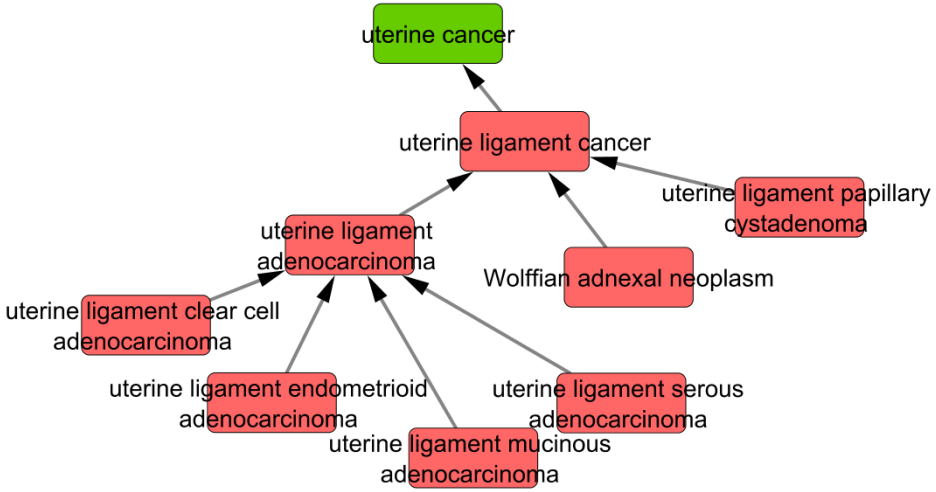

specific morphology

# *Results* - Characterizing unmapped concepts (Distribution)

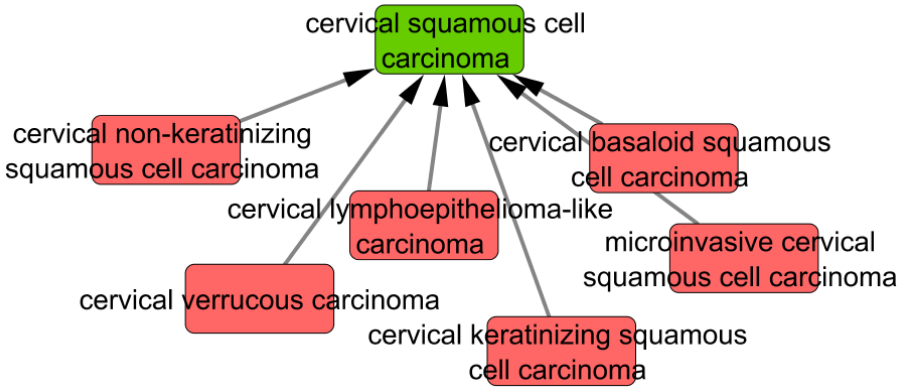


# *Results* - Characterizing unmapped concepts (Structural)

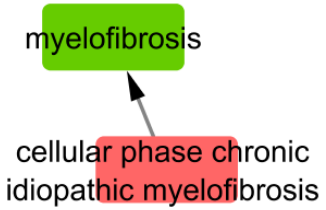
- Patterns among connected components



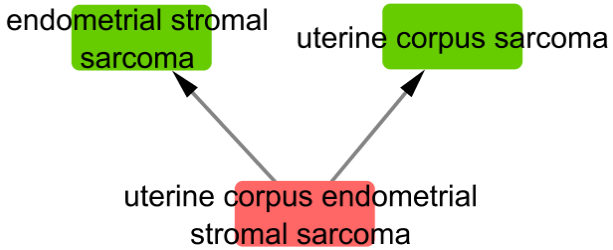

743 concepts (30.3%) within entirely unmapped subhierarchies of a mapped parent




1064 (43.4%) unmapped leaves of a single parent

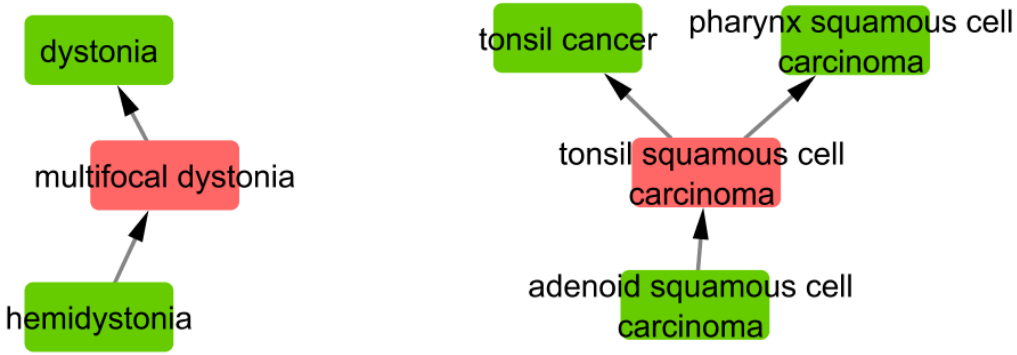



214 (9%) single parents




187 (8%) multiple parents

Single unmapped leaves




246 unmapped intermediary concepts

# *Results* - Characterizing unmapped concepts (Semantic)

- 2207 concepts - sub-hierarchies and leaves only (90% of 2453 unmapped)

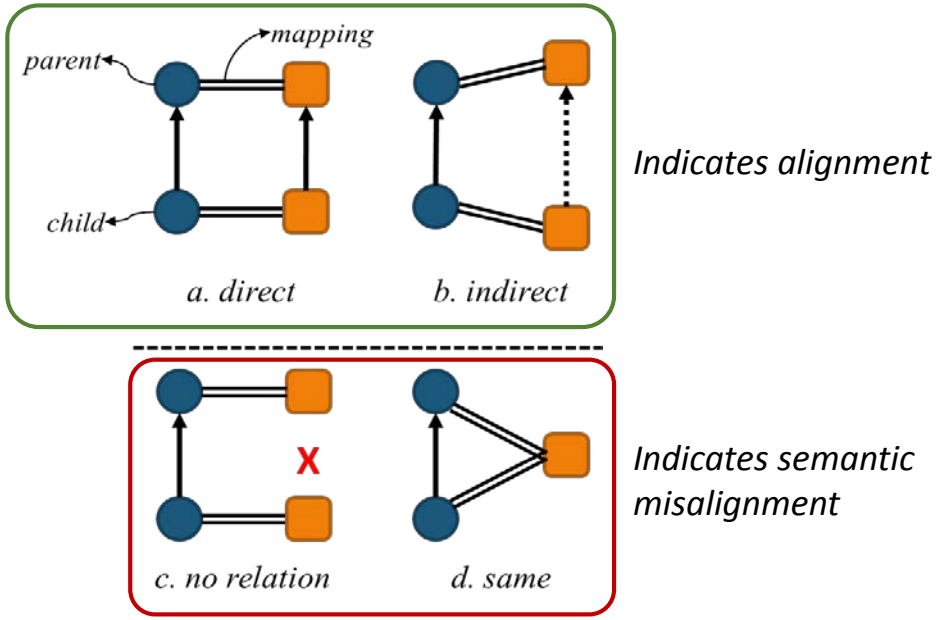
| Type of semantic granularity | Count | % |
|---|---|---|
| **specific morphology** (e.g. ***follicular*** dendritic cell sarcoma) | 831 | 37.65 |
| **morphology and anatomic site** | 520 | 23.56 |
| **specific subtype X** (e.g. spinocerebellar ataxia ***type 1***) | 253 | 11.46 |
| **anatomic site** (e.g. ***intramuscular*** hemangioma) | 147 | 6.66 |
| **morphology and period of onset** | 61 | 2.76 |
| **period of onset** (e.g. ***pediatric*** osteosarcoma) | 45 | 2.04 |
| **chromosomal location and anomaly** | 45 | 2.04 |
| **complex syndrome** (e.g. agnathia-otocephaly complex) | 42 | 1.90 |
| **other generic subtypes** | 42 | 1.90 |
| **organism** (e.g. ***screw worm*** infectious disease) | 30 | 1.36 |
| ***others*** | *191* | *8.65* |
| Total | 2207 | |

# *Methods* - Overview

- Establishing a reference set of mappings
  - Apply semantic constraints on existing mappings from DO to SNOMED CT
  - Find additional mappings lexically
- Characterizing DO concepts not mapped to SNOMED CT
  - Distribution of mapped vs. unmapped concepts by top-level hierarchies in DO
  - Analysis of connected components of unmapped concepts
  - Manual review of semantic “differentia” for unmapped concepts
- Comparing the hierarchical organization based on mapped concepts

# *Methods* - Comparing hierarchical organization

- Identify differences in the hierarchical organization of SNOMED CT and DO based on the mapped concepts
- Compare the parent-child pairs with relation between corresponding mapped concepts
- Applied in both directions (i.e. from DO to SNOMED CT and vice versa)

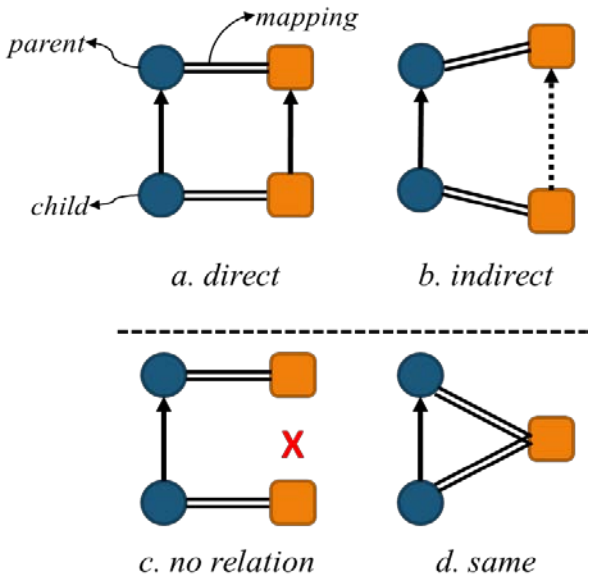



*Indicates alignment*

*Indicates semantic misalignment*

# *Results* - Comparing hierarchical organization

- There is a lot of “confusion” regarding the hierarchical organization of concepts
- Multiple mappings are largely responsible




| Mapping direction | Type (a) or (b) only | Type (c) or (d) only | (a or b) & (c or d) | Total pairs |
|---|---|---|---|---|
| DO to SNCT | 1198 [28%] | 1075 [25%] | 1978 [48%] | 4233 |
| SNCT to DO | 1842 [32%] | 2792 [48%] | 1138 [20%] | 5772 |

# Discussion

- DO has 2453 potentially “new” concepts for UMLS
  - Adding some semantic differentia
- Pre- vs. Post-coordination of concepts
  - Responsible for some of the concepts being unmapped in DO
- Multiple mappings are not good for interoperability
  - Further work needed to formulate rules to resolve such mappings
- Characterizing concepts based on hierarchical relation between the two ontologies
  - Further work needed to identify specific causes for semantic misalignment

***Limitation***

- Inclusion in the “Clinical Findings” hierarchy was the only validation criteria for existing mappings

# Conclusion – Practical Contribution

- When using both DO and SNOMED CT
  - “Better” set of mappings
    - Removal of invalid mappings
    - Additional mappings through lexical match
- Choosing between DO and SNOMED CT
  - Characterization of specific content in DO
    - Semantic categorization
    - Hierarchical organization

# Publications/presentations

- **Raje S, Bodenreider O.** Investigating the coverage of diseases across biomedical research and clinical ontologies. [abstract]. *Proceedings of the AMIA Joint Summits on Translational Science 2017*

- **Raje S, Bodenreider O.** Interoperability of Disease Concepts in Clinical and Research Ontologies – Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT. *Stud Health Technol Inform (Proc Medinfo) 2017 (selected)*

# Other projects

# List of other projects

- Leveraging Lexical Features of Concept Terms for Quality Assurance in SNOMED CT using Description Logics
- Harmonizing User-defined Phenotypic Variables using Latent Semantic Analysis (LSA) to Improve Data Discoverability in dbGaP

*NCBI Hackathons:*

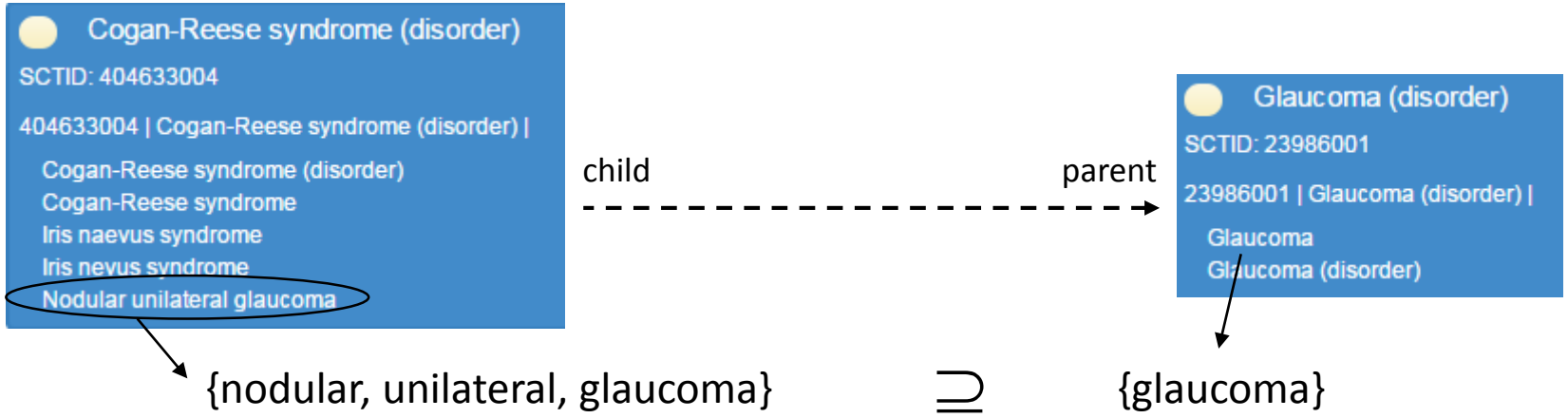
- MeSHgram: An Open Source Tool to Visually Browse Co-occurrence of MeSH Terms in PubMed
- A search tool to extend TCGA queries to automatically identify analogous genomic data from dbGaP

# Leveraging Lexical Features using Description Logics

- Identify potential missing hierarchical relation in SNOMED CT

**Bodenreider O.** Identifying missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. *Proc. of 6th Intl. Conf. on Bio. Ontology (ICBO 2016)*

- Lexical semantics




- Extended to include synonyms and lexical constraints based on shallow parsing
- Applied to *almost* all of SNOMED CT

# Leveraging Lexical Features using Description Logics

- Create logical definitions for concepts based on lexical features
- Represent using Description Logics (DL)
- Infer hierarchy “automatically” using ELK reasoner

Chronic pansinusitis (disorder)
SCTID: 88850006
88850006 | Chronic pansinusitis (disorder) |
Chronic pansinusitis
Chronic pansinusitis (disorder)

child → parent

Pansinusitis (disorder)
SCTID: 195790000
195790000 | Pansinusitis (disorder) |
Pansinusitis
Pansinusitis (disorder)

{chronic, pansinusitis} $\supseteq$ {pansinusitis}

Annotations
rdfs:label
Pansinusitis (disorder)

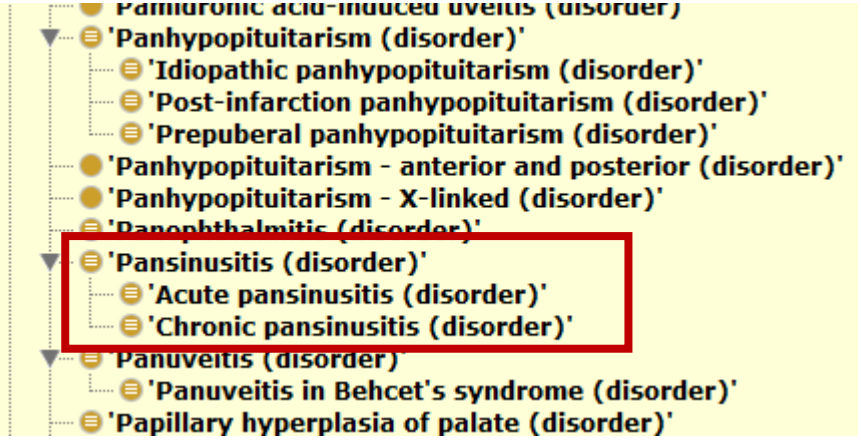
Description: 'Pansinusitis (disorder)'

Equivalent To
T_disorder
and (has_head some pansinusitis)

Annotations
rdfs:label
Chronic pansinusitis (disorder)

Description: 'Chronic pansinusitis (disorder)'

Equivalent To
T_disorder
and (has_head some pansinusitis)
and (has_mod some chronic)



# Leveraging Lexical Features using Description Logics

- Filter out relations present in original hierarchy of SNOMED CT (transitively closed)
- Remaining relations are identified as potentially missing

| SNCT ID | Label | FSN only | | | with synonyms | | |
|---|---|---|---|---|---|---|---|
| | | Total relation inferred | Not present in SNOMED CT | % | Total relation inferred | Not present in SNOMED CT | % |
| **123037004** | **Body Structure (body structure)** | 24835 | 8666 | 34.89 | 45316 | 15170 | 33.48 |
| **404684003** | **Clinical Finding (finding)** | 56698 | 10674 | 18.83 | 79306 | 18235 | 22.99 |
| **71388002** | **Procedure (procedure)** | 31818 | 6981 | 21.94 | 48873 | 13467 | 27.56 |

- Preliminary study completed for hierarchies “Disorder of head (disorder)” and “Operative procedure on head (procedure)”

**Raje S, Bodenreider O.** Identifying potentially missing hierarchical relations in SNOMED CT based on lexical features – Impact of synonyms and lexico-syntactic constraints. [abstract] *AMIA Annu Symp Proc 2017 (submitted)*

# Harmonizing user-defined phenotypic variables in dbGaP

- NCBI's Database of Genotypes and Phenotypes (dbGaP) – Largest collection of genomic and epigenomic data
- Over 220,000 variables in dbGaP (as per latest stats https://www.ncbi.nlm.nih.gov/projects/gap/summaries/cgi-bin/summary.cgi?)
- Group user-defined phenotypic variables based on the variable descriptions
  - "weight", "weight of subject", "weight of participant", "how much does the patient weigh?"
  - "participant gender", "gender", "gender: male, female", "male or female?", "sex"
  - "how many cigarettes per day did you average over all the time you smoked?", "cigarettes: average # smoked per day", "average number of cigarettes smoked per day"

- **Reused data from previous NCBI study**
  - 20635 total variables (all variables associated with the PhenX toolkit)
  - PhenX toolkit is "a catalog of recommended, standard measures of phenotypes"
  - Manually identified PhenX IDs and LOINC codes for each variable

# Harmonizing user-defined phenotypic variables in dbGaP

- Simple lexical similarity based methods (such as Hamming or Levenshtein distance) was not enough. (Good precision but very poor recall)
- Used Latent Semantic Analysis (LSA) to compute pairwise similarity score between all variables.

- Manually evaluated identified "duplicates" for 10% of variables (randomly chosen).
- 96% precision when all 3 evaluators mark valid (kappa = 0.07)
- Compared our groupings with groups based on existing LOINC and PhenX IDs
  - Confusion due to semantic versus lexical information
  - "sex" and "gender"; "age of participant" and "age of participant at first diagnosis"

**Amos L, Raje S, Kimura M, *et. al.*** Harmonizing User-defined Phenotypic Variables using Latent Semantic Analysis (LSA) to Improve Data Discoverability in dbGaP [poster]. *AMIA 2017 (submitted)*

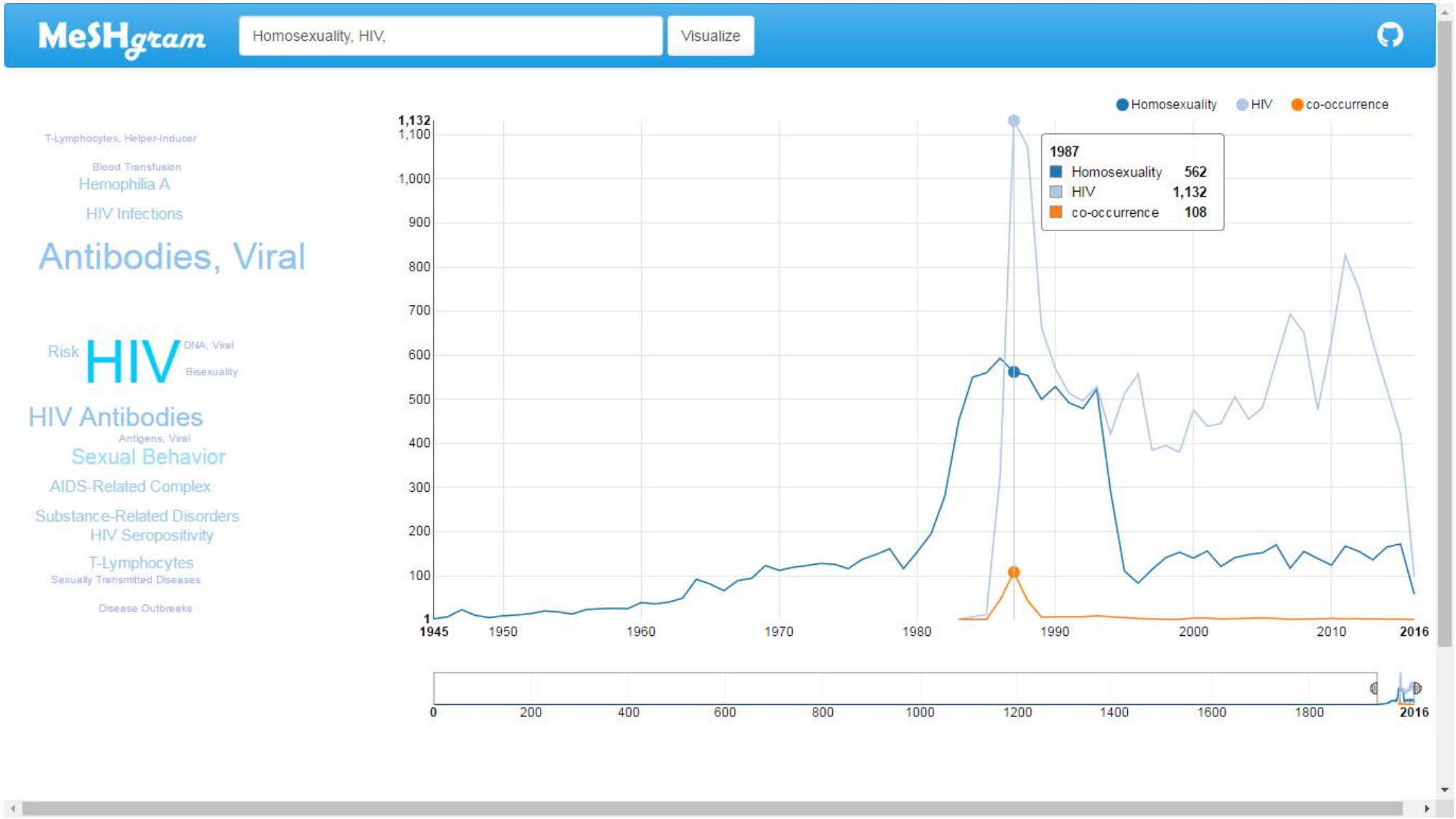
# MeSHgram

- A tool to visually browse co-occurrence of MeSH terms in PubMeb
- Use the MeSH terms associated with PubMed articles to visualize co-occurrence over time
  - Processed NLM PubMed corpus (approx. 24.5 mil. citations) from 1809 to 2016.
  - Extracted the MeSH terms for all citations
  - Ignore MeSH “stop words” when calculating co-occurence
- *Applications:*
  - Visual browsing / querying of PubMed
  - Support meta-analysis
  - Hypotheses generation, medical research activity, etc..

**Raje S, Bhupatiraju RT, Hosny A, Busby B.** Investigating the coverage of diseases across biomedical research and clinical ontologies. [poster] *AMIA Annu Symp Proc 2017 (submitted)*

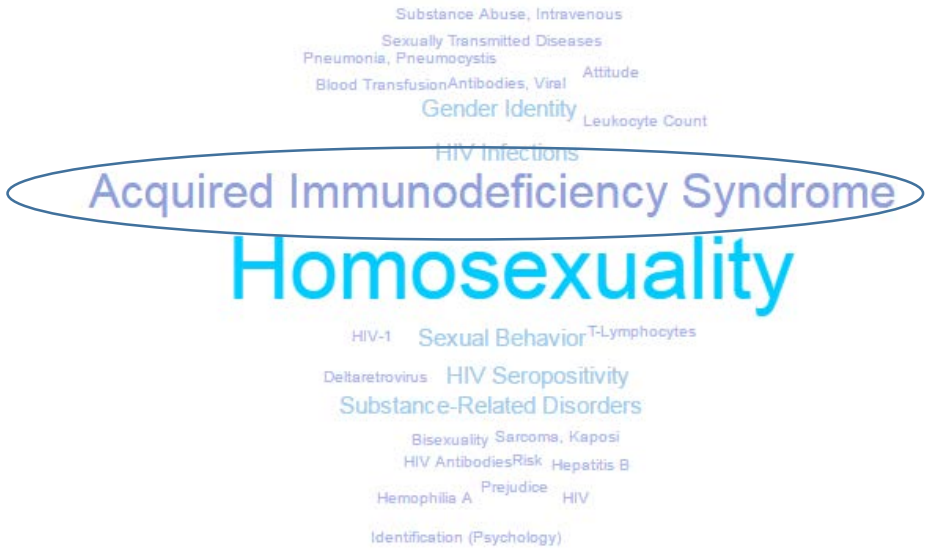
https://github.com/NCBI-Hackathons/Visualizing_MeSH_Term_Interaction_Over_Time

# MeSHgram – Case-study of HIV and Homosexuality



# MeSHgram – Homosexuality through the ages

1960-1970 – Psychology → 1980-1995 – Classical Pathology → 2000-2010 – Sociology


# Extending TCGA queries to fetch analogous data from NCBI databases

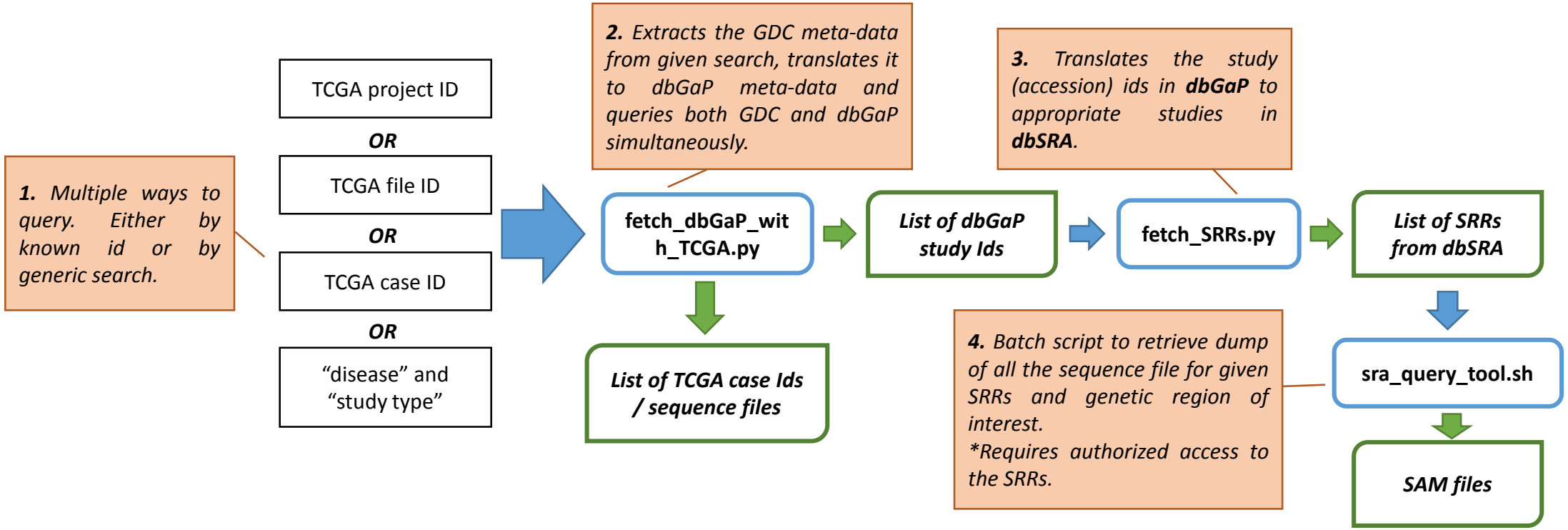
- TCGA: The Cancer Genome Atlas (now part of Genomic Data Commons) from NCI
  - Over 270,000 samples
- dbGaP: Database of Genotypes and Phenotypes
- SRA: Sequence Read Archive stores the sequence data
  - Over 2 million samples *(not all human)*

- Software toolkit to find analogous genomic data (sequence files) using TCGA queries

**Wagner EK, Raje S, Amos L *et al.*** Extending TCGA queries to automatically identify analogous genomic data from dbGaP. *F1000Research* 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)

https://github.com/NCBI-Hackathons/TCGA_dbGaP

# Extending TCGA queries to fetch analogous data from NCBI databases




**Wagner EK, Raje S, Amos L *et al.*** Extending TCGA queries to automatically identify analogous genomic data from dbGaP. *F1000Research* 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)

https://github.com/NCBI-Hackathons/TCGA_dbGaP

satyajeet.raje@gmail.com

Medical Ontology Research

## *Thank You*

*Dr. Olivier Bodenreider*

*Dr. Paul Fontelo*

*Cognitive Science Branch*

*Dr. Clem McDonald*

*& LHC*

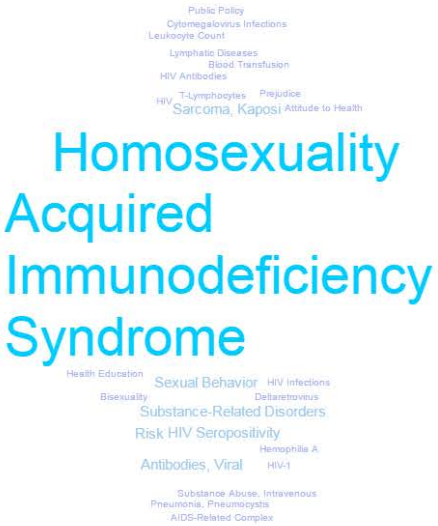
*Liz Amos, Library Operations*

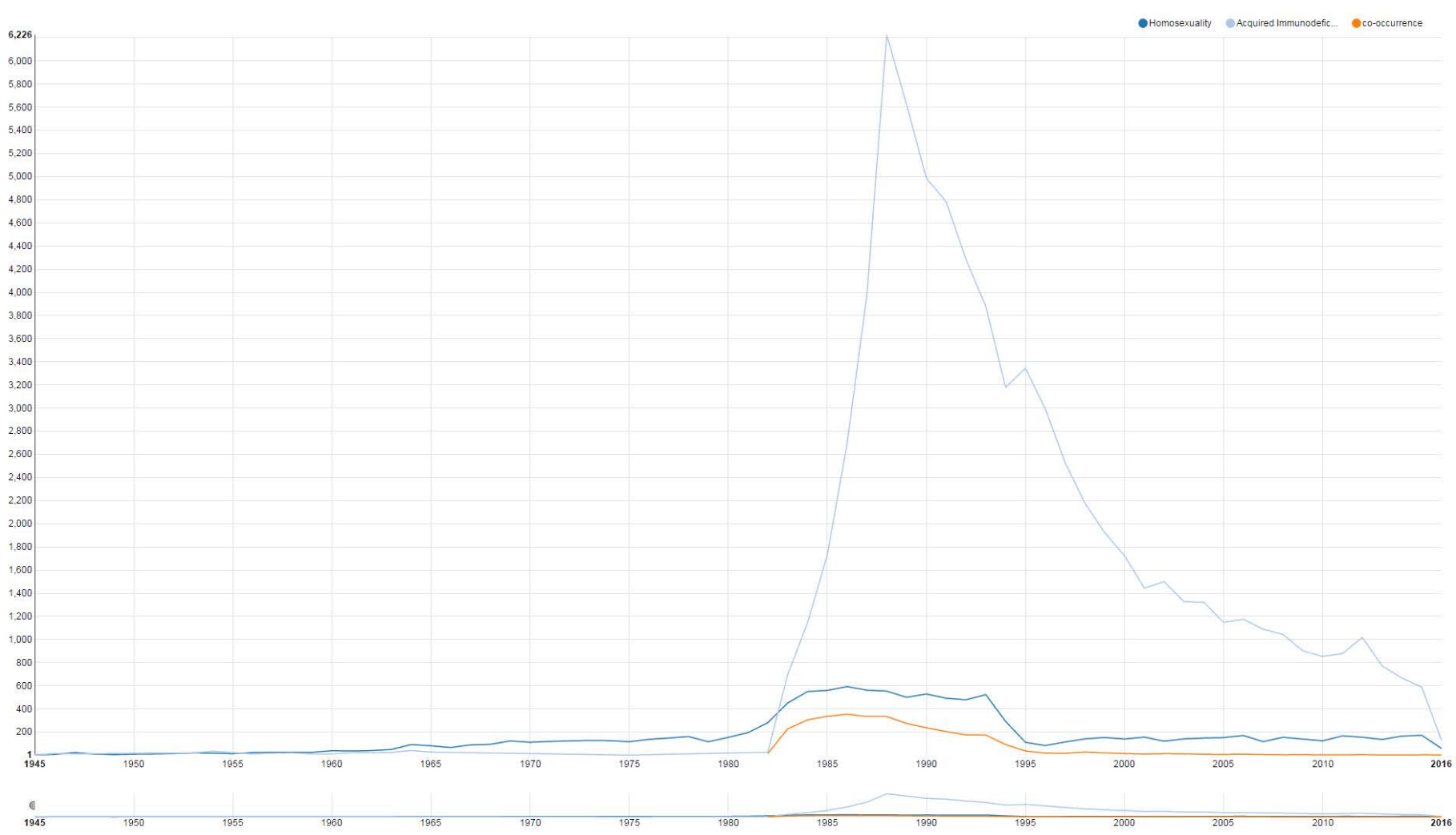
*Ben Busby, NCBI*

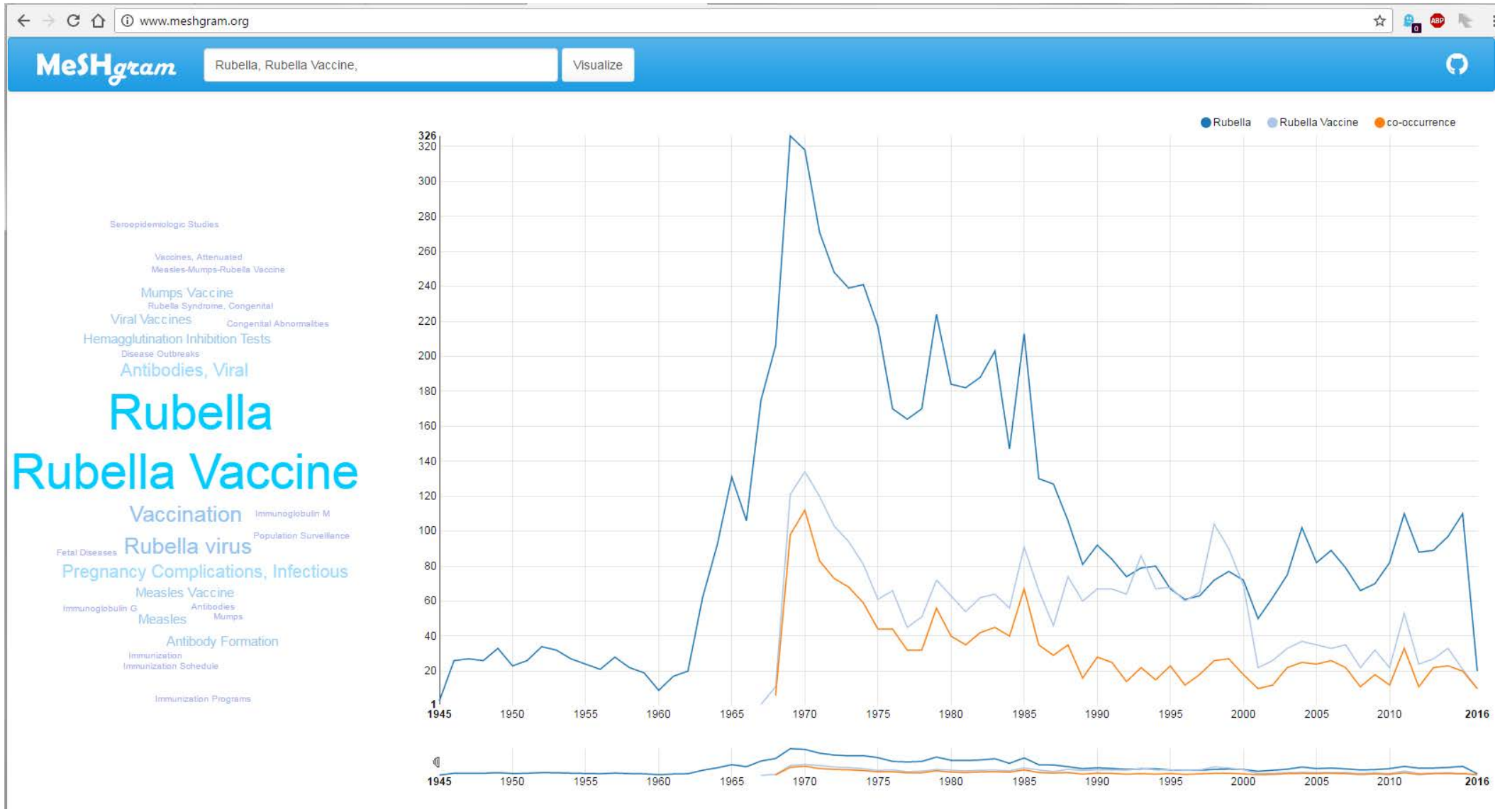
**Medical Informatics Postdoctoral Research Fellowship**

This work is supported by intramural funding and in part by the Intramural Research Program of the U.S. National Library of Medicine (NLM) and an appointment to the NLM Research Participation Program administered by ORISE through an interagency agreement between the U.S Dept. of Energy and the NLM.

U.S. National Library of Medicine


Public Policy
Cytomegalovirus Infections
Leukocyte Count
Lymphatic Diseases
Blood Transfusion
HIV Antibodies
HIV
T-Lymphocytes
Prejudice
Sarcoma, Kaposi
Attitude to Health
Homosexuality
Acquired Immunodeficiency Syndrome
Health Education
Sexual Behavior
HIV Infections
Bisexuality
Deltaretrovirus
Substance-Related Disorders
Risk
HIV Seropositivity
Hemophilia A
Antibodies, Viral
HIV-1
Substance Abuse, Intravenous
Pneumonia, Pneumocystis
AIDS-Related Complex


Homosexuality
Acquired Immunodefic...
co-occurrence
6,226
6,000
5,800
5,600
5,400
5,200
5,000
4,800
4,600
4,400
4,200
4,000
3,800
3,600
3,400
3,200
3,000
2,800
2,600
2,400
2,200
2,000
1,800
1,600
1,400
1,200
1,000
800
600
400
200
1
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2016


www.meshgram.org
MeSHgram
Rubella, Rubella Vaccine,
Visualize
Seroepidemiologic Studies
Vaccines, Attenuated
Measles-Mumps-Rubella Vaccine
Mumps Vaccine
Rubella Syndrome, Congenital
Viral Vaccines
Congenital Abnormalities
Hemagglutination Inhibition Tests
Disease Outbreaks
Antibodies, Viral
Rubella
Rubella Vaccine
Vaccination
Immunoglobulin M
Population Surveillance
Fetal Diseases
Rubella virus
Pregnancy Complications, Infectious
Measles Vaccine
Immunoglobulin G
Antibodies
Measles
Mumps
Antibody Formation
Immunization
Immunization Schedule
Immunization Programs
Rubella
Rubella Vaccine
co-occurrence
326
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
1
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2016