What End Users Need from the Ontology Community

- *Experience from NCPI, FDA, and COVID-19 Ontologies Harmonization effort.*

Asiyah Yu Lin, DATA scholar @ NHGRI/NIH

WSBO2021

July 15, 2021

End users: which ontology do I use?

- End user :
  - developer, domain expert, project manager
  - non-ontologist, not involved in ontology development
- MONDO, DO, HPO? Or SNOMED CT, MeSH…
- PubChem or ChEBI?
- OBI or BAO?

FDA GSRS Use Case

Which disease ontology to use for annotating clinicaltrial.gov data?

By Alex Welsh and Larry Callahan (FDA)

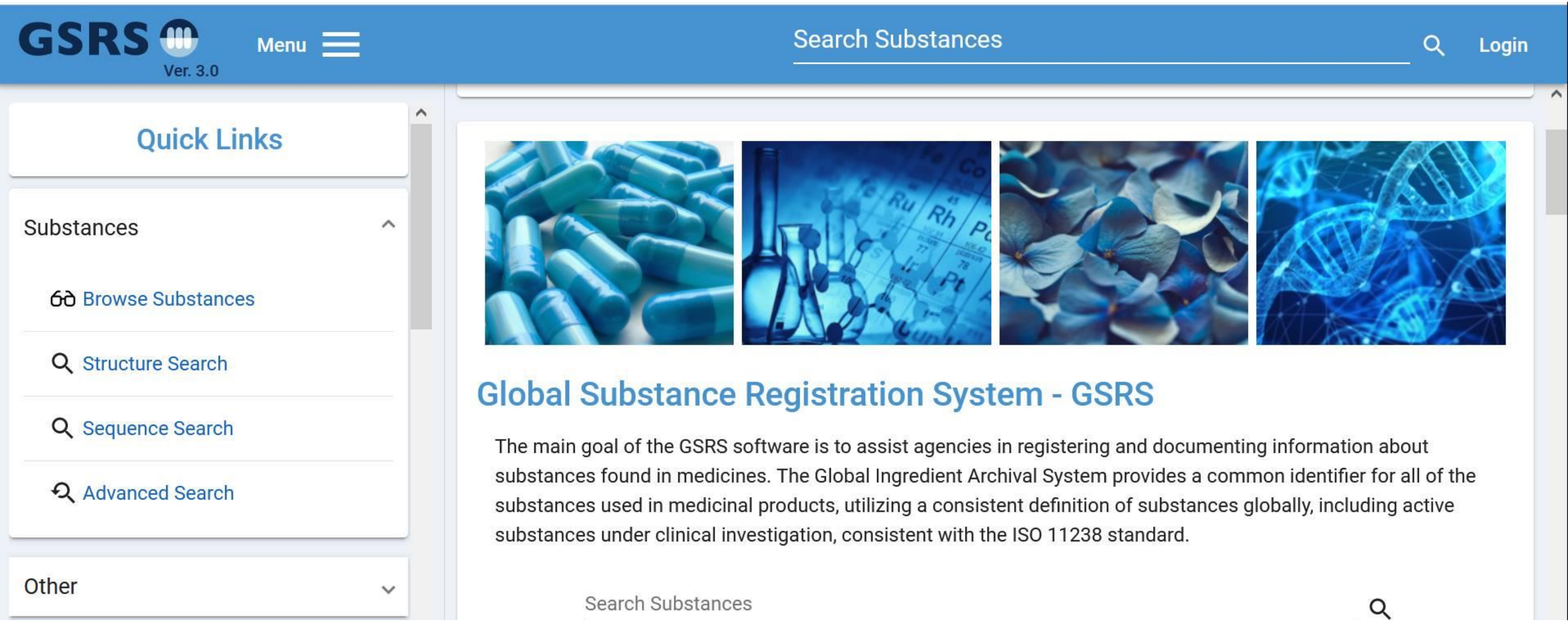
FDA's Global Substance Registration System

Home of the Unique Ingredient Identifier (UNII)

Share | Tweet | Linkedin | Email | Print

Substance Name:
Diethylene Glycol
UNII:61BR964293

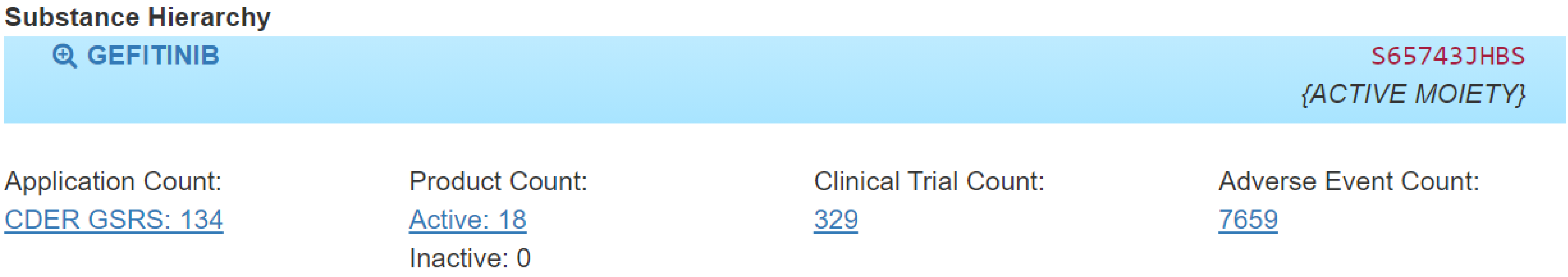
- **Substance**: Any matter of defined composition that has discrete existence, whose origin may be biological, mineral or chemical. (ISO 11238)
- International collaborative.
- NCATS: GINAS
- ~300K substance to clinical trial relationships via "intervention" in clinicaltrials.gov



Global Substance Registration System - GSRS

The main goal of the GSRS software is to assist agencies in registering and documenting information about substances found in medicines. The Global Ingredient Archival System provides a common identifier for all of the substances used in medicinal products, utilizing a consistent definition of substances globally, including active substances under clinical investigation, consistent with the ISO 11238 standard.

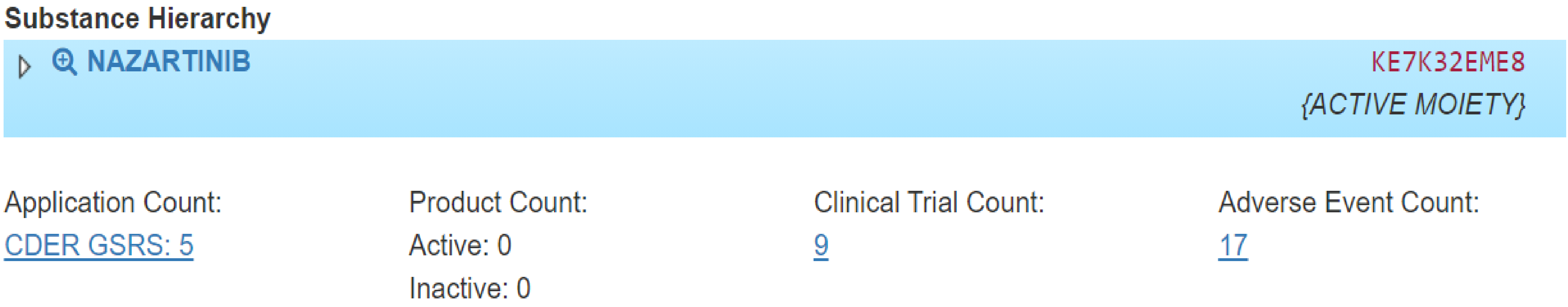
Mapping to Clinical Trial enables links from application to trials



Substance Hierarchy

GEFITINIB S65743JHBS *{ACTIVE MOIETY}*

Application Count:	Product Count:	Clinical Trial Count:	Adverse Event Count:
CDER GSRS: 134	Active: 18	329	7659
	Inactive: 0		

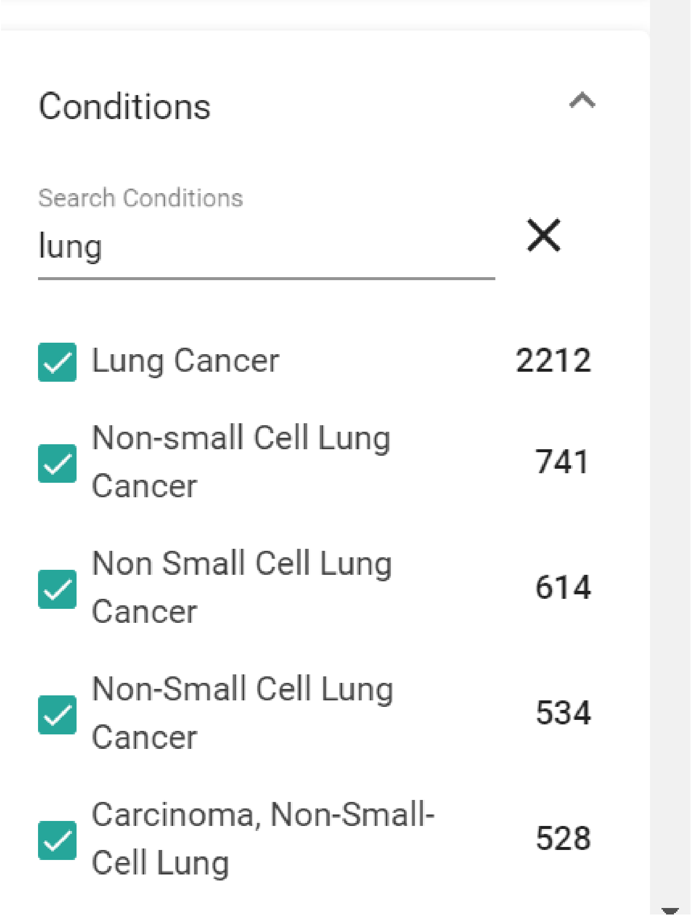


Substance Hierarchy

NAZARTINIB KE7K32EME8 *{ACTIVE MOIETY}*

Application Count:	Product Count:	Clinical Trial Count:	Adverse Event Count:
CDER GSRS: 5	Active: 0	9	17
	Inactive: 0		

Linking conditions to substance and trials is challenging!

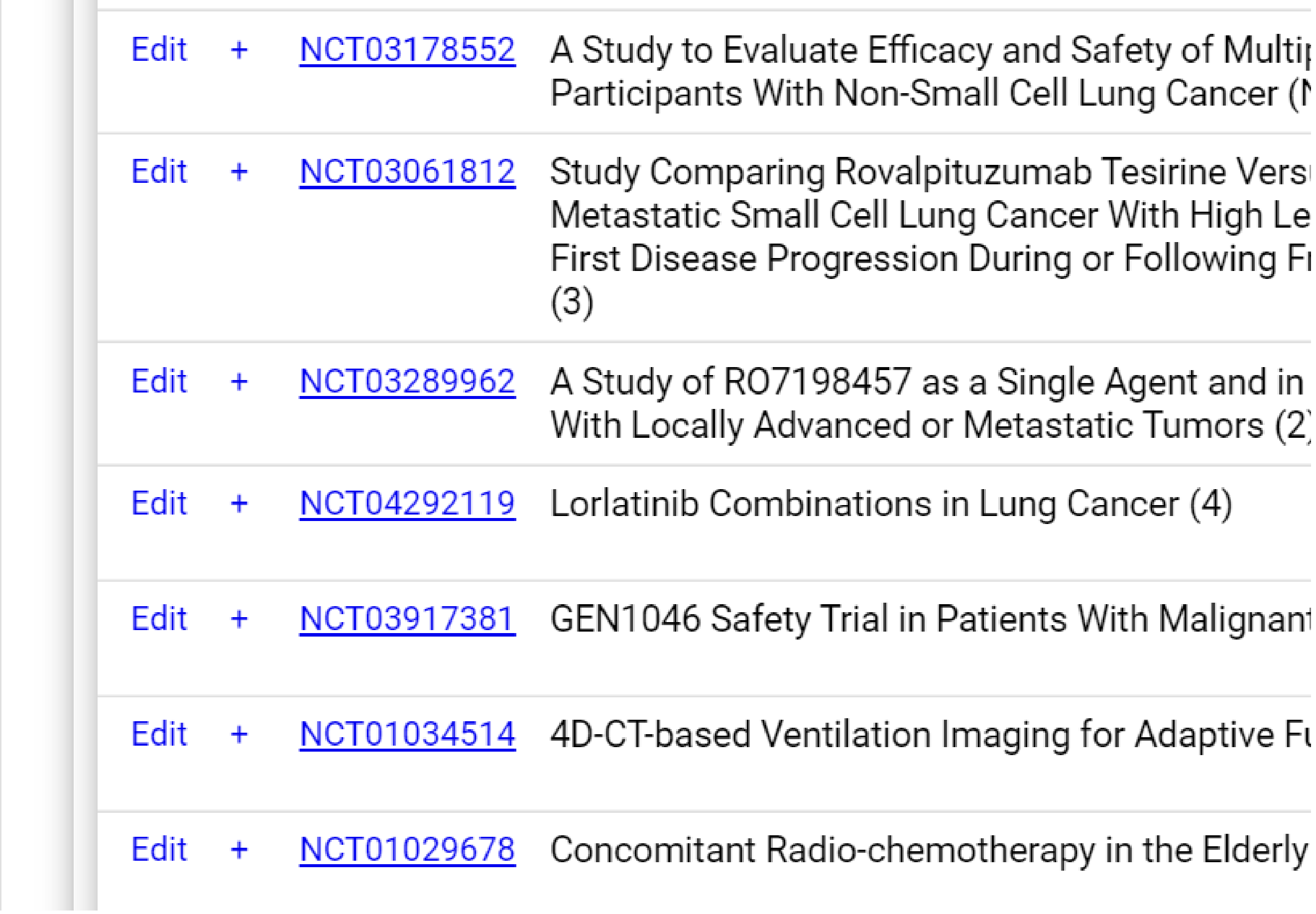


Conditions

Search Conditions

lung

☑	Lung Cancer	2212
☑	Non-small Cell Lung Cancer	741
☑	Non Small Cell Lung Cancer	614
☑	Non-Small Cell Lung Cancer	534
☑	Carcinoma, Non-Small-Cell Lung	528

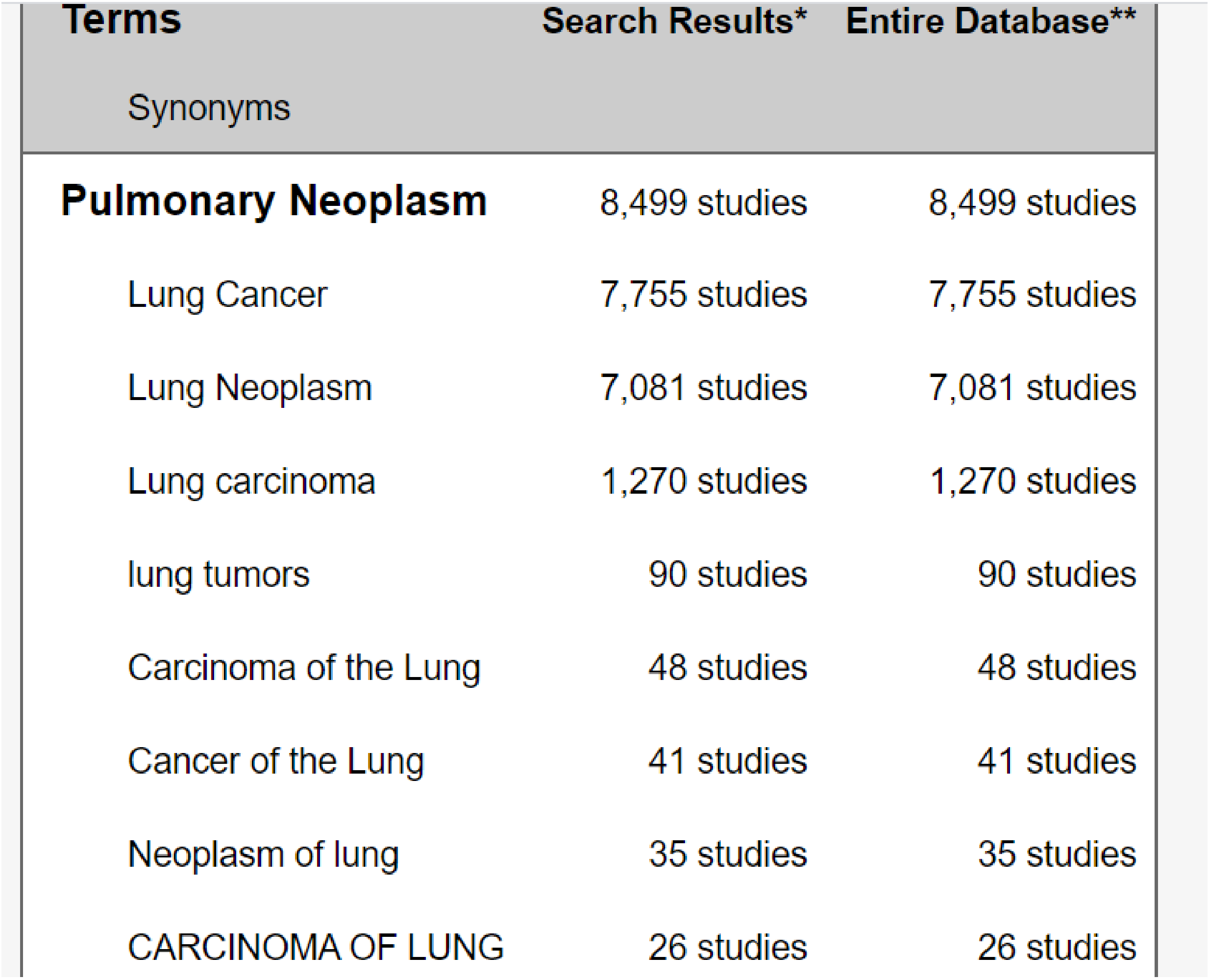


Edit	+	NCT03178552	A Study to Evaluate Efficacy and Safety of Multi Participants With Non-Small Cell Lung Cancer (N
Edit	+	NCT03061812	Study Comparing Rovalpituzumab Tesirine Vers Metastatic Small Cell Lung Cancer With High Le First Disease Progression During or Following Fi (3)
Edit	+	NCT03289962	A Study of RO7198457 as a Single Agent and in With Locally Advanced or Metastatic Tumors (2
Edit	+	NCT04292119	Lorlatinib Combinations in Lung Cancer (4)
Edit	+	NCT03917381	GEN1046 Safety Trial in Patients With Malignant
Edit	+	NCT01034514	4D-CT-based Ventilation Imaging for Adaptive Fu
Edit	+	NCT01029678	Concomitant Radio-chemotherapy in the Elderly

Possible solutions

- Make use of ClinicalTrials.gov strategies to categorize trials by condition, inside the GSRS.
- Use NLP strategies to classify raw clinical trial conditions text into a **broad** set of organ::disease terms for easier searching and faceting.

ClinicalTrials.gov



Terms Synonyms	Search Results*	Entire Database**
Pulmonary Neoplasm	8,499 studies	8,499 studies
Lung Cancer	7,755 studies	7,755 studies
Lung Neoplasm	7,081 studies	7,081 studies
Lung carcinoma	1,270 studies	1,270 studies
lung tumors	90 studies	90 studies
Carcinoma of the Lung	48 studies	48 studies
Cancer of the Lung	41 studies	41 studies
Neoplasm of lung	35 studies	35 studies
CARCINOMA OF LUNG	26 studies	26 studies

NCPI Use Case

Which ontology to use for disease hierarchy in facet search display?

What is NCPI?

The NIH Cloud Platform Interoperability (NCPI) effort aims to establish and implement guidelines and technical standards to empower end-user analyses across participating NIH cloud platforms, to facilitate the realization of a trans-NIH, federated data ecosystem.

Established in late 2019 as a coalition of independently funded NIH IC cloud-based data platforms, with additional support from ODSS

https://anvilproject.org/ncpi

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

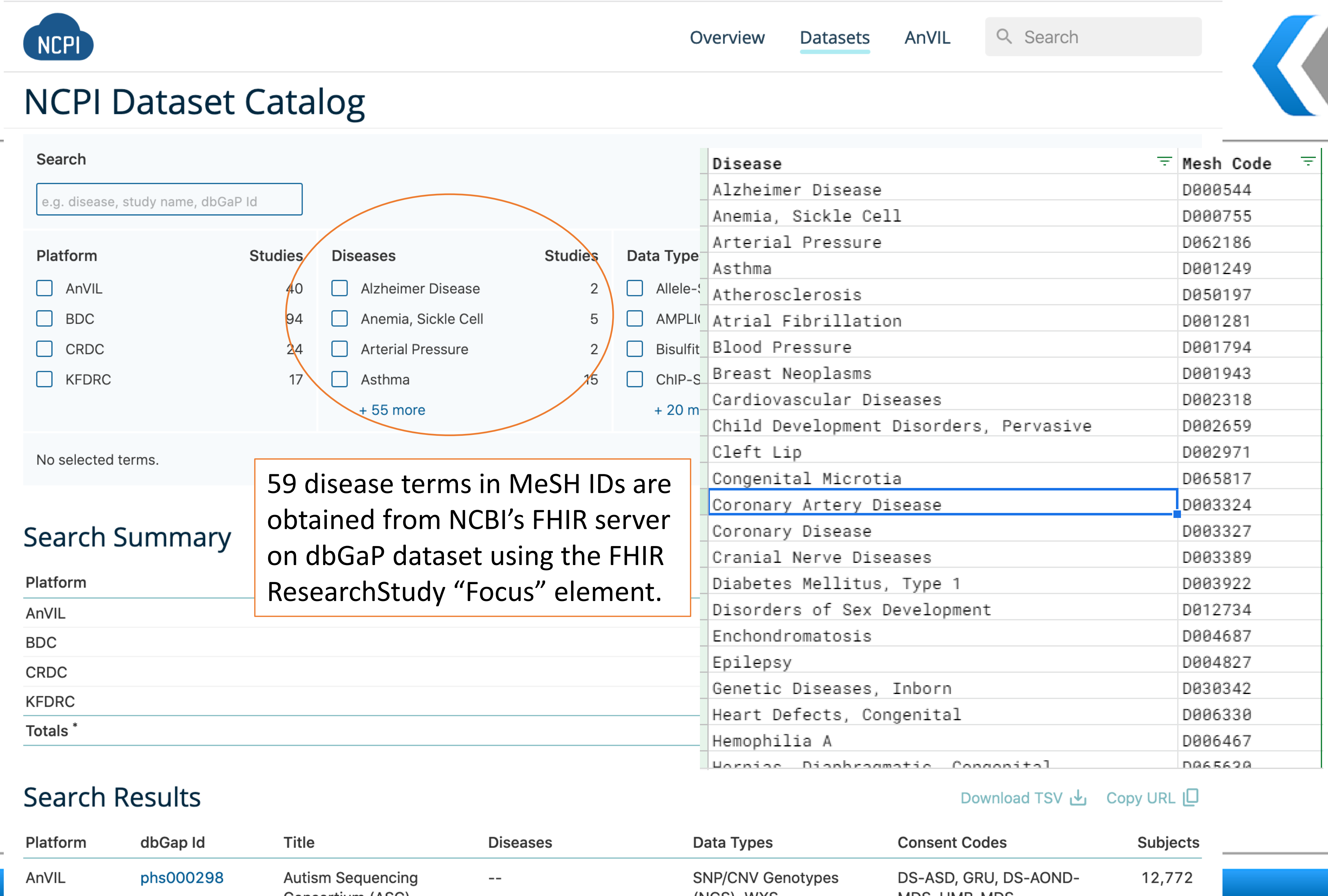
AnVIL

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

NIH National Library of Medicine National Center for Biotechnology Information

By Valentina di Francesco (NHGRI)



NCPI Dataset Catalog

Search

e.g. disease, study name, dbGaP Id

Platform	Studies
□ AnVIL	40
□ BDC	94
□ CRDC	24
□ KFDRC	17

Diseases	Studies
□ Alzheimer Disease	2
□ Anemia, Sickle Cell	5
□ Arterial Pressure	2
□ Asthma	15

+ 55 more

Data Type

□ Allele-S

□ AMPLI(

□ Bisulfit

□ ChIP-S

+ 20 m

No selected terms.

59 disease terms in MeSH IDs are obtained from NCBI's FHIR server on dbGaP dataset using the FHIR ResearchStudy "Focus" element.

Disease	Mesh Code
Alzheimer Disease	D000544
Anemia, Sickle Cell	D000755
Arterial Pressure	D062186
Asthma	D001249
Atherosclerosis	D050197
Atrial Fibrillation	D001281
Blood Pressure	D001794
Breast Neoplasms	D001943
Cardiovascular Diseases	D002318
Child Development Disorders, Pervasive	D002659
Cleft Lip	D002971
Congenital Microtia	D065817
Coronary Artery Disease	D003324
Coronary Disease	D003327
Cranial Nerve Diseases	D003389
Diabetes Mellitus, Type 1	D003922
Disorders of Sex Development	D012734
Enchondromatosis	D004687
Epilepsy	D004827
Genetic Diseases, Inborn	D030342
Heart Defects, Congenital	D006330
Hemophilia A	D006467
Hernias, Diaphragmatic, Congenital	D065630

Search Summary

Platform
AnVIL
BDC
CRDC
KFDRC
Totals *

Search Results

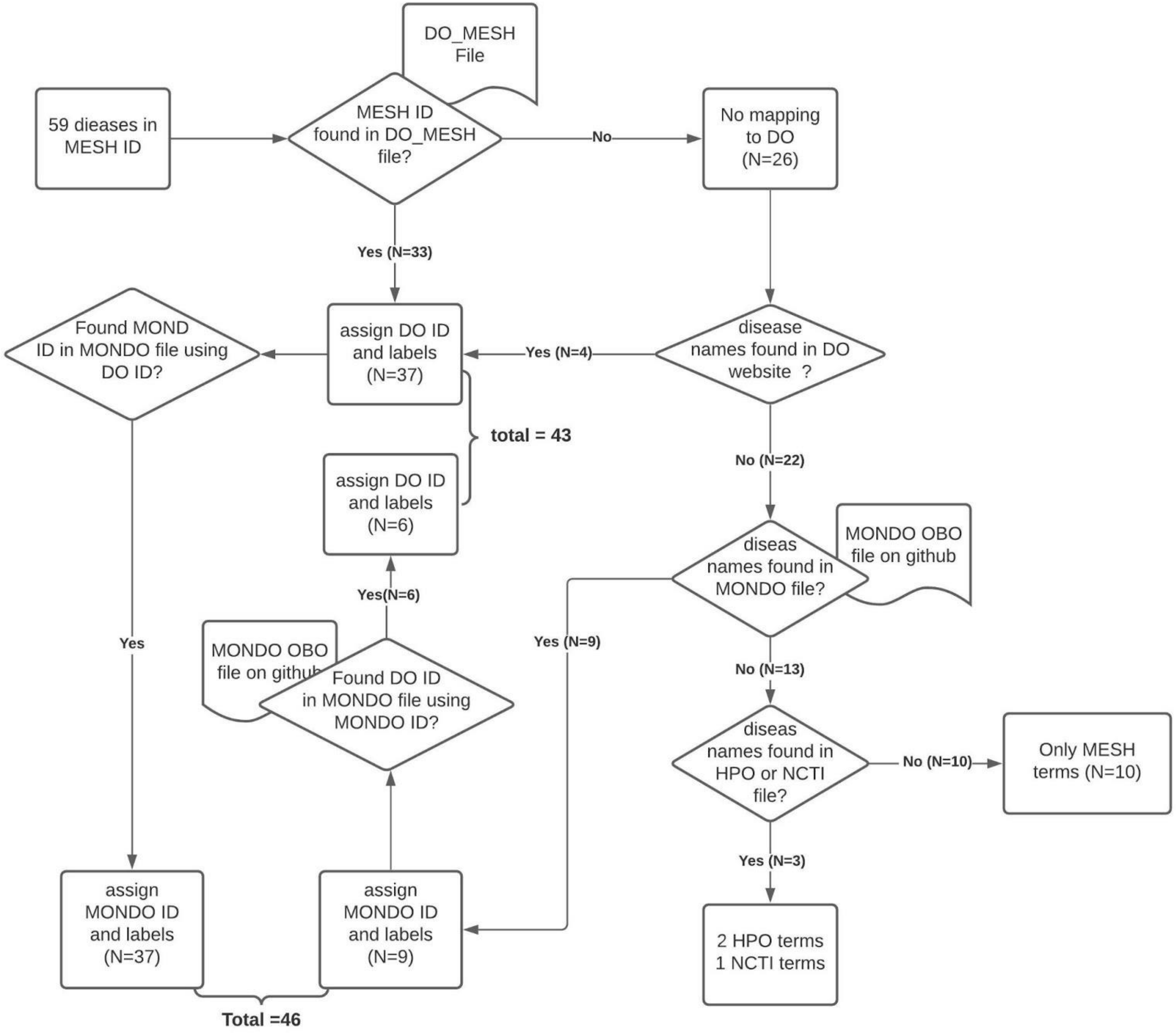
Download TSV

Copy URL

Platform	dbGap Id	Title	Diseases	Data Types	Consent Codes	Subjects
AnVIL	phs000298	Autism Sequencing Consortium (ASC)	--	SNP/CNV Genotypes (NGS), WXS	DS-ASD, GRU, DS-AOND-MDS, HMB-MDS	12,772

Mapping the diseases to MONDO/DO/HPO

- Only look for “exact match” or “equivalent match”
- MESH IDs to DO IDs using DO-MESH mapping file provided by DO(Lynn Schriml), then manually evaluated using manual search on DO.
- MONDO mapping: using DO mapped IDs to find MONDO IDs, then search the disease names for MONDO IDs.
- If neither MONDO or DO, then search HPO. If not HPO, then search NCIT.



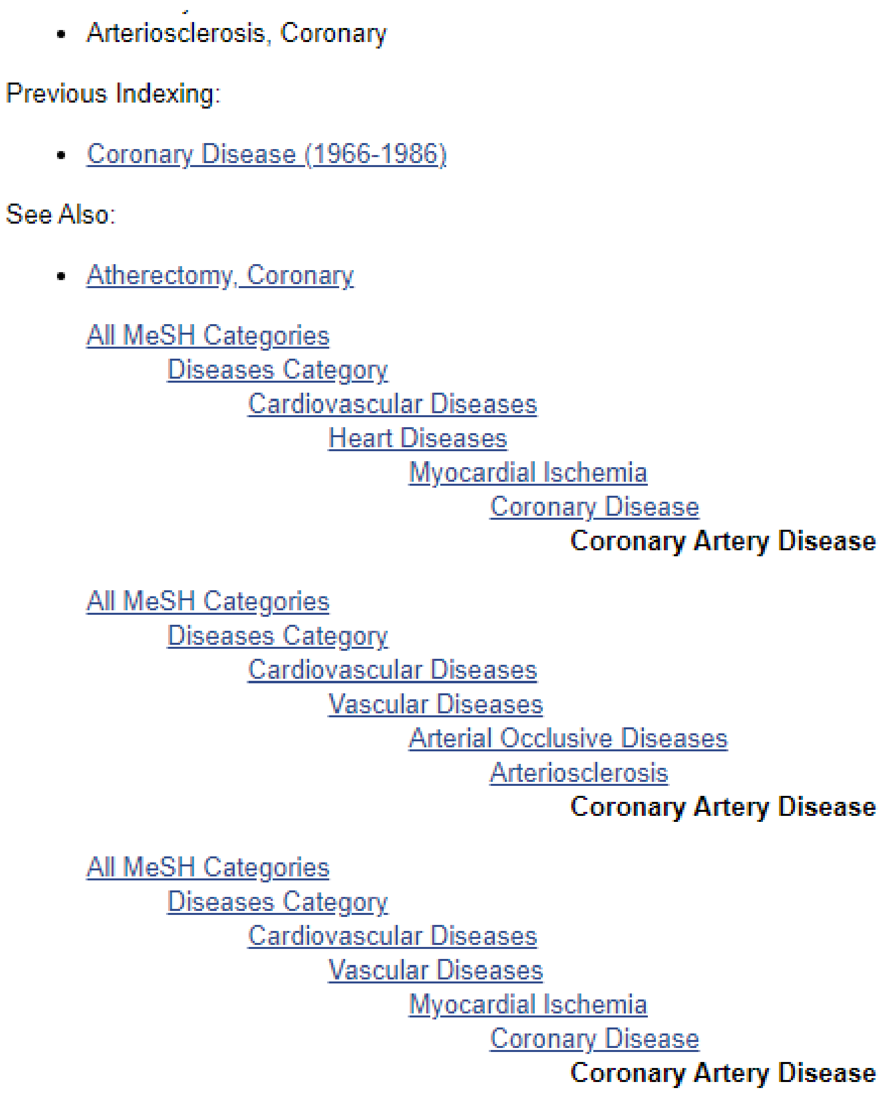
DO_MESH File
59 dieases in MESH ID
MESH ID found in DO_MESH file?
No
No mapping to DO (N=26)
Yes (N=33)
assign DO ID and labels (N=37)
Yes (N=4)
disease names found in DO website ?
Found MOND ID in MONDO file using DO ID?
total = 43
No (N=22)
assign DO ID and labels (N=6)
MONDO OBO file on github
diseas names found in MONDO file?
Yes(N=6)
Yes
Yes (N=9)
MONDO OBO file on github
Found DO ID in MONDO file using MONDO ID?
No (N=13)
diseas names found in HPO or NCTI file?
No (N=10)
Only MESH terms (N=10)
Yes (N=3)
assign MONDO ID and labels (N=37)
assign MONDO ID and labels (N=9)
2 HPO terms 1 NCTI terms
Total =46

Mapping Results:

- **73% (43/59)** is mapped to DO IDs
- **78% (46/59)** mapped to MONDO IDs (includes all 43 DO IDs)
- HPO only term (2):
  Venous thrombosis, Left ventricular hypertrophy
- NCIT only term (1):
  Prostatic Neoplasms, Castration-Resistant
- 17% (10/59) terms are not mapped (**non disease terms**):
  Arterial Pressure, Blood Pressure, Lipids, Mendelian Conditions, Metabolomics, Platelet Aggregation, Population, Reference Values, Women's Health, Xenograft Model Antitumor Assays

Decisions for the NCPI dataset catalog

- Remain using the MeSH terms.
- Plan to use MeSH hierarchy.
- Non disease terms can not be covered by any of the candidate ontologies.
- Switch to display as “disease/focus” as the same in dbGaP.



- Arteriosclerosis, Coronary

Previous Indexing:

- Coronary Disease (1966-1986)

See Also:

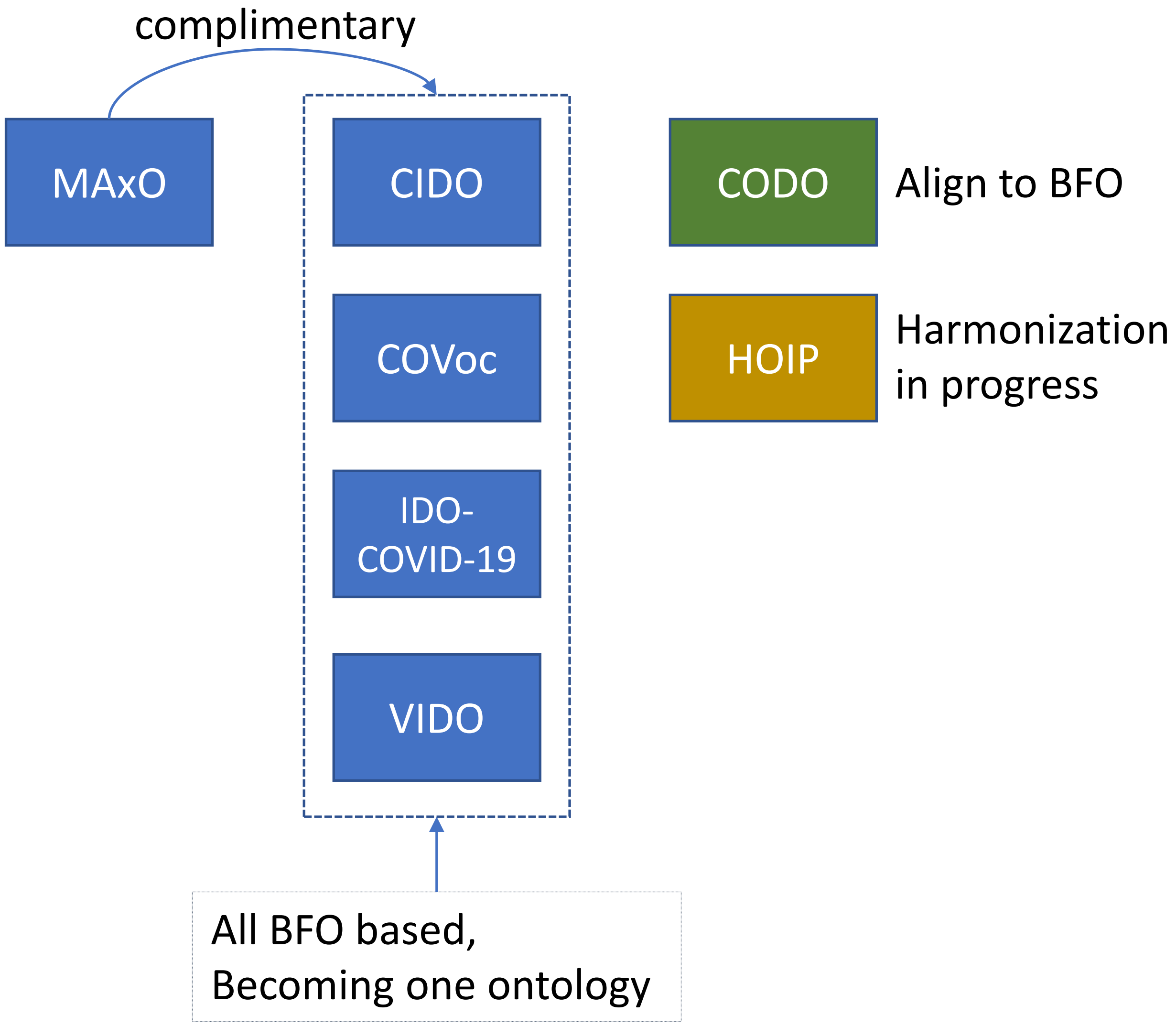
- Atherectomy, Coronary

```
All MeSH Categories
    Diseases Category
        Cardiovascular Diseases
            Heart Diseases
                Myocardial Ischemia
                    Coronary Disease
                        Coronary Artery Disease

All MeSH Categories
    Diseases Category
        Cardiovascular Diseases
            Vascular Diseases
                Arterial Occlusive Diseases
                    Arteriosclerosis
                        Coronary Artery Disease

All MeSH Categories
    Diseases Category
        Cardiovascular Diseases
            Vascular Diseases
                Myocardial Ischemia
                    Coronary Disease
                        Coronary Artery Disease
```

What will end users need?

- A single system encompass MONDO, DO, HPO, MeSH, and NCIT
- Mapping to other standard systems, such as OMOP vocabulary, SNOMED CT, MedDRA, etc.
- Reliable “exact match” and “equivalent match” synced in all relevant ontologies.
- Weighted matches for user’s references.
- A unified translator mid-layer to point to this single system.
- Channels to feedback to ontology developers, and more importantly, **knowing** that end users can provide feedback to ontology community and submit new terms.

COVID-19 Ontology Harmonization Effort



complimentary

MAxO

CIDO

COVoc

IDO-COVID-19

VIDO

All BFO based,
Becoming one ontology

CODO

Align to BFO

HOIP

Harmonization in progress

1. Ontology of Coronavirus Infectious Disease (CIDO) - University of Michigan: Oliver He and *et.al.*
2. Controlled Vocabulary for COVID-19 (COVoc) – EBI : Zoë May Pendlington and Paola Roncaglia
3. COVID-19 Infectious Disease Ontology (IDO-COVID-19) – Northwestern University: John Beverley
4. Virus Infectious Disease Ontology (VIDO) – Northwestern University: John Beverley
5. Ontology for collection and analysis of COviD-19 data (CODO) - India: Biswanath Dutta and Michael DeBellis
6. Homeostasis imbalance process ontology (HOIP) – Japan RIKEN: Yuki Yamagata and *et.al.*
7. Medical Action Ontology (MAxO) – Jackson Lab: Leigh C. Carmody

Possible solutions

- Ontology harmonization: COVID-19 ontology harmonization as an example.
    - Pro: less ontology mapping, one term for a concept
    - Cons: time consuming, requires higher coordination of ontology developers, may not be realistic
- Ontology mappings: standardized metrics?
- Tools support : enhance end users and developer’s interactions.
- Unified interface to interact with end users, build a business service model to respond to end user’s request, educational material to lower the entrance bar for users.

Thank You!

asiyah.lin@nih.gov

Acknowledgement: The Office of Data Science Strategy, NIH, provided funding for AYL as a Data and Technology Advancement (DATA) National Service Scholar.