

How many (rare) diseases are there?

A feeble attempt at systematic disease
harmonization across ontologies

Dac-Trung Nguyen Qian Zhu

NCATS Informatics

Jan 6, 2020

How dare you?

- ▶ No agreed upon definition for what a “disease” is
- ▶ *disease* \rightleftharpoons *syndrome* \rightleftharpoons *condition* \rightleftharpoons *indication* ...
 - ▶ Andersen Syndrome (C1563715) vs. Andersen's Disease (C0017923); both are cross referenced by Orphanet 367.
- ▶ Numerous ontologies: MeSH, OMIM, Disease Ontology, Orphanet, GARD, UMLS, HPO, MONDO, ...
- ▶ Entity resolution/normalization and ontology matching are still unsolved problems

Baby steps...

1. Build a comprehensive knowledge graph based on available ontologies: [Ontology Lookup Service](#)
2. Build on our previous effort for [drug harmonization](#) using in-house [stitcher](#) codebase
3. Focus on the GARD subset

Disease knowledge graph

- ▶ Available as a Neo4j database at
<https://disease.ncats.io/browser>
 - ▶ No username/password required
 - ▶ Use `disease.ncats.io:80` as the **Host**
- ▶ 1,806,027 entities spanning diseases, genes, proteins, drugs, phenotypes, tissues, FDA orphan designations.
- ▶ 473,382,893 relationships that span ontological (e.g., `subclassOf`), phenotypic (e.g., `manifestation_of`), disease association (e.g., `is_associated_disease_of`), etc.
- ▶ Accessible programmatically via numerous Neo4j drivers
 - ▶ Python 2.7 example <https://github.com/ncats/gitlab-snippets/blob/master/neo4j-disease.py>

Disease knowledge graph data sources

Data Source	Entities
GARD	6,763
BRENDA	5,902
ISSUE	1,287
GHR	12,694
DOID	17,387
HPO	2,238
MEDLINEPLUS	275,329
MESH	110,363
MONDO	102,715
OMIM	15,909
UBERON	13,871
ORDO	49,290
GO	69,688
OGG	315,937
PR	91
OGMS	2,730
PATO	130,403
CHEBI	6,074
FDAOrphanGARD_20190216.txt	19,817
RANCHO-DISEASE-DRUG_2018-12-18_13-30	164,448
HPO_ANNOTATION_100918	216,700
MEDGEN	94,645
ICD10	144,695
NCI Thesaurus	26,300
EFO	

A closer look at GARD

GARD diseases by content feature

Content Feature	Count
Prevalence	95
Diagnosis	577
Inheritance	702
Treatment	1037
Cause	873
Symptoms	835
Prognosis	568

Total 6,763 diseases, of which 6,504 are considered rare. Out of 6,504 rare diseases, there are 480 that do not map to anything (e.g., *Hillig syndrome*). It turns out that about 58 of these do map to UMLS exactly; e.g., *Webster Deming syndrome* (GARD 428) maps to *Craniofrontonasal dysplasia with Poland anomaly syndrome* (C4303859).

GARD disease categories

Category	Count
Congenital and Genetic Diseases	3040
Nervous System Diseases	1254
Musculoskeletal Diseases	652
Skin Diseases	591
Eye diseases	573
Rare Cancers	532
Metabolic disorders	509
Blood Diseases	321
Kidney and Urinary Diseases	290
Endocrine Diseases	263
Digestive Diseases	248
Ear, Nose, and Throat Diseases	242
Mouth Diseases	210
Heart Diseases	176
Chromosome Disorders	151
Immune System Diseases	148
Lung Diseases	137
Female Reproductive Diseases	89
Newborn Screening	84
Male Reproductive Diseases	70
Bacterial infections	58
Viral infections	39
Parasitic diseases	33
Hereditary Cancer Syndromes	26
Connective tissue diseases	22
Fungal infections	12
Autoimmune / Autoinflammatory diseases	9
Behavioral and mental disorders	7
Nutritional diseases	3
Environmental Diseases	2

Disease resources

GARD	https://rarediseases.info.nih.gov/
Orphanet	https://www.orpha.net
GHR	https://ghr.nlm.nih.gov/
Disease Ontology (DO)	http://disease-ontology.org/
OMIM	https://omim.org/
MeSH	https://meshb.nlm.nih.gov
MEDLINE+	https://medlineplus.gov/
MONDO	http://monarchinitiative.org/
HPO	https://hpo.jax.org
NORD	https://rarediseases.org
MEDGEN	https://www.ncbi.nlm.nih.gov/medgen/
NCI	https://ncit.nci.nih.gov/
ICD10	https://www.icd10data.com/
EFO	https://www.ebi.ac.uk/efo/

Disease overlap matrix

Direct match by synonyms or identifiers

	GARD	Orphanet	GHR	DO	OMIM	MeSH
GARD	6,504	4,377	837	2,760	3,597	4,017
Orphanet	4,299	9,290	616	3,301	4,427	3,613
GHR	829	616	1,287	620	556	876
DO	2,506	3,554	617	8,699	3,473	3,295
OMIM	4,850	7,313	883	5,842	12,883	8,505
MeSH	3,889	3,330	836	3,140	5,871	8,892
MEDLINE+	310	414	99	966	495	743
MONDO	6,573	11,856	1,245	10,776	8,696	8,555
HPO	730	868	143	1,510	1,458	1,382
NORD	1,049	617	364	741	703	937
MEDGEN	4,896	7,753	989	6,796	7,147	9,400
NCI	1,465	1,435	501	2,319	1,633	2,430
ICD10	963	1,252	167	3,015	1,258	2,394

The matrix is asymmetric due to $1 - n$ and/or $m - 1$ mappings; e.g., there are 4,377 GARD diseases that mapped to 4,299 diseases in Orphanet. These numbers improve as we extend the matching indirectly to two-, three-, four-neighbor.

Disease overlap matrix (cont'd)

	MEDLINE+	MONDO	HPO	NORD	MEDGEN	NCI	ICD10
GARD	257	5,796	715	1,153	4,075	1,410	964
Orphanet	280	8,973	858	642	5,952	1,439	1,245
GHR	99	1,125	143	365	879	490	187
DO	722	8,679	1,502	810	4,417	2,108	2,525
OMIM	664	11,374	2,491	1,053	9,010	2,657	2,029
MeSH	653	7,882	1,153	1,003	7,357	2,074	1,896
MEDLINE+	2,238	662	440	202	801	765	1,228
MONDO	645	21,826	2,385	1,633	11,672	3,377	3,273
HPO	379	2,237	13,725	277	2,160	1,072	1,695
NORD	190	1,163	249	1,251	938	622	371
MEDGEN	668	13,221	2,934	1,361	37,546	3,632	4,727
NCI	593	3,329	1,092	703	3,450	5,323	1,880
ICD10	1,213	3,379	1,820	368	4,301	2,024	16,066

So how many rare diseases are there?

According to MONDO,¹ there are currently **10,393** rare diseases. This number includes diseases that are (i) not considered as rare by GARD (e.g., *Klinefelter syndrome*) and (ii) having UMLS semantic type other than *Disease or Syndrome* (e.g., *ovary leiomyosarcoma* is considered as *Neoplastic Process*).

New rare diseases are being added on a regular basis; e.g., a new entry is being considered for GARD at this moment:

- ▶ *SLC6A1 Epileptic Encephalopathy*
- ▶ *SLC6A1-related Disorders*

Stay tuned...

¹How many rare diseases are there?

Disease harmonization challenges

How many “distinct” diseases are here?



A strategy for harmonization

1. Generate *strongly connected components*
 - ▶ Perform transitive closure on nearest (weighted) neighbors
2. For each strongly connected component, do the following:
 - 2.1 Calculate pair-wise similarities for all synonyms and identifiers based on Jaccard metric

$$\text{sim}(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

- 2.2 Group synonyms and identifiers based on $\text{sim}(x, y) \geq \delta$
($\delta = 0.4$ gives reasonable grouping)
- 2.3 Each such synonym/identifier grouping serves as an initial seed for which entities are then merged/split

Next steps

1. Ongoing evaluation of the proposed harmonization strategy
2. How best to normalize entities?
 - ▶ Identify preferred synonyms and identifiers
3. Support manual curation and harmonization cycle

Example cypher queries on disease knowlege graph

Data source listing

```
match(n:DATASOURCE) return n.name,n.instances
```

OMIM disease count

```
match(n:`S_OMIM`:T047) return count(n)
```

GARD disease category breakdown

```
match(n:`S_GARD`)-[]-(m:DATA) where m.is_rare=true  
with distinct labels(n) as t, count(n) as cnt  
unwind t as Category  
return Category,sum(cnt) as Count order by Count desc
```

Example of overlap query

Orphanet and MONDO

```
match(n:`S_ORDO`)-[]-(m:DATA)
where not exists(m.symbol)
and not exists(m.reason_for_obsolescence)
with n match(n)-[:N_Name|
    :I_CODE*1]-(a:`S_MONDO`)-[]-(b:DATA)
where exists(b.label) and not exists(a.status)
return count(distinct n) as `Orphanet`,
       count(distinct a) as `MONDO`
```