LHNCB Medical Informatics Training Program

Final Report

Title: Analyzing rare diseases terms in biomedical terminologies

Mentor: Dr. Olivier Bodenreider

Training program: October, 2010 – December, 2010

Introduction

A rare disease is a pathological condition with low prevalence and incidence. There are between 6000 and 8000 rare diseases. Many rare diseases are sparsely distributed in some geographic areas and more frequent in others, for reasons linked to genetic factors and environmental conditions that influence the spread of pathogens and the life habits. Thalassemia, for example, is a relatively common genetic disease in the Mediterranean basin (very common in Southern Italy) and is rare in the United States.

A disease or disorder is defined as rare in Europe when it affects less than 5 in 10,000[1]. One rare disease may affect only a handful of patients in the European Union (EU), and another touch as many as 245,000. Overall, rare diseases may affect 30 million EU-citizens. In the United States a rare (or orphan) disease is defined as having a prevalence of fewer than 200,000 affected individuals[2]. Many diseases are much rarer, reaching a rate of one case per 100,000 persons or more.

Rare disease patients too often face common problems, including the lack of access to correct diagnosis, delay in diagnosis, lack of quality information on the disease, lack of scientific knowledge of the disease, inequities and difficulties in access to treatment and care.

These things could be changed by implementing a comprehensive approach to rare diseases, increasing international cooperation in scientific research, by gaining and sharing scientific knowledge about all rare diseases, not only the most "frequent" ones, and by developing tools for extracting and sharing knowledge.

---

[1]<http://ec.europa.eu/health-eu/health_problems/rare_diseases/index_en.htm> (last revision on 10 December 2010)

[2]< http://www.nlm.nih.gov/medlineplus/rarediseases.html> (last revision on 10 December 2010)

Organizations such as the National Institute of Health (NIH) Office of Rare Diseases Research (ORDR), the National Organization for Rare Disorders (NORD) and Orphanet provide information to patients and physicians and facilitate the exchange of information among different actors involved in this field by fostering standardization in clinical terminologies, key factors in information retrieval and information exchange.

The ORDR was established in 1993 within the Office of the Director of the NIH, the Federal point of biomedical research in the U.S. The aim of ORDR is to coordinate and support rare disease research, respond to research opportunities and provide information, as well as promote international collaboration and interoperation.

Orphanet, on the other hand, was established in 1997 by the French Ministry of Health (Direction Générale de la Santé) and the INSERM (Institut National de la Santé et de la Recherche Médicale). Orphanet maintains a database of information on rare diseases and orphan drugs for all publics and aims to contribute to the improvement of the diagnosis, care and treatment of patients with rare diseases.

It includes a Professional Encyclopedia which is a comprehensive collection of review articles on rare diseases, author-based and peer-reviewed, a Patient Encyclopedia and a Directory of expert Services. This Directory includes information on relevant clinics, clinical laboratories, research activities and patient organizations.

The NORD was founded in 1983 by patients and families who worked together to get the Orphan Drug Act passed. This legislation provides financial incentives to encourage the development of new treatments for rare diseases.

The purpose of NORD is to supply information about rare diseases, referrals to patient organizations, and research grants and to serve rare-disease patients and their families. NORD is a non-profit voluntary health agency. Its primary sources of funding are contributions through membership fees.

Objectives

The aim of this project is to analyze a specific area of biomedical terminologies, namely rare disease terms. We examine the representation of rare diseases terms in biomedical terminologies such as MeSH, ICD-10, SNOMED CT and OMIM, leveraging the fact that these terminologies are integrated in the UMLS. More specifically, we want to analyze the overlap among sources and the presence of rare diseases terms in target vocabularies included in UMLS, working at the term and concept level. We also expect to find additional terms and concepts for rare diseases in target terminologies in order to enrich the sources.

Sources of rare disease terms

For the purpose of this project we have acquired terminological resources from the ORDR, Orphanet and NORD. To obtain the needed data it was asked directly to people involved in the organizations in order to be allowed to use them for research purpose.

Office of Rare Diseases Research (ORDR). We received a flat list of 6,857 preferred terms and 11,803 synonyms. The total number of concepts is of 6,857.

Orphanet. We received a flat list of 7,715 preferred terms and 5,224 synonyms and, in addition, a list of Orphanet concepts with the corresponding links to OMIM and ICD10 codes. The total number of concepts is 7,715.

National Organization of Rare Disorders (NORD). We have acquired the list of terms directly from NORD website after obtaining the authorization to do it. We acquired a list of 1,236 preferred terms and 4,562 synonyms and, in addition, a list of 1,283 disorders subdivision. The total number of concepts is 1,236.

Methods

1) Mapping to UMLS;
2) Coverage in target vocabularies and overlap among sources;
3) Enrichment with additional synonyms and descendants.

1) Mapping to UMLS.

The Unified Medical Language System (UMLS) Metathesaurus integrates terms from over 100 biomedical terminologies and groups synonymous terms into concepts.

We utilized the UMLS to map the concepts from the sources to the other biomedical terminologies. The mapping was performed first by Exact Match (EM) and then after normalization against the Normalized String Index (NSI).

> Ex. *Glycogen storage disease type 4* → C0017923 (Exact Match);
> Ex. *Isolated growth hormone deficiency type IA* → C1849790 (*IGHD IA*) (Normalized String).

To validate the results we applied a filter at the semantic level, extracting concepts having "Disorders" as semantic group.

The mapping results can be classified as follows:

1. *Unambiguous Concepts*: all the terms of a given concept map to the same UMLS CUI.
2. *Ambiguous Concepts*: the terms of a given concept map to several different UMLS CUIs

   a. *Ambiguity due to granularity:* the terms of a given concept map to more than one UMLS CUIs, but these UMLS CUIs are hierarchically related.
   b. *Ambiguity due not to granularity:* the terms of a given concept map to more than one UMLS CUIs, and these UMLS CUIs are not hierarchically related.
3. *Unmapped concepts*: no term of a given concept maps any Disorders in UMLS

2) Coverage in target vocabularies and overlap among sources

We analyzed the presence/absence of source rare disease concepts in target vocabularies from the UMLS in order to assess how well these target vocabularies cover the rare diseases terminology. In particular we focused the attention to vocabularies as MeSH (thesaurus used for indexing biomedical terminologies), ICD9/SNOMED CT (vocabularies used for clinical purposes e.g. EHR), OMIM (vocabulary used in genetic databases).

We also analyzed the overlap among the sources of rare disease terms to investigate how many discrepancies / concordances there are among them.

3) Enrichment with additional synonyms and descendants

After analyzing the coverage in target vocabularies, we looked for additional information (when provided) in target vocabularies, in order to acquire additional synonyms and descendants to enrich the starting sources.

Results

1) Mapping to UMLS

The first results of the mapping from the sources to UMLS could be summarized in three categories:

1. Unambiguous concepts.
   All the terms of a given concept map to the same Concept Unique Identifiers (CUI).
   Ex. ORD00117 (*Acrodysostosis*) → C0220659 (*Acrodysostosis*);
   Ex. ORPHA001248 (*Maxillo-nasal dysplasia*) → C0220692 (*MAXILLONASAL DYSPLASIA, BINDER TYPE*);
   Ex. NORD00312 (*Conn Syndrome*) → C1384514 (*Conn Syndrome*).

2. Ambiguous concepts.

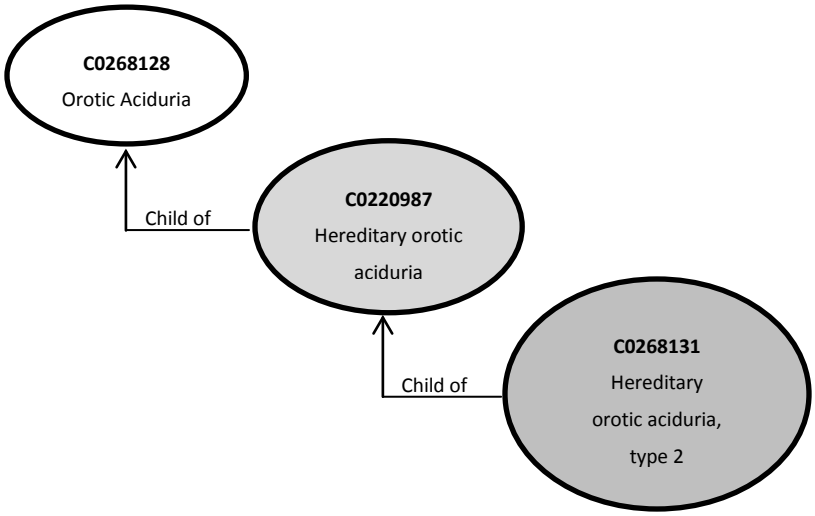
The majority of terms of a given concept map to more than one CUIs. There are two more sub-categories:

- *Ambiguous concepts related to granularity issue*:

ORPHA000030	**CUI 1** **C0268128**	**CUI 2** **C0220987**	**CUI 3** **C0268131**
Oroticaciduria	Orotic aciduria		
Orotic aciduria hereditary		Hereditary orotic aciduria	
Orotidylic decarboxylase deficiency			Hereditary orotic aciduria, type 2
Uridine monophosphate synthetase deficiency	---	---	---

Table 1 . Example of an ambiguous concept related to granularity issue.

As shown in table 1, from a given Orphanet concept, three terms map to three different CUIs and one maps to nothing. In this specific case Orphanet grouped together what SNOMED CT organizes in a hierarchy:



- *Ambiguous concept not related to granularity issue:*

ORPHA000016	CUI1 C0339537	CUI2 C1844778
Blue cone monochromatism	Blue cone monochromatism	
Achromatopsia incomplete, X-linked		Achromatopsia, incomplete, x-linked
Achromatopsia, atypical, X linked	---	---
S-cone monochromatism	---	---

Table 2 . Example of an ambiguous concept not related to granularity issue.

As shown in Table 2, from a given Orphanet concept, the terms map to several CUIs, but from UMLS perspective we don't have any additional information because both terms come from OMIM, so we don't have any information about hierarchical relations.

3. Unmapped Concepts.

There are some terms from the sources that have no mapping in target vocabularies in UMLS:

- Lateral body wall complex
- Levy-Yeboa Syndrome

The possible explanation for that could be because these are extremely rare diseases (e.g. Lateral body wall complex, approximately 250 cases have been reported in the literature so far) or recently discovered (e.g. Levy-Yeboa Syndrome, discovered in June 2006).

2) Coverage in target vocabularies

a) Overall results

The table below (table 3) shows a part of the overall representation in target vocabularies in the UMLS. On the total number of concepts mapped to UMLS (8,435), we noticed a good representation in the sources we investigated:

1. MeSH 5,663 (67%);
2. SNOMEDCT 4,192 (50%);
3. OMIM 3,802 (45%);
4. ICD10 1,029 (12%)

CUI	ORPHANET	NORD	ORD	ICD10	MDR	MSH	NAN	NCI	OMIM	PSY	QMR	RAM	RCD	SNM	SNMI	SNOMEDCT	ULT	UWDA	WHO
	4400	2567	6250	1029	2624	5663	1	2039	3802	180	205	46	3456	2058	2977	4192	1	3	505
C0000744	1	1	1			1		1	1				1	1	1	1			
C0000833		1	1		1	1							1	1	1	1			1
C0000880	1				1	1		1					1			1			
C0000889	1	1	1	1	1	1		1	1				1	1	1	1			
C0001079	1	1	1	1	1			1					1	1	1	1			
C0001080	1	1	1	1	1	1		1	1				1	1	1	1			
C0001126			1		1	1		1	1				1	1	1	1			1
C0001175		1			1	1		1		1		1	1		1	1			1
C0001193	1	1	1		1	1			1					1	1	1			
C0001197			1		1	1		1							1	1			1
C0001206	1	1	1		1	1		1	1		1		1	1	1	1			1
C0001231		1	1	1	1	1		1			1		1	1	1	1			
C0001261			1	1	1	1		1					1	1	1	1			
C0001403	1	1	1	1	1	1		1	1	1			1		1	1			1
C0001429			1			1		1					1	1	1	1			
C0001519		1	1		1	1		1	1					1	1	1			
C0001529	1	1	1		1	1		1	1				1	1	1	1			1
C0001622	1		1		1	1			1				1	1	1	1			1
C0001623			1		1	1		1	1				1		1	1			1
C0001624	1				1	1		1					1			1			
C0001627	1		1		1	1		1	1				1	1	1	1			
C0001733			1		1	1		1	1				1		1	1			
C0001768	1	1			1	1			1				1		1	1			
C0001815	1	1	1			1		1	1			1	1	1	1	1			1
C0001816			1	1	1	1		1	1	1			1	1	1	1			
C0001824			1	1	1	1		1	1				1	1	1	1			1

Table 3. Overlap among sources and representation in target vocabularies .

As shown in table 3, the blank columns represent those sources that have a very small number of mappings (only one or two). This is because some of them were created for a specific purpose, e.g.:

- NANDA nursing diagnoses: definitions & classification (NAN);
- Ultrasound Structured Attribute Reporting (ULT);
- Foundational Model of Anatomy (FMA)

b) Overlap among sources:

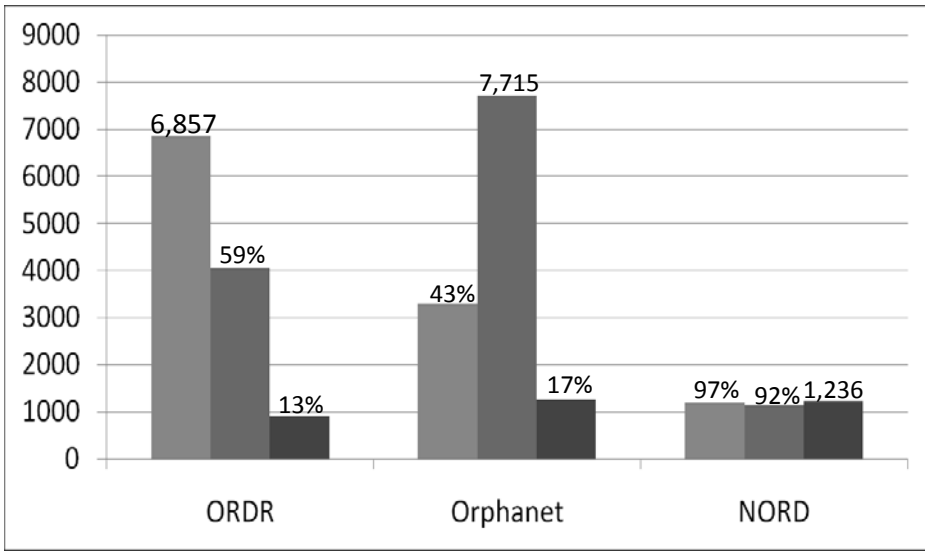


Table 4. Overlap among sources

Table 4 shows the representation of the overlap among sources. From the ORDR perspective there is 59% of common concepts with Orphanet and 13% with NORD; from Orphanet perspective there is the 43% of common concepts with ORDR and 17% with NORD; and from NORD perspective, there is the 97% of common concepts with ORDR and 92% with Orphanet.

3) Enrichment with additional synonyms and descendants

Among the objectives of this work we set out to find, where provided, additional information for the given concepts from the rare diseases sources. After analyzing the representation in the target terminologies, we went deeper in details to find synonyms and more specific terms from target vocabularies. As shown in the example below, from a given concept common to the starting sources, we found that SNOMED CT can provide additional synonyms and descendants:

Cryptococcosis:

- Torulosis

- Busse-Buschke's disease

- European blastomycosis

- European Blastomycosis

- Busse-Buschke disease

Additonal synonyms provided by SNOMED CT:

- European cryptococcosis

- Infection by Cryptococcus neoformans

- Torula

Additional descendants provided by SNOMED CT:

Systemic cryptococcosis

Cryptococcal gastroenteritis

Cryptococcosis associated with AIDS

Cryptococcus infection of the central nervous system

Disseminated cryptococcosis

Hepatic cryptococcosis

Mucocutaneous cryptococcosis

Ocular cryptococcosis

Osseous cryptococcosis

Pulmonary cryptococcosis

Conclusion

1) Mapping to UMLS

   We found a good coverage in UMLS, especially in the sources we analyzed in details. Sometimes there are differences due to the different ways concepts are organized in target vocabularies. This is partly because each source is originally built for a specific purpose and the concepts are organized following specific principles.

   As presented in the results, there are also some unmapped concepts. This is partly because some diseases are extremely rare, but in the case of Orphanet it could be a problem of overestimation of unmapped concepts. The Orpahnet database has a hierarchy among

concepts, and there is a difference among *grouper concepts* : e.g. “rare genetic skin disease” and *leaf concepts*: e.g. “xeroderma pigmentosum”. While leaf concepts are expected to be represented in target vocabularies, such as SNOMED CT, grouper concepts are likely to be specific to Orphanet.

In addition, it isn’t possible to validate results with other sources because there is no ontological consistency.

2) Coverage in target vocabularies and overlap among sources

   The sources are quite different among them, but in general the representation of rare diseases is organized in the same way. There are some differences among them: if we compare NORD and ORDR, for example, there is an important difference in the number of concepts, but it is partly explained because NORD is a subset of ORDR and it is also more patient-friendly oriented.

3) Enrichment with additional synonyms and descendants.

   We found additional synonyms and descendants in target vocabularies. In this case we could enrich our starting sources to better provide information about rare diseases.

We will share our results and plan further collaboration with ORDR and Orphanet. Our work will contribute to better harmonization between these sources and better integration of these sources in the UMLS. These organizations will also help us validate our results with clinical experts.

Finally, discrepancies in the grouping of rare disease terms into concepts will be shared with the UMLS team and may help detect missed synonymy in the UMLS.