JLIS.it

Analyzing rare diseases terms in biomedical terminologies

Erika Pasceri

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, 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 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 EU, and another touch as many as 245,000. Overall, rare diseases may affect 30 million European Union 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

[1] http://ec.europa.eu/health-eu/health_problems/rare_diseases/index_en.htm.

[2] http://www.nlm.nih.gov/medlineplus/rarediseases.html.

JLIS.it. Vol. 3, n. 1 (Giugno/June 2012).
DOI: 10.4403/jlis.it-4783

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.

Organizations such as the National Institute of Health (NIH), Office of Rare Diseases Research (ORDR), 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 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. The aim of ORDR is to coordinate and support rare disease research, responding to research opportunities and providing information, promoting international collaboration and interoperation.

Orphanet, on the other hand, was established in 1997 by the French Ministry of Health (Direction Générale de la Santé)[3] and the Institut National de la Santé et de la Recherche Médicale (INSERM).[4] 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 compre-

[3] http://www.sante.gouv.fr.
[4] http://www.inserm.fr.

hensive 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 development of new treatments of rare diseases. The purpose of NORD is to supply information about rare diseases, referrals to patient organizations, research grants and all those people that have interest in rare disease. The purpose of NORD is to supply information about rare diseases, referrals to patient organizations, research grants and all those people that have interest in rare disease. It isn't a government agency; it is a non-profit voluntary health agency that exists to serve rare-disease patients and their families. Its primary sources of funding are contributions membership fees.

Objective

The aim of this project is to analyze a specific area of biomedical terminologies, namely rare disease terms. The representation of rare diseases terms has been analyzed in biomedical terminologies such as Medical Subject Headings (MeSH), International Classification of Diseases (ICD)-10, Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) and Online Mendelian Inheritance in Man (OMIM), leveraging the fact that these terminologies are integrated in the Unified Medical Language System (UMLS). It has been analyzed the overlap among sources and the presence of rare diseases terms in target sources included in UMLS, working at the term and concept level.

Material

In this section the resources used in this study are briefly described: the two sources of rare disease terms (ORDR and Orphanet), the four target terminologies (ICD, MeSH, OMIM, and SNOMED-CT) and the UMLS.

The **UMLS**® is a terminology integration system developed at the National Library of Medicine. The UMLS Metathesaurus® integrates almost 160 biomedical vocabularies, including the four target vocabularies under investigation (ICD-10, MeSH, OMIM and SNOMED-CT). Synonymous terms from the various source vocabularies are grouped into one concept. Additionally, the Metathesaurus records the relations asserted among terms in the source vocabularies, including hierarchical, associative and mapping relations. Version 2010AB of the UMLS is used in this study. This version contains approximately 2.4 M concepts and 40 M relations.

Source terminologies

The **ORDR**[5] publishes a list of rare diseases. This resource does not represent any relations among rare diseases, but groups all the synonyms of a given disorder into a single concept. It maintains a list of 6,857 rare disease concepts (and 11,803 synonyms) on its Web site of which about 800 have extensive information on resources relating to questions by the public. The rare disease concepts are either diseases for which information requests have been made to directly to the Office of Rare Diseases Research, the Genetic and Rare Diseases Information Center (GARD) which is funded by theORDR and the National Human Genome Research Institute (NHGRI), or NHGRI directly; or (2) diseases from various data sources and those

[5] http://rarediseases.info.nih.gov.

that over the last 10 years have been suggested as being rare. The purpose of the Rare Diseases and Related Terms list is to facilitate the distribution of information.

Orphanet[6] provides information about 5,954 rare diseases. Orphanet diseases are organized into a Directed Acyclic Graph. In the Orphanet database, diseases are linked to external reference terminologies, such as ICD10 and OMIM. The Orphanet list of rare diseases comprises 7,715 concepts. We acquired a list of 7,715 preferred terms and 5,224 synonyms. Additionally, Orphanet shared with us the correspondence they established between rare disease concepts and OMIM and ICD10 codes.

Target terminologies

The **ICD** is the international standard diagnostic classification for all general epidemiological, many health management purposes and clinical use. It is used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and health records. In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by World Health Organization World Health Organization (WHO) Member States. The 10th revision of ICD (ICD-10) is used in this study. It is included in UMLS.

The **MeSH** is a controlled vocabulary developed by the U.S. National Library of Medicine for the indexing and retrieval of the biomedical literature, especially in the MEDLINE bibliographic database. It consists of sets of terms naming some 25,000 descriptors

[6] http://www.orpha.net.

in a hierarchical structure that permits searching at various levels of specificity. Version 2011 of MeSH is used in this study. Of note, this version provides partial coverage for the rare disease terms from ORDR. MeSH is one of the terminologies in the UMLS.

The ***OMIM*** is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes developed at John Hopkins University. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. Its terminological component – including clinical synopses – is available through the UMLS.

The ***Systematized Nomenclature of Medicine*** (**SNOMED-CT**) is the world's largest clinical terminology developed by the International Health Terminology Standard Development Organization (IHTSDO) for use in electronic health records. It covers most areas of clinical information such as diseases, findings, procedures, microorganisms, pharmaceuticals etc. SNOMED-CT provides a consistent way to index, store, retrieve, and aggregate clinical data across specialties and sites of care. It also helps organizing the content of medical records, reducing the variability in the way data is captured, encoded and used for clinical care of patients and research. The version of SNOMED-CT used in this study is dated July 31, 2010 and is integrated in the UMLS.

In the remainder of this paper, for simplification purpose, ORDR and Orphanet will be named as *sources* and SNOMED-CT, MeSH, OMIM and ICD10 as the *targets*.

Method

UMLS has been used in various data creation, indexing and encoding systems. It accomplishes this by conjoining the sets of

synonyms and concept relationships in its multiple constituent terminologies (Merabti et al.). In this study rare disease terms from the two sources were mapped to the corresponding UMLS concept(s) using an exact match or after normalization. Normalization abstracts away from such unessential differences as case, punctuation, and inflectional variants (e.g., singular vs. plural) and stop words in terms:

Ex. *Glycogen storage disease type 4* → C0017923 (Exact Match);

Ex. *Isolated growth hormone deficiency type IA* → C1849790 (Normalized String).

Because the terms from ORDR and Orphanet are all expected to name (rare) disorders, we restricted the UMLS concepts mapped to disorder concepts through a filter based on the Semantic Group *Disorders* (including such semantic types as*Disease or Syndrome* and *Congenital Abnormality*). This simple filter provides some level of word sense disambiguation.

Results

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 only one 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*:

ORPHA0000	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 match to three different CUIs and one match to nothing. In this specific case Orphanet grouped together what SNOMED-CT put in hierarchy:

   - *Ambiguous concept not related to granularity issue:*

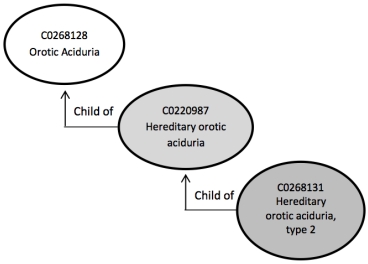

Figure 1

As shown in table 2 on the facing page, from a given Orphanet concept, the terms match to several CUIs, but from UMLS

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.

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. ***Unmatched Concepts***
   There are some terms from the sources that have no mapping in target sources 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).

Overall representation in targets

The figure 2 on page 11 shows a part of the overall representation in target sources in the UMLS. On the total number of concepts mapped to UMLS (8,435), we noticed a good representation in the sources we focused the attention:

1.	*MeSH*	5,663 (67%)
2.	*SNOMED-CT*	4,192 (50%)
3.	*OMIM*	3,802 (45%)
4.	*ICD10*	1,029 (12%)

As shown in figure 2 on the facing page, 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 context, e.g.:

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

Overlap among sources

Figure 3 on page 12[7] 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.

Additional information for a given concept from sources

Among the objectives of this work we set out to find, where provided, additional information for the given concepts from rare dis-

[7]For better details, see http://leo.cilea.it/index.php/jlis/article/downloadSuppFile/4783/5747.

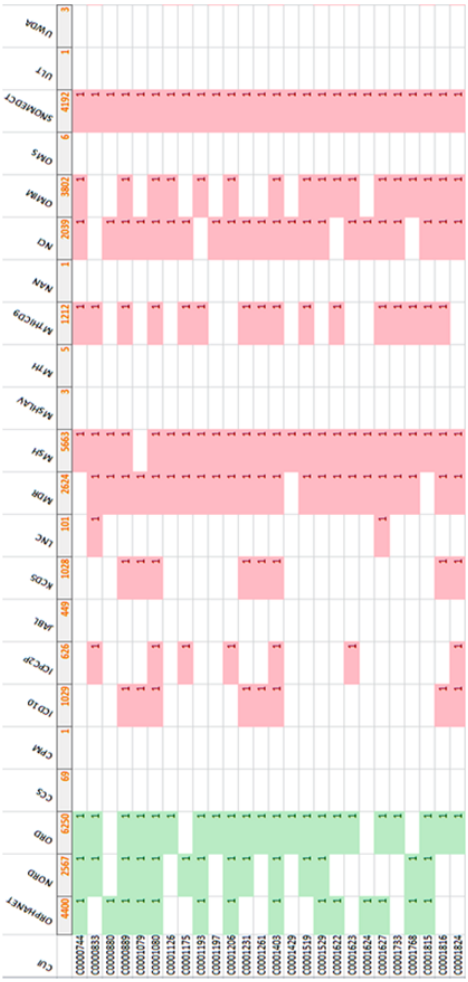

Figure 2: Overlap among sources and representation in targets

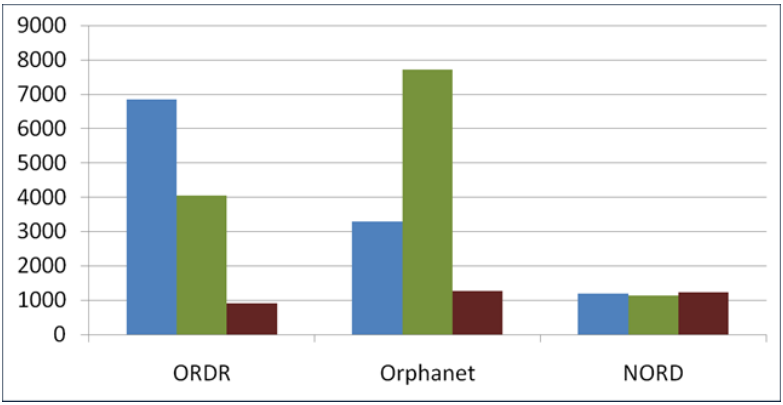

Figure 3: Overlap among sources

eases lists. After analyzing the representation in the target sources, 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 in 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

Additional 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

Limitations

In some cases we didn't find any correspondence of terms or concepts in UMLS. This is partly because everything is seen from UMLS perspective; which is because the target sources organize in different ways the terms from their perspectives that makes the difference among the several vocabularies included in UMLS. We also noticed that some concepts not present in UMLS, but probably because there are some diseases that are extremely rare and also because some of them have been recently discovered. If we focus the attention only on Orphanet, maybe we overestimated the percentage of unmapped concepts because in the list of terms there are some that are very general terms as "*rare genetic skin disease*" versus what we have in target sources really specific as "*xeroderma pigmentosus*".

Conclusion

Rare diseases are insufficiently and inconsistently represented in medical terminologies. More than 50% of rare diseases concepts are still not aligned. Automatic approaches can be used to create a draft of the alignment and facilitate the work of domain experts. We found a good representation in target sources in UMLS, especially in the sources where we focused the attention; we also found additional information for the rare diseases concepts. We will share the result with the organizations that work in this particular field so that to enhance the information retrieval. They will provide to review all data with the supervision of clinical experts. This work could be also a feedback to UMLS, for those terms that ORDR, Orphanet and NORD grouped together and UMLS doesn't.

Works cited

Aymé, Ségolène, et al. "Information on rare diseases: the Orphanet project". *La Revue de médecine interne* 19. (1998).

Merabti, Tayeb, et al. "Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH". *Ingénierie et Recherche Biomédicale/BioMedical Engineering and Research* 31.4. (2010): 221–225. (Cit. on p. 7).

Zhang, Songmao and Olivier Bodenreider. "Alignment of multiple ontologies of anatomy: deriving indirect mappings from direct mappings to a reference". *AMIA Annu Symp Proc*. (2005): 865–868.

ERIKA PASCERI, Università di Udine.
erika.pasceri@unical.it

Pasceri, E."Analyzing rare diseases terms in biomedical terminologies". *JLIS.it* Vol. 3, n. 1 (Giugno/June 2012): 4783-1–4783-15. DOI: 10.4403/jlis.it-4783. Web.

ABSTRACT: Rare disease patients too often face common problems, including the lack of access to correct 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 and by developing tools for extracting and sharing knowledge. A significant aspect to analyze is the organization of knowledge in the biomedical field for the proper management and recovery of health information. For these purposes, the sources needed have been acquired from the Office of Rare Diseases Research, the National Organization of Rare Disorders and Orphanet, organizations that provide information to patients and physicians and facilitate the exchange of information among different actors involved in this field. The present paper shows 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. At the first level, it was analyzed the overlap among sources and at a second level, the presence of rare diseases terms in target sources included in UMLS, working at the term and concept level. We found that MeSH has the best representation of rare diseases terms.
KEYWORDS: Rare Deseases; MeSH; Terminology; Text mining; Thesauri
ACKNOWLEDGMENT: This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine (NLM) and in part by the University of Udine. I would like to thank Olivier Bodenreider and Bastien Rance from NLM for technical support and Maurella Della Seta from the Istituto Superiore di Sanità (ISS) that has made this experience possible.

Submitted: 2012-02-10
Accepted: 2012-03-02
Published: 2012-06-01