Coverage of Rare Disease Names in Clinical Coding Systems and Ontologies and Implications for Electronic Health Records-Based Research

Rachel Richesson
Duke University School of Nursing
Durham, NC USA
rachel.richesson@dm.duke.edu

Kin Wah Fung
National Library of Medicine
Bethesda, MD, USA

Olivier Bodenreider
National Library of Medicine
Bethesda, MD, USA

***Abstract*—This poster will present the completeness of coverage of rare disease names in standard coding systems, including the International Classification of Diseases (ICD) and SNOMED CT, and ontologies such as the Orphanet Rare Diseases Ontology (RDO). Using use cases and a set of 45 rare diseases for the national Patient Centered Outcomes Research Network (PCORnet), the poster will describe the current capacity and implications for electronic health records-based research on these diseases. Authors will provide suggestions on how clinical coding systems and ontologies can be used in a coordinated approach to support the use of electronic health record data for various types of research related to rare diseases.**

Keywords—rare diseases; clinical classifications; ontologies; biomedical research; electronic health records

I. Introduction

Rare diseases are defined in the US as conditions that affect less than 200,000 Americans and in the European Union as those with a prevalence of 5 per 10,000 or less.[1,2] The NIH Office of Rare Diseases Research recognizes 6,485 rare diseases.[3] Although each rare disease is uncommon, collectively they constitute a significant burden to the health care system. One estimate suggests that 1 in 10 Americans are affected by a rare disease.[2] Consequently 'rare diseases' have emerged as priority topics in public health and research. Rare disease names are included, at different levels of completeness and granularity, in a number of clinical coding systems that are embedded in electronic health record (EHR) systems, and in a number of ontologies designed to support the diagnosis rare diseases and investigation of their causes and treatments.[4]

With increased adoption and "meaningful use" of EHRs, there is renewed effort in leveraging EHRs for research. In the U.S., the national Patient Centered Outcomes Research Network (PCORnet) was funded this year from the Affordable Care Act to examine real-world treatment decisions, and is specifically tasked to conduct observational and interventional research on the comparative effectiveness of various treatments, using distributed and heterogeneous EHR systems.[5] The PCORnet research portfolio currently includes 45 rare diseases (in addition to approximately 20 more common conditions). The objective of this poster is to determine the coverage of these rare disease names in standard coding systems and explore the current capacity and implications for EHR-based research on these and other rare diseases.

II. Methods

In this poster we present an inventory of various clinical coding systems and ontologies that are relevant to rare diseases, and summarize their coverage of rare disease names from previous studies. We match rare diseases names and synonyms from the Office of Rare Disease Research (ORD) and Orphanet (RDO) to the Unified Medical Language System (UMLS) Metathesaurus and identify maps to SNOMED CT and other terminologies. To characterize the coverage of rare diseases studied in PCORnet, we estimate the number of precise and equivalent matches in the three clinical classifications (ICD-9-CM, ICD-10-CM, and SNOMED CT) for a set of 45 rare diseases studied in PCORnet. Finally, we present the likely use of existing classifications, ontologies, mappings, and tools to support the research process, from the collection of data in clinical settings to their use in various types of EHR-based research.

III. RESULTS

SNOMED CT has the highest coverage of rare disease names among clinical terminologies in UMLS, and covers 44% of the 6,485 diseases (19,504 terms) recognized by the Office of Rare Diseases (ORD), and 48% of the 6,750 diseases (15,585

terms) diseases listed in the Orphanet Rare Disease Ontology. 25% (1,611) of ORD and 14% (1,592) RDO disease names have bi-directional one-to-one maps to SNOMED CT. The rest are one-to-many or many-to-one maps. Two terminologies have higher coverage than SNOMED CT. Medical Subject Headings (MeSH) covers 75% and 70%, while Online Mendelian Inheritance in Man (OMIM) covers 49% and 57%, of ORD and RDO respectively. Overall, the UMLS covers 82% of ORD-recognized and 84% of RDO-recognized rare diseases.

All of the rare diseases studied in PCORnet were included in the UMLS and its source terminologies. 8 diseases did not have any match to SNOMED CT, ICD-9-CM or ICD-10-CM. The 45 rare diseases studied in PCORnet yielded multiple matches to terms in clinical coding systems; i.e., many PCORnet rare disease names matched to more than one (term) code in a coding system, and many codes from clinical coding systems matched more than one rare disease name. Of 55 ICD-9-CM codes that matched to a PCORnet rare disease, 7 were matched to multiple rare diseases. Of 47 matched ICD-10-CM codes, 4 matched to multiple rare diseases, and of 59 matched SNOMED CT codes, one SNOMED CT code matched to multiple PCORnet rare diseases. The proportions of matched codes that were considered equivalent matches (rather than broader matches or related terms) were 25%, 45% and 94% for ICD-9-CM, ICD-10-CM and SNOMED CT respectively.

IV. Conclusions

The coverage and quality (i.e., precision and equivalence) of terms for rare diseases in clinical coding systems is less than ideal, but is markedly improved with SNOMED CT in comparison to ICD 9 and 10 classifications. The lack of precise and complete coverage of rare disease names in clinical coding systems will inhibit the automated identification patients with rare diseases from EHR data for clinical trial recruitment or observational research. The coverage of rare disease names in specialized ontologies (e.g., OMIM) is higher, but these are not designed for use in clinical EHR systems.

Given the intended purpose for each classification and ontology and the completeness and coverage of rare disease names, we propose how these various clinical coding systems, ontologies, and UMLS mappings can be leveraged to support an efficient national research infrastructure and learning healthcare system. The UMLS is a vital tool to support the linkage across clinical coding systems and specialized ontologies that will be essential for a national EHR-based rare diseases research infrastructure.

Ontologies can support advances in understanding disease etiology and potential treatments. Specialized ontologies, such as OMIM, RDO, and others (such as the Human Phenotype Ontology) can provide the vocabulary for detailed clinical documentation , or “deep phenotyping”, of genetic diseases (e.g., in the NIH Undiagnosed Diseases Network), and complement clinical terminologies and administrative classifications widely used in EHRs. This poster will include an illustrative representation of the collection of rare disease-specific data in dedicated ontologies to support diagnosis, and the use of mappings to standardized clinical terminologies or classifications as needed for clinical documentation, data exchange, billing and public health reporting.

Acknowledgment

This work was partly supported by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine. This work was also supported in part by PCORnet, funded by the Patient Centered Outcomes Research Institute (PCORI).

References

[1] Orphanet. *Orphanet Rare Disease Ontology (ORDO)*. 2014 [cited 2014 March 14]; Available from: http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php.

[2] NORD. *Rare Disease Information*. 2014 [cited 2014 March 14]; Available from: http://www.rarediseases.org/rare-disease-information.

[3] NIH. *Office of Rare Diseases Research (ORDR) Brochure*. 2009 [cited 2010 20/08/2010]; Available from: http://rarediseases.info.nih.gov/asp/resources/ord_brochure.html.

[4] Fung, K.W., R.L. Richesson, and O. Bodenreider, Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research, in Paper accepted for presentation: American Medical Informatics Association Annual Symposium2014: Washington, D.C.

[5] PCORI. *PCORnet: The National Patient-Centered Clinical Research Network*. 2014 [cited 2014 March 13]; Available from: http://www.pcori.org/funding-opportunities/pcornet-national-patient-centered-clinical-research-network/.

Coverage of Rare Disease Names in Clinical Coding Systems and Ontologies and Implications for Electronic Health Records-Based Research

Rachel Richesson[1], Kin Wah Fung[2], Olivier Bodenreider[2] | [1]Duke University School of Nursing, Durham, NC, USA; [2]National Library of Medicine, Bethesda, MD, USA

ABSTRACT

This poster highlights clinical coding systems and ontologies relevant to rare diseases, including the International Classification of Diseases (ICD) and SNOMED CT, and ontologies such as the Human Phenotype Ontology (HPO) and the Orphanet Rare Diseases Ontology (ORDO). Using use cases and a set of rare diseases for the national Patient Centered Outcomes Research Network (PCORnet), the poster will describe the current capacity and implications for EHR-based research on these diseases. Authors will provide suggestions on where mappings across classifications and ontologies are needed to support the use of EHR data for various types of research related to rare diseases.

BACKGROUND

Rare diseases are defined in the US as conditions that affect less than 200,000 Americans and in the European Union as those with a prevalence of 5 per 10,000 or less. There is no globally authoritative list of rare diseases, but there are several online disease catalogues developed by reliable sources.[1-3] Although each rare disease is uncommon, collectively they are more common, and consequently 'rare diseases' have emerged as priority topics in public health and research.

Table 1. Sources of Rare Disease Names

Source	# of rare diseases
NCATS, Office of Rare Diseases Research (United States) [1]	6,485
Orphanet Rare Diseases Ontology [3]	6,750

Table 2. Terminologies, Coding Systems, and Ontologies with Coverage of Rare Diseases

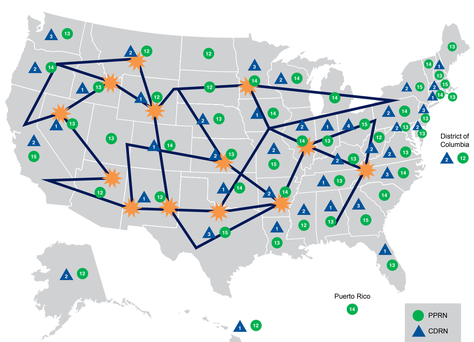
Terminology or Coding System	Sponsor	Intended Purpose	Estimated Coverage
International Classification of Diseases version 10 (ICD-10)	World Health Organization	Disease Surveillance; Mortality	12% [4]
International Classification of Diseases Clinical Modifications (ICD-CM, versions 9 and 10)	World Health Organization; national government/public health sponsors by country	Medical Billing	*Using UMLS-based method [5]:* 13% for ICD-9-CM 26% for ICD-10-CM
Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)	International Standards Development Organization http://www.ihtsdo.org/	Coding the clinical content of electronic health records to support patient care and other secondary data uses.	50 - 53% [4, 5]
Medical Dictionary for Regulatory Activities (MedDRA)	International Federation of Pharmaceutical Manufacturers and Associations (IFPMA); maintained and supported by MSSO	Adverse event reporting; regulatory submissions for new drugs and devices.	Unknown
Medical Subject Headings (MeSH)	U.S. National Library of Medicine	To index article topics for the published medical literature.	67% [4]
Online Mendelian Inheritance in Man (OMIM)	Distributed by the U.S. National Center for Biotechnology Information; Authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine	A catalog of human genes and genetic disorders and traits, with focus on the molecular relationship between genetic variation and phenotypic expression; considered a "phenotypic companion" to the Human Genome Project.	45% [4]
Human Phenotype Ontology (HPO)	Various translational and genetics research collaborators http://www.human-phenotype-ontology.org/	"Deep phenotyping" for EHRs in genetics and specialty clinics; support interoperability between current major genetics databases.	Unknown; presumably complete; 54% of HPO terms are in the UMLS [6]
Orphanet Rare Disease Ontology (ORDO)	Orphanet	A research resource for computational analysis and data mining/knowledge discovery for rare diseases. Supports editorial procedures of Orphanet knowledge bases and services.	100%
PhenX	U.S. National Human Genome Research Institute	Standard questions for clinical phenotyping data to support de novo data collection in GWAS studies.	N/A *(PhenX is for risk factors and environmental exposures)*
Unified Medical Language System (UMLS)	U.S. National Library of Medicine	The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.	Contains 8,435 rare disease names

Rare disease names are included, at different levels of completeness and granularity, in a number of clinical coding systems that are embedded in electronic health record (EHR) systems, and in a number of ontologies designed to support the diagnosis of rare diseases and investigation of genetic causes and treatments.

As was shown in Table 2, a range of coverage for rare disease names across coding systems has been reported using a variety of methods. In 2010, the NLM mapped 8,435 rare disease names (collected from ORDR, Orphanet, and the National Organization for Rare Disorders, a patient advocacy and voluntary health organization in the US) to the UMLS, and found different levels of coverage for Medical Subject Headings (MeSH) (5,663 ; 67%), Online Mendelian Inheritance in Man (OMIM) (3,802 ; 45%), SNOMED CT (4,192 ; 50%), and ICD-10 (1,029 ;12%).[4] More recently, we used the UMLS and the published maps from SNOMED CT to ICD-9-CM (developed by IHTSDO) and ICD-10-CM (developed by NLM).[5]

With increased adoption and "meaningful use" of EHRs, there is renewed effort in leveraging EHRs for research. The national Patient Centered Outcomes Research Network (PCORnet) was funded from the Affordable Care act to examine real-world treatment decisions, and is specifically tasked to conduct observational and interventional research on the comparative effectiveness of various treatments, using distributed and heterogeneous EHR systems. The PCORnet research portfolio currently includes 48 rare diseases and conditions.

Figure 1. The Patient Centered Outcomes Research Network (PCORnet) of Networks



METHODS

- We match rare diseases names and synonyms from the Office of Rare Disease Research and the Orphanet Rare Disease Ontology (ORDO) to the Unified Medical Language System (UMLS) Metathesaurus and identify maps to SNOMED CT and other terminologies.
- We estimate the number of precise and equivalent matches in the three clinical terminologies (ICD-9-CM, ICD-10-CM, and SNOMED CT) for a set of 48 rare diseases studied in PCORnet.
- We assess the precision of mapping by looking at the number of rare disease names that map to distinct codes in each terminology or coding system, and the equivalence by characterizing the semantic nature of the maps to determine whether the mapped term was broader, narrower, or equivalent to the PCORnet rare disease name.

RESULTS

- SNOMED CT has the highest coverage among clinical coding systems, and covers 44% of the 6,485 diseases recognized by the Office of Rare Diseases, and 28% of the 6,750 diseases that are listed in the Orphanet Rare Disease Ontology (ORDO).

Table 3. Coverage of Rare Diseases from 2 Sources by Coding Systems

Coding System		% coverage of 6,485 diseases from US NCATS/ORDR	% coverage of 6,750 diseases from Orphanet ORDO
UMLS		82%	62%
MeSH		75%	52%
OMIM		52%	41%
SNOMED CT	Clinical coding systems	44%	36%
ICD-9-CM		11%	7%
ICD-10-CM		21%	16%

- Overall, the UMLS covers 82% of ORD and 62% of ORDO-recognized rare diseases.
- Two terminologies have higher coverage than SNOMED CT: Medical Subject Headings (MeSH) covers 75% and 52%, while Online Mendelian Inheritance in Man (OMIM) covers 52% and 41%, of ORD and ORDO respectively.

- SNOMED CT covers 44% of ORD and 36% of ORDO-recognized rare diseases.
- 25% (1,611) of ORD and 14% (1,592) ORDO disease names have bi-directional one-to-one maps to SNOMED CT.
- The rest are one-to-many or many-to-one maps.

Examples

Many rare diseases map to the following codes:

759.89 Other specified congenital anomalies (ICD-9-CM)
Q82.8 Other specified congenital malformations of skin (ICD-10-CM)
→ **These are not "high precision" mappings**

Pseudoneonatal adrenoleukodystrophy maps to:
(238069004) Acyl-CoA oxidase deficiency (disorder) in SNOMED CT and is the only rare disease that does.
→ **This is a "high precision" map**

Joubert Syndrome maps to:

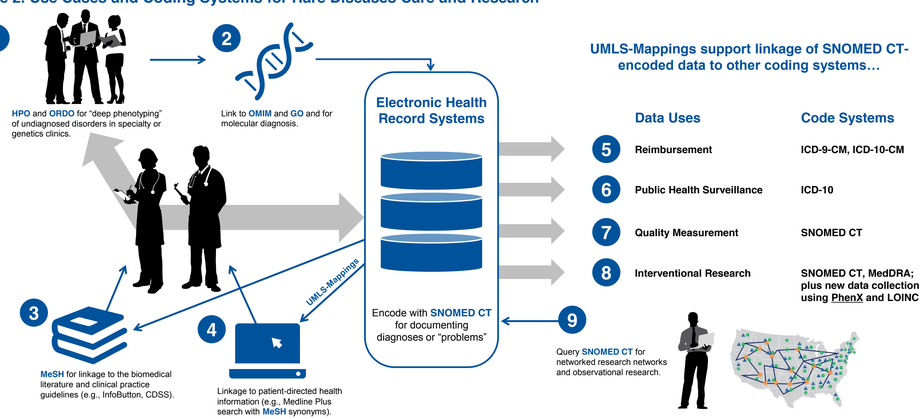
742.4 Other specified congenital anomalies of brain (ICD-9-CM) → **Not semantically "equivalent"**
742.9 Unspecified congenital anomaly of brain, spinal cord, and nervous system (ICD-9-CM) → **Not semantically "equivalent"**
Q04.3 Other reduction deformities of brain (ICD-10-CM) → **Not semantically "equivalent"**
253175003 Familial aplasia of the vermis (disorder) (SNOMED CT) → **Semantically "equivalent"**

- As shown in Table 4, of the 48 rare diseases studied in PCORnet, the number of rare diseases with one and only one match to coding system term (considered "high precision") was 87% for ICD-9-CM, 91% for ICD-10-CM, and 98% for SNOMED CT.
- Authors (RR, KWF) assessed the semantic nature of the maps to determine whether the mapped term was broader, narrower, or equivalent to the PCORnet rare disease name. The proportions of equivalent matches were 25%, 45% and 94% for ICD-9-CM, ICD-10-CM and SNOMED CT, respectively.

Table 4. Precision of Coverage of PCORnet Rare Diseases in Different Clinical Coding Systems

Coding System	% of PCORnet rare disease codes that did not map to other rare diseases (1-1 map) (Precision)	% of PCORnet rare disease codes considered an equivalent map (Equivalence)
ICD-9-CM	87%	25%
ICD-10-CM	91%	45%
SNOMED CT	98%	94%

Figure 2. Use Cases and Coding Systems for Rare Diseases Care and Research



DISCUSSION

The diagram in Figure 2 includes the collection of rare disease-specific data in dedicated ontologies to support diagnosis, the use of mappings to standardized clinical terminologies or classifications as needed for clinical documentation, data exchange, billing and public health reporting.

The current coverage of rare disease names in standard coding systems can support a number of use cases, including: the identification of rare disease patients from EHR data for research (#8 and #9 on figure), and the identification of appropriate rare diseases information, including published medical literature, clinical practice guidelines for providers (#3 on figure) and authoritative consumer-directed information for patients (#4 on figure), using coded data from EHRs.

Advances in the discovery of genetic causes and possible treatments can be supported by specific ontologies to the extent that they can be used in cooperation with EHR data coded in clinical coding systems (#1 and #2 on figure).

The differences in the coverage, intended purpose, and granularity of different coding systems can impact how EHRs can support the consistent and reliable identification of rare disease patients, to enable evidence-based care and multi-site research. The lines in the figure describe where mappings and linkages across terminologies are needed to support various use cases.

CONCLUSIONS

- The coverage of terms for rare diseases in clinical terminologies is highest with SNOMED CT in comparison to ICD 9 and 10 classifications.
- SNOMED CT has the greatest proportion of high-precision mappings and equivalent mappings in our sample of PCORnet rare diseases.
- The coverage of rare disease names in specialized ontologies is higher, but these are not designed for use in clinical EHR systems.
- **Understanding the intended purpose of each classification and ontology and the coverage of rare disease names can facilitate an efficient national research infrastructure and learning healthcare system.**
- Further, ontologies can support advances in understanding disease etiology and potential treatments.
- The UMLS is a vital tool to support the linkage across clinical coding systems and specialized ontologies that will be essential for a national EHR-based rare diseases research infrastructure.

REFERENCES

1. Orphanet. *The portal for rare diseases and orphan drugs*. 2014 March 12, 2014 [cited 2014 March 14]; Available from: http://www.orpha.net/consor/cgi-bin/index.php.
2. NIH. *Office of Rare Diseases Research (ORDR) Brochure*. 2009 [cited 2010 20/08/2010]; Available from: http://rarediseases.info.nih.gov/asp/resources/ord_brochure.html.
3. NORD. *Rare Disease Information*. 2014 [cited 2014 March 14]; Available from: http://www.rarediseases.org/rare-disease-information.
4. Pasceri, E. *Analyzing rare diseases terms in biomedical terminologies. (LHNCB Medical Informatics Training Program Final Report; Dr. Olivier Bodenreider, Mentor)*. 2010 [cited 2014 July 21]; Available from: http://mor.nlm.nih.gov/pubs/alum/2010-pasceri.pdf.
5. Fung, K.W., Richesson R.L., and O. Bodenreider, *Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research*, in *American Medical Informatics Association Annual Symposium*. 2014: Washington, D.C.
6. Winnenburg, R. and O. Bodenreider. *Coverage of phenotypes in standard terminologies*. Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day" 2014:41-44.

ACKNOWLEDGEMENTS

The authors thank PCORnet collaborators and staff for their support of strategies for using electronic health data to advance rare diseases research. This paper was supported by grants from the Patient Centered Outcomes Research Institute (PCORI) (P122013-499A) and the NIH Collaboratory (5 U54 AT007748-02). This work was partly supported by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine. The views expressed do not necessarily represent the views of the NLM, NIH, or PCORI.

Duke University School of Nursing

NIH U.S. National Library of Medicine