Integrating ontologies of rare diseases and radiological diagnosis

Charles E Kahn, Jr.

ABSTRACT

Purpose The author sought to integrate an ontology of rare diseases with a large ontological model of radiological diagnosis.

Materials and Methods The Orphanet Rare Disease Ontology (ORDO) comprised 6794 rare diseases. The Radiology Gamuts Ontology (RGO) incorporated 16 197 terms and 53 425 causal relations linking disorders to imaging manifestations. Semi-automated string-matching was used to match ORDO terms to RGO terms.

Results Of 6794 ORDO terms, 1587 (23.3%) were matched to RGO terms. An additional 700 ORDO terms whose names were hyphenated lists of phenotypic features were added to RGO with causal links from the disease name to the various features. Matched terms were more likely to have higher disease prevalence.

Conclusions Integrating these ontologies expanded the set of terms and scope of knowledge available for radiological differential diagnosis, and can support translational rare-disease research by linking knowledge of genetics and imaging phenotypes.

INTRODUCTION

The United States defines a rare disease, also called an “orphan disease,” as any disease or condition that affects fewer than 200 000 Americans,1 which corresponds to a prevalence of less than 1 in 1500 (0.06%). The European Commission on Public Health defines rare diseases those affecting fewer than 1 in 2000 people. Approximately 6800 rare diseases are known today.3 In aggregate, these disorders affect 25–30 million people in the United States—almost 1 in 10—and engender significant medical, social, and economic burdens.4–6 Radiology can play an important role in the diagnosis, classification, and understanding of these diseases; to that end, we sought to integrate knowledge models of rare diseases and radiological diagnosis.

To manage and apply evolving biomedical knowledge, scientists have turned to ontologies—formal knowledge models that comprise a set of concepts and their relationships to one another.7 Biomedical ontologies have been applied to knowledge management, data integration, and automated reasoning.8–10 More than 300 biomedical ontologies span basic, translational, and clinical science.11 For example, the Gene Ontology encodes knowledge of gene products with respect to their molecular function, cellular component, and biological role.12 In radiology, the RadLex ontology describes imaging procedures, diseases, radiological anatomy, and imaging observations.13

One of many strengths of the ontology framework is the ability to join related ontologies by mapping identical or related terms. In so doing, one is able to integrate the knowledge of the individual models into a broader, more useful knowledge resource. This article describes the efforts to integrate an ontology of rare diseases with an ontology of disorders and their radiological manifestations. This integration provides a foundation to integrate rare diseases into broader radiological differential-diagnosis listings and potentially to diagnose rare disorders more effectively.

MATERIALS AND METHODS

The Orphanet Rare Disease Ontology (ORDO), derived from the Orphanet database (www.orpha.net), integrates information about rare (“orphan”) diseases, genes, epidemiology, and orphan drugs.14,15 ORDO defines semantic relationships between genes and diseases: causal germ-cell or somatic mutations, modifier genes, susceptibility genes, fusion genes involved in causation of tumors, and genes with a major role in the phenotype of chromosomal abnormalities. ORDO version 1.0.20 (modified December 14, 2012) and a list of diseases in alphabetical order were downloaded November 18, 2013 from the www.orpha.net website. ORDO contained 6794 disease entries, which included information such as estimated prevalence, mode of inheritance, age of onset, and age of death.

The Radiology Gamuts Ontology (RGO) expresses relationships between diseases and their imaging manifestations.16 In radiology, the word “gamuts” has been used to denote comprehensive differential-diagnosis listings.17 RGO includes names of disorders (e.g., Chagas disease), interventions (e.g., steroid therapy), and imaging observations (e.g., splenomegaly). In addition to the conventional subsumption (is_a) relation, RGO defines a causal (may_cause) relation and its inverse (may_be_caused_by) to link disorders and their imaging manifestations. As of April 7, 2014, the Gamuts ontology, obtained with permission from the gamuts.net website, incorporated 16 197 terms, synonyms, and abbreviations, 1422 subsumption relations, and 53 425 causal relations.

A semi-automated process was developed to map ORDO terms to RGO. ORDO and RGO concepts were stored using a MySQL relational database system (version 5.4, Oracle Corporation, Redwood Shores, CA, USA) and the names of RGO terms were indexed as set of words using the FULLTEXT indexing method. A specialized program was created in the PHP programming language (version 5.4.19, The PHP Group) to present each ORDO term with up to 12 potentially matching RGO entities. A natural-language search method was applied, which excluded common stopwords (words such as “a,” “of,” and “the”). The words “disease” and “syndrome” also were excluded from the search process. The software allowed one to select a matching RGO entity or to mark the target term as unmatched. A physician with more than 20 years of experience in radiology and informatics reviewed each potential match and made the determination to assign a match or mark the ORDO term as unmatched. The matching algorithm was applied to each ORDO term.
The matching process included a search of known synonyms in both ORDO and RGO. For terms where no match was identified, references were consulted to identify any additional synonyms that could be used to identify a match. ORDO and RGO terms were matched if they were determined to designate the same disease, regardless of minor variations in the name (e.g., Pyle disease and Pyle dysplasia) or word order. If the terms were designated to match, the ORDO term and its synonyms were added as synonyms to RGO.

We separately identified ORDO terms in which the disorder was named by its features (e.g., *Agnathia – holoprosencephaly – situs inversus*) and tallied the number of matched and unmatched terms. For unmatched terms, a custom script was created to add the ORDO terms to RGO, and to assert causal relationships to the entities in each term’s name.

To test the hypothesis that matched ORDO entities would have greater prevalence than unmatched entities, we analyzed the frequency that matched and unmatched ORDO terms had a known prevalence value in the OrphaNet database. For the purposes of this analysis, we defined “unspecified” as those ORDO terms with prevalence values of “unknown” or “no data available” in their corresponding OrphaNet entry. For those ORDO terms with a known prevalence value, we tallied the number where that value was greater than or less than one in 1 million ($10^{-6}$).

**RESULTS**

Of the 6794 ORDO terms, 1587 (23.3%) matched an RGO term. There were 1062 lexically identical matches, such as acromicric dysplasia, autosomal recessive polycystic kidney disease, brittle cornea syndrome, congenital muscular dystrophy, gangliogioma, IMAGe syndrome, MURCS association, oral-facial-digital syndrome type 4, and Robinow syndrome. An additional 520 ORDO terms had a lexically inexact match to an RGO term. For these terms, the ORDO term and its synonyms were added as synonyms to RGO.

The remaining 700 “hyphenated” ORDO terms were added to RGO with may_cause relationships to existing RGO terms. For example, the ORDO term *Agnathia – holoprosencephaly – situs inversus* (Orpha Number 990) was added to RGO along with may_cause relationships to existing RGO terms *agathia, holoprosencephaly*, and *situs inversus*. Some features that comprised the “hyphenated” ORDO disease names were added to RGO as well. For example, the term *hemoglobin E disease* was added to RGO, as well as a may_cause relationship from *Sickle cell – hemoglobin E disease* (Orpha Number 251375). A total of 2297 ORDO disease names (33.8%) were mapped to RGO. Where necessary, the names of the imaging manifestations were normalized to assure consistency within RGO. For example, the ORDO term *Cataracts – aberrant oral frenula – growth retardation was linked to cataract* (rather than *cataracts*). The terms unusual facies and peculiar facies, which appeared in the names of four ORDO terms, were added to RGO as synonyms for facial dysmorphism.

The remaining 4,502 ORDO terms (66.3%) were marked as unmatched (Table 2). For example, the system attempted to match the ORDO term *Isolated 3-methylcrotonyl-CoA carboxylase deficiency* (Orpha Number 6) with RGO terms such as *pyruvate carboxylase deficiency, propionyl-CoA carboxylase deficiency, and isolated radial deficiency* based on substring similarity, but the ORDO term was designated as unmatched because none of the candidate RGO terms was synonymous.

Of 1587 matched ORDO terms, 779 (47.4%) had prevalence data; in contrast, 1662 (36.9%) of 4502 unmatched terms had prevalence data. Of the 779 matched ORDO terms with prevalence data, 369 (23.9%) had prevalence >$10^{-6}$, compared to 354 of 1662 (7.9%)
unmatched terms with prevalence data. Results are summarized in Figure 1. Chi-square tests with 1 degree of freedom showed both pairs of values to be significantly different ($P < .001$).

**DISCUSSION**

Integrating these two ontologies builds bridges across domains. ORDO is linked to terminologies such as Medical Subject Headings, Systematized Nomenclature of Medicine Clinical Terms, and the Unified Medical Language System. ORDO provides connections to scientific databases, such as Online Mendelian Inheritance in Man (OMIM) and the UniProt database, and disease classifications such as International Classification of Diseases, version 10, Clinical Modification (ICD-10-CM). RGO terms have been mapped to the RadLex radiology lexicon, and are supplemented with definitions from Wikipedia and by exemplary images retrieved by ARRS GoldMiner, a specialized biomedical image search engine. Work is underway to integrate RGO’s imaging findings with the Human Phenotype Ontology, a model created to provide abstraction about phenotypic manifestations of disorders specified by OMIM.

The 2287 ORDO disease names mapped to RGO expanded the set of terms and the scope of knowledge available through RGO. As part of the integration, RGO’s vocabulary was harmonized with ORDO. For example, RGO included the term *mental retardation* and its acronym *MR*, which were used in the names of several syndromes from RGO’s reference sources. To better harmonize the names of various findings, RGO was amended to use *intellectual deficit* as its preferred term, which matched the naming of ORDO syndromes more closely. The results of the current work also were used to update the gamuts.net website to incorporate links to matched ORDO terms (Figure 2). A potential limitation of the current study is that there may have been ORDO terms that were unmatched due to spelling variation or the absence of a synonym within one of the ontologies. There were no identical ORDO and RGO terms whose definitions differed. Periodic manual review will be needed to reconcile deletions or changes to mapped terms.

Integration of these knowledge resources allows RGO applications to apply ORDO knowledge by reference. One use-case of this integration is to enhance an interactive diagnostic decision support tool based on RGO with prior probabilities of diseases from ORDO to determine the most likely diagnoses and identify the most informative imaging features to discriminate among possible diagnoses. Links from ORDO to other knowledge resources, such as OMIM, allow RGO to incorporate knowledge about a disorder’s mode of inheritance and age of onset to better inform diagnostic decision making; such information need not be added directly into RGO, but rather can be applied by reference to mapped ORDO terms.

Investigators have used semantic features to identify relationships among disorders and have used ontology-based knowledge for diagnostic decision support. The availability of large datasets combining electronic health record data and gene sequence or expression data provides opportunities to use for associations between phenotypic data (including imaging phenotypes) to discover affected but undiagnosed subjects with rare diseases. One could consider using ORDO’s

![Figure 1: Histogram of prevalence of ORDO entities, grouped by those that were matched or unmatched to corresponding RGO terms. Matched ORDO entities were more likely to have higher prevalence (greater than 1 per million) and were less likely to have an unspecified prevalence value.](image)

**Table 2: Example ORDO terms not matched to RGO**

<table>
<thead>
<tr>
<th>Orpha Number</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>276066</td>
<td>Bile acid CoA ligase deficiency and defective amidation</td>
</tr>
<tr>
<td>228003</td>
<td>Severe combined immunodeficiency due to CORO1A deficiency</td>
</tr>
<tr>
<td>1773</td>
<td>Sacrococcygeal dysgenesis association</td>
</tr>
<tr>
<td>97563</td>
<td>Pauci-immune glomerulonephritis with ANCA</td>
</tr>
<tr>
<td>280663</td>
<td>Hermansky-Pudlak syndrome type 9</td>
</tr>
<tr>
<td>238475</td>
<td>Familial hypercholesterolemia</td>
</tr>
<tr>
<td>247868</td>
<td>NLRP12-associated hereditary periodic fever syndrome</td>
</tr>
<tr>
<td>99085</td>
<td>Coronary artery intramyocardial course</td>
</tr>
<tr>
<td>93220</td>
<td>Sporadic idiopathic steroid-resistant nephrotic syndrome with diffuse mesangial sclerosis</td>
</tr>
<tr>
<td>268388</td>
<td>Leptomyelolipoma</td>
</tr>
<tr>
<td>75326</td>
<td>Retinal arterial tortuosity</td>
</tr>
<tr>
<td>3417</td>
<td>Van den Bosch syndrome</td>
</tr>
</tbody>
</table>
Radiological diagnosis of rare disorders is challenging because the diseases and their manifestations may be unfamiliar to most radiologists. The large number of rare disorders means that although each disease is indeed rare, patients with a rare disease are quite common. Mapping the Orphanet Rare Disease Ontology to the RGO allows cross-domain knowledge interchange. Because morphological and genetic data are used for classification and diagnosis of rare diseases, it is hoped that the integration of these ontologies will improve clinical diagnosis and advance translational research.

FUNDING
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
COMPETING INTERESTS

None.

CONTRIBUTORS

The author is solely responsible for this work.

REFERENCES


AUTHOR AFFILIATION

Department of Radiology, Perelman School of Medicine, University of Pennsylvania, 3400 Spruce Street, 1 Silverstein, Philadelphia, PA 19104, USA, Tel: +1 215-662-3032, Fax: +1 215-662-7011