Proposal of a core set of terms to describe human phenotypes by the International Consortium of Human Phenotype Terminologies

Introduction

Collectively rare diseases constitute a substantial healthcare burden, although difficult to quantify as most coding nomenclatures do not include specific codes for these diseases. Although each individual disease is rare (defined in the EU as affecting <1 in 2000 individuals), the summative impact of these disorders on patients, families and healthcare systems must be huge because there are several thousands of such diseases. The inventory of rare diseases maintained by Orphanet includes over 6,000 rare clinical entities, of which most are Mendelian disorders. OMIM catalogs about 7500 Mendelian phenotypes, including disorders and traits, such as drug metabolism, not all of which are rare. OMIM also splits diseases by molecular basis, e.g., Noonan syndrome is represented once in Orphanet, but with 8 separate entries in OMIM, split by the gene responsible.

Genome-wide sequencing is a disruptive technology that is transforming our understanding of the genetic basis of many rare diseases. To capitalize on this technology, it is important that researchers and clinicians speak interoperable languages when describing phenotypes that are observed in genetic and rare diseases. This will facilitate the data-exchange and meta-analysis that underpins collaborative research to improve human health.

Standards have become an indispensible feature of biomedical research on almost every level, ranging from mutation nomenclature (cite HGVS), exchange formats for high-throughput data, and ontologies for a wealth of domains in biomedical research (OBO Foundry paper). Nonetheless, it has been substantially more difficult to develop and disseminate standards for the description of human phenotypic abnormalities in genetic and rare diseases that would enable exchange of data, sophisticated search algorithms for biomedical literature, and the comprehensive and accurate documentation of phenotypic abnormalities in hospital information systems. There are many reasons for this, including first and foremost the complexity of the topic, the disparate needs of different stakeholders (patients, researchers, clinicians, insurers, and public health officials), to say nothing of the fact that different clinical subspecialties examine, and therefore describe the human phenotype in different ways.

Current developments in biomedical research and the health care system of many countries have led to a massive increase in the role of informatics at many levels. With the advent of next-generation sequencing, medicine, and especially human genetics, is becoming a data-driven field and many efforts are instituting large-scale sequencing programs with the goal of identifying the genetic etiology of all forms of Mendelian diseases.

On the other hand, electronic health records (EHR) are being widely adopted to increase efficiency and improve quality of care, and the potential of intelligent EHR systems to provide physicians with information on which to base clinical care decisions is enormous. The next edition of the International Classification of Diseases (ICD-11) is in preparation, and the representation of rare diseases in ICD-11 will influence the way physicians, but also hospital administrators and national policy makers, view rare diseases for decades to come (REF ICD11 paper).

Therefore, there is now a window of opportunity for introducing a terminological representation for the field of genetic and rare diseases into health IT systems. To identify the most realistic and acceptable solution, a group of interested-parties set up an International Consortium to develop a way forward to best serve the needs of patients as well as those of specialist and non-specialist physicians, researchers, and policy makers. This work is now complete and its outcome is presented in this article.

Methods

The consortium was set up by inviting to a workshop, in Paris in September 2012, the project leaders of terminologies in use among the genetic and rare diseases community of researchers and expert clinicians. The purpose was to explore the current state of the terminologies then to discuss the possible options for these terminologies to converge toward a common core terminology.

Given the multitude of needs and applications in the field of genetic and rare diseases, it is not currently realistic or even desirable to have one terminology for all applications in the way that, say, Gene Ontology has become the lingua franca for describing functions, processes, and localizations of gene products. Prominent terminologies for genetic medicine have different focuses and user bases. The Orphanet thesaurus of signs and symptoms is intended for use by clinicians who are not necessarily geneticists, while the London Dysmorphology Database and POSSUM have terminologies used to index pictorial atlases of genetic disease. The Human Phenotype Ontology has been developed to enable computational analysis of human disease manifestations. PhenoDB was intended to enable quick entry of phenotypic features by clinicians (or allied health care providers). The Elements of Morphology group has developed a glossary of state of the art definitions for about 750 phenotypic features related to dysmorphology. Generalist systems, such as the ICD and SNOMED CT, do not currently comprise many of these features, meaning that they are generally coded, if at all, using more general concepts.

A consensus strategy was defined and its deployment was assigned to one of the represented group (Orphanet).

The consensus strategy was the following:

1- Given the multitude of needs and applications in the field of rare diseases, it is not currently realistic or even desirable to have one terminology for all applications. Prominent terminologies have different focuses and user bases. The Orphanet thesaurus of signs and symptoms is intended for use by clinicians who are not necessarily geneticists. The Human Phenotype Ontology has been developed to enable computational analysis of human disease manifestations. PhenoDB is intended to enable quick entry of phenotypic features by clinicians (or health care providers). The Elements of Morphology is a glossary of state of the art definitions for phenotypic features. Generalist systems such as the ICD and SNOMED CT do not currently comprise many of these features, meaning that they are coded, if at all, using more general concepts;

2- The expert group will agree upon a core set of about 2 000 terms that represent the major phenotypic abnormalities encountered in persons with rare diseases which will be cross-matched with the available terminologies. This core set of terms will be recommended for use in any new information system intended to collect phenotypic data, either for research or clinical purposes. They will be published together with definitions;

3- The core set of phenotypic terms will be set up by comparing the different terminologies, considering that terms used by the majority of them are likely to constitute the candidates for standard terms. The Orphanet team will carry out the preparatory work and the expert group will act as reviewers and decision makers to ensure that there is a good coverage of all body systems for which descriptors are needed;

4- This set of terms will be proposed for inclusion in SNOMED CT and ICD-11;

5- As there is a need to continuously revise the proposal, the expert group proposes to set up an International Consortium of Human Phenotype Terminologies. Therefore, the core set of terms will be named ICHPT codes.

A second workshop was organized in Boston in October 2013 to review in detail the proposal prepared. When the list of terms to be included in any nomenclature was finalized, it was sent to every participating terminology for inclusion, to allow cross-referencing terminologies for inter-operability between phenotype databases.

Results

The groups responsible for the above named terminology have agreed upon a core set of terms that represent the major phenotypic abnormalities encountered in persons with rare diseases.

The corresponding terms in our ontologies and terminologies have been supplied with cross references such that it will now be a trivial task to integrate data from many different kinds of databases. The core terms are identified by generating mappings between the various terminologies using text mining followed by manual curation. The mappings are based on term names and where available synonyms and definitions. In this way, all of the terms from all of the terminologies are ranked according to the number of terminologies in which they were found. We reasoned that terms found in multiple terminologies were likely to be used by a wide spectrum of stakeholders and therefore deserve to be represented in the core set of terms. These terms were reviewed by hand, and a preferred name, synonyms and definition were agreed upon in a series of telephone conferences and meetings. Our goal was to choose a set of core terms that would allow a "balanced" description of the features of rare disease, and the final list of terms was determined based on both the frequency of use of the terms in the various terminologies, and also, subjectively, to provide terms to describe arbitrary phenotypes at a level of granularity that is optimal for communication between medical specialties (for instance, a retinal specialist might describe an abnormality of the retinal artery differently to a geneticist than to another retinal specialist).

We are working to get this set of terms included in SNOMED CT and ICD-11. Once this has been done, then information flow from EHRs to researchers and back, and between different databases in human genetics and other fields of medicine, will be dramatically improved. For the first time, there will be an adequate representation of genetic and rare diseases in hospital IT systems, which ideally will lead policy makers to appropriate resources according to the true frequency of rare diseases amongst patients. Although different terminologies will continue to have different representations of many specialized areas of the human phenotype, the agreement about a core set of terms will mean that it is possible to integrate data from specialist and generalist databases for research purposes, e.g., to search for additional patients carrying a mutation in a suspected novel disease gene in several databases. If researchers require resources offered by only one of the terminologies that have agreed on the core set, they can use their preferred terminology but still be able to integrate their data at a certain level of granularity simply by subsuming annotations to the level of the common terms.

Table 1: List of terminologies which were considered to establish the core set of terms to describe phenotypes of patients with genetic and rare diseases

Terminology name	Representatives
Human Phenome Ontology	Peter Robinson
Orphanet terms	Ana Rath and Ségolène Aymé
PhenoDB terms	Ada Hamosh
SNOMED CT	Jan-Eric Slot
ICD10	Robert Jakob
LDDB	Raoul Hennekam
POSSUM	Catherine Rose
Elements of Morphology	Raoul Hennekam
UMLS	-