Phenotypes in Standard Terminologies

Bridging the gap between clinical practice and research

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA

June 10, 2015
Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.
Outline

- Phenotype terms in the Human Phenotype Ontology (HPO)
- Biomedical terms in the Unified Medical Language System
- 3 approaches to mapping phenotype terms to SNOMED CT
  - Looking for equivalent terms
    - Pre-coordinated terms in UMLS
    - Post-coordinated terms in SNOMED CT
  - Looking for broader terms
- Challenges
Examples

HPO

- Multicystic kidney dysplasia
- Hypoplasia of the 3\textsuperscript{rd} metatarsal
- \textbf{Recurrence} urinary tract infection

SNOMED CT

- Multicystic kidney
- Based on “Congenital hypoplasia of kidney”
- Urinary tract infection [partial match]
Loosely based on 3 papers


Introduction

◆ Phenotyping is crucial to understanding how genetic variation relates to clinical manifestations
  ● Precise phenotyping is required for the study of rare syndromes
  ● Poor interoperability of phenotypic data
    ▪ Across clinical data repositories
    ▪ Between research and clinical data repositories
Phenotype terminologies
Human Phenotype Ontology

- Developed collaboratively
  - Coordination: Peter Robinson
- Nightly builds
- Distributed as an OWL file
- 10,589 classes (as of Jan. 21, 2015)
- 16,608 names for phenotype
  - One preferred term for each class
  - 6019 exact synonyms
- Cross-references to standard terminologies
# Human Phenotype Ontology

## Label

<table>
<thead>
<tr>
<th>Name</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicystic kidney dysplasia</td>
<td>Multicystic dysplastic kidney</td>
</tr>
<tr>
<td></td>
<td>Multicystic kidneys</td>
</tr>
<tr>
<td></td>
<td>Multicystic renal dysplasia</td>
</tr>
</tbody>
</table>

## Identifier

<table>
<thead>
<tr>
<th>Primary ID</th>
<th>Alternative IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP:0000003</td>
<td>HP:0004715</td>
</tr>
</tbody>
</table>

## Definition

**Textual Definition:**

Multicystic dysplasia of the kidney is characterized by multiple cysts of varying size in the kidney and the absence of a normal pelvocaliceal system. The condition is associated with ureteral or ureteropelvic atresia, and the affected kidney is nonfunctional.

**Logical Definition:**

- `intersection_of: PATO:0002089`
- `intersection_of: inheres_in UBERON:0002113`
### Annotation of phenotypes in OMIM and OrphaNet

<table>
<thead>
<tr>
<th>Condition</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>#508836 CARNITINE PALMITOYLTRANSFERASE II DEFICIENCY, LETHAL NEONATAL; CARNITINE PALMITOYLTRANSFERAS...</td>
<td>CPT2</td>
</tr>
<tr>
<td>#512513 CHROMOSOME 2P16.1-P15 DELETION SYNDROME</td>
<td>-</td>
</tr>
<tr>
<td>#514209 MECKEL SYNDROME, TYPE 9; MKS9</td>
<td>B9D1</td>
</tr>
<tr>
<td>#514527 CHROMOSOME 17Q12 DELETION SYNDROME</td>
<td>-</td>
</tr>
<tr>
<td>MOSAIC VARIEGATED ANEUPLOIDY SYNDROME</td>
<td>CEP57; BUB1B; BUB1; BUB3</td>
</tr>
<tr>
<td>BOR SYNDROME</td>
<td>SIX5; SIX1; EYA1</td>
</tr>
<tr>
<td>BARDET-BIEDE SYNDROME</td>
<td>MKKS; SDCCAG8; WDPCP; BBS5; BBS1; TRIM32; BBS2; IFT27; ARL6; BBS4; CEP290; BBS12; LZTFL1; MKS1; BBS10; BBIP1; NPHP1; BBS7; IFT172; BBS9; TTC8</td>
</tr>
<tr>
<td>SHORT RIB-POLYDACTYLY SYNDROME</td>
<td>-</td>
</tr>
<tr>
<td>CARNITINE PALMITOYLTRANSFERASE II DEFICIENCY</td>
<td>-</td>
</tr>
<tr>
<td>NON-RHIZOMELIC CHONDRODYSPLASIA PUNCTATA</td>
<td>-</td>
</tr>
<tr>
<td>RENAL DYSPLASIA - MEGALOCYSTIS - SIRENOMELIA</td>
<td>-</td>
</tr>
<tr>
<td>INDOMETHACIN EMBRYOCFETOPATHY</td>
<td>-</td>
</tr>
</tbody>
</table>
Annotation of phenotypes in OrphaNet

**ORPHA110**

**Synonym(s):** BBS

**Prevalence:** 1/8 / 1 000 000

**Inheritance:** Oligogenic or Autosomal recessive

**Age of onset:** Infancy Neovatal Antenatal

**ICD-10:** OMM:

**Q87.6**

209901 [1]

60151 [1]

60523 [1]

615981 [2]

615982 [1]

615983 [1]

615984 [1]

615985 [1]

615986 [1]

615987 [1]

615988 [1]

615989 [1]

615990 [1]

615991 [1]

615992 [1]

615993 [1]

615994 [1]

615995 [1]

615996 [1]

**UMLS:**

C0752166

**MeSH:**

D020788

**MedDRA:**

10056715

**SUMMARY**

Bardet-Biedl syndrome (BBS) is a ciliopathy with multisystem involvement. Its prevalence in Europe is estimated at between 1/125,000 and 1/175,000. This disorder is characterized by a combination of clinical signs: obesity, pigmentary retinopathy, post-anal polydactyly, polycystic kidneys, hypogonadism, and learning disabilities, many of which appear several years after disease onset. Clinical expression is variable but most patients manifest the majority of clinical signs during the disease course. Pigmentary retinopathy is the only constant clinical sign after childhood. BBS may also be associated with several other manifestations including diabetes, hypertension, congenital cardiac hypertrophy and Hirschsprung disease (see this term). The wide clinical spectrum observed in BBS is associated with significant genetic heterogeneity. The disorder is transmitted mainly in an autosomal recessive manner but oligogenic inheritance has been reported in some cases. To date, mutations in 12 different genes (BBS1 to BBS12) have been identified as being responsible for this phenotype. These genes code for proteins involved in the development and function of primary cilia. Absence or dysfunction of BBS proteins results in ciliary anomalies in organs such as the kidney or eye. However, the relationship between symptoms and ciliary dysfunction remains obscure for some of the clinical manifestations of BBS. Recognition of the clinical picture is important as the diagnosis can be confirmed by molecular analysis, allowing appropriate genetic counseling for family members and possible prenatal diagnosis. The differential diagnosis should include the Alström, McKusick-Kaufmann and Meckel-Gruber syndromes (see those terms). Patients with BBS will need multidisciplinary medical care. The renal abnormalities are the main life-threatening manifestations because they can lead to end-stage renal failure and require renal transplantation. Progressive vision loss due to retinal dystrophy, together with moderate intellectual deficit (when present), behavioral anomalies, hypomimia and obesity will affect the social life of these patients.
Biomedical terminologies
Unified Medical Language System (UMLS)

- Terminology integration system
- Developed by NLM
- Integrates many (140) standard biomedical terminologies
  - SNOMED CT
  - MeSH
  - International Classification of Diseases
  - MedDRA
- 3M concepts
- 8M normalized terms
Integrating subdomains

Clinical repositories

Genetic knowledge bases

SNOMED CT

OMIM

Biomedical literature

HPO

MeSH

Genome annotations

NCBI Taxonomy

FMA

GO

Anatomy

Phenotypes

Model organisms

UMLS
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Phenotypes
Terminology integration

Male pseudohermaphroditism (111332007)

Phenotypes

Male pseudohermaphroditism (HP_0000037)

Clinical repositories

SNOMED CT

OMIM

Biomedical literature

Male Pseudohermaphroditism (D058490)

Genetic knowledge bases

Genome annotations

Clinical repositories

Model organisms

Anatomy

GO

FMA

NCBI Taxonomy

UMLS

C0238395

Male pseudohermaphroditism (111332007)

Biomedical literature

Male pseudohermaphroditism (HP_0000037)
SNOMED CT

- Developed by the International Health Terminology Standard Development Organization
- Description logics formalism
  - Supports post-coordination
- Broad coverage of clinical medicine
  - ~300,000 concepts
- Clinical findings
  - ~100,000 concepts
  - 169,000 names
- Integrated in the UMLS
SNOMED CT browser
Post-coordinated expression

32659003  
Congenital hypoplasia of kidney (disorder)

44513007  
Congenital anomaly of the kidney (disorder)

246454002  
Occurrence (attribute)

255399007  
Congenital (qualifier value)

116676008  
Associated morphology (attribute)

55199003  
Hypoplasia (morphologic abnormality)

363698007  
Finding site (attribute)

64033007  
Kidney structure (body structure)
Coverage of human phenotypes in biomedical terminologies

(a) Finding equivalent concepts in biomedical terminologies (pre-coordinated terms)
Methods

HPO

terms

xrefs

UMLS

SNOMED CT

MeSH

...
Male pseudohermaphroditism (HP_0000037)

**HPO**

- terms
- xrefs

oboInOwl:hasDbXref
umm:0238395

**UMLS**

C0238395

Snomed CT

- Male pseudo-hermaphroditism (11132007)

MeSH

- Male Pseudo-hermaphroditism (D058490)

...
HPO terms covered by SNOMED CT (xref)

- Male pseudohermaphroditism (HP_0000037)
  - Lexical UMLS: C0238395
  - Xref UMLS: C0238395
  - Mapping to SNOMED CT: Male pseudo-hermaphroditism (111332007)
  - Xref SNOMED CT: Male pseudo-hermaphroditism (111332007)

- Spinal canal stenosis (HP_0003416)
  - Lexical UMLS: C0039144
  - Xref UMLS: C0039144
  - Mapping to SNOMED CT: Spinal stenosis (76107001)
  - Xref SNOMED CT: Spinal stenosis (76107001)
HPO terms covered by SNOMED CT (no xref)

- **Atrial fibrillation (HP_0005110)**
  - Lexical UMLS: C0004238
  - Xref UMLS: C0004238
    - Mapping to SNOMED CT: Atrial fibrillation (49436004)
  - Xref SNOMED CT: none

- **Inlet ventricular septal defect (HP_0011622)**
  - Lexical UMLS: C0221215
    - Through MeSH, SNOMED CT, ICD10-CM
    - Mapping to SNOMED CT: Common atroventricular canal (360481003)
  - Xref UMLS: none
  - Xref SNOMED CT: none
HPO terms NOT covered by SNOMED CT

- **Palmoplantar keratoderma (HP_0000982)**
  - Lexical UMLS: C0022596
    - Through MeSH, MedDRA, ICD10-CM
  - Xref UMLS: none
  - Xref SNOMED CT: none

- **Hypoplastic nasal septum (HP_0005104)**
  - Lexical UMLS: C1861328
    - Through OMIM
  - Xref UMLS: C1861328
    - Mapping to SNOMEDCT: none
  - Xref SNOMED CT: none
HPO terms NOT covered by UMLS

- **Oval transradiancy** (humeral) (HP_0003877)
  - Lexical UMLS: none
  - Xref UMLS: none
  - Xref SNOMED CT: none

- **Lower limb peromelia** (HP_0009820)
  - Lexical UMLS: none
  - Xref UMLS: none
  - Xref SNOMED CT: none
Results

- UMLS: 54
- SNOMED CT: 30
- Consumer Health Vocabulary: 24
- MedDRA: 24
- MeSH: 19
- NCI thesaurus: 16
- ICD-10-CM: 15
- ICD-9-CM: 9
- Omim: 9
- MedlinePlus: 6
- ICD-10: 5
- ICD-9-CM: 0

% HPO concepts covered and % HPO concepts with Cross-references
Summary

◆ **Limited coverage of HPO phenotypes**
  - In UMLS: 50%
  - In specific biomedical terminologies
    - Highest is SNOMED CT with 30%

◆ **Insufficient for interoperability between research and healthcare datasets**

◆ **Integrating HPO into UMLS would benefit**
  - UMLS by increasing its coverage of phenotypes
  - HPO by increasing/keeping up to date its set of cross-references
Coverage of human phenotypes in biomedical terminologies

(b) Expressing HPO concepts through post-coordinated expressions in SNOMED CT
SNOMED CT already provides pre-coordinated concepts (and logical definitions) for some 3000 phenotypes
- Which can be used as templates for other phenotypes

Some phenotype concepts are more specialized than corresponding concepts in SNOMED CT
- By removing modifiers in a controlled fashion, we map demodified phenotype terms to SNOMED CT

Demodified HPO terms can be associated with SNOMED CT phenotype templates to create post-coordinated expressions
Methods Overview

- Transformation rules for phenotype terms
  - Replace
  - Split
  - Demodification to disorder (D0)
  - Demodification to anatomy, physiology or chemical substance (D1)
- Templates for phenotype concepts
- Mapping transformed HPO terms to SNOMED CT
- Creating post-coordinated expressions for HPO terms
1. Replace

- Replace the abbreviations for ordinals by their expanded form

- **Example**
  - “Aplasia of the phalanges of the 4th toe”
  - → “Aplasia of the phalanges of the fourth toe”
2. Split

◆ Split expressions with words concatenated by "/" into two individual phrases

◆ Example
  ● “aplasia/hypoplasia of the thymus”
  ● → 2 phrases
    ▪ “aplasia of the thymus”
    ▪ “hypoplasia of the thymus”
3. Demodification to disorder (D0)

- HPO contains modifiers that specialize disorders
- By removing these modifiers, we create a more general disorder concept
- For example
  - “bilateral congenital cataract”
    - Modifier “congenital”
  - → “bilateral cataract”
4. Demodification to other types (D1)

- Modifiers denoting a disorder of a specific anatomical structure, physiological process or chemical substance
- By removing these modifiers, we extract the object of the disorder
- Examples
  - “Abnormality of the lip” → “lip”
  - “Abnormality of intracranial pressure” → “intracranial pressure”
  - “Abnormality of prothrombin” → “prothrombin”
Logical definitions for phenotype concepts in SNOMED CT

◆ “Congenital hypoplasia of kidney”

◆ 'Disease (disorder)' and

  ● ('Role group (attribute)' some (  
    ● ('Associated morphology (attribute)'  
      — some 'Hypoplasia (morphologic abnormality)'  
    )  
    and ('Occurrence (attribute)'  
      — some 'Congenital (qualifier value)'  
    )  
    and ('Finding site (attribute)'  
      — some 'Kidney structure (body structure)'  
    )))
Template for phenotype concepts in SNOMED CT

◆ “Congenital hypoplasia of <ANATOMY>”

◆ 'Disease (disorder)' and

  ● ('Role group (attribute)' some (  
    ● ('Associated morphology (attribute)'
      ● some 'Hypoplasia (morphologic abnormality)')
    ● and ('Occurrence (attribute)'
      ● some 'Congenital (qualifier value)')
    ● and ('Finding site (attribute)'
      ● some '<ANATOMY> (body structure)'))


## Templates for phenotype concepts in SNOMED CT

<table>
<thead>
<tr>
<th>Template</th>
<th>Logical definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>{abnormal, abnormality of, abnormality of the, abnormality involving the}</td>
<td>'Disease (disorder)' and ('Role group (attribute)' some ( &quot;Associated morphology (attribute)&quot; some 'Developmental anomaly (morphologic abnormality)') and ('Occurrence (attribute)' some 'Congenital (qualifier value)') and ('Finding site (attribute)' some <code>&lt;ANATOMICAL STRUCTURE&gt;</code>))</td>
</tr>
<tr>
<td>&lt;ANATOMICAL STRUCTURE&gt; {abnormal}</td>
<td></td>
</tr>
<tr>
<td>{aplastic, aplasia of, aplasia involving the, absence of</td>
<td>'Disease (disorder)' and ('Role group (attribute)' some (&quot;Associated morphology (attribute)&quot; some 'Congenital absence (morphologic abnormality)') and ('Occurrence (attribute)' some 'Congenital (qualifier value)') and ('Finding site (attribute)' some <code>&lt;ANATOMICAL STRUCTURE&gt;</code>))</td>
</tr>
<tr>
<td>ANATOMICAL STRUCTURE&gt; {agensis, aplasia, absent}</td>
<td></td>
</tr>
<tr>
<td>{congenital absence of, congenital aplasia of }&lt;ANATOMICAL STRUCTURE&gt;</td>
<td></td>
</tr>
<tr>
<td>{hypoplastic, hypoplasia of, hypoplasia of the, hypoplasia involving the,</td>
<td>'Disease (disorder)' and ('Role group (attribute)' some (&quot;Associated morphology (attribute)&quot; some 'Hypoplasia (morphologic abnormality)') and ('Occurrence (attribute)' some 'Congenital (qualifier value)') and ('Finding site (attribute)' some <code>&lt;ANATOMICAL STRUCTURE&gt;</code>))</td>
</tr>
<tr>
<td>hypoplasia involving the, hypoplasia affecting the} &lt;ANATOMICAL STRUCTURE&gt;</td>
<td></td>
</tr>
<tr>
<td>{congenital hypoplasia of} &lt;ANATOMICAL STRUCTURE&gt;</td>
<td></td>
</tr>
<tr>
<td>{duplication of, duplication of the, duplication involving} &lt;ANATOMICAL</td>
<td>'Disease (disorder)' and ('Role group (attribute)' some (&quot;Associated morphology (attribute)&quot; some 'Double structure (morphologic abnormality)') and ('Occurrence (attribute)' some 'Congenital (qualifier value)') and ('Finding site (attribute)' some <code>&lt;ANATOMICAL STRUCTURE&gt;</code>))</td>
</tr>
<tr>
<td>STRUCTURE&gt; {duplication}</td>
<td></td>
</tr>
<tr>
<td>{complete duplication of, complete duplication of the} &lt;ANATOMICAL</td>
<td>'Disease (disorder)' and ('Role group (attribute)' some ... <em>D1 logical definition here ...</em> and ('Finding site (attribute)' some 'left &lt;ANATOMICAL STRUCTURE&gt;') and ('Finding site (attribute)' some 'right &lt;ANATOMICAL STRUCTURE&gt;'))</td>
</tr>
<tr>
<td>STRUCTURE&gt;</td>
<td></td>
</tr>
<tr>
<td>{bilateral, D1} &lt;ANATOMICAL STRUCTURE&gt; examples:</td>
<td></td>
</tr>
<tr>
<td>{bilateral aplasia} &lt;ANATOMICAL STRUCTURE&gt; {aplasia}</td>
<td></td>
</tr>
<tr>
<td>{bilateral absence of} &lt;ANATOMICAL STRUCTURE&gt;</td>
<td></td>
</tr>
</tbody>
</table>
Post-coordinated expressions created

- 12 templates involving 10 $D1$ modifiers
  - and 3 additional $D0$ modifiers removed when necessary
- Post-coordinated expressions for 1,617 HPO classes

- HPO “Macular hypoplasia” [HPO:HP_0001104]
  - $D1$ modifier “hypoplasia”
  - template “Congenital hypoplasia of <ANATOMICAL STRUCTURE>”
  - Post-coordinated expression created by substituting “Macula lutea structure (body structure)”
Summary

- With 12 post-coordination templates, we generated post-coordinated mappings to SNOMED CT for 1617 HPO concepts
- In complement to the ~3000 HPO concepts for which there is a pre-coordinated mapping to SNOMED CT
- Template-based mappings are usually high-quality
- Additional templates could be investigated
Coverage of human phenotypes in biomedical terminologies

(b) Finding more general concepts in SNOMED CT
Methods

Intuition

- HPO phenotype concepts are often more specialized than corresponding concepts in SNOMED CT
  - By removing modifiers in a controlled fashion, we map demodified phenotype terms to SNOMED CT
- Demodified HPO terms can be associated with SNOMED CT pre-coordinated concepts
- Linguistically-motivated approach (shallow parsing)
Example

Lexico-syntactic analysis
- Multicystic\textsubscript{mod} kidney\textsubscript{mod} dysplasia\textsubscript{head}

Remove modifiers
- 1 modifier: Multicystic dysplasia
- 1 modifier: kidney dysplasia
- 2 modifiers: dysplasia

Map demodified terms to UMLS/SNOMED CT
- Multicystic dysplasia $\rightarrow$ no mapping
- kidney dysplasia $\rightarrow$ Renal dysplasia
- dysplasia $\rightarrow$ [Dysplasia]
Frequent lexico-syntactic profiles

<table>
<thead>
<tr>
<th>Lexico-syntactic profile</th>
<th>Freq.</th>
<th>%</th>
<th>Mapping</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[MOD – HEAD]</td>
<td>4722</td>
<td>28</td>
<td>2898</td>
<td>61</td>
</tr>
<tr>
<td>[MOD – MOD – HEAD]</td>
<td>2416</td>
<td>15</td>
<td>1095</td>
<td>45</td>
</tr>
<tr>
<td>[HEAD]</td>
<td>2294</td>
<td>14</td>
<td>1989</td>
<td>87</td>
</tr>
<tr>
<td>[HEAD] [PREP – DET – HEAD]</td>
<td>767</td>
<td>5</td>
<td>163</td>
<td>21</td>
</tr>
<tr>
<td>[MOD – MOD – MOD – HEAD]</td>
<td>549</td>
<td>3</td>
<td>149</td>
<td>27</td>
</tr>
<tr>
<td>[HEAD] [PREP – MOD – HEAD]</td>
<td>432</td>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>[MOD – HEAD] [PREP – HEAD]</td>
<td>392</td>
<td>2</td>
<td>151</td>
<td>39</td>
</tr>
<tr>
<td>[HEAD] [PREP – HEAD]</td>
<td>383</td>
<td>2</td>
<td>83</td>
<td>22</td>
</tr>
</tbody>
</table>
Frequent modifiers and head nouns

◆ Modifiers
  - Abnormal
  - Increased
  - Decreased
  - Absent

◆ Head nouns
  - Hypoplasia
  - Abnormality
  - Atrophy
  - Weakness
  - Ossification
Lexical mappings between HPO and SNOMED CT

- No mapping: 5228
- Complete mapping: 3090
- Partial mapping (level 1): 1811
- Partial mapping (level 2): 258
- Partial mapping (level 3): 49
- Partial mapping (level 4+): 18

5226 (50%)
2136 (20%)
Summary

- We found partial mappings for ~2000 HPO concepts for which no pre-coordinated mapping exists.
- In complement to the ~3000 HPO concepts for which there is a pre-coordinated mapping to SNOMED CT.
Ontological issues

◆ Anatomical structures as phenotypes
  ● Small kidney vs. Hypoplasia of kidney
    ◦ Small kidney isa Kidney (anatomical structure)
    ◦ Hypoplasia of kidney isa Hypoplasia (clinical finding)
  ● Ductus arteriosus
    ◦ Syn. For Patent ductus arteriosus (condition)
    ◦ Ductus arteriosus is a normal anatomical structure in the fetus
    ◦ Its persistence after birth is abnormal
Conclusions

- Need for greater interoperability between terminologies used for research and for healthcare
- Need for greater collaboration between the developers of OBO-style ontologies and clinical terminologies

- HPO being integrated into the UMLS
  - Increase the coverage of phenotypes in UMLS
  - Facilitate the development and maintenance of cross-references in HPO
Medical Ontology Research

Contact: olivier@nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA