

Implementation SIG

March 8, 2016



Interoperability between phenotypes in research and healthcare terminologies

Investigating partial mappings between HPO and SNOMED CT



Olivier Bodenreider

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



U.S. National Library of Medicine



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.



Reference

Dhombres and Bodenreider *Journal of Biomedical Semantics* (2016) 7:3
DOI 10.1186/s13326-016-0047-3

Journal of
Biomedical Semantics

RESEARCH

Open Access



Interoperability between phenotypes in research and healthcare terminologies—Investigating partial mappings between HPO and SNOMED CT

Ferdinand Dhombres and Olivier Bodenreider*



Additional references

- ◆ Winnenburg R, Bodenreider O. Coverage of phenotypes in standard terminologies. Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day" 2014:41-44.
- ◆ Dhombres F, Winnenburg R, Case JT, Bodenreider O. Extending the coverage of phenotypes in SNOMED CT through post-coordination. Stud Health Technol Inform (Proc Medinfo) 2015:795-799.



Introduction

Introduction Phenotypes

- ◆ Phenotype: observable characteristics of an organism (anatomy, physiology, behavior)
- ◆ 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



Introduction Mapping characteristics

- ◆ Complete vs. partial mappings
 - Complete – equivalence
 - Partial – subclass relation
- ◆ Basis for creating mappings
 - Lexically – through the lexical properties of phenotype names
 - Logically – through the logical definitions and the hierarchical arrangement of phenotype concepts



Introduction Mappings

	Complete	Partial
Lexical	<ul style="list-style-type: none">• Exact and normalized matches between existing (“pre-coordinated”) concept names• Equivalence relations• 30% of HPO concepts map to SNOMED CT	<ul style="list-style-type: none">• (Controlled) incomplete matches between existing (“pre-coordinated”) concept names• Subclass relations
Logical	<ul style="list-style-type: none">• Equivalence between logical definitions• Equivalence relations• 15% of HPO concepts (with no complete lexical mapping) map to SNOMED CT	<ul style="list-style-type: none">• An ancestor of the source HPO concept is equivalent to some SNOMED CT concept• Subclass relations

Introduction Mappings

	Complete	Partial
Lexical	<ul style="list-style-type: none">• Exact and normalized matches between existing (“pre-coordinated”) concept names• Equivalence relations• 30% of HPO concepts map to SNOMED CT	<ul style="list-style-type: none">• (Controlled) incomplete matches between existing (“pre-coordinated”) concept names• Subclass relations

Multicystic dysplastic kidney [HP:0000003]
Multicystic renal dysplasia [SCTID:204962002]
(through synonymy in UMLS)



Introduction Mappings

	Complete	Partial
Lexical	<p><i>Aplastic clavicle</i> [HP:0006660]</p>	<p><i>Disease</i></p> <ul style="list-style-type: none"> • <i>Associated morphology</i> some <i>Hypoplasia</i> • <i>Occurrence</i> some <i>Congenital</i> • <i>Finding site</i> some <i>Clavicle</i>
Logical	<ul style="list-style-type: none"> • Equivalence between logical definitions • Equivalence relations • 15% of HPO concepts (with no complete lexical mapping) map to SNOMED CT 	<ul style="list-style-type: none"> • An ancestor of the source HPO concept is equivalent to some SNOMED CT concept • Subclass relations



Introduction Mappings

	Complete	Partial
Lexical	<ul style="list-style-type: none"> Exact and normalized matches <p><i>Bilateral renal atrophy</i> [HP:0012586] <i>Atrophy of kidney</i> [SCTID:197659005]</p>	<ul style="list-style-type: none"> (Controlled) incomplete matches between existing (“pre-coordinated”) concept names Subclass relations
Logical	<ul style="list-style-type: none"> Equivalence between logical <p><i>Oral cleft</i> [HP:0000202] <i>Abnormality of the mouth</i> [HP:0000153] <i>Congenital anomaly of mouth</i> [SCTID:128334002]</p>	<ul style="list-style-type: none"> An ancestor of the source HPO concept is equivalent to some SNOMED CT concept Subclass relations

Objectives

- ◆ To investigate and contrast lexical and logical approaches to deriving partial mappings between HPO and SNOMED CT
 - Partial lexical mappings
 - Partial logical mappings

Background

Related work **Ontology matching**

- ◆ **Ontology matching**
 - Schema matching
 - Lexical approaches – leverage the labels
 - Structural approaches – leverage the relations
- ◆ **Partial mappings**
 - Especially interesting when one ontology is finer-grained than the other
- ◆ **General ontology matching systems tend to be suboptimal for specialized ontologies**



Related work Partial mappings

- ◆ Partial lexical mappings
 - Compositional properties of ontologies
 - Complex terms derived from simpler terms through addition of modifiers
 - e.g., Gene Ontology
- ◆ Partial logical mappings
 - Partial mappings for descendants of an ancestor for which there is a complete mapping
 - e.g., used to create partial mappings between MedDRA and SNOMED CT

Resources HPO and SNOMED CT

HPO

Human Phenotype Ontology

Developed collaboratively
(coordination: Peter Robinson)

Specialized terminology

phenotypes for clinical genetics

10,491 classes for phenotype
16,414 terms for phenotype
(one preferred term for each class,
5,923 exact synonyms)

Concept names + synonyms

Hierarchical relations

Description logic formalism

Textual and logical definitions



for most concepts

SNOMED CT

Developed by the International Health
Terminology Standard Development
Organization

General terminology

broad coverage of Clinical Medicine

~300,000 concepts

clinical findings ~100,000 concepts
~169,000 names

Concept names + synonyms

Hierarchical relations

Description logic formalism

Logical definitions

for most pre-coordinated concepts

Resources UMLS

- ◆ Terminology integration system
- ◆ Integrates ~170 source vocabularies
 - SNOMED CT
 - [but not HPO*] *at the time of this investigation
- ◆ Used as a reference for synonymy
- ◆ Also provides lexical tools for mapping terms to UMLS concepts
 - Exact / normalized matches



Methods

Methods Overview

- ◆ Extracting phenotype terms
- ◆ Identifying complete lexical mappings
- ◆ Deriving partial lexical mappings
 - Identifying modifiers through lexico-syntactic analysis
 - Demodifying phenotype terms
 - Mapping demodified terms through UMLS
- ◆ Deriving partial logical mappings



Extracting phenotypes terms

◆ HPO

- *Phenotypic abnormality* [HP:0000118]
- and all its descendants
- with all preferred terms and synonyms

◆ SNOMED CT

- *Clinical Findings* [SCTID:404684003]
- and all its descendants
- with all preferred terms and synonyms



Identifying complete lexical mappings

Renal hypoplasia [HPO:HP_0000089]

*exact
match*

- [UMLS:C0266295]
- Congenital hypoplasia of kidney*
 - Hypoplasia of kidney*
 - Hypoplasia of kidneys*
 - Hypoplastic kidney*
 - kidney hypoplasia*
 - Renal hypoplasia*
 - [...]

Congenital hypoplasia of kidney [SCTID:32659003]



Deriving partial lexical mappings

Bilateral renal atrophy [HP:0012586]

[MOD][MOD][HEAD]



Partial lexical mappings

remove modifiers

Derivified HPO terms

Complete lexical mapping through UMLS

SNOMED CT Clinical finding term

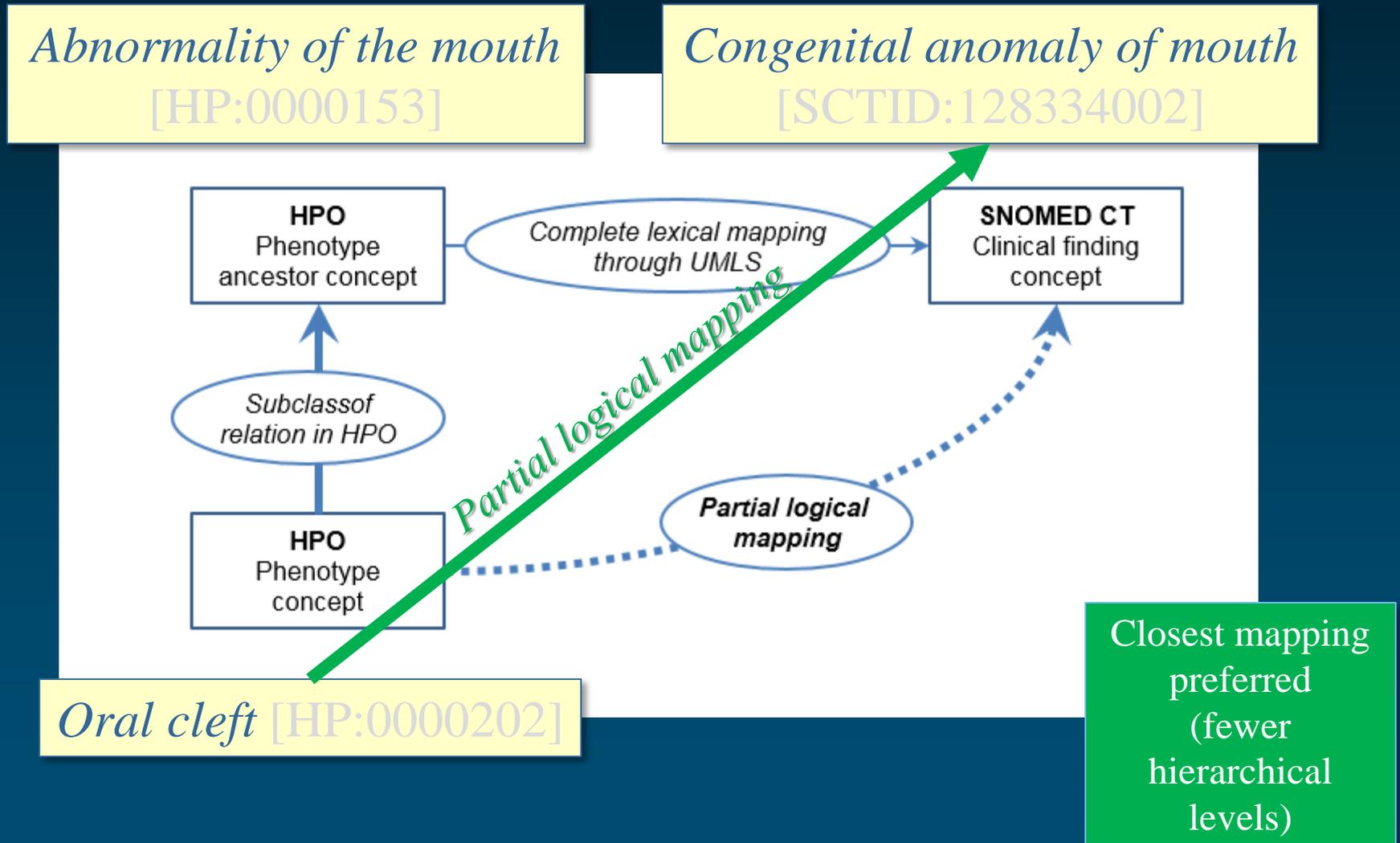
Partial lexical mapping through UMLS

- *Bilateral renal atrophy*
- *Bilateral renal atrophy*
- *Bilateral renal atrophy*

Atrophy of kidney [SCTID:197659005]

Closest mapping preferred (fewer modifiers removed)

Deriving partial logical mappings



Evaluation

◆ Quantitative evaluation

- Number of HPO concepts with
 - Partial lexical mappings
 - Partial logical mappings

◆ Qualitative evaluation

- Manual review
 - Random subset of 10% of the partial lexical mappings
 - 25 mappings per level among the partial logical mappings
- 2 criteria
 - Ontologically valid = consistent with a subclass relation
 - Clinical relevant = useful for cohort selection



Results

Extracting phenotypes terms

◆ HPO

- *Phenotypic abnormality* [HP:0000118]
and all its descendants
 - 10,454 concepts
 - 16,612 terms
 - 10,454 preferred terms
 - 6158 synonyms

◆ SNOMED CT

- *Clinical Findings* [SCTID:404684003]
and all its descendants
 - 103,748 concepts
 - 167,491 terms
 - [103,748 fully specified names]
 - 167,491 synonyms



Identifying complete lexical mappings

- ◆ 10,454 phenotype concepts in HPO
 - 3096 HPO concepts (30%) with complete lexical mapping to clinical findings in SNOMED CT
 - 7358 concepts with no with complete lexical mapping
 - 10,631 terms
 - Used for identifying partial mappings lexically and logically



Deriving partial lexical mappings

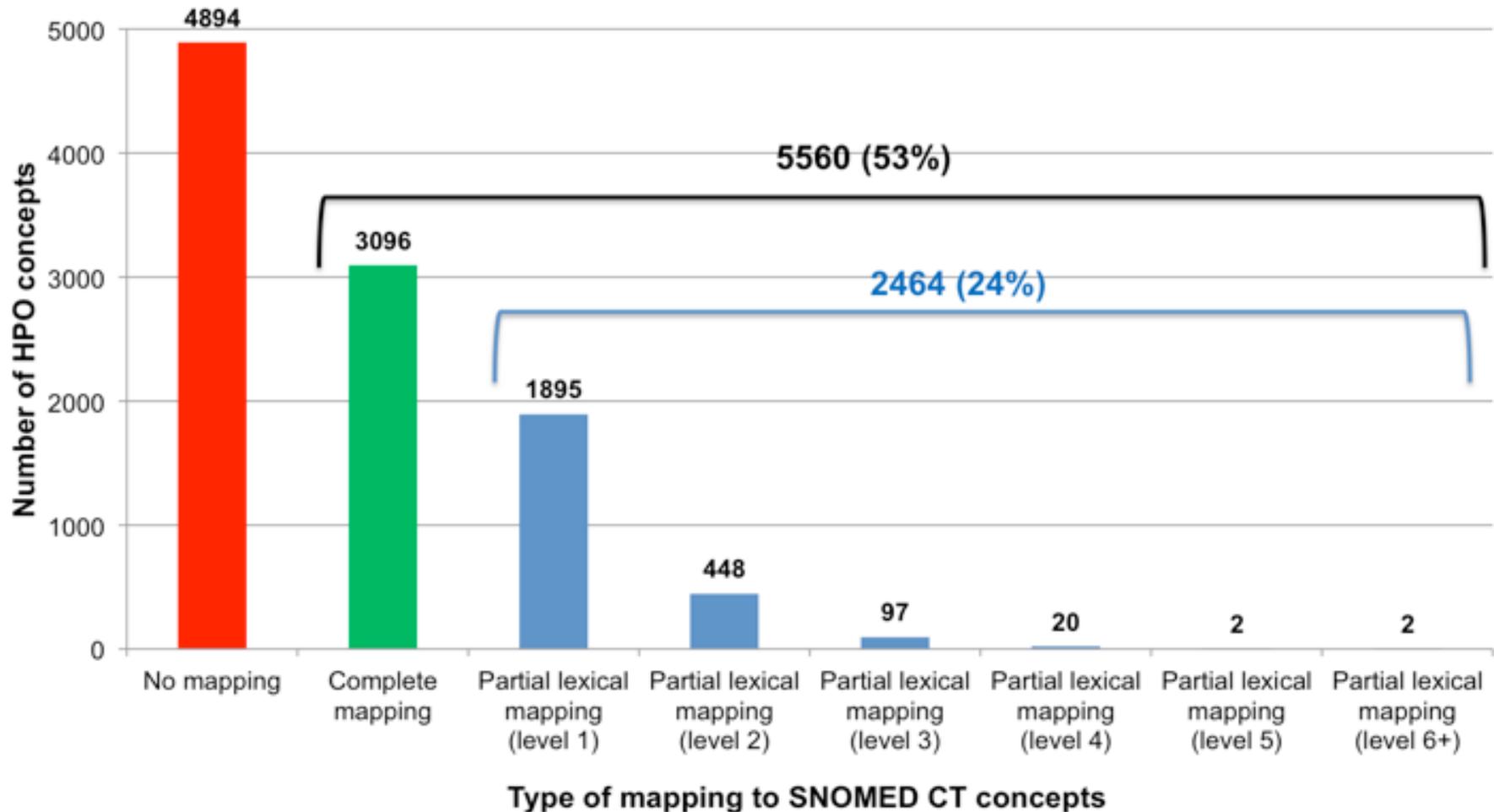
- ◆ 494 distinct lexico-syntactic profiles for 10,631 HPO terms
 - Amenable to demodification – 6959 terms (65%)
 - Not amenable to demodification – 3672 terms (35%)
 - Too simple ([HEAD]) – 218 terms
 - Too complex (e.g., multiple prepositions) – 3454 terms

Lexico-syntactic profile	Terms	(%)	Examples of HPO terms
[MOD—HEAD]	2478	(23 %)	<i>Oral cleft, Aplastic clavicles, Abnormal philtrum</i>
[MOD—MOD—HEAD]	1811	(17 %)	<i>Asymmetric limb shortening, Multicystic kidney dysplasia</i>
[HEAD] [PREP—DET—HEAD]	536	(5 %)	<i>Abnormality of the philtrum, Polydactyly of the foot</i>
[MOD—MOD—MOD—HEAD]	478	(4 %)	<i>Small proximal femoral epiphyses, Increased cup disc ratio</i>
[HEAD] [PREP—MOD—HEAD]	386	(4 %)	<i>Delay in motor development, Abnormality of renal excretion</i>
[MOD—HEAD] [PREP—HEAD]	321	(3 %)	<i>Hypertensive disorder of pregnancy, Coronal cleft of vertebrae</i>
[HEAD] [PREP—HEAD]	259	(2 %)	<i>Abnormality of upper lip, Tremor at rest, Tetralogy of Fallot</i>
[HEAD]	218	(2 %)	<i>Gastroschisis, Polydactyly, Pre-eclampsia</i>
[HEAD] [PREP—DET—MOD—HEAD]	209	(2 %)	<i>Abnormality of the paralabial region, Fragmentation of the metacarpal epiphyses</i>
[MOD—HEAD] [PREP—DET—HEAD]	202	(2 %)	<i>Downturned corners of the mouth, IgA deposition in the glomerulus</i>
top 10	6898	(65 %)	

Deriving partial lexical mappings

- ◆ 23,936 demodified terms created from the 6959 original terms
- ◆ 7358 HPO concepts with no complete mapping
 - 2464 (33%) with partial lexical mapping to SNOMED CT
- ◆ A majority of the partial mappings occurred after removing a single modifier (“level 1”)

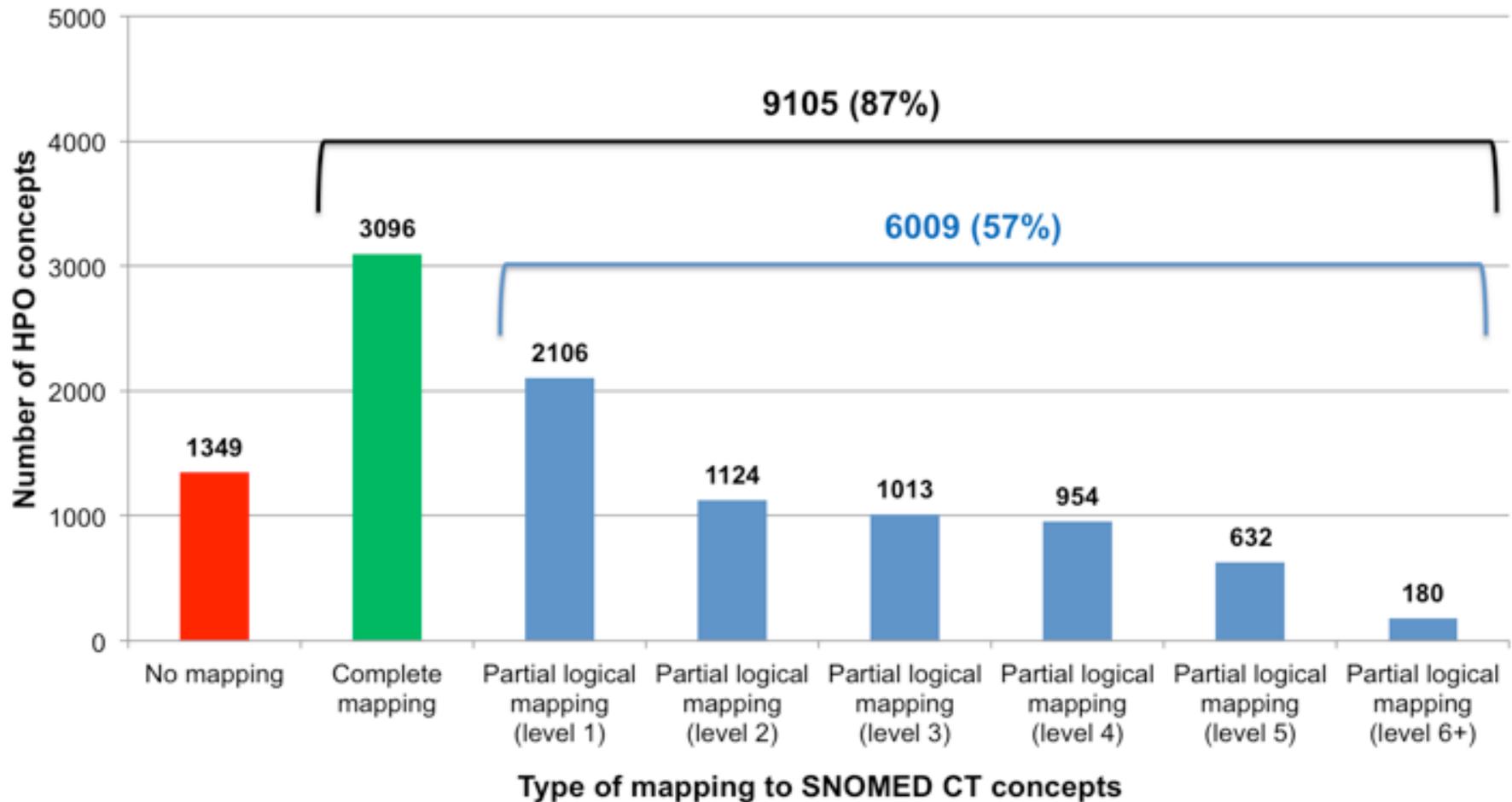
Lexical mappings



Deriving partial logical mappings

- ◆ 7358 HPO concepts with no complete mapping
 - 6009 (82%) with partial logical mapping to SNOMED CT
- ◆ Levels
 - Level 1: 35%
 - Level 1-4: 86%

Logical mappings



Evaluation

◆ Quantitative evaluation

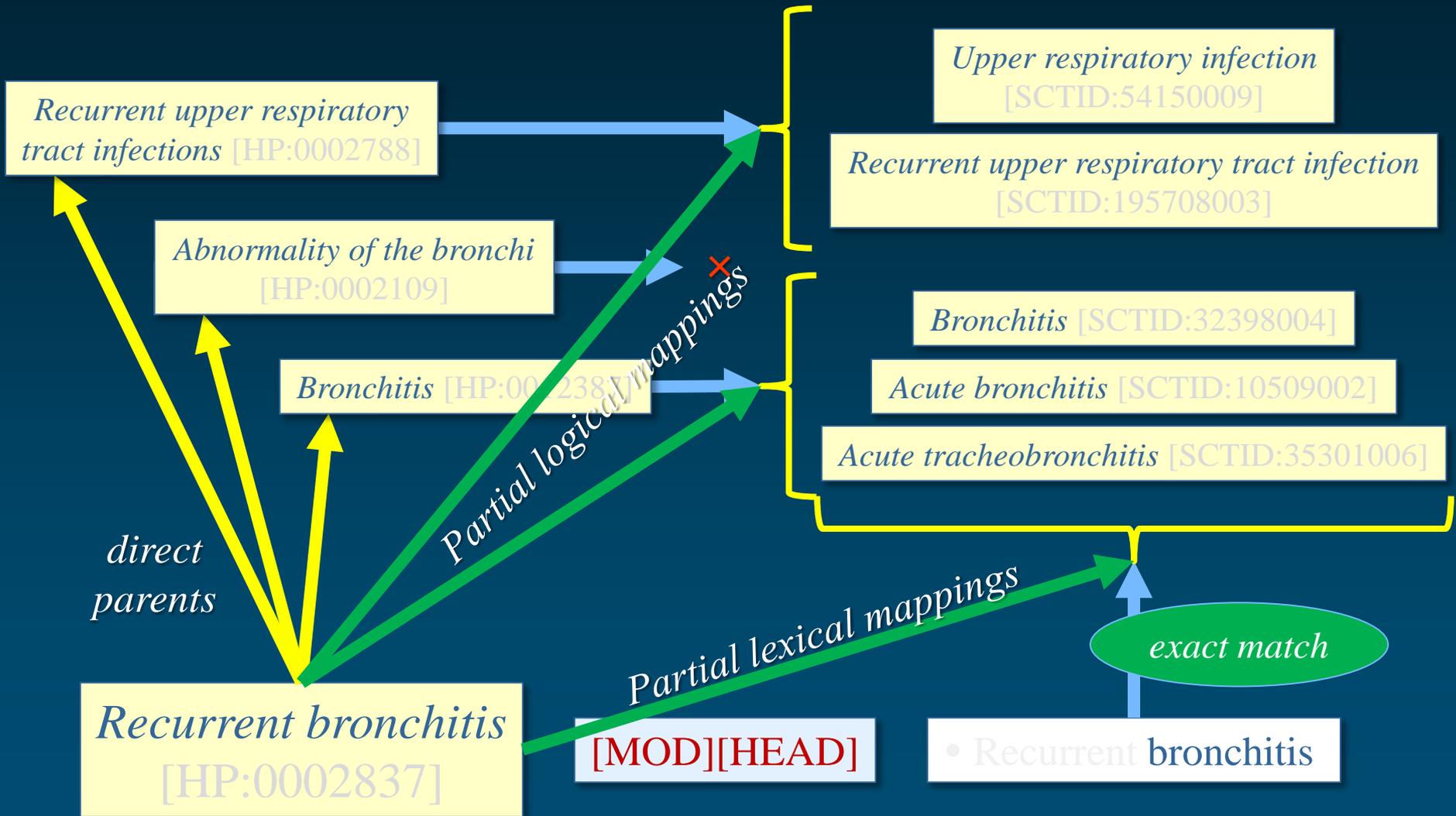
- Number of HPO concepts (10,454) with
 - Complete lexical mappings 3096 (30%)
 - Partial lexical mappings 2464 (24%)
 - Partial logical mappings 6009 (57%)
 - Lexical or logical 6474 (62%)

◆ Qualitative evaluation

- Lexical mappings (247)
 - Ontologically valid 62%
 - OV and Clinical relevant 49%
- Logical mappings (125)
 - Ontologically valid 71%
 - OV and Clinical relevant 67%



Extended example



Discussion

Discussion

- ◆ Enhanced mapping of phenotype concepts
- ◆ Lexical vs. logical techniques for partial mappings
- ◆ Failure analysis
- ◆ Implicit congenitality
- ◆ Limitations and future work

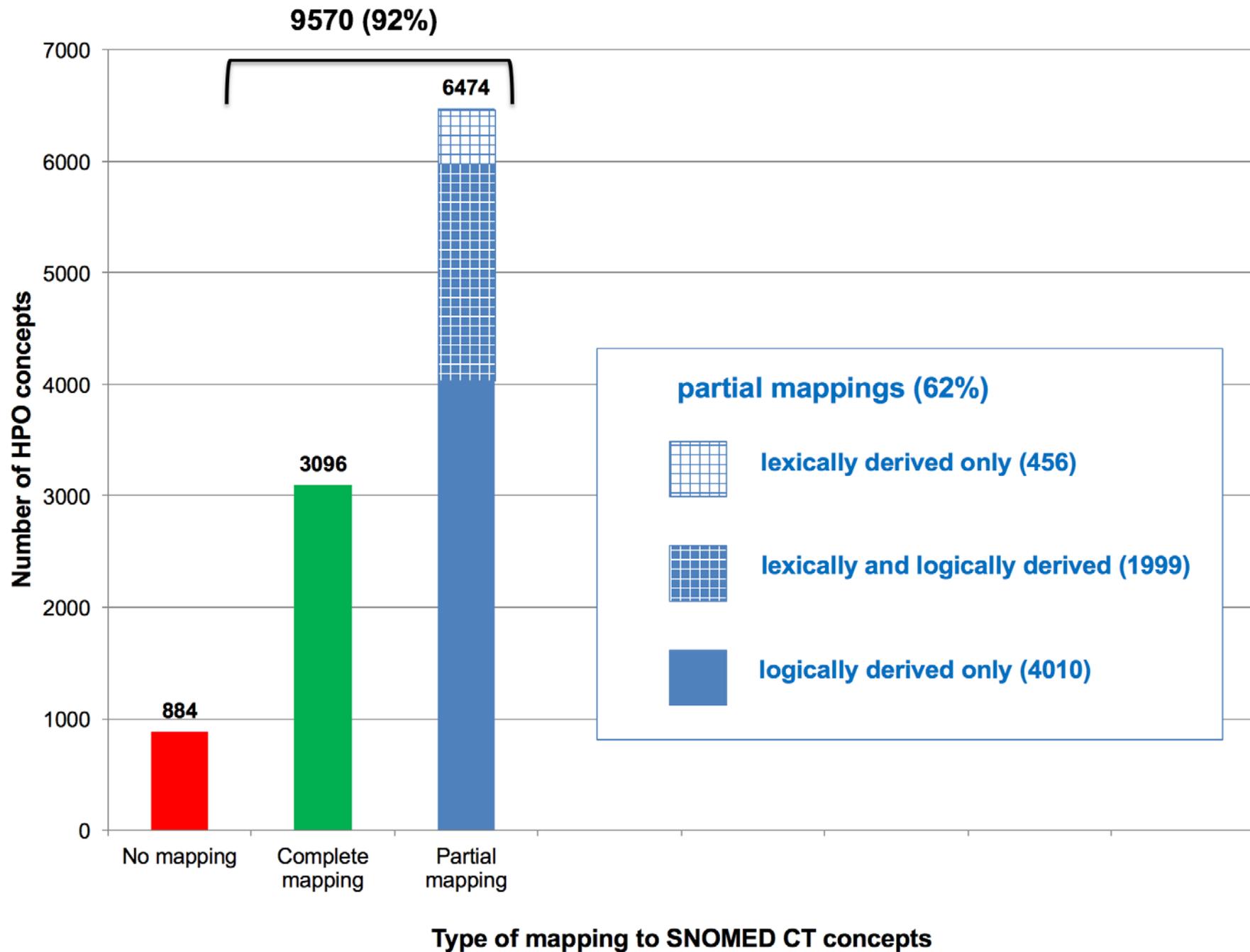
Enhanced mapping of phenotype concepts

- ◆ Mapping of SNOMED CT for 92% of HPO concepts
 - 30% complete lexical mappings
 - 62% partial (lexical or logical) mappings



Lexical vs. logical techniques for partial mappings

- ◆ Overall for the 7358 HPO concepts with no complete mapping to SNOMED CT
 - 82% partial logical mappings
 - 33% partial logical mappings
- ◆ By level: partial mappings at level 1 or 2
 - 95% partial logical mappings
 - 54% partial logical mappings
- ◆ Overlap lexical / logical
 - Limited overlap (31%) – because fewer lexical mappings
 - Only 7% of the lexical mappings are not covered by logical mappings
- ◆ Partial logical mappings are more often ontologically valid and clinically relevant



Failure analysis

◆ Partial lexical mappings

- Head noun outside the domain of disorders
 - *Hypoplastic sacrum* [HP:0004590]
- Complex lexico-syntactic patterns
 - *Complete duplication of the proximal phalanx of the 5th toe* [HP:0100415]
 - [MOD-HEAD][PREP-DET-MOD-HEAD][PREP-DET-MOD-HEAD]
- Complex lexical items identified as HEAD
 - *Pyruvate dehydrogenase complex deficiency* [HP:0002928]

◆ Partial logical mappings

- No ancestor with mapping to SNOMED CT through UMLS
 - *Absent sternal ossification* [HP:0006628]



Implicit congenitality

◆ Congenitality

- Expressed explicitly in **SNOMED CT**
- Often implicit in **HPO**
- Often considered equivalent in **UMLS**

[UMLS:C0266295]

Congenital hypoplasia of kidney

Renal hypoplasia

[...]



[UMLS:C0026633]

Congenital anomaly of mouth

Abnormality of the mouth

[...]



◆ Impact on partial mappings



Limitations and future work

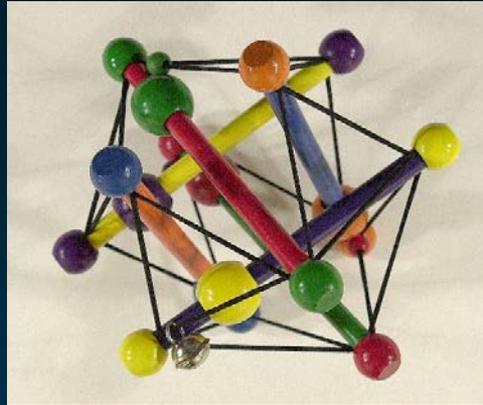
- ◆ Direction of mappings
 - SNOMED CT to HPO not considered
- ◆ Partial lexical mappings
 - Ignored complex lexico-syntactic profiles
- ◆ UMLS
 - HPO was not integrated in UMLS at the time of this investigation
 - HPO is now integrated in UMLS (as of UMLS 2015AB)
 - Curated UMLS mappings should provide a better foundation for deriving partial mappings



Conclusions

- ◆ 92% of the 10,454 HPO concepts can be mapped to SNOMED CT
 - 30% complete and 62% partial
- ◆ Partial mappings provide a next-best approach for traversing between the two systems
- ◆ Lexical and logical mapping techniques are complementary to each other
- ◆ This work highlighted
 - Interesting properties (both lexical and logical) of HPO and SNOMED CT
 - Limitations of mapping through UMLS





Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: <http://mor.nlm.nih.gov>



Olivier Bodenreider

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