FETAL MEDICINE INFORMATICS RESEARCH

EXPLORING PHENOTYPE REPRESENTATION, TERATOGENIC AGENTS DISPENSED DURING PREGNANCY AND FETAL MEDICINE TERMINOLOGY

Dr. Ferdinand C Dhombres, MD, PhD

Mentor: Dr. Olivier Bodenreider, MD, PhD

Fellowship: November 2014 — August 2016

Medical Ontology Research Group, Cognitive Science Branch (CgSB),
Lister Hill National Center for Biomedical Communications, NLM, NIH

Contents

Introduction

My general research interest in Medical Informatics is medical ontology design, evaluation and applications in the field of Fetal Medicine. My motivation is that there is currently no reference terminology dedicated to my clinical field of expertise which is Fetal Medicine. Rare Diseases (RD) and their related knowledge are massively involved and overlap with Fetal Medicine: prenatal and postnatal imaging semiology (ultrasound, MRI), "omics" data, prenatal and postnatal phenotypes, prognosis of disorders, etc. There is a critical need for a good representation of these RDs (ie. RD diagnosis and phenotypes) in terminologies in order to support clinical decision support in Fetal Medicine.

My overall objective is to establish resources and methods for the fetal medicine domain in order to support the representation of fetal disorders, fetal phenotypes, and drugs with teratogenic potential. Accordingly, in my two-year fellowship in the Cognitive Science Branch of the Lister Hill National Center for Biomedical Communications, I conducted research on (i) human phenotype representation, (ii) teratogenic agent in prescription drugs and (iii) fetal medicine terminology. These three aspects are developed in the three main sections of this report. The abstracts[1] of the articles corresponding to these sections are presented in the appendix.

Phenotype representation

I joined the Medical Ontology Research Group in November 2014 and immediately started to work on the coverage of phenotypes in SNOMED CT. In this work, we aimed to extend the coverage of phenotypes in SNOMED CT through post-coordination. Toward this end, we identified frequent modifiers in terms extracted from the Human Phenotype Ontology (HPO), which we associated with templates for post-coordinated expressions in SNOMED CT. We identified 176 modifiers, created 12 templates, and generated 1,167 post-coordinated expressions. Through this novel approach, we increased the current number of mappings by 50%. This work was presented at MEDINFO 2015 and won the "best paper award" at the conference (1). It was also accepted for presentation[2] at the SNOMED CT Implementation Showcase in 2015 and at the NLM Training Conference in 2015.

The interoperability between HPO and SNOMED CT can be addressed in several complementary ways, through lexical mappings (complete or partial) and by leveraging the logical definitions of phenotypes. In our MEDINFO 2015 paper, we investigated complete lexical mappings and mappings leveraging the logical definitions of phenotypes. I pursued this work on phenotype representation by

[1] The full text of all articles is available online (cf. links in the appendix). It will also be attached to this report as supplementary material.
[2] https://mor.nlm.nih.gov/pubs/pdf/2015-snomedct_expo-fd-abstract.pdf

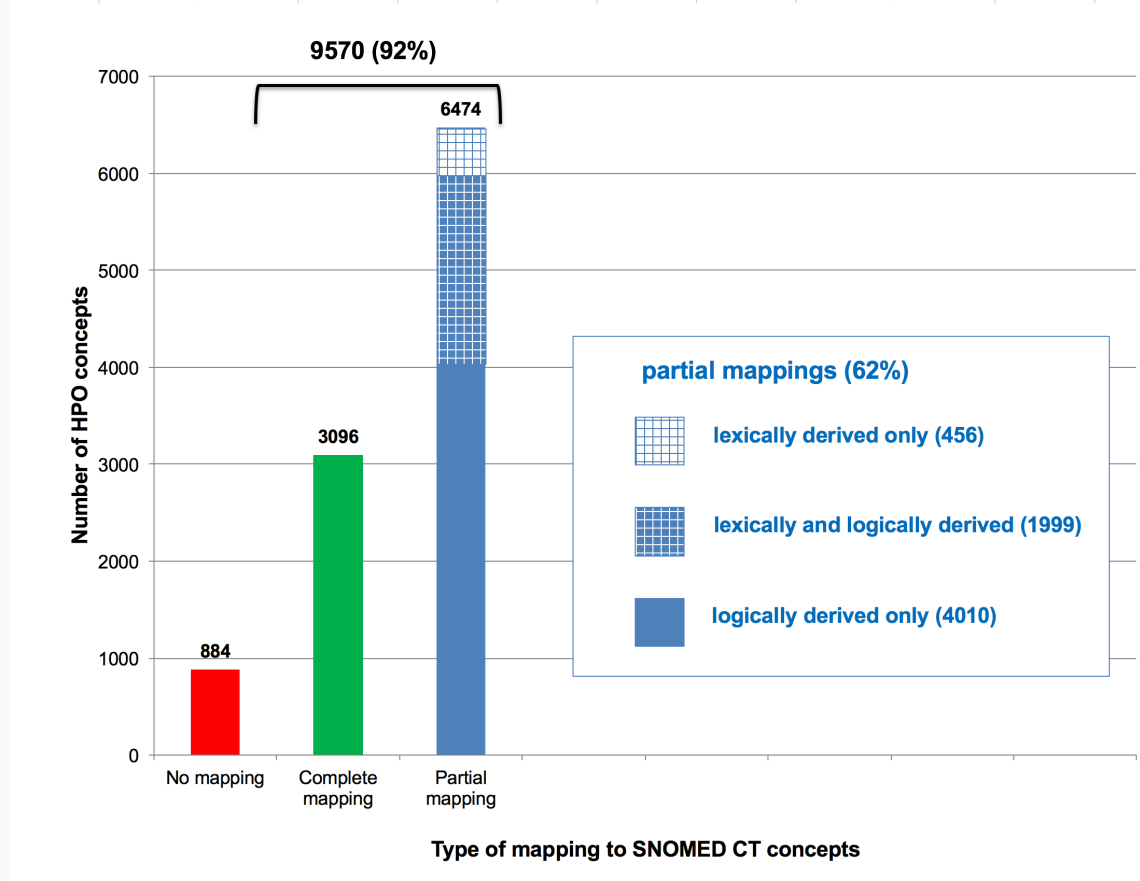
exploring partial lexical mappings between HPO terms and SNOMED CT. Identifying partial mappings between two terminologies is of special importance when one terminology is finer-grained than the other, as is the case for HPO and SNOMED CT. Our approach is similar to identifying matches with complete lexical mappings, but allows some words from the HPO terms to be omitted from the mapping to SNOMED CT. Such mappings denote subsumption (subclass) relations between the more specific HPO concept and the more general SNOMED CT concept mapped to. A complete lexical mapping was identified for 30% of HPO classes and a partial lexical mapping was identified for 20% additional classes. This work on partial mappings was presented at the Joint Bio-Ontologies and BioLINK ISMB'2015 SIG session "Phenotype Day" (2) and at the SNOMED CT Expo 2016[3].

We extended the work on partial lexical mappings to the partial logical mappings between HPO and SNOMED CT. We investigated and contrasted lexical and logical approaches to deriving partial mappings between HPO and SNOMED CT.

1) Lexical approach-We identify modifiers in HPO terms and attempt to map demodified terms to SNOMED CT through UMLS;
2) Logical approach-We leverage subsumption relations in HPO to infer partial mappings to SNOMED CT;
3) Comparison-We analyze the specific contribution of each approach and evaluate the quality of the partial mappings through manual review.

There are 7358 HPO concepts with no complete mapping to SNOMED CT. We identified partial mappings lexically for 33% of them and logically for 82%. We identified partial mappings both lexically and logically for 27%. The clinical relevance of the partial mappings (for a cohort selection use case) is 49% for lexical mappings and 67% for logical mappings.

[3] https://mor.nlm.nih.gov/pubs/pdf/2016-snomedct_expo-fd-abstract.pdf



In conclusion, through complete and partial mappings, 92% of the 10,454 HPO concepts can be mapped to SNOMED CT (30% complete and 62% partial), as shown in the graph above. Equivalence mappings between HPO and SNOMED CT allow for interoperability between data described using these two systems. However, due to differences in focus and granularity, equivalence is only possible for 30% of HPO classes. In the remaining cases, partial mappings provide a next-best approach for traversing between the two systems. Both lexical and logical mapping techniques produce mappings that cannot be generated by the other technique, suggesting that the two techniques are complementary to each other. Finally, this work demonstrates interesting properties (both lexical and logical) of HPO and SNOMED CT and illustrates some limitations of mapping through UMLS. These results were published in the Journal of Biomedical Semantics (3).

Teratogenic agents.

Over eighty percent of pregnant women in the United States are prescribed at least one drug during pregnancy. And precise information about the adverse effects of drugs on fetal development is sparse and scattered. A drug can cause congenital malformations when the exposure occurs at a specific time in pregnancy and at a given dose. The definition of teratogenic exposure includes the teratogenic agent

(the drug), the dose and the time in pregnancy. I conducted research to assess the extent to which teratogenic drugs are prescribed during pregnancy. The approach used was to establish a reference list of teratogenic prescription drugs, then to extract drugs dispensed to pregnant women from a clinical dataset, and finally to compare the drugs dispensed to pregnant women to the reference list of teratogenic drugs.

Our objective was to assess the potential risk in drug prescriptions during pregnancy, with respect to the new FDA standard of June 2015. Our approach can be summarized as follows:

1) **Drug risk categories**. As a proxy for the FDA standard, we used the "pregnancy recommendations" from a reference textbook (Briggs, 10th ed. 2015). For each ingredient, it provides the level of risk (contraindicated, high risk, moderate risk, low risk, probably compatible and compatible with pregnancy), the source of evidence, if any (human or animal data), and other information as appropriate (trimester, dose and drug association restrictions). When an ingredient was associated with more than one category (e.g., to account for risk variation based on dose, length of exposure, or associated comorbidities), we used the highest risk category.
2) **Prescription data processing**. We analyzed patient-level, de-identified claims data of a privately insured population of 159.7M patients from 2003 to 2014 provided by the IMEDS Research Lab. We relied on procedure codes for delivery to identify pregnant women (13 CPT (Current Procedural Terminology v4) codes covering all vaginal deliveries and Caesarean sections). We considered a period of 270 days prior to delivery or C-section for drugs dispensed during pregnancy. We used the RxNorm API to relate drugs from claims data to the reference. We derived the risk and supporting evidence associated with each drug, taking the highest risk in case of multi-ingredient drugs. We restricted our analysis to systemic drugs, because topical drugs generally pose a much lower risk. We counted prescriptions by category, using the new standard (level of risk and source of evidence) recommended by the FDA.

A total of 3,741,743 pregnant women were selected, to which 19,654,083 prescription drugs were dispensed (15,815,624 systemic drugs). The level of risk was defined using the classification extracted from Briggs for 14,719,736 prescriptions (93%). Overall, 40.2% of the prescriptions were "compatible" with pregnancy and 1.2% were "probably compatible". The prescriptions were contraindicated in 2.8%. There was a potential risk in 8,191,485 prescriptions (55,6%). And evidence based on human data is available for 91.85% of all prescriptions.

This investigation demonstrates the feasibility of assessing the potential risk in drug prescriptions during pregnancy from a large claims dataset using RxNorm and the Briggs reference with respect to the new FDA standard. It had already been demonstrated that pregnant women are commonly prescribed drugs associated with fetal risk. However, supporting evidence was not previously reported. This work was accepted for presentation at the upcoming AMIA annual symposium in 2016 (4) and a full paper will be submitted by the end of 2016.

Fetal Medicine Terminology.

The third objective of my research was to assess the need for a standard terminology for Fetal Medicine. I started by establishing a corpus of fetal medicine articles in order to support term extraction. This corpus is a collection of full text articles from the journal Prenatal Diagnosis; 10 years of the journal archive were used to build this corpus.

In a preliminary analysis, I performed an evaluation of the coverage of the corpus terms in standard terminologies (the assessment of the quality of terms was done by manual review assisted by natural language processing tools). The corpus for this analysis included 2,4K articles from the journal Prenatal Diagnosis (390K sentences and 6M words). From this corpus, the term extractor Termine4 provided 4.1K high-frequency terms (1.3K normalized terms), of which 70% were deemed relevant to fetal medicine. Only 38% of these terms could be mapped to the UMLS (essentially to disorder and procedure concepts). This investigation suggests that there is a significant body of fetal medicine terms. These terms are generally poorly covered in standard terminologies. In order to support interoperability and data exchange in the domain of fetal medicine, these terms would benefit from being integrated into existing clinical terminologies or organized into a specific terminology.

In a complementary analysis, we used this corpus to identify trends in Fetal Medicine. More precisely, we conducted a bibliometric analysis of all full-text articles published in Prenatal Diagnosis from January 1, 2006 to December 31, 2015. Our approach can be summarized as follows. We extracted salient terms from the articles; we calculated their frequencies over time; and we established evolution profiles for the most frequent terms, from which we derived falling, stable and rising trends. The main results of this analysis are as follow:

- We identified 3598 salient medical terms. On average, the terms occurred in 101.9 ± 5.2 articles over the decade. Our manual review rescued 231 (2.7%) of the 8637 terms that had been inappropriately filtered out, including “prenatal ultrasound”, “maternal plasma”, “fetal

[4] Available through a web-service from team of the National Centre for Text Mining (http://nactem.ac.uk)

nuchal translucency" and "cell-free DNA". Those terms are not represented in standard terminologies.

- We identified 618 terms with decreasing frequencies over time (falling trend), 2142 stable terms, and 839 terms with increasing frequencies (rising trend). Not surprisingly, while stable terms occur in a large number of articles, terms with decreasing or increasing frequencies occur in fewer articles.
- We were able to identify trends in Fetal Medicine over the past 10 years with consistent result in comparison with the editorials provided by the experts of the journal.

This work in currently under revision for publication in Prenatal Diagnosis. Overall, this research confirmed the need for a better coverage of Fetal Medicine terms in standard terminologies.

Conclusion

During these two years at the NLM, I was able to address three issues in Fetal Medicine informatics research, related to human phenotype representation, to teratogenic agent in prescription drugs and to fetal medicine terminology. Additionally, I learned to use NLM informatics tools (in particular, the Unified Medical Language System and its APIs) and I also completed my informatics training (in programming with JAVA and R, and in SPARQL and RDF).

In the future, I will keep collaborating with the NLM and the CgSB as a special volunteer, in particular for the Fetal Medicine Terminology project and the Teratogenic Drugs project.

Publications

1. Dhombres F, Winnenburg R, Case JT, Bodenreider O, editors. Extending the coverage of phenotypes in SNOMED CT through post-coordination. MEDINFO2015; 2015 2015. São Paulo, Brazil.
2. Dhombres F, Bodenreider O. Investigating the lexico-syntactic properties of phenotype terms – Application to interoperability between HPO and SNOMED CT. Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2015 SIG session "Phenotype Day" 2015:8-11.
3. Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies – Investigating partial mappings between HPO and SNOMED CT. J Biomed Semantics. 2015.
4. Dhombres F, Huser V, Rodriguez LM, Bodenreider O. Assessing the potential risk in drug prescriptions during pregnancy. AMIA Annu Symp [Podium Abstract]. 2016.

Appendix

Extending the coverage of phenotypes in SNOMED CT through post-coordination.

Objectives: To extend the coverage of phenotypes in SNOMED CT through post-coordination.

Methods: We identify frequent modifiers in terms from the Human Phenotype Ontology (HPO), which we associate with templates for post-coordinated expressions in SNOMED CT.

Results: We identified 176 modifiers, created 12 templates, and generated 1,617 post-coordinated expressions.

Conclusions: Through this novel approach, we can increase the current number of mappings by 50%.

Dhombres F, Winnenburg R, Case JT, Bodenreider O. Extending the coverage of phenotypes in SNOMED CT through post-coordination. MEDINFO2015; 2015 2015. São Paulo, Brasil

Full text at https://mor.nlm.nih.gov/pubs/pdf/2015-medinfo-fd.pdf

Investigating the lexico-syntactic properties of phenotype terms – Application to interoperability between HPO and SNOMED CT

Objective: To investigate the lexico-syntactic properties of clinical phenotype terms in order to identify partial lexical mappings between HPO and SNOMED CT.
Methods: We identify modifiers HPO terms and attempt to map demodified terms to SNOMED CT through UMLS.
Results: We identified partial mappings to SNOMED CT for 20% of HPO concepts with no complete mapping to SNOMED CT.
Conclusions: Through complete and partial mappings, 50% of the HPO concepts can be mapped to SNOMED CT.

Dhombres F, Bodenreider O. Investigating the lexico-syntactic properties of phenotype terms – Application to interoperability between HPO and SNOMED CT. Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2015 SIG session "Phenotype Day" 2015:8-11

Full text at http://mor.nlm.nih.gov/pubs/pdf/2015-phenoday-fd.pdf

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

Background. Identifying partial mappings between two terminologies is of special importance when one terminology is finer-grained than the other, as is the case for the Human Phenotype Ontology (HPO), mainly used for research purposes, and SNOMED CT, mainly used in healthcare.
Objectives. To investigate and contrast lexical and logical approaches to deriving partial mappings between HPO and SNOMED CT.
Methods. 1) Lexical approach - We identify modifiers in HPO terms and attempt to map demodified terms to SNOMED CT through UMLS; 2) Logical approach - We leverage subsumption relations in HPO to infer partial mappings to SNOMED CT; 3) Comparison - We analyze the specific contribution of each approach and evaluate the quality of the partial mappings through manual review.
Results. There are 7358 HPO concepts with no complete mapping to SNOMED CT. We identified partial mappings lexically for 33% of them and logically for 82%. We identified partial mappings both lexically and logically for 27%. The clinical relevance of the partial mappings (for a cohort selection use case) is 49% for lexical mappings and 67% for logical mappings.
Conclusions. Through complete and partial mappings, 92% of the 10,454 HPO concepts can be mapped to SNOMED CT (30% complete and 62% partial). Equivalence mappings between HPO and SNOMED CT allow for interoperability between data described using these two systems. However, due to differences in focus and granularity, equivalence is only possible for 30% of HPO classes. In the remaining cases, partial mappings provide a next-best approach for traversing between the two systems. Both lexical and logical mapping techniques produce mappings that cannot be generated by the other technique, suggested that the two techniques are complementary to each other. Finally, this work demonstrates interesting properties (both lexical and logical) of HPO and SNOMED CT and illustrates some limitations of mapping through UMLS.

Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies – Investigating partial mappings between HPO and SNOMED CT. J Biomed Semantics. 2015

Full text at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748471/

Assessing the potential risk in drug prescriptions during pregnancy

This investigation demonstrates the feasibility of assessing the potential risk in drug prescriptions during pregnancy from a large claims dataset (14.7M prescriptions; 3,7M pregnant women) using RxNorm and the Briggs reference (new FDA regulation of June 2015), with finer-grained recommendations compared to the old FDA categories, as well as stronger evidence. In our cohort, there is human data evidence for 87.8% of the prescriptions for drugs with potential risk.

Dhombres F, Huser V, Rodriguez LM, Bodenreider O. Assessing the potential risk in drug prescriptions during pregnancy. AMIA Annu Symp [Podium Abstract]. 2016

Full podium abstract at https://mor.nlm.nih.gov/pubs/pdf/2016-amia-fd-abstract.pdf

Trends in Fetal Medicine: A 10-year bibliometric analysis of Prenatal Diagnosis

Objective: To automatically identify trends in Fetal Medicine over the past 10 years through a bibliometric analysis of the articles published in Prenatal Diagnosis, using text mining techniques.
Method: We processed 2423 full-text articles published in Prenatal Diagnosis between 2006 and 2015. We extracted salient terms, calculated their frequencies over time, and established evolution profiles for the terms occurring at least 10 times in one year during the decade, from which we derived falling, stable and rising trends.
Results: We identified 618 terms with decreasing frequencies over time (falling trend), 2142 stable terms, and 839 terms with increasing frequencies (rising trend). Many terms related to Cytogenetics exhibit a falling trend. Terms with increasing frequencies include those related to statistics and medical study design. The most recent of these terms also reflect the new opportunities of next-generation sequencing. Many terms identified by the journal editors as reflecting advances in Fetal Medicine were captured by our approach among the recent terms exhibiting a rising trend.
Conclusion: A bibliometric analysis based on text mining effectively supports the identification of trends over a long period of time. This scalable approach is complementary to analyses based on metadata or expert opinion.

[THIS ARTICLE IS UNDER REVISION]