2021 AMIA Virtual Joint Summits

Submission Type: Clinical Research Informatics

Keywords: Secondary use of EHR data; Ontologies; Knowledge representation, management, or engineering

Abstract (74/75 words): Common data models have solved many challenges of utilizing electronic health records, but have not yet meaningfully integrated clinical and molecular data. Aligning clinical data to open biological ontologies (OBOs), which provide semantically computable representations of biological knowledge, requires extensive manual curation and expertise. To address these limitations, we introduce OMOP2OBO, a health system-scale, disease-agnostic methodology to create interoperability between standardized clinical terminologies and semantically encoded OBOs and present results demonstrating the utility within two health systems.

After attending this talk, the learner should be better able to:

- Describe the pros and cons of different clinical concept mapping strategies on downstream computational analysis
- Recognize different approaches used in the field for mapping clinical concepts from an EHR to Open Biomedical Ontology concepts
- Hypothesize different translational research use cases for leveraging mappings that integrate biomedical knowledge and patient-level clinical data

Suggested Citation:

Callahan TJ, Wyrwa JM, Vasilevsky NA, Robinson PR, Haendel MA, Hunter LE, Kahn MG. Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality. Podium abstract; 2021 Virtual Joint Summits of the American Medical Informatics Association.

Ontologizing Health Systems at Scale: Making Translational Discovery a Reality

Tiffany J. Callahan, MPH[1], Jordan M. Wyrwa, DO[1], Nicole A Vasilevsky, PhD[2], Peter N. Robinson, MD, PhD[3], Melissa A Haendel, PhD[4], Lawrence E. Hunter, PhD[1], Michael G. Kahn, MD, PhD[1]

[1]University of Colorado Anschutz Medical Campus, Aurora, CO, USA; [2] Oregon Health Sciences University, Portland, OR, USA; [3]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; [4]Oregon State University, Corvallis, OR, USA

Introduction

A significant promise of electronic health records (EHRs) lies in the ability to perform large-scale investigations of mechanistic drivers of complex diseases. Despite significant progress in biomarker discovery, this promise remains largely aspirational due to its disconnectedness from biomedical knowledge[1]. Linking molecular data to clinical data will enable biologically meaningful analysis by integrating knowledge about biology and pathophysiology from multiple ontologies. Similar to clinical terminologies, ontologies are classification systems that provide detailed representations of a specific domain of knowledge consisting of a set of concepts and logically defined relationships[2]. Unlike most clinical terminologies, ontologies are computable and interoperable, which means they can be easily integrated with other data including data from basic science and clinical research. The usefulness of normalizing (i.e. mapping or annotating) clinical data to ontologies, like those in the Open Biomedical Ontology (OBO) Foundry, has been recognized as a fundamental need for the future of deep phenotyping[1]. Existing work has largely focused on using ontologies to improve phenotyping in specific diseases[3] and for the enhancement of specific biological and clinical domains[4,5]. Prior work has largely been limited to one-to-one mappings and rarely includes robust evaluation or external validation. Unfortunately, learning algorithms are not yet able to capture the complex semantics underlying these concepts and their relationships. Until a comprehensive resource that includes mappings between multiple clinical domains and ontologies is created and validated, automatic generation of inference between patient-level clinical observations and biological knowledge will not be possible.

To address these limitations, we have developed OMOP2OBO, the first health system-wide integration and alignment between Observational Medical Outcomes Partnership (OMOP) standardized clinical terminologies and eight OBO ontologies. To verify that the mappings are both clinically and biologically meaningful, we have performed extensive validation with assistance from multiple domain experts. Here, we present preliminary results examining the coverage of the mappings in two institutions' EHR data.

Methods

Standard clinical terminology concepts were extracted from a PEDSnet OMOP v5 de-identified instance of the Children's Hospital Colorado EHR. Clinical concepts (and their ancestor concepts) included all OMOP standard terminology identifiers from the Condition Occurrence, Drug Exposure, and Measurements tables. Additional metadata for each concept identifier included source codes, labels, and synonyms at both the concept and concept ancestor levels. Ontologies were selected under the advice of domain experts and included diseases, phenotypes, anatomical entities, cell types, organisms, chemicals, hormones, metabolites, vaccines, and proteins. Clinical data use was approved by the Colorado Multiple Institutional Review Board (#15-0445). Additional details, data, and code are available on GitHub: https://github.com/callahantiff/OMOP2OBO.

Mapping OMOP Concepts to OBO Concepts. Clinical concepts were mapped at the concept and ancestor level, drug exposure concepts were mapped at the ingredient level, and measurement concepts were mapped at the result level according to their LOINC scale type. One-to-one and one-many mappings were created using a combination of automatic and manual strategies, for each clinical concept to applicable concepts in each ontology. The automatic approach employed database cross-reference mapping, exact string mapping (using concept labels and synonyms), and word embedding-based cosine similarity scoring (using all clinical and ontology concept labels, synonyms, and definitions). All concepts unable to be mapped automatically were manually mapped. For all mappings, evidence

was generated and includes the mapping source, metadata/provenance (e.g. cross-referenced identifiers, exact match strings), and validation source (e.g. expert review). Mappings were converted to Resource Description Framework (RDF) and logically validated by running a deductive logic reasoner. Additionally, a random 20% sample of the most challenging manual condition, drug exposure, and measurement mappings were verified by a panel of domain experts spanning molecular biology, clinical pharmacology, pediatric and adult medicine, and biomedical ontology curation. Several iterations of review were performed until reaching a consensus.

Results

The full set of mapped clinical concepts included 29129 condition concepts, 1697 unique drug exposure concepts, and 4083 measurement concepts. For conditions, 20850 concepts were mapped to 4661 phenotypes and 3614 diseases. For drug ingredients, 1574 concepts were mapped to 1422 chemicals, 91 proteins, 39 organisms, and 54 vaccines. Expanding measurement concepts by result type yielded 11072 results which mapped to over 920 phenotypes, 25 anatomical entities, 27 cell types, 338 chemicals/hormones/metabolites, 194 organisms, and 113 proteins. The ratio of automatic to manual mappings differed by clinical domain and ontology with conditions and drug ingredients having more automatic mappings than measurements. These findings are likely a result of the ontologies for these domains providing significantly more metadata($\chi^2(14) = 2,664,853.82, p < 0.0001$). Agreement between the domain experts and the mapping annotators was moderate to excellent with 90.9% on measurements, 75% on drug ingredients, and 82.5% on conditions. Coverage analysis of the OMOP2OBO concepts on clinical data obtained from two independent health systems revealed 80-92% coverage of condition occurrence concepts, 91-96% coverage of drug exposure concepts, and 50-55% coverage of measurement concepts. Finally, the RDF version of the mappings was found to be logically consistent by a deductive logic reasoner.

Discussion and Conclusion

OMOP2OBO is the first health system-wide resource to provide interoperability between clinical EHR concepts and OBO ontologies. OMOP2OBO presents unprecedented opportunities to improve clinical decision making and computational phenotyping by providing additional insight into the molecular mechanisms underlying each patient's unique set of observations at hospital scale. Currently, OMOP2OBO contains 23824 standardized OMOP clinical terminology concepts and 42249 concepts in eight OBO biomedical ontologies. Although evaluation is still ongoing, preliminary results suggest excellent coverage of OMOP2OBO condition and drug concepts and excellent coverage of measurement concepts when examined in two health systems. It is important to note that the frequently updated ontologies which also contained detailed metadata on each concept (e.g. labels, definitions, synonyms, and database cross-references) tended to result in a larger number of accurate automatic mappings. These types of ontologies were also easier for both the researchers and domain experts to navigate and utilize when performing manual annotation. Additionally, it appears that concepts which are frequently used in clinical practice may also be more likely to be represented by an existing ontology. We will be exploring both of these observations further in follow-up experiments. Additional work currently underway includes expanding mapping provenance, conducting an expanded coverage study on 24 national and international hospital databases and health systems, and developing a novel machine learning pipeline with the goal of improving the accuracy of automatic annotation.

References

1. Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality-towards scalability, portability, and interoperability. J Biomed Inform. 2020;105:103433.
2. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med. 2018;379:1452-62.
3. Thompson R, Papakonstantinou NtalisA, Beltran S, Töpf A, de Paula Estephan E, et al. Increasing phenotypic annotation improves the diagnostic rate of exome sequencing in a rare neuromuscular disorder. Hum Mutat. 2019;40:1797-812.
4. Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med. 2019;2.
5. Raje S, Bodenreider O. Interoperability of disease concepts in clinical and research ontologies: Contrasting coverage and structure in the Disease Ontology and SNOMED CT. Stud Health Technol Inform. 2017;245:925-9.