OMOP2OBO: Semantic Integration of Standardized Clinical Terminologies to Power Translational Digital Medicine Across Health Systems

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

Abstract

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. *Detailed documentation and code are openly available at https://github.com/callahantiff/OMOP2OBO.*

Research Category: Observational data standards and management

Background

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[1,2]. Linking molecular data to clinical EHR data will support biologically meaningful analysis of these data, which can be achieved by integrating knowledge from multiple ontologies. Similar to clinical terminologies, ontologies are classification systems that provide detailed representations of a specific domain 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 logically verified using description logics and easily integrated with other data from basic science and clinical research[2].

The usefulness of mapping or annotating clinical data to ontologies, like those in the Open Biomedical Ontology (OBO) Foundry, is fundamental for the future of deep phenotyping[1]. Existing work has largely focused on using ontologies to improve phenotyping in specific diseases[4-5], for the enhancement of specific biological and clinical domains[6-7], has been largely limited to one-to-one mappings (e.g. mapping a single clinical term to a single ontology concept), and rarely includes external validation. Unfortunately, learning algorithms are not yet able to capture the complex clinical and biological semantics underlying these concepts and their relationships. Until a comprehensive resource that includes mappings between multiple clinical domains and OBO ontologies is available, automatic 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 OMOP standardized clinical terminologies and OBO ontologies spanning diseases, phenotypes, anatomical entities, cell types, organisms, chemicals, metabolites, hormones, vaccines, and proteins. This work has been extensively validated with assistance from domain experts spanning molecular biology, clinical pharmacology, pediatric/adult medicine, and ontology curation. We present preliminary findings examining the coverage of the mappings in two institutions' EHR data.

Methods

Standard clinical terminology concepts were extracted from a PEDSnet (v3.0) OMOP (v5.0) de-identified version of the Children's Hospital Colorado EHR. Clinical concept lists consisted of all OMOP standard terminology concept identifiers and source codes from the Condition Occurrence, Drug Exposure, and Measurements tables. All concepts and concept's ancestors from the OMOP concept ancestor table were included. Additional metadata for each concept identifier included source codes mapped to and from standard concept identifiers, labels, and synonyms at both the concept and concept ancestor-level. Ontologies were selected under the advice of several domain experts and included diseases, phenotypes, anatomical entities, cell types, organisms, chemicals, hormones, metabolites, vaccines, and proteins. This study was approved by the Colorado Multiple Institutional Review Board (#15-0445). See GitHub for additional details: https://github.com/callahantiff/OMOP2OBO.

Mapping OMOP Standard Clinical Terminologies to OBO Concepts

Clinical concepts were mapped at the concept and ancestor level, drug exposures concepts were mapped at the ingredient level, and measurement concepts were mapped according to their LOINC scale and result types. One-to-one and one-many mappings were created using a combination of automatic and manual strategies, for each clinical concept to concepts in each applicable ontology. The automatic approach consisted of ontology database cross-reference (dbXRef) mapping, exact string mapping, and Bag-of-Words cosine similarity scoring with Term Frequency Inverse Document Frequency weighting. For dbXRefs, all clinical concept and ancestor source codes were aligned to ontology concept dbXRefs. Exact string mapping was performed using all clinical and ontology concept labels and synonyms. Cosine similarity scoring was performed using all clinical and ontology concept labels, synonyms, and definitions. Concepts with no automatic mapping were manually mapped. For all mappings, evidence was generated and includes the mapping source, metadata/provenance (e.g. dbXRefs and exact labels/synonyms matches), and validation source (e.g. manual expert review). Based on these attributes, a composite score (0-1) was generated to reflect the level of confidence underlying each mapping.

Mapping Validation

Mappings were converted to Resource Description Framework and logically validated by running a deductive logic reasoner. Additionally, a random 20% sample of the most challenging mappings from each domain were verified by a panel of clinical and molecular domain experts. Several iterations of review were performed and only completed when a consensus was reached.

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 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. Agreement between the domain experts and the mapping annotators was moderate to excellent with 91.6% on measurements, 75.8% on drug ingredients, and 73.8% on conditions. Coverage analysis of the OMOP2OBO concepts on clinical data obtained from two independent health systems revealed 80-92% for condition occurrence concepts, 91-96% for drug exposure concepts, and 50-55% for measurement concepts.

Discussion/Conclusions

We introduce OMOP2OBO, the first health system-wide resource to provision interoperability between 23824 standardized OMOP clinical terminology concepts and 42249 concepts in eight OBO biomedical ontologies. Although the evaluation is still ongoing, preliminary results suggest excellent coverage of OMOP2OBO condition, drug ingredient, and measurement concepts when examined in two independent health systems. Work underway includes expanding mapping provenance and conducting an extensive coverage study, which includes 22 national and international hospital databases and health systems.

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. Kafkas Ş, Abdelhakim M, Hashish Y, Kulmanov M, Abdellatif M, et al. PathoPhenoDB, linking human pathogens to their phenotypes in support of infectious disease research. Sci Data. 2019;6:79.
4. Thompson R, Papakonstantinou Ntalis A, Beltran S, Töpf A, de Paula Estephan E, Polavarapu K, et al. Increasing phenotypic annotation improves the diagnostic rate of exome sequencing in a rare neuromuscular disorder. Hum Mutat. 2019;40:1797-812.
5. 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.
6. 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.