Ontologizing Health Systems Data At Scale: Making Translational Discovery A Reality. TJ Callahan, (Ph.D., GS), JM Wyrwa, NA Vasilevsky, PN Robinson, MA Haendel, LE Hunter, and MG Kahn, Computational Bioscience Program, Department of Pharmacology, Graduate School, University of Colorado Anschutz Medical Campus, Aurora, CO.

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 the disconnectedness of EHR data and biomedical knowledge. Linking molecular data to EHR data will support biologically meaningful analysis and can be achieved by integrating biomedical knowledge from multiple ontologies. Similar to clinical terminologies, computational ontologies are classification systems that provide detailed representations of a specific domain of knowledge. The usefulness of mapping 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. Existing work has largely focused on using ontologies to improve phenotyping in specific diseases and for the enhancement of specific biological and clinical domains. Until a comprehensive resource that includes mappings between multiple clinical domains and ontologies is created and validated, automatic inference between patient-level clinical observations and biological knowledge will not be possible.

Study Purpose: We developed OMOP2OBO, the first health system-wide integration and alignment between the Observational Medical Outcomes Partnership (OMOP) standardized clinical terminologies and eight OBO biomedical ontologies spanning diseases, phenotypes, anatomical entities, cell types, organisms, chemicals, metabolites, hormones, vaccines, and proteins. To verify the mappings, we performed extensive validation with assistance from multiple domain experts.

Methods: Clinical terminology concepts were extracted from the Children's Hospital Colorado EHR. Additional metadata included source codes, labels, and synonyms. 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. EHR use was approved by the Colorado Multiple Institutional Review Board (#15-0445). Condition concepts were mapped at the concept level, drugs were mapped at the ingredient level, and measurements were mapped at the result level. Mappings were created using automatic and manual strategies, for each clinical concept to concepts in each applicable ontology. The automatic strategy consisted of ontology database cross-reference mapping, exact string mapping, and cosine similarity scoring. All concepts unable to be mapped automatically were manually mapped. For all mappings, evidence was generated and includes the mapping source, provenance, and validation source. A random 20% sample of the most challenging mappings for each clinical domain were verified by clinical and molecular domain experts.

Results: OMOP2OBO mappings clinical concepts included 92367 condition concepts, 8615 unique drug exposure ingredients, and 11072 measurement results. Agreement between the domain experts and the mapping annotators was 75% on drug ingredients, 82.5% on conditions, and 90.9% on measurements. Coverage analysis on clinical data obtained from 24 independent health systems revealed OMOP2OBO included 99.2% of conditions, 96% of drug exposure ingredients, and 70% of measurements.

Discussion and Conclusion: OMOP2OBO is the first health system-wide resource to provision interoperability between 105020 OMOP clinical concepts and 142249 concepts in eight OBO ontologies. Preliminary results suggest excellent coverage of clinical concepts when examined in 24 health systems.