# OMOP-2-OPMI: Ontologization of OMOP CDM using OPMI to support clinical data interoperability and analysis

Long Tran [1], and Yongqun He [1]

[1] *University of Michigan, Ann Arbor, MI, USA*

**Abstract**

The OMOP Common Data Model (CDM) has been widely used as an open community data standard in observational data integration and analysis. However, it still has its drawbacks including weak semantics and interoperability with other CDMs. In this study, we report our ontologization of the OMOP CDM elements and the semantic relations among the elements using the Ontology of Precision Medicine and Investigation (OPMI). A total of 165 terms from 15 OMOP CDM tables has been mapped to OPMI, with 46 terms newly generated with OPMI namespace and the other terms reported from OBO reference ontologies. An Omop2Opmi.owl file was also generated by extracting the OMOP CDM related terms and relations from OPMI. Three categories of use cases are reported, using the ontology-level OMOP CDM element standardization and data integration, adverse event (AE) modeling, and COVID-19 clinical data studies. Following the Ontology of Adverse Events (OAE) definition, we developed a generalizable OMOP-AE model that transforms the OMOP data to systematically define, identify, and analyze specific adverse events following some medical interventions that include Drug/Device Exposure and Procedure Occurrence in OMOP. Overall, OMOP-2-OPMI complements and empower OMOP CDM for enhanced clinical data standardization, sharing, interoperability, and analysis.

**Keywords**

OMOP, Common Data Model, ontology, OPMI, adverse events, COVID-19.

## 1. Introduction

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open community data standard that aims to allow for systematic analysis of disparate observational databases [1]. With the CDM, the data contained in those databases can be transformed into a common format with a common representation. OMOP CDM has been widely used to support the standardization of various electronic medical records (EMR) and administrative claims within and outside the United States. Billions of patient records have been standardized using OMOP CDM. Recently, OMOP CDM has become an established data model used by the National COVID Cohort Collaborative (N3C, https://ncats.nih.gov/n3c). As of May 2022, the N3C data enclave has stored the records of 14 million persons, including over 5 million COVID+ cases. Based on the N3C data use design, the COVID-19 clinical data warehouse data dictionary used in N3C is based on OMOP CDM, and the other data formats need to be aligned with the OMOP CDM in order to be entered and used in the N3C data enclave. Therefore, the OMOP CDM has clearly played a significant role in the data standardization and integration.

Still the OMOP CDM has its own drawbacks [2, 3]. One drawback is its weak semantics in that OMOP CDM does not provide robust semantic relations among CDM elements. Basically, the OMOP CDM provides the schema structure of a

ICBO 2022, September 25-28, 2022, Ann Arbor, USA
EMAILs: longtr@umich.edu (A.1 ); yongqunh@med.umich.edu (A. 2). ORCID: 0000-0002-5735-7540 (A. 1); 0000-0001-9189-9661 (A. 2)

standardized a relational database that includes over 10 tables, which has an inherent weakness in terms of representing the relations among terms from different tables. As a result, the layout of OMOP and how it is set up to document patients' conditions could lead to ambiguities, inaccurate representations and erroneous counting [2]. Another drawback is that OMOP CDM does not inherently provide systematic interoperability with other CDMs such as National Patient-Centered Clinical Research Network (PCORnet) [4] and Clinical Data Interchange Standards Consortium (CDISC) [5]. In the N3C data integration, the COVID-19 data formulated with other CDMs are required to be harmonized based on OMOP CDM version 5.3 [6], which is separately conducted and difficult to achieve robust interoperability and scalability.

Ontology can be a solution to solve the above drawbacks [3, 7]. In the 2018 OHDSI Symposium, we proposed a strategy of ontological representation of the OMOP CDM using the OBO framework [3]. In addition to the core OMOP CDM model, the OMOP system also includes many standardized clinical terminologies that can be used under the OMOP CDM framework to collaboratively support observational data standardization and integration. In the 2020 OHDSI Symposium, Callahan et al. reports their development of the OMOP2OBO, a health system-wide program of the integration and alignment between OMOP's standardized clinical terminologies and eight OBO biomedical ontologies spanning diseases, phenotypes, anatomical entities, cell types, organisms, chemicals, metabolites, hormones, vaccines, and proteins [7]. As of the end of May 2022, the OMOP2OBO mapping program has collected 92,367 OMOP Conditions, 8,615 Drug Exposure ingredients, and 3,827 Measurements (10,673 measurement test results) terms [8]. OMOP2OBO allows its users to construct their own sets of omop2obo mappings.

Among >100 ontologies in the Open Biomedical Ontology (OBO) library, the Ontology of Precision Medicine and Investigation (OPMI) is an ontology in the domain of precision medicine and investigation [9, 10]. Following the OBO ontology principles (e.g., openness and collaboration, OPMI reuses many terms of existing reference ontologies and include many of its own terms in the field of clinical and translational precision medicine, supporting non-redundant and interoperable ontology development [11]. OPMI has been developed and used to support the Kidney Precision Medicine Project [9, 10]. We have been using the OPMI to model and represent the core OMOP CDM elements and relations among the elements [3].

This manuscript reports our usage and extension of the OPMI to ontologize the OMOP CDM elements and the relations among these elements, and how such OMOP-2-OPMI ontologization supports systematic clinical data interoperability, sharing, and integration.

## 2. Methods

### 2.1. OMOP CDM resource used in the study

The OMOP version 5.4 was used in our OPMI mapping. First, we obtained terms and their annotations from the OMOP CDM version 5.4 resource [12]. The Athena software program (https://athena.ohdsi.org/) is the tool used to search OMOP CDM terms and related terms from OMOP-associated terminologies.

### 2.2. OMOP-2-OPMI development strategy

The OPMI ontology is used as the default ontology platform for the ontology mapping and new term generation of the OMOP CDM elements and semantic relations among the elements. In general, the eXtensible Ontology Development (XOD) strategy [13], including the methods of ontology term reuse, semantic alignment, ontology design pattern, and community extensibility, were used for the OPMI mapping. Specifically, all those OMOP CDM element terms were first searched in Ontobee [14]. For those terms existing in reference OBO ontologies that map to the OMOP CDM elements, Ontofox [15] was used to import those terms to OPMI (if the import has not been done before). For those OMOP elements that cannot be mapped to any OBO reference ontology, we generated new terms and defined them with OPMI namespace based on specific ontology design patterns. The OPMI ontology editing was performed using Protege-OWL editor [16], and the ontology reasoning was conducted using the Hermit reasoner [17]. All the terms are aligned under the upper-level Basic Formal Ontology (BFO) [18]. Meanwhile, we have discussed our project design in different scenarios, and community feedback and

comments were obtained to adjust our definitions and design.

### 2.3. Download and license

The OMOP-2-OPMI GitHub web page is: https://github.com/OPMI/OMOP-2-OPMI. The source code of the Omop2Opml.owl file is openly available at this GitHub website for downloading. The OWL file is generated primarily by extracting the OMOP CDM-related terms and associated relations from the OPMI using Ontofox [15]. Considering the usage of OPMI as the platform for the OMOP CDM mapping, the OMOP-2-OPMI source page is designated as a repository under the general OPMI organization in GitHub.

Meanwhile, the OMOP-2-OPMI repository has also stored related data files including our cleanup spreadsheets of the mapping details available at: https://github.com/OPMI/OMOP-2-OPMI/tree/main/docs.

### 2.4. Use case studies

Three use cases are developed and discussed in this study. Specifically, the first use case is about the OMOP data standardization and inference. The second use case is the development of an adverse event model based on the OMOP CDM logic and available data formats. The third use case is the usage of OMOP-2-OPMI to study N3C COVID-19 related clinical data.

## 3. Results

### 3.1. General OMOP CDM ontologization architecture

Figure 1 represents the hierarchical structure of the OMOP-2-OPMI, which is the ontologization of the OMOP CDM using the OPMI as the ontology platform. Specifically, all the terms are aligned under the Basic Formal Ontology (BFO) [18], an ISO-approved upper level ontology [19]. BFO includes two branches: continuants and occurrents. Continuants cover time-independent entities including material entities, quality, realizable entities such as disposition, and information content entities. Occurrents are time-dependent entities including temporal region and processes. All the OMOP CDM elements can be categorized under these two categories (Figure 1). BFO has been used by over 300 ontologies. The alignment with BFO allows us to integrate our ontology with the large number of other ontologies, supporting data interoperability.

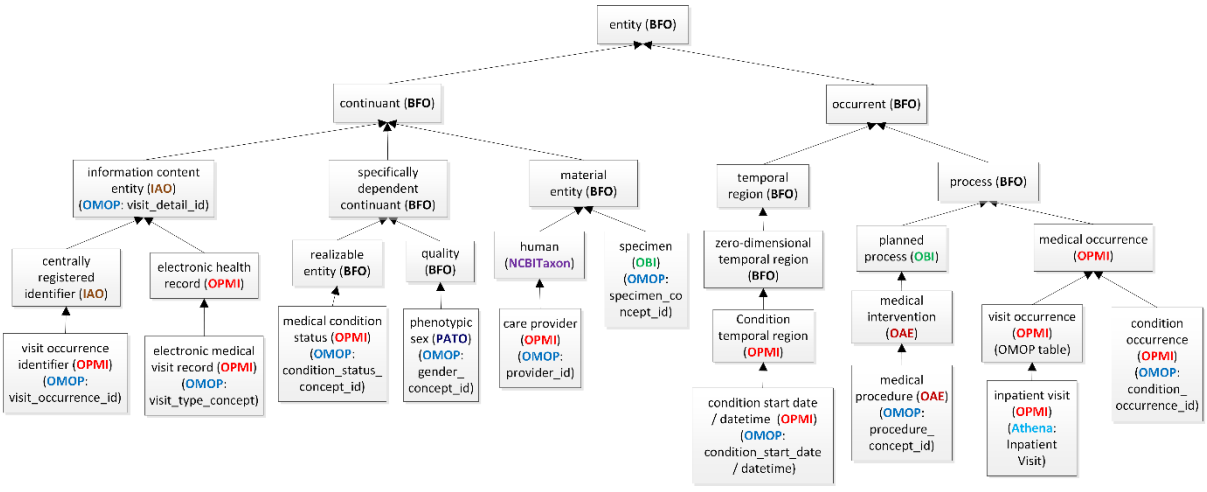


**Figure 1**: OMOP-2-OPMI top level hierarchical structure and representative terms. Ontology names are highlighted with different colors. Ontology-mapped OMOP terms are also provided.

Figure 2 is a simplified high level OMOP-2-OPMI ontology design pattern (ODP) that covers the major elements in 11 OMOP tables. Specifically, the person (usually here it refers to patient in OMOP) is centric to the ODP. The person participates in five medical occurrences (i.e., visit/condition/procedure occurrences, and drug/device exposure) and the observation process, which are all under BFO:process (Figure 1). The observation happens during a specific observation period. The person is also the target of measurement. A specimen derives from some organ or tissue of the person. The person has

different phenotypes, and death is a specific phenotype (Figure 2).

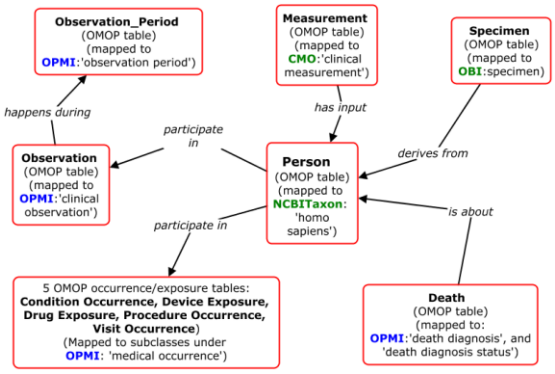


**Figure 2**: General ontology design pattern that links CDM elements from 11 OMOP tables. Note one box covers five OMOP occurrence/exposure tables. Mapped ontology terms are also labeled.

## 3.2. OMOP-2-OPMI statistics

A total of 165 terms from 15 OMOP CDM tables has been mapped to OPMI, with 46 terms newly generated with OPMI namespace and the other terms reported from OBO reference ontologies. In addition to the 11 tables listed in Figure 2, the other four tables are Care Site, Payer Plan Period, Episode, and Location, which are not included in Figure 2 to simplify that figure. Table 1 lists ontology mapped CDM element terms from 10 representative OMOP tables.

Our current mapping primarily covers those clinical data tables and health system data tables. We have not yet included the Metadata Tables, Vocabulary Tables, Standardized Derived Tables except for Episode, and the Cost table which belongs in the Health Economics Data Tables category. These missing tables do not directly involve clinical investigation, which is our current focus. Also as shown in Table 2, many terms are not mapped to ontology. Most of these missing terms are various "source value" or source concept ID terms. Throughout OMOP CDM, there are similar terms representing various source concepts and source values. In the OMOP structure, a source concept set organizes terms into groups called source value sets. A value set (e.g., 'procedure_source_value') is a set of codes whose context and usage are defined by one or more code systems in which the clinical data came from. However, the organization of value sets is not often ontology-based. In most cases, we have decided to not incorporate terms for "source concept" and "source value" sets until we figure out a place for these terms to make sense ontologically within OPMI. In our ontologization, we have also included specific source value terms as seen in Table 1 and detailed later in the manuscript.

**Table 1. CDM terms from 10 representative OMOP tables mapped to OPMI**

| Selected OMOP tables | Mapped OMOP terms | Mapped Ontology Term Examples |
|---|---|---|
| PERSON | 13/19* | person ID (OPMI_0000470), gender (PATO_0001894), year of birth (OPMI_0000473), race (NCIT_C17049) |
| PROVIDER | 9/13 | care provider (OPMI_0000163), National Provider Identifier (OPMI_0000503), DEA identifier (OPMI_0000504) |
| SPECIMEN | 6/15 | specimen ID (OBI_0001616), date of specimen collection (OBIB_0000714), anatomical structure (UBERON_0000061) |
| VISIT OCCURRENCE | **26**/17 | visit occurrence (OPMI_0000482), visit start date (OPMI_0000487), preceding visit occurrence (OPMI_0000492) |
| PROCEDURE OCCURRENCE | 13/16 | procedure (NCIT_C25218), procedure start date (OPMI_0000508), procedure end date (OPMI_0000510) |
| DRUG EXPOSURE | 18/23 | drug exposure (OPMI_0000572), drug product (DRON_00000005) drug exposure start time (OPMI_0000565) |
| CONDITION OCCURRENCE | **38**/16 | condition occurrence (OPMI_0000527), medical condition status (OPMI_0000533), admission diagnosis status (OPMI_0000542) |
| DEVICE | 7/15 | device exposure (OPMI_0000554), device (OBI_0000968), device exposure |

| | | |
|---|---|---|
| EXPOSURE | | start date (OPMI_0000562) |
| MEASUREMENT | 11/20 | clinical measurement identifier (OPMI_0000582), measurement time (OPMI_0000579), measurement unit label (IAO_0000003) |
| OBSERVATION PERIOD | 5/6 | observation period start date (OPMI_0000577), observation period end date (OPMI_0000578), |

Note: *13/19 represents that 13 out of 19 OMOP CDM terms in the specific category have been mapped to terms in the OPMI ontology. The unmapped terms are primarily those terms related to "source value". More terms in the visit/condition occurrences are mapped because some specific source value terms are ontologized.

In addition to source values or source concept IDs, there are also many terms in OMOP CDM not yet ontologized. The reasons of such imcompleteness include the lack of necessity of many terms, and the complexity of many other terms in terms of ontology modeling. We will continue this work later, ideally by involving more collaboration and discussion with the ontology and clinical informatics communities.

**Table 2. Ontology mapping of OMOP CDM terms by element types**

| types | OMOP terms | OMOP mapped | percent mapped |
|---|---|---|---|
| _id | 23 | 19 | 82.61% |
| _date | 34 | 27 | 79.41% |
| _concept_id | 41 | 29 | 70.73% |
| _concept_name | 30 | 16 | 53.33% |
| _source_concept_id | 17 | 1 | 5.88% |
| _source_value | 34 | 1 | 2.94% |
| Total | 179 | 93 | 51.96% |

Next we will focus on a few major ontology modeling topics to show how we model and ontologize the OMOP CDM elements.

## 3.3. Ontologization of OMOP medical occurrences

By examining the OMOP CDM elements, we found that five OMOP tables can be categorized under an ontology class called 'medical occurrence', which is defined as a process event that a patient experiences over a period of time (Figure 3). These five OMOP tables are: 'condition occurrence', 'device exposure', 'drug exposure', 'procedure occurrence', and 'visit occurrence' (Figure 3).

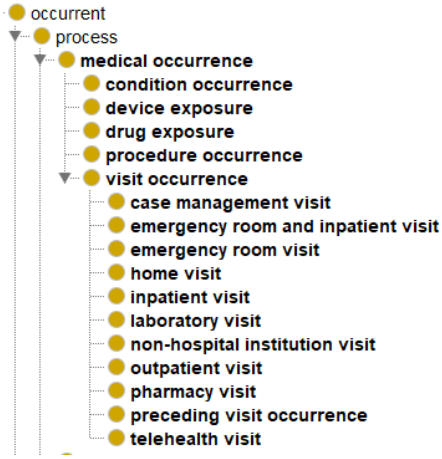


**Figure 3**: Modeling of 5 medical occurrence categories and 11 specific visit occurrences.

In two of the five OMOP tables, Visit Occurrence and Condition Occurrence, in addition to mapping the elements in original tables (Table 1), we also added some terms from the supporting OMOP vocabularies for developing a complete semantic model. In the case of Visit Occurrence, the extra terms are due to the ontologization of 11 types of visit occurrences (e.g., 'emergency room visit', 'home visit') that are originally not defined in OMOP's CDM model and instead are from the supporting OMOP vocabularies identified on the Athena program. We have ontologized such terms under 'visit occurrence' (OPMI) (Figure 3). These terms represent the overarching types of encounters between a person and the healthcare system, which are adopted in most healthcare systems worldwide.

In the case of Condition Occurrence, the extra 22 terms come from the incorporation of medical condition statuses (e.g., 'admission diagnosis', 'cause of death', and 'confirmed diagnosis'), which were defined by OMOP and searchable in Athena. In OMOP, a medical condition status denotes the stages of a patient's diagnosis, not the actual state of the disease by itself. OPMI represents these medical condition statuses in two

strategies. First, OPMI includes a term called 'medical condition status' under the 'status' term, which is a subclass of BFO:'realizable entity'. In this classification, a medical diagnosis status, such as admission diagnosis, represents a patient diagnosis status such as the status of diagnosis at the time when the patient is admitted to the hospital.

We have also adopted the OGMS:diagnosis classification and defines various diagnosis types under the OGMS:diagnosis (Figure 4). According to the Ontology for General Medical Science (OGMS), diagnosis (OGMS_0000073) is a subclass of clinical data item and represents the conclusion of a diagnostic process. Based on the OMOP classification, OPMI has defined different categories of diagnosis, including 'admission diagnosis', 'primary diagnosis', 'secondary diagnosis', and 'death diagnosis', etc. (Figure 4). These specific diagnosis types are commonly used at the clinical setting. The classification of these diagnosis types facilitates the clinical data annotations.

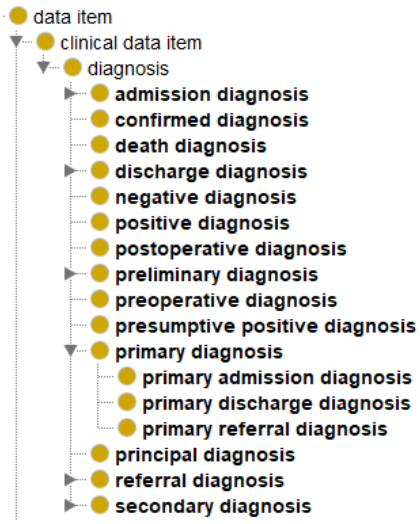


**Figure 4**: Modeling of different medical diagnosis under the OGMS:diagnosis, which is a subclass of clinical data item.

As OPMI separates diagnosis clinical data type vs the diagnosis medical condition status, we can define different diagnoses and diagnosis statuses. For example, 'discharge diagnosis status', 'referral diagnosis status', and 'admission diagnosis status' are realizable entities, and 'discharge diagnosis, 'referral diagnosis status', and 'admission diagnosis' are data items. The main benefit of separate representation of status and data is the semantic separation and clarity. The medical condition status represents the current status of the patient at a specific stage. For example, 'admission diagnosis status' represents the status at which a person is diagnosed at the admission stage. On the other hand, as the data item, the 'admission diagnosis' indicates the conclusion or outcome of the diagnosis process at the stage of patient admission. A diagnosis conclusion made at the admission or discharge stage may be the same or different.

Meanwhile, the diagnosis clinical data type vs the diagnosis medical condition status are closely related. In OPMI, we propose to generate a relation term called 'has status content', which represents a relation between a status and an information content entity where the status has its content information defined by the information content entity. For example, we can define an axiom that links a diagnosis status to a diagnosis data item:

*'admission diagnosis status': 'has status content' some 'admission diagnosis'*

However, such duplicated representation may not be needed. It is possible to just define 'admission diagnosis status' and remove the term 'admission diagnosis'. We will examine more use cases and discuss with the ontology and medical informatics communities on this regard.

### 3.4. Ontologization of temporal date/time in OMOP

To ontologically represent various entities denoting time that can be found throughout OMOP, we have mapped 24 temporal terms from 6 tables. The OMOP tables that have temporal terms ontologized are Visit Occurrence, Device Exposure, Drug Exposure, Procedure Occurrence, Condition Occurrence, and Person. For all tables but Person, the entities are ontologized with temporal terms for -start date, -start datetime, -end date, and -end datetime. Meanwhile, temporal terms related to the Person table are instead ontologized with more familiar terms which are 'birth datetime', 'day of birth', 'month of birth', and 'year of birth'. All temporal terms are grouped under a higher level term for a better organizational purpose (e.g., 'visit start date/datetime', 'end date/datetime' are all grouped under 'visit temporal region') (Figure 5).

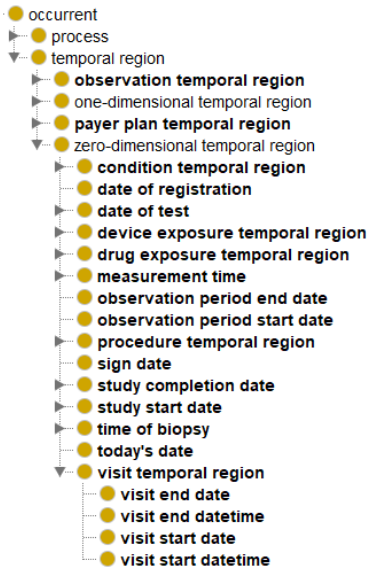


**Figure 5**: OPMI modeling of date and time used in OMOP CDM.

## 3.5. Ontologization of entity identifiers in OMOP CDM

In OMOP, fields with the suffix "_id_" usually denote identifiers, which function as primary keys in their respective OMOP tables along with other supporting entities (e.g., person_id in Person table). These identifiers can also be used as foreign keys to connect other related OMOP tables (e.g., person_id to connect Provider and Care Site tables).

OPMI has ontologized OMOP CDM related identifiers under the class of 'centrally registered identifier', a subclass under 'information content entity'. Example identifiers defined include 'person ID', 'care site identifier', 'clinical measurement identifier', 'DEA identifier' and 'National Provider Identifier'. These identifiers identify assets belonging to different but centrally registered local databases.

## 3.6. Ontologization of provenance records in OMOP

In OMOP, most entities from various tables have their own "type_concept" terms, which indicate the provenance, or the source of the record in which it comes from. For instance, drug exposure entries could come from either prescriptions list or self-reported by patients, the provenance of which can differ from a patient's measurement records.

In OPMI, type_concepts are mapped as various terms under 'provenance of record', a class under 'information content entity'. So far, we have generated 12 terms for the provenance of records for 12 corresponding entities of OMOP CDM tables. The provenance of records is dedicated for each corresponding OMOP entity since the sources of the entries can vary across different fields.

Meanwhile, OPMI also defines most of the records for the OMOP provenance purposes under 'electronic health record', such as 'electronic medical visit record,' 'electronic death record,' 'electronic device record,' etc. (Figure 6). The users can choose the usage of these electronic health records as the sources of the data collected to the OMOP database. Note that not all the provenance records are electronic health records (EHR). For example, in addition to the record from an EHR system, the measurement record might also come from an insurance claim, registry, or other sources.

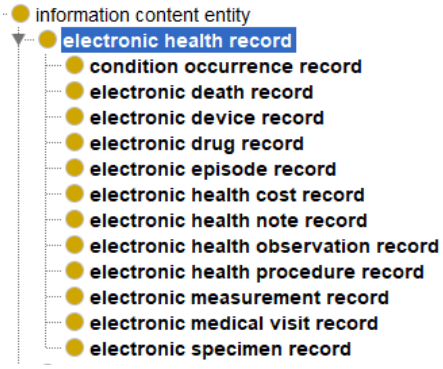


**Figure 6**: OPMI modeling of different records used as data provenance in OMOP CDM.

Next, we will focus on the description of three use cases of the OMOP-2-OPMI approach.

## 3.7. Use case 1: Ontology-level data standardization

The first use case is rooted in the nature of ontology. As an open access ontology following the OBO ontology development principles, OMOP-2-OPMI provides the standard representation and definitions of the OMOP CDM mapped terms and the axioms among these terms. The OMOP-2-OPMI ontology terms can be used to support standardized clinical data representation and annotation. The semantic relations among the OMOP CDM terms and their associated other terms provide solid semantic

associations, which addresses the OMOP CDM drawback of weak semantics.

The ontologized terms are also interoperable. For example, the Coronavirus Infectious Disease Ontology (CIDO), a biomedical ontology in the domain of coronavirus diseases [20], has imported the OMOP-2-OPMI ontology contents. The contents of OMOP-2-OPMI fit seamlessly with the other CIDO contents, providing another demonstration of the ontology-supported knowledge and data interoperability, sharing, and integration. It is also possible to use the some ontology terms for mapping to the other CDMs such as PCORnet [4] and CDISC [5], which will be explored in the future.

Such interoperable ontology representation also supports data and knowledge inferencing. This is also rooted from the nature of ontology. The following two other use cases provide such demonstrations.

## 3.8. Use case 2: Adverse event modeling and analysis

Another use case of the OMOP CDM ontologization is the modeling of adverse events (AEs) post medical intervention. The OMOP CDM does not include AE per se. However, by specific modeling, we can find the OMOP CDM data can be processed to support specific AE identification and analysis.

Figure 7 is a general OMOP-AE ontology design pattern, which follows the AE definition by the Ontology of Adverse Events (OAE) [21]. According to the OAE, an adverse event (AE) is a pathological bodily process that occurs following some medical intervention [21]. In order to model AEs with OMOP data, we need to identify the medical intervention vs. adverse events to be mapped in OMOP. By examining all the five medical occurrence types defined in OMOP, only three of them are considered as medical interventions: Drug Exposure, Device Exposure, and Procedure Occurrence (e.g., surgical procedure). Vaccination can be considered as a special drug exposure.

Note that the visit occurrence and condition occurrence are regarded as natural occurrence events without medical intervention. Based on the AE definition, contracting a natural infection is not an AE since the patient does not receive an adverse outcome after a medical intervention. However, the condition occurrence may include conditions of different phenotypes that are the outcomes of specific adverse events (Figure 7).

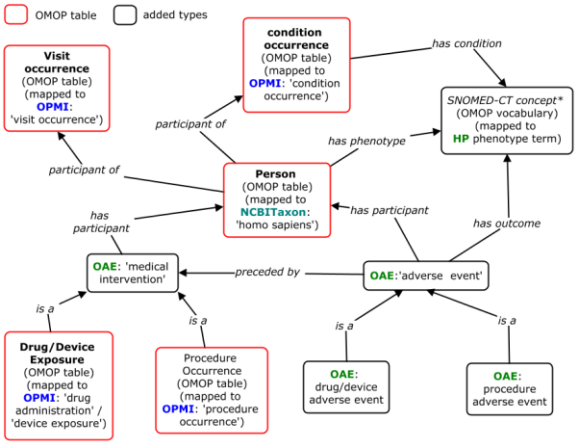


**Figure 7**: General OMOP-AE model based on OMOP-2-OPMI. The red boxes represent OMOP tables and their mapped ontology terms. The black boxes are added ontology representation to fill up the gaps for adverse event modeling. *, OMOP uses SMOMED-CT concepts for disease or symptom representation. These can be mapped to Human Phenotype Ontology (HP) terms.

Our original OPMI conference proceeding paper presented a use case study of identifying and analyzing the acute kidney injury (AKI) AE following heart surgery [9]. Using OHDSI data provided by the IQVIA Pharmetric Plus database, our OHDSI cohort study identified a total of 15,548 patients that fulfilled our predefined model of AKI AE following heart surgery. Specific patterns were identified. For example, 72% of the identified patients were male and 28% were female patients. Over 78% of these AE cases occurred in patients aged greater than 55 years old. Many phenotypes, such as coronary arteriosclerosis, kidney disease, pain, dyspnea, hyperlipidemia, and Type II diabetes, were found in these patients as well [9].

Our OMOP AE model is a very general model in that it can be used to study specific adverse event profiles following various medical interventions including different drug/medicine exposure and procedure occurrence. We are currently applying such a strategy to design a pattern for identifying and analyzing the vaccine and drug AEs in COVID-19 patients using the N3C data. Note that if a patient contracted COVID-19 in a natural environment, the patient has a condition, which is not an adverse event (because an AE is always associated with medical intervention). However, the occurrence of new

phenotypes after medical treatment on these COVID-19 patients are considered AEs.

### 3.9. Use case 3: COVID-19 clinical data standardization, modeling, and analysis

In addition to the import of the OMOP-2-OPMI to CIDO and the study of COVID-19 associated AE modeling and analysis as described above, we are also applying the OMOP-2-OPMI for more COVID-19 clinical data modeling and analysis. Two data resources for our OMOP-2-OPMI based studies are the literature reports and N3C clinical data.

One specific use case is the study of the relation between the COVID-19 infection and the increased risk for kidney diseases. For example, acute kidney injury (AKI) is a significant complication of COVID-19. The incidence of AKI in hospitalized patients varies from 0.5% to 75%. The mortality rate for patients with kidney disease is also significantly higher than the general infected population. However, the big variation of AKI incidence in COVID-19 patients appears to depend on many factors such as race, region, and disease severity. The N3C cohort data is being used to detect, compare, and analyze the occurrences of kidney disease following COVID-19 infection. The OMOP-2-OPMI model, together with the OMOP2OBO, can be used to support data modeling, integration, and analysis. The integrated data can also be further used for machine learning tool development for kidney disease prediction following COVID-19 prediction. We have registered for an N3C program to perform related research.

Another use case in this category is the application of OMOP-2-OPMI and CIDO for secondary literature data analysis and knowledge representation. There have been a big number of COVID-19 studies reported in the literature, many of which involve the usage of OMOP CDM model. For example, one study examined the association between immune dysfunction and COVID-19 breakthrough infection after SARS-CoV-2 vaccination in the US using N3C data [22]. The N3C data and the results out of the data analysis can both be modeled, annotated, and represented using ontology including our OMOP-2-OPMI and CIDO.

The above two studies are currently ongoing and we expect to have more specific results available in near future.

## 4. Discussion

This manuscript has made two main contributions. First, we report our systematic survey and ontologization of the OMOP CDM elements using the OPMI ontology. The Omop-2Opmi.owl file is the OWL file that includes only the OMOP CDM-related ontology terms, their directly associated terms (e.g., their parent terms), and the semantic relations between these terms that are presented as ontology axioms. Second, we presented three categories of use cases of our OMOP CDM ontologization, including ontology-level OMOP CDM element standardization and inferencing, adverse event modeling and analysis, and COVID-19 clinical data studies. Overall, our systematic ontologization of the OMOP CDM complements and empowers the OMOP CDM system, providing a new way of supporting systematic clinical data interoperability, sharing, and integration.

A similar and related system is OMOP2OBO, a systematic mapping tool that maps OMOP related terms to OBO ontologies [7]. The terms mapped in OMOP2OBO cover 8 OBO ontologies, including Cell Ontology (CL), ChEBI chemical entity ontology, Human Phenotype Ontology (HP), MONDO disease Ontology, NCBI Taxonomy Ontology (NCBITaxon), Protein Ontology (PR), Uberon anatomy ontology, and Vaccine Ontology (VO). While OMOP2OBO includes the mapping of over 100,000 terms in the OMOP terminology system, it does not cover the OMOP CDM elements in the over 10 basic OMOP tables. Instead, OMOP-2-OPMI focuses on the core OMOP CDM level mapping and representation. In addition to ontology term mapping, since many high level terms in OMOP CDM are not yet represented in OBO ontologies, we have taken extensive effort to generate many new terms in OPMI. We have also generated ontological relations among these OMOP CDM elements using the OPMI ontology platform. Overall, OMOP2OBO and OMOP-2-OPMI are complementary in that they map and integrate OMOP data from different aspects.

There are still many issues to consider in our ontologization. For example, we presented two types of methods for representing medical condition statuses and two types of methods of

representing provenance records in our work. Since most medical condition statuses are different types of diagnosis, such status representations can be defined under "status", which is defined as a BFO:'realizable entity', or under OGMS:diagnosis, which is basically a type of clinical data item. Similarly, for the provenance records, they can be represented under provenance itself or under electronic health record. The ICBO-2022 conference will provide us a discussion platform to discuss the pros and cons of different representation styles.

Several use cases are introduced in this article. We demonstrated the development of a new OMOP-based adverse event model based on the OMOP CDM data structure. Such an OMOP-AE model can be used to support various specific AE studies, including the modeling of adverse event cases post COVID-19 vaccination (or drug admin) using N3C data. In addition to the AKI AE study following heart surgery [9], we are currently applying the OMOP AE model for more COVID-19 related AE studies. Furthermore, we can develop new models to apply OMOP CDM to study other topics such as long COVID and the effects of different variables to the disease outcomes.

One future project is to map the CDM terms from other systems, including PCORnet [4] and CDISC [5], to the OPMI ontology using the same OMOP-2-OPMI development strategy. These different CDMs are overlapped. For example, There are similarities between the organizations of OMOP and PCORnet CDMs, evidenced by the overlaps of certain tables such as Demographic, Procedures, or Condition [23]. When all these CDM elements and relations are mapped to the same OPMI structure, we can integrate all the data using different CDMs, leading to compatible and interoperable clinical and observational data standardization and integration. A recent study reports the development of an ETL tool for converting the PCORnet CDM into OMOP CDM to facilitate the COVID-19 data integration [24]. It is possible to apply our ontology approach to enhance such an ETL tool.

## 5. Acknowledgements

We acknowledge the Kidney Precision Medicine Project (KPMP) project supported by NIH-NIDDK grant: 1U2CDK114886, and a COVID-19 research grant from the Michigan Medicine–Peking University Health Sciences Center Joint Institute for Clinical and Translational Research (U072807). We appreciate the discussion and comments from the ontology and OMOP societies including Dr. Asiyah Yu Lin and Dr. Andrew Williams.

## 6. References

[1] E. A. Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M. Schuemie, *et al.*, "Feasibility and utility of applications of the common data model to multiple, disparate observational health databases," *J Am Med Inform Assoc*, vol. 22, pp. 553-64, May 2015.

[2] W. Ceusters and J. Blaisure, "A Realism-Based View on Counts in OMOP's Common Data Model," *Stud Health Technol Inform*, vol. 237, pp. 55-62, 2017.

[3] Y. He, E. Ong, and J. Zheng, "Ontological representation of OMOP CDM using the OBO framework," presented at the 2018 OHDSI Symposium, Bethesda North Marriott, Bethesda, MD, 2018.

[4] F. S. Collins, K. L. Hudson, J. P. Briggs, and M. S. Lauer, "PCORnet: turning a dream into reality," *J Am Med Inform Assoc*, vol. 21, pp. 576-7, Jul-Aug 2014.

[5] S. Hume, J. Aerts, S. Sarnikar, and V. Huser, "Current applications and future directions for the CDISC Operational Data Model standard: A methodological review," *J Biomed Inform*, vol. 60, pp. 352-62, Apr 2016.

[6] *COVID-19 Clinical Data Warehouse Data Dictionary Based on OMOP Common Data Model Specifications Version 5.3* Available: https://ncats.nih.gov/files/OMOP_CDM_COVID.pdf

[7] T. J. Callahan, J. M. Wyrwa, N. A. Vasilevsky, and P. N. Robinson, "OMOP2OBO: Semantic Integration of Standardized Clinical Terminologies to Power Translational Digital Medicine Across Health Systems," in *2020 OHDSI Symposium*, Virtual meeting, 2022.

[8] T. J. Callahan. (2022). *OMOP2OBO*. Available: https://github.com/callahantiff/OMOP2OBO

[9] Y. He, E. Ong, J. Schaub, F. Dowd, J. F. O'Toole, A. Siapos, *et al.*, "OPMI: the

Ontology of Precision Medicine and Investigation and its support for clinical data and metadata representation and analysis," in *The 10th International Conference on Biomedical Ontology (ICBO-2019), July 30 - August 2*, Buffalo, NY, USA., 2019, pp. 1-10.

[10] E. Ong, L. L. Wang, J. Schaub, J. F. O'Toole, B. Steck, A. Z. Rosenberg, *et al.*, "Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project," *Nat Rev Nephrol*, vol. 16, pp. 686-696, Nov 2020.

[11] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotechnol*, vol. 25, pp. 1251-5, Nov 2007.

[12] *OMOP CDM version 5.4* Available: http://ohdsi.github.io/CommonDataModel/cdm54.html

[13] Y. He, Z. Xiang, J. Zheng, Y. Lin, J. A. Overton, and E. Ong, "The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability," *J Biomed Semantics*, vol. 9, p. 3, Jan 12 2018.

[14] E. Ong, Z. Xiang, B. Zhao, Y. Liu, Y. Lin, J. Zheng, *et al.*, "Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration," *Nucleic Acids Res*, vol. 45, pp. D347-D352, Jan 04 2017.

[15] Z. Xiang, M. Courtot, R. R. Brinkman, A. Ruttenberg, and Y. He, "OntoFox: web-based support for ontology reuse," *BMC Res Notes*, vol. 3:175, pp. 1-12, 2010.

[16] M. A. Musen, "The Protégé project: A look back and a look forward. AI Matters. ," *Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, vol. 1, p. DOI: 10.1145/2557001.25757003., 2015.

[17] *Hermit OWL reasoner*. Available: http://hermit-reasoner.com/

[18] R. Arp, B. Smith, and A. D. Spear, *Building Ontologies with Basic Formal Ontology*. MIT Press: Cambridge, MA, USA, 2015.

[19] *ISO/IEC 21838-2:2021. Information technology — Top-level ontologies (TLO) — Part 2: Basic Formal Ontology (BFO)*. Available: https://www.iso.org/standard/74572.html

[20] Y. He, H. Yu, E. Ong, Y. Wang, Y. Liu, A. Huffman, *et al.*, "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis," *Sci Data*, vol. 7, p. 181, Jun 12 2020.

[21] Y. He, S. Sarntivijai, Y. Lin, Z. Xiang, A. Guo, S. Zhang, *et al.*, "OAE: The Ontology of Adverse Events," *J Biomed Semantics*, vol. 5, p. 29, 2014.

[22] J. Sun, Q. Zheng, V. Madhira, A. L. Olex, A. J. Anzalone, A. Vinson, *et al.*, "Association Between Immune Dysfunction and COVID-19 Breakthrough Infection After SARS-CoV-2 Vaccination in the US," *JAMA Intern Med*, vol. 182, pp. 153-162, Feb 1 2022.

[23] *PCORnet Common Data Model (CDM) Specification, Version 6.0*. Available: https://pcornet.org/wp-content/uploads/2022/01/PCORnet-Common-Data-Model-v60-2020_10_221.pdf

[24] Y. Yu, N. Zong, A. Wen, S. Liu, D. J. Stone, D. Knaack, *et al.*, "Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration," *J Biomed Inform*, vol. 127, p. 104002, Mar 2022.