OPMI: the Ontology of Precision Medicine and Investigation and its support for clinical data and metadata representation and analysis

Yongqun He[1], Edison Ong[1], Jennifer Schaub[1], Frederick Dowd[2], John F. O'Toole[3], Anastasios Siapos[4], Christian Reich[4], Sarah Seager[4], Ling Wan[1,5], Hong Yu[6], Jie Zheng[7], Christian Stoeckert[7], Xiaolin Yang[8], Sheng Yang[8], Becky Steck[1], Christopher Park[2], Laura Barisoni[9], Matthias Kretzler[1], Jonathan Himmelfarb[2], Ravi Iyengar[10], Sean D. Mooney[2], for the Kidney Precision Medicine Project Consortium

[1] University of Michigan Medical School, Ann Arbor, MI 48109, USA; [2] University of Washington, Seattle, WA 98195, USA; [3] Cleveland Clinic, Cleveland, OH, USA; [4] IQVIA, Brighton, UK; [5] OntoWise, Nanjing, Jiangzu, China; [6] Department of Pulmonary and Critical Care Medicine, Guizhou Provincial People's Hospital, Guiyang, Guizhou 550002, China; [7] University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA; [8] Institute of Basic Medical Science, Chinese Academy of Medical Sciences, Beijing, China; [9] Duke University, NC, USA; [10] University Icahn School of Medicine at Mount Sinai, NY 10029, USA.

***Abstract*—Consortia conducting precision medicine studies face a major challenge of integrating big data including clinical and biomedical data. In this study, we report our development of the community-driven Ontology of Precision Medicine and Investigation (OPMI) and its applications in clinical data and metadata representation. OPMI has been used to represent the common data model (CDM) of the Observational Health Data Sciences and Informatics (or OHDSI) program. It has also been used to represent approximately 30 case report forms defined by the NIH-supported Kidney Precision Medicine Project (KPMP). Our case studies showed that OPMI is able to semantically and precisely represent the OHDSI CDM, various KPMP clinical forms, and their associated data and metadata. Such ontological representations support standardized data representation, sharing, recording, integration, and advanced analysis.**

***Keywords*— Common data model; kidney; case report form.**

I. Introduction

Precision medicine is an emerging medical approach for disease prevention and treatment that takes into account individual variability in genes, environment, and lifestyle. An example of a study in precision medicine is the Kidney Precision Medicine Project (KPMP; http://kpmp.org), a large NIH/NIDDK-funded consortium project with the aim of understanding and treating human kidney diseases. With a focus on human studies, the KPMP project covers clinical recruitment, clinical study, biopsy, pathology, molecular data and Omics data analysis. With the large amounts of data generated, we will identify how to systematically collect, represent, integrate, and analyze and make use of the big data with the help of ontologies.

Precision medicine faces the challenge of big data. Big data represents the data characterized with the 5 Vs: volume, veracity, velocity, variety, and value [1], which requires specific technology and analytical methods for its transformation into meaningful knowledge.

In precision medicine, basic research results, such as Omics study results, are affected by many clinical factors. Clinical factors (e.g., biological sex and age) are generally poorly recorded and studied. Before investigators can deeply and accurately analyze precision medicine data, the clinical data need to be captured and modeled systematically and robustly. For example, to achieve this goal, KPMP investigators created over 30 case report forms (CRFs), which are being used across many institutes. These clinical forms cover over 2000 questions and hundreds of clinical factors. Each of the clinical factors may affect the phenotype or omics analysis outcomes.

To support clinical data collection and analysis, there have exist many common data models (CDMs), including the CDMs of the OHDSI Observational Medical Outcomes Partnership (OMOP) [2], the Patient-Centered Outcomes Research Network (PCORnet) [3], the healthcare management organizations' research network (HMORN) virtual data warehouse [4], and the Study Data Tabulation Model (SDTM) of the Clinical Data Interchange Standards Consortium (CDISC) [5]. One issue is that these CDMs are often not interoperable at the semantic level. We hypothesized that an ontological representation of the OMOP CDM (and other CDMs) would better semantically represent and standardize the data formatted based on the CDM and support better data analysis. As an example, the OMOP CDM is a relational database model that supports interoperable analyses of disparate observational databases [2]. The OMOP CDM has been widely adopted to support the accommodation of observational medical data from disparate data sources. However, the terms in the OMOP CDM lacks strong semantic relations. For example, the "Condition" in the OMOP CDM could be a natural disease or an adverse event following a surgery or drug administration. The usage of ontology makes it possible to better differentiate the two types of conditions and support better data representation and analysis.

A formal biomedical ontology is a human-comprehensible and computer-interpretable set of terms and relations that represent entities in a specific domain and their relationships to each other. The Open Biological/Biomedical Ontology (OBO) community [6] has developed over 150 biomedical ontologies that support alignment with each other. Most current OBO ontologies cover basic research domains. Our proposed Ontology of Precision Medicine and Investigation (OPMI) has recently been included in the OBO library ontology list, which aims to focus on the representation of entities and relations in the domain of precision medicine and its investigation.

In this study, we report the OPMI development strategy and results with a focus on its supporting clinical studies. OPMI has been used to ontologize OMOP CDM and CRFs and to further support the KPMP precision medicine study.

II. Methods

A. OPMI ontology development methods

OPMI is developed as a community-based open source biomedical ontology by following the OBO Foundry ontology development principles such as openness and collaboration [6]. The eXtensive Ontology Development (XOD) strategy [7] was applied for the ontology development. Specifically, OPMI reuses many terms and relations from existing ontologies, including the Ontology of General Medical Science (OGMS) [8], Ontology for Biomedical Investigations (OBI) [9, 10], Human Phenotype Ontology (HP) [11], Uberon multi-species anatomy ontology (UBERON) [12], Ontology of Adverse Events (OAE) [13], and Informed Consent Ontology (ICO) [14]. The tool Ontofox (http://ontofox.hegroup.org) [15] was used to extract and reuse terms from these existing ontologies.

OPMI-specific terms were assigned new identifiers using the prefix "OPMI_" followed by auto-generated seven-digit numbers. The Protégé OWL editor (http://protege.stanford.edu/) was used for the OPMI visualization and manual term editing. The Hermit reasoner (http://hermit-reasoner.com/) inside the Protégé OWL editor was applied for ontology consistency checking and inferencing.

B. OPMI representation and analysis of OHDSI CDM

We used OPMI to ontologically model the OMOP CDM used in the OHDSI program. As the underlying data standard of OHDSI, the OMOP CDM allows for interoperable analyses of disparate observational databases. To demonstrate the usage of OPMI to study OMOP CDM, we used the data extracted from the IQVIA Pharmetric Plus database data (https://www.iqvia.com), which had already been converted into the OMOP CDM format. In this study, kidney disease data were extracted from the database based on the OPMI data model. Supported by this model, we developed an algorithm to identify the concept IDs that covered the correct conditions of interest. Once identified, we extracted the patients who initially did not have acute kidney injury (AKI), then were treated with heart surgery, and diagnosed with AKI with 14 days after the surgery. The SNOMED concept term "Acute renal failure syndrome" and 62 other associated concept terms were used. The conditions within 30 days before the heart surgery were extracted and mapped to the Human Phenotype Ontology (HP) [11]. To better analyze the subset of related HP terms, the tool Ontofox [15] was used to extract these HP terms and their associated upper level terms, and the Protégé OWL editor tool [16] was used to display the structure.

C. OPMI representation of KPMP case report forms and their contents using CRF-Question-Entity model

The KPMP CRFs were extracted, modeled, and analyzed using the OPMI platform. The CRFs and the contents defined in CRFs were represented using a newly designed "CRF-Question-Entity" model. Based on this model, OPMI generates specific ontology terms to represent various CRFs in the ontology. Each CRF usually includes many textual questions, e.g., "Are you aged less than 18 years old?" OPMI also represents such textual questions, and also identifies the entities in reality (e.g., age and its value of less than 18 years old) that are referred to by the questions. Many of these entity terms are imported from existing ontologies. All the labels, synonyms and definitions of the CRFs and CRF-related terms were carefully evaluated by the KPMP community and domain experts in the field.

D. OPMI format, source code, license, and deposition

Formatted in the W3C standard Web Ontology Language (OWL2), the OPMI source code is open and freely available at GitHub: https://github.com/OPMI/opmi. The OPMI uses the open Creative Commons CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

The OPMI ontology is deposited in several well recognized ontology repositories, including the Ontobee [17] website: http://www.ontobee.org/ontology/OPMI, NCBO BioPortal website: https://bioportal.bioontology.org/ontologies/OPMI, as well as OLS: https://www.ebi.ac.uk/ols/ontologies/opmi.

E. OPMI query and analysis

To demonstrate the usage of OPMI, we developed SPARQL scripts to query OPMI using Ontobee's SPARQL query endpoint (http://www.ontobee.org/sparql), and DL (description logic) query using the Protégé OWL editor.

III. Results

A. OPMI design and top level structure

Fig. 1 illustrates selected key OPMI terms and top level hierarchical structure. OPMI adopts the Basic Formal Ontology (BFO) [18, 19] as its upper level ontology. The BFO:continuant branch represents entities (e.g., 'material entity' which endure through time. The BFO:occurrent branch represents entities that are temporal (e.g., temporal region) and which occur over time (e.g., 'process'). As the default upper level ontology in the OBO ontology community, BFO has been adopted by many ontologies. The alignment with the BFO structure makes OPMI interoperable with a large number of other ontologies, including those OBO ontologies.

OPMI imports and semantically links terms from many existing biomedical ontologies, such as OGMS [8], OBI [9, 10], HP [11], UBERON [12], and ICO [14] (Fig. 1). There are many reasons to choose these ontologies. First, the importing and reusing of these reliable precision medicine-related ontology terms avoids the reinvention of the wheel and also provides a good starting point for OPMI development. Second, all these ontologies are reliable OBO library ontologies (http://obofoundry.org/) and can all be aligned with the same upper level ontology BFO. Such alignments allow the interoperability among these reused terms with the same semantic relations. The semantic alignments and interoperability also make it efficient to build up OPMI. It is noted that the OBO Foundry aims to form a non-redundant set of ontologies to cover different biological and biomedical areas, the terms imported from the other ontologies are designed to be unique and do not overlap with terms from other OBO library ontologies.

OPMI also includes many OPMI-specific precision medicine-related terms such as ‘precision medicine investigation’. The newly added OPMI terms also includes those CRF terms, textual questions used in CRFs, the question-related entities in reality, clinical metadata terms related to precision medicine studies, and terms related to clinical and health-related CDMs.

The most important reason why OPMI focuses on ontologization of CRFs and CRF questions is that the CRF development is critical to clinical studies and a lot of questions are frequently reused. But it is time consuming to build up new CRFs from the ground, and it is difficult to compare the questions and results from different CRFs. To make more efficient CRF design and usage, it would be important to standardize CRF components. Textual questions are the key components of CRFs. The same questions (e.g., age and biological sex questions) may appear in different CRFs. Therefore, the standardization of the questions becomes essential to the whole CRF standardization process. Meanwhile, the same textual question may be expressed in different ways. From a scientific research standpoint, we should more focus on what each question is really about in reality, i.e., the entities or metadata types behind each question rather than how a question is expressed. Accordingly, we developed the CRF-Question-Entity strategy with the aim to standardize CRF questions, entities (or metadata types) and answers under these questions, leading to the standardization and efficient analysis of different CRFs. While the KPMP project will learn a lot from the ontologization of KPMP CRFs and their contents, many of benefits will go to future CRF studies that do not have to go over the CRF generation from scratch as KPMP has done.

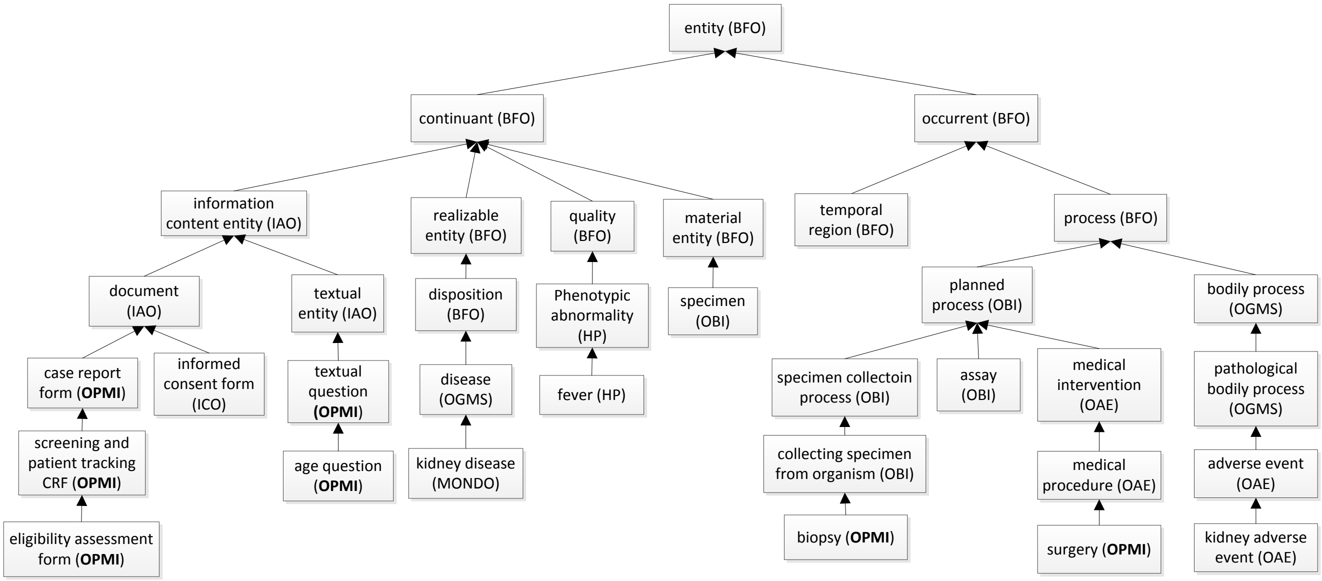


Fig. 1. OPMI top level hierarchical structure and representative terms. All terms are aligned together under the BFO structure.

B. *OPMI ontology design pattern to support OMOP CDM*

Figure 2 represents the overall layout of OPMI ontological representation of OMOP CDM. OPMI ontology unambiguously represents the clinical terms defined in OMOP CDM and the relations among these terms. Established on a realism-based view [20], OPMI treats ‘visit occurrence’ as a process and ‘visit detail’ as information content entity. Many other processes, including ‘procedure occurrence’ and ‘device exposure’ but not necessarily ‘drug exposure’, are ‘part of’ the visit occurrence process. OPMI separates ‘condition occurrence’ into different scenarios including disease course, symptom phenotype, and drug/surgery adverse events. To support specimen-focused precision medicine investigations, OPMI also includes additional terms such as ‘specimen collection’ and ‘specimen assay’, which are linked to OMOP elements (e.g. specimen and measurement).

The OPMI model clearly shows the differences between natural disease courses and adverse events. A disease course is a pathological bodily process that produces specific signs or symptoms at a specific location of a patient. An adverse event is a pathological bodily process that occurs after a medical intervention such as a drug exposure or a surgery procedure [13]. According to the FDA standards, it is not necessary to have a causal relation between the medical intervention and the adverse event outcome. However, the main aim of adverse event study is to identify potential causal relations. To identify whether a surgery adverse event occurs, we need to ensure that an abnormal medical condition occurs after a surgery instead of before it. Such a strategy was then used in our kidney adverse event use case study as described below.

In OMOP CDM-based database schema, foreign keys are used to link different tables. In OPMI, the relations among these entities are more clearly represented using well-defined relations that are commonly used among OBO ontologies. New relations are also generated (Figure 2).

Note that such a class level ontology design pattern (Figure 2) can also be used to represent instance level data, which can be stored in a RDF triple store and subject to SPARQL queries and analyses.

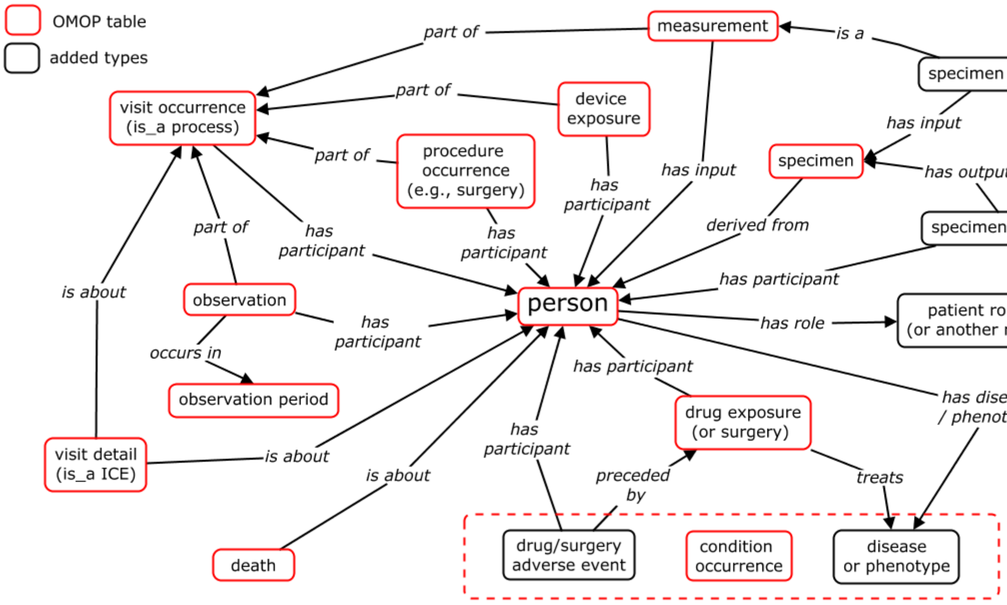


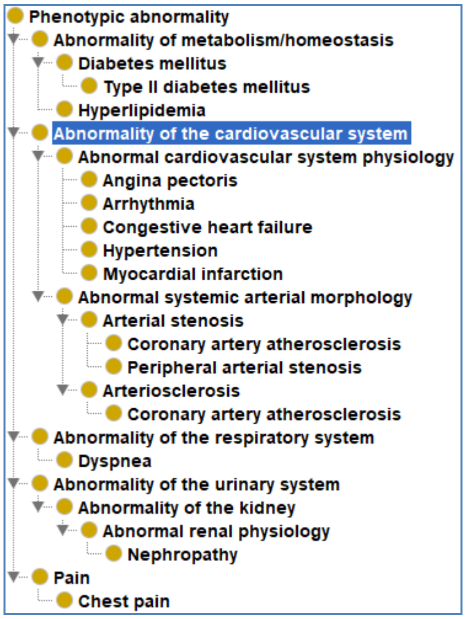
Fig. 2. OPMI ontological representation of OMOP CDM elements and their relations. The terms highlighted in red boxes are table names in OPMI CDM that are also represented as OPMI ontology terms. The terms in black boxes represent ontology terms in OPMI to add values to the OMOP CDM. The lines with text in the middle represent the relations (i.e., object properties) between different terms. OMOP model uses relational database primary keys and foreign keys to make links between different CDM elements. In contrast, OPMI uses the ontology relations to more explicitly represent the linkages between terms. Such ontology relations have the advantage of logically defining the relation meanings and directions with input and output. ICE: information content entity.

C. OHDSI kidney data analysis using OPMI stratregy

An important precision medicine application is related to the precision medical intervention to reduce the occurrence of various adverse events, especially severe adverse events. It is possible that the occurrences of these adverse events are due to various genetic, health or environmental conditions. If we can identify important conditions that are correlate with the adverse events, we can then design rational tests to reduce the threats of adverse events and support public health.

In this study, we hypothesize that ontology-based semantic modeling, together with the usage of ontologies, including Human Phenotype Ontology (HP) and Ontology of Adverse Events (OAE), could help clarify different conditions in OMOP CDM-compatible database, and better understand the contributions of different factors to the presence of specific adverse events. In the area of kidney adverse event research, surgery and drug-induced kidney injury is common, well recognized and an important public health problem. For example, heart surgeries are often followed with AKI adverse events [21]. The incidence of AKI among patients after cardiac surgery can be up to 30-50% [21, 22]. Many risk factors are associated with AKI after cardiac surgery, for example, advanced age, female gender, hypertension, hyperlipemia, diabetes, surgery types, etc. [21, 23]. Therefore, the study of this highly prevalent and prognostically important AKI adverse event after heart surgery is very needed to the public health. The knowledge learned from this study may also later help the study of drug-associated kidney adverse events.

Based on the Fig. 2 OPMI modeling, we developed an algorithm to differentiate surgery adverse events from natural diseases. Specifically, our algorithm identifies and treats the heart surgery time as the index time. To be qualified as an AKI adverse event following heart surgery, the patient should not have AKI during a period before the index time, and have AKI during a short period after the index time. We then used ontologies to represent the phenotypes, heart surgeries, and adverse events systematically, with the aim to identify insightful patterns.



- Phenotypic abnormality
  - Abnormality of metabolism/homeostasis
    - Diabetes mellitus
      - Type II diabetes mellitus
    - Hyperlipidemia
  - Abnormality of the cardiovascular system
    - Abnormal cardiovascular system physiology
      - Angina pectoris
      - Arrhythmia
      - Congestive heart failure
      - Hypertension
      - Myocardial infarction
    - Abnormal systemic arterial morphology
      - Arterial stenosis
        - Coronary artery atherosclerosis
        - Peripheral arterial stenosis
      - Arteriosclerosis
        - Coronary artery atherosclerosis
  - Abnormality of the respiratory system
    - Dyspnea
  - Abnormality of the urinary system
    - Abnormality of the kidney
      - Abnormal renal physiology
        - Nephropathy
  - Pain
    - Chest pain

Fig. 3. Identification of conditions associated with the heart surgery and the following AKI adverse event using OHDSI data and OPMI ontololgy modeling. The condition terms are represented using HPO.

We used OHDSI data provided by the IQVIA Pharmetric Plus database. Our OHDSI cohort study identified a total of 15,548 patients that fulfilled our selection criteria. These patients were categorized as having a heart surgery-associated AKI adverse event.

Our demographic study of the cohort data showed that among all the identified 15,548 patients, 72% are male and 28% are female patients. The patient groups aged greater than 55 years old occupied 78.5% of the AKI adverse event cases. The high incidence in advanced age group is consistent with the previous report [21]. Different from the previous reports of higher risk in female patients [21, 24], our study showed a much higher incidence (18:7) in male patients than in female patients. The underlying reasons deserve further investigation.

The conditions during 30 days before the heart surgery associated with AKI adverse events are represented and classified using the Human Phenotype Ontology (HPO) (Fig. 3). The largest group of phenotype conditions is the abnormality of the cardiovascular system. Many of these conditions might be reasons for heart surgery, and some of them might have higher chance to causally link with the AKI occurrence. For example, our study found that 8433 patients (54%) had coronary arteriosclerosis. The identified patients were also associated with other phenotypes including kidney disease, pain, dyspnea, hyperlipidemia, and Type II diabetes (Fig. 3). Our cohort includes 7,546 patients with hypertensive disorder, 4,684 with kidney disease, 5,121 with hyperlipidemia, 4,561 with Type 2 diabetes, and 4,523 with dyspnea.

Specific surgery types were also identified. For example, our cohort study found that many patients underwent different types of valvular procedures, which were previously found to be associated with a higher risk [23].

D. *OPMI representation of KPMP case report forms*

Figure 4 demonstrates the representative list of KPMP CRFs. In total, KPMP includes approximately 30 CRFs used in different stages of clinical study. These stages cover the screening and patient tracking, enrollment, pre-biopsy, biopsy, post-biopsy, and pathology test, etc. Overall, these CRFs cover over 2,800 questions. Each question is about some specific entities related to the clinical study. Note that for the US Food and Drug Administration (FDA), a case report form often means the cases of adverse events. However, in clinical trials or clinical studies, a case report form means any form that related to clinical study, which has a broader coverage.

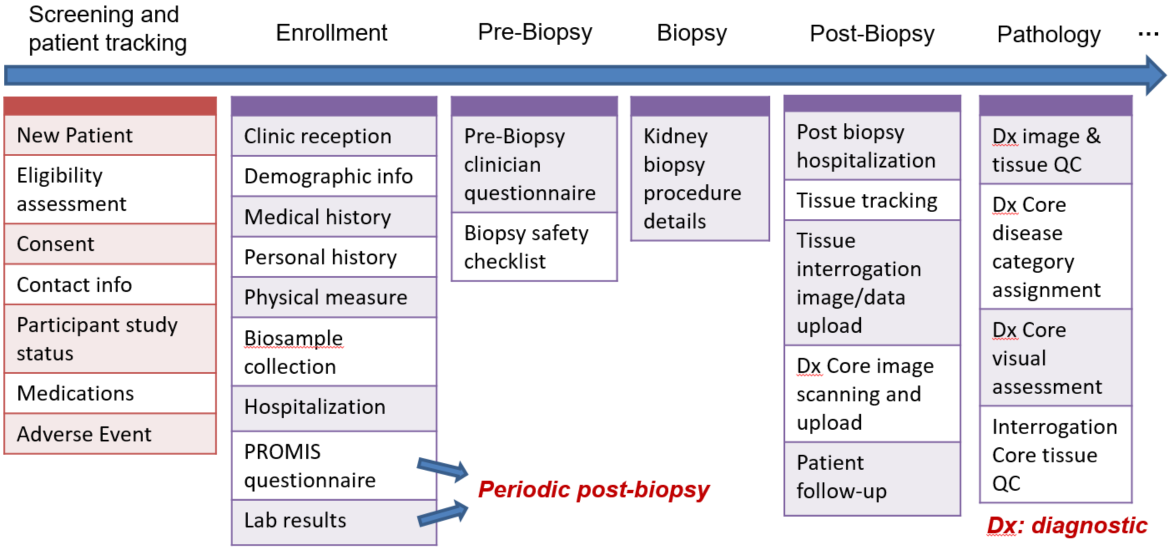
Our OPMI strategy of representing these CRFs can be summarized as "CRF-Question-Entity" (Fig. 5A). In this strategy, each CRF includes one or more questions, and each question is about some entity or entities, and different entities are connected using semantic relations in ontology. The questions in the strategy are essential since they link CRF and entities. While CRFs for a particular project may be very specific and cannot be reused, the questions are often similar among projects and can be reused. It is also noted that the same question may be expressed in different words, for example, the questions "Are you aged less than 18 years old" and "Are you aged 18 years or younger?" are essentially the same question. Once we model the entity or content behind the question, we do not need to worry about different expression formats.

Fig. 5B provides an example on how the "CRF-Question-Entity" can be used. This example illustrates the KPMP eligibility assessment form, which includes different questions. We defined two specific types of questions: exclusion question and inclusion question. An exclusion question is a question where a positive answer of the question would lead to the exclusion of the participant candidate from the specific clinical study. For example, if a person is aged 17 years, he or she will answer Yes to a "Whether age less than 18 years" question. These questions are explicitly asked in the CRFs for IRB and legality requirement which are frequently asked in other clinical studies besides KPMP. These questions are also often time anchored in multiple CRF forms at different stage of the studies. Even though these questions may not be necessarily important to the scientific interests, they are important in the context of precision medicine studies to enroll participants. In this example, the age can be calculated from the date of birth recorded in the database or retrieved from other questions. However, the definition of the concepts in the ontology enables us to raise questions from different angles and with additional information. Since this is an exclusion question that defines an exclusion criterion, the person's positive answer will indicate that he or she is ineligible for the KPMP study. This specific question *is about* the entity term '*age less than 18 years*', and then we can logically define this term as a subclass of 'age', which is a physical quality by itself. Furthermore, we can define this specific age quality with a specific measured value:

'quality measured by year' max 17

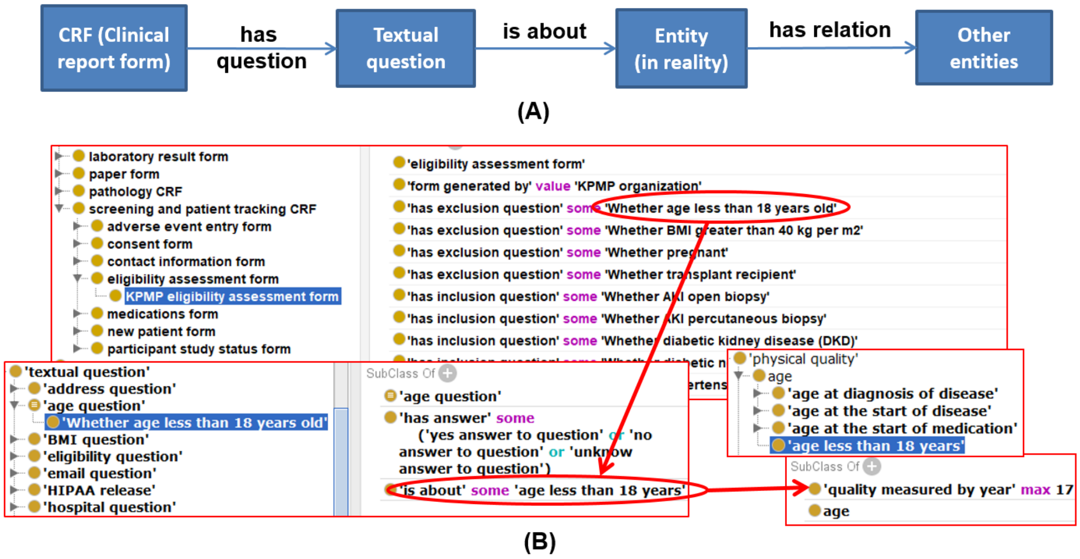
Such a logical definition can be parsed and understood by computers. Therefore, our strategy successfully defines the question, what the question is about, and how the question is used in the eligibility assessment CRF.

One use of such strategy is the interoperability of CRFs and CRF questions. For example, some new European precision medicine project may quickly sum up their CRFs using the questions defined in OPMI. Their specific questions can differ, and their ways to express their questions can differ. However, as long as their questions can be mapped to the OPMI question IDs, OPMI will be able to provide the underlying entities and their relations. This way can help support the CRF and clinical data standardization, sharing, and cross-institute data analysis.



Screening and patient tracking	Enrollment	Pre-Biopsy	Biopsy	Post-Biopsy	Pathology
New Patient	Clinic reception	Pre-Biopsy clinician questionnaire	Kidney biopsy procedure details	Post biopsy hospitalization	Dx image & tissue QC
Eligibility assessment	Demographic info	Biopsy safety checklist		Tissue tracking	Dx Core disease category assignment
Consent	Medical history			Tissue interrogation image/data upload	Dx Core visual assessment
Contact info	Personal history			Dx Core image scanning and upload	Interrogation Core tissue QC
Participant study status	Physical measure			Patient follow-up	
Medications	Biosample collection				
Adverse Event	Hospitalization				
	PROMIS questionnaire				
	Lab results				

Fig. 4. CRFs developed in the KPMP project.



laboratory result form
paper form
pathology CRF
screening and patient tracking CRF
adverse event entry form
consent form
contact information form
eligibility assessment form
KPMP eligibility assessment form
medications form
new patient form
participant study status form

'eligibility assessment form'
'form generated by' value 'KPMP organization'
'has exclusion question' some 'Whether age less than 18 years old'
'has exclusion question' some 'Whether BMI greater than 40 kg per m2'
'has exclusion question' some 'Whether pregnant'
'has exclusion question' some 'Whether transplant recipient'
'has inclusion question' some 'Whether AKI open biopsy'
'has inclusion question' some 'Whether AKI percutaneous biopsy'
'has inclusion question' some 'Whether diabetic kidney disease (DKD)'

'textual question'
'address question'
'age question'
'Whether age less than 18 years old'
'BMI question'
'eligibility question'
'email question'
'HIPAA release'
'hospital question'

SubClass Of
'age question'
'has answer' some ('yes answer to question' or 'no answer to question' or 'unknow answer to question')
'is about' some 'age less than 18 years'

'physical quality'
age
'age at diagnosis of disease'
'age at the start of disease'
'age at the start of medication'
'age less than 18 years'

SubClass Of
'quality measured by year' max 17
age

Fig. 5. OPMI design pattern of representing CRFs. (A) General "CRF-Question-Entity" design pattern; (B) Example of eligibiilty assement CRF. This form includes many questions such as "Whether age less than 18 years old", which is about the age quality that has a measured value of less than 18 years old. All these are logically represented in OPMI.

E. OPMI representaiton of clinical metadata

The follow-up Omics and pathology studies in KPMP would generate a lot of genes up- or down-regulated given different conditions. The clinical variables become a big pool of conditions that would influence the data analysis of the follow-up data analysis. The conditions are essentially reflected by the "entity" part laid out in the "CRF-Question-Entity" strategy as described above. In addition, these clinical variables can be represented as metadata, i.e., "data about data", which sum up the clinical variable types to be studied in KPMP and other studies. These ontologically represented clinical variables will later be useful in systematic Omics data analysis by providing possible reasons for some statistically identified Omics data analysis results.

Table 1 provides a set of representative metadata types that are derived from the entities referred by the KPMP CRF questions, which are defined in the ~30 KPMP CRFs.

TABLE I. REPRESENTATIVE KPMP CLINICAL METADATA TYPES

Metadata types	***Metadata Examples***
Quality and measurements	Measurement protocol details (e.g., arm and stand/sit/lay position in blood pressure measurement)
Health conditions	Comorbities, pregnancy, adverse events
Medical interventions	drug medication, prior surgeries transplantation, dialysis, biopsy, transplantation
Substances exposed to	Additional prescription drugs, recreation drugs, cigarettes, and alcohols
Socioeconomic factors	employment status, race, ethnicity, education level, income, health insurance status
Environmental	county, state, country, hospital, primary care location
Biosample	collection time, processing time, transportatoin tracking, biopsy location, storage location, storage time
Patient reported outcomes	patient experience, quality of life, pain, axiety, complicatoin, likert scale
Patient study status tracking	pass or fail screening, whether informed consent signed, is active in study? is live?
Electronic health record (EHR)	source of EHR, record availability, processing/harmonization method

F. *OPMI statistics*

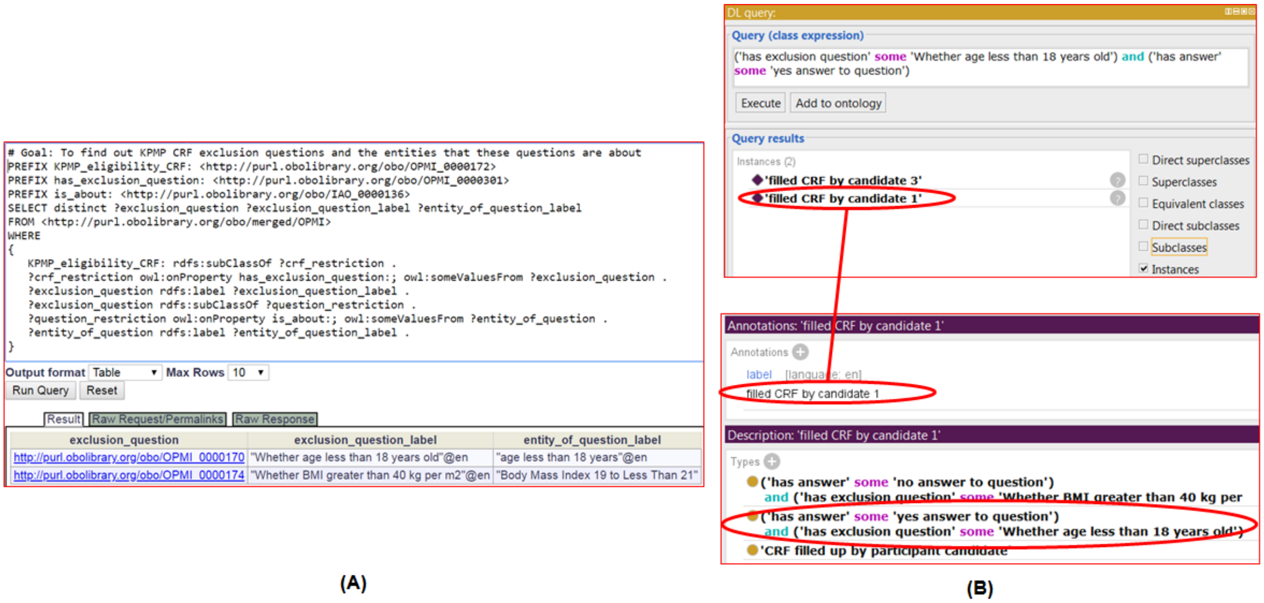
The latest release of OPMI contains a total of 2,958 terms, including 2,701 classes, 124 object properties, 2 data properties, and 118 annotation properties. Among these terms, 340 terms have OPMI_ namespace, and the other terms were imported from over 30 existing ontologies. The full ontology statistics of OPMI can be found on the Ontobee ontology statistics website at: http://www.ontobee.org/ontostat/OPMI.

G. *OPMI-based data query and analysis*

The OPMI ontology is being developed with many applications in mind. Here we demonstrate the usage of the OPMI information for querying for two important questions.

The first example is to use SPARQL to query what questions are exclusion questions in the KPMP eligibility assessment form and what entities these questions are about (Fig. 6A). With only a few lines, this query easily identified those exclusion questions and the entities to which the questions refer.

Based on the exclusion question setting and participant candidates' answers, we can identify which candidates are ineligible. We generated a use case demonstration to illustrate such an application (Fig. 6B). In our sandbox study, there are 3 candidates who provided different answers to a list of eligibility questions. These candidates and their provided answers can be represented as instances of OPMI classes. A DL (description logic) query can be used developed to query the data. Let us assume the 3 clinical study participant candidates came from 2 different recruitment sites (e.g., UT Southwestern and Yale University). Since we used the same ontology and terminology, we can query across different sites. As shown in Fig. 6B, we could identify that two of the participants answered yes to the 'Whether age less than 18 years' question. Based on the exclusion rule, this candidate is not qualified for participating in the KPMP project.



```
# Goal: To find out KPMP CRF exclusion questions and the entities that these questions are about
PREFIX KPMP_eligibility_CRF: <http://purl.obolibrary.org/obo/OPMI_0000172>
PREFIX has_exclusion_question: <http://purl.obolibrary.org/obo/OPMI_0000301>
PREFIX is_about: <http://purl.obolibrary.org/obo/IAO_0000136>
SELECT distinct ?exclusion_question ?exclusion_question_label ?entity_of_question_label
FROM <http://purl.obolibrary.org/obo/merged/OPMI>
WHERE
{
   KPMP_eligibility_CRF: rdfs:subClassOf ?crf_restriction .
   ?crf_restriction owl:onProperty has_exclusion_question:; owl:someValuesFrom ?exclusion_question .
   ?exclusion_question rdfs:label ?exclusion_question_label .
   ?exclusion_question rdfs:subClassOf ?question_restriction .
   ?question_restriction owl:onProperty is_about:; owl:someValuesFrom ?entity_of_question .
   ?entity_of_question rdfs:label ?entity_of_question_label .
}
```

Output format: Table; Max Rows: 10

exclusion_question	exclusion_question_label	entity_of_question_label
http://purl.obolibrary.org/obo/OPMI_0000170	"Whether age less than 18 years old"@en	"age less than 18 years"@en
http://purl.obolibrary.org/obo/OPMI_0000174	"Whether BMI greater than 40 kg per m2"@en	"Body Mass Index 19 to Less Than 21"

('has exclusion question' some 'Whether age less than 18 years old') and ('has answer' some 'yes answer to question')

'filled CRF by candidate 3'
'filled CRF by candidate 1'

filled CRF by candidate 1

('has answer' some 'no answer to question')
 and ('has exclusion question' some 'Whether BMI greater than 40 kg per
('has answer' some 'yes answer to question')
 and ('has exclusion question' some 'Whether age less than 18 years old')
'CRF filled up by participant candidate

Fig. 6. OPMI query examples. (A) SPARQL query of exclusion questions and the entities that the questions are about as defined in KPMP eligibility assessment form. Ontobee SPARQL (http://www.ontobee.org/sparql) was used for this query. (B) DL (description logic) query of who are ineligible based on an exclusion question. This sandbox example includes three patients, each of which provided some answers to CRF questions. The DL query was conducted using the Protégé OWL editor.

IV. Discussion

To support challenging precision medicine studies, we can greatly benefit from ontologies to represent, standardize, share, and integrate various clinical and biomedical big data. Similar to other big data domains, the big data in precision medicine have features of high volume, high variety, high velocity, and high veracity. As an open source ontology in the domain of precision medicine, OPMI is a timely community-based effort to systematically represent various precision medicine-related entities and how these entities are related. Our use case studies demonstrate that OPMI, together with other existing OBO ontologies, is able to support OHDSI CDM and OHDSI data analysis, as well as KPMP CRF and associated content representation and analysis, leading to valuable clinical and scientific insights.

The ontology representation of different common data models (CDMs) may provide a feasible way to semantically integrate the different CDM systems. The CDMs, like OMOP CDM, provides a robust platform to standardize data from different databases and clinical studies. The OMOP relational database CDM is easy to be interpreted by humans. The relations between elements in different tables can be linked and queried through relational database primary keys and foreign keys. However, the CDM relations are indirect (through foreign keys instead of direct linkages), and the representation is difficult to be interpreted by machines without human operation. Meanwhile, the CDM model is overall a high level design and may not be used to handle deep granularity as ontology can do. Our OPMI modeling (Fig. 2) shows that the CDM elements and their relations can be logically represented using ontology. The OHDSI-based kidney adverse event data analysis (Fig. 3) further demonstrated that the ontological modeling and application can support practical research studies. In this case, OMOP Condition cannot differentiate adverse events as a consequence from a medical intervention (e.g., surgery or drug treatment) from the symptoms or abnormal phenotypes of an on-going disease. However, based on the adverse event definition, we can design a method to perform such a differentiation in ontology level i.e., that an adverse event is an abnormal condition that occurs after a medical intervention. In our study, we only considered AKI adverse event that did not occur within 30 days before heart surgery but did occur after the heart surgery. The representation and analysis of the conditions before heart surgery using the Human Phenotype Ontology (HPO) (Fig. 3) allowed us to have a clear idea on how the patients' information (e.g., age and symptoms) and heart surgery are associated with the AKI adverse event. However, even though the ontology can help better represent and interpret the adverse event definitions, the ontology by itself does not directly handle large volumes of big data well, for which OMOP is good at. Therefore, our ontology representation can be used as a complementary method to support OMOP data analysis. Furthermore, the logic generated by ontology can be used to support CDM description and harmonize the integration of data from different CDM systems. While the current study focuses on OHDSI OMOP CDM, we plan to study other CDMs and test how OPMI can be used to harmonize different CDMs at a semantic ontology representation level.

The follow-up KPMP study provides a more systematic and integrated use case to study the kidney disease precision medicine. Over 20 universities and institutes will participate in the KPMP, recruiting individuals with various forms of acute kidney injury (AKI) and chronic kidney disease (CKD). Each participant will be biopsied, and the kidney tissue samples will be divided for assays including RNA-seq, proteomics, metabolomics, pathology, and histological studies. To better analyze the basic assay data, we will need to fully capture the clinical data types and all instance data from each patient given different conditions. With this information, we can then analyze whether an Omics finding is related to a clinical variable (such as age or biological sex).

Our CRF-Question-Entity strategy is a new way to capture the CRF contents and their associated entities. CRFs are commonly used. It is time consuming to generate CRFs. Once generated and used for a specific study they are then archived, but not reused for similar studies. To support efficient CRF generation and reuse, our ontology-based strategy systematically record CRFs, their associated questions, and the question-referred entities. Although specific CRFs may not be reused, the questions often reappear in different forms. Although many questions are expressed differently, they are designed to capture the same concepts. Through modeling and representation of the underlying concepts, we are able semantically define questions, which then further help define the CRFs. We believe that such a strategy can help automate the process of digitalizing and processing CRFs, supporting clinical research.

To the best of our knowledge, such a CRF-Question-Entity strategy is first proposed and implemented in this study. This strategy was inspired by our own previous ontology representation and analysis of 12 informed consent forms from pharmacies and local governments [25]. The representation of those forms allowed us to compare different questions in different forms. However, that study did not emphasize the representation of the concepts in reality that the questions are designed determine. Abidi et al. presented a framework to semi-automatically extract medical entities from referral letters, classifying the unstructured referral letters according to their semantic types based on SNOMED-CT [26], and transcribe CRFs based on the extracted information from the referral letters. Such a strategy does not result in ontology representation of CRFs. However, the semi-automatic extraction of medical entities from text is a valuable way to improve the speed of ontology development. Lin et al. presented a multi-technique approach to facilitate electronic

CRF (eCRF) design by adopting common data element standards and ontology-based knowledgebase [27]. It is likely that our OPMI CRF-Question-Entity representation will indeed support eCRF development. OPMI will be able to provide a pool of questions for eCRF designers to choose and use. Once a set of questions are defined, our system will be able to allow users to automatically identify the concepts in reality behind these questions and the semantic relations between the entities.

We presented the OPMI and its CRF-Question-Entity strategy in the Seventh Clinical and Translational Science Ontology Workshop, Orlando, Florida, on February 20 2019. This workshop had the theme of "Ontology for Precision Medicine: From Genomes to Public Health". Our presentation and another one-hour discussion on this topic in the next day were well-received. While there were efforts to record CRF questions and answers, our strategy of ontological modeling of the underlying semantic meanings of CRF questions was generally considered novel. Constructive and insightful comments were also received, for example, how to properly represent the reality of 'unknown answer to question'. These comments are being carefully considered in our OPMI development.

OPMI is a community effort. Its initial development came from the development of the Ontology of Respiratory Disease Investigation (ORDI), which ontologically represented many clinical terms frequently used in the respiratory disease studies [28]. Respiratory diseases are among the leading causes of death worldwide. It remains a challenge to standardize, integrate, and analyze high volume and heterogeneous respiratory disease investigation data for deep mechanism understanding and rationale treatment design. One study surveyed hundreds of residents from the urban and suburb communities associated with various variables and different respiratory diseases [28].

Another use case is the application of OPMI to support the National Physique and Health Database in China (http://cnphd.bmicc.cn/chs/en/), which was initiated in 2001, and is being maintained by the Biologic Medicine Information Center of China (BMICC, http://www.bmicc.org), Institute of Basic Medical Sciences (IBMS), Chinese Academy of Medical Sciences, Beijing, China. The database contains the physical and health data of over 160,000 Chinese from different locations, genders, and ages. Over 200 parameters, related to morphology, function and physical capacity of an individual body, were identified and used in the database. In addition, more data will be collected and added to this database in the future. OPMI is being applied to standardize and analyze the data in the database and make the data more accessible and useful by others.

The ClinEpiDB project, launched in February 2018, is an open-access online resource enabling investigators to maximize the utility and reach of their clinical epidemiology data and to make optimal use of the data released by others (https://clinepidb.org). With a focus on diarrheal and infectious disease epidemiology, ClinEpiDB datasets involve many clinical epidemiology-related questions from CRFs. Representing these requires many clinical terms that overlap with the coverage of OPMI and represents one area of potential collaboration. It will also be interesting to compare the commonalities and differences between the CRFs in ClinEpiDB and KPMP, and provide template CRFs for other clinical projects.

In addition, OPMI is also being explored to support many other community-based precision medicine projects, including the representation of clinical trial terms as seen in ClinicalTrials.gov, a database of clinical studies conducted around the world (https://clinicaltrials.gov/). The ClinicalTrials.gov database defines many clinical trial related terms (https://prsinfo.clinicaltrials.gov/definitions.html). We are currently collaborating with the researchers in the US National Institute of Health (NIH) and model and represent these terms in OPMI.

Acknowledgment

This KPMP project is supported by the NIH National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) U2C Project #: 1U2CDK114886-01. We appreciate Dr. Deborah Hoshizaki's discussion and support during the ontology development and applications. We also appreciate the discussion and feedback provided by the attendees (including Matthias Brochhausen, Peter Elkin, William Hogan, etc.) in the Seventh Clinical and Translational Science Ontology Workshop.

Address for correspondence

Please contact YH from the University of Michigan, Ann Arbor, MI, USA. Email address: yongqunh@med.umich.edu. Telephone: +1-734-615-8231.

References

[1] R. Higdon, W. Haynes, L. Stanberry, E. Stewart, G. Yandl, C. Howard, *et al.*, "Unraveling the Complexities of Life Sciences Data," *Big Data,* vol. 1, pp. 42-50, Mar 2013.

[2] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, *et al.*, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers," *Stud Health Technol Inform,* vol. 216, pp. 574-8, 2015.

[3] 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.

[4] T. R. Ross, D. Ng, J. S. Brown, R. Pardee, M. C. Hornbrook, G. Hart, *et al.*, "The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration," *EGEMS (Wash DC),* vol. 2, p. 1049, 2014.

[5] T. Souza, R. Kush, and J. P. Evans, "Global clinical data interchange standards are here!," *Drug Discov Today,* vol. 12, pp. 174-81, Feb 2007.

[6] 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.

[7] 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.

[8] *The Ontology for General Medical Science (OGMS)*. Available: https://github.com/OGMS/ogms

[9] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos*, et al.*, "The Ontology for Biomedical Investigations," *PLoS One,* vol. 11, p. e0154556, 2016.

[10] R. R. Brinkman, M. Courtot, D. Derom, J. M. Fostel, Y. He, P. Lord*, et al.*, "Modeling biomedical experimental processes with OBI," *J Biomed Semantics,* vol. 1 Suppl 1, p. S7, 2010.

[11] T. Groza, S. Kohler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel*, et al.*, "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease," *Am J Hum Genet,* vol. 97, pp. 111-24, Jul 2 2015.

[12] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, "Uberon, an integrative multi-species anatomy ontology," *Genome Biol,* vol. 13, p. R5, 2012.

[13] 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.

[14] Y. Lin, M. R. Harris, F. J. Manion, E. Eisenhauer, B. Zhao, W. Shi*, et al.*, "Development of a BFO-based Informed Consent Ontology (ICO)," in *The 5th International Conference on Biomedical Ontologies (ICBO)*, Houston, Texas, USA, October 8-9, 2014, 2014.

[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] D. L. Rubin, N. F. Noy, and M. A. Musen, "Protege: a tool for managing and using terminology in radiology applications," *J Digit Imaging,* vol. 20 Suppl 1, pp. 34-46, Nov 2007.

[17] 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.

[18] P. Grenon and B. Smith, "SNAP and SPAN: Towards Dynamic Spatial Ontology," *Spatial Cognition and Computation,* vol. 4, pp. 69-103, 2004.

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

[20] W. Ceusters and J. Blaisure, "A realism-based view on counts in OMOP's common data model," 2017, pp. 1-8. DOI: 10.3233/978-1-61499-761-0-55.

[21] J. B. O'Neal, A. D. Shaw, and F. T. t. Billings, "Acute kidney injury following cardiac surgery: current understanding and future directions," *Crit Care,* vol. 20, p. 187, Jul 4 2016.

[22] M. G. Lagny, F. Jouret, J. N. Koch, F. Blaffart, A. F. Donneau, A. Albert*, et al.*, "Incidence and outcomes of acute kidney injury after cardiac surgery using either criteria of the RIFLE classification," *BMC Nephrol,* vol. 16, p. 76, May 30 2015.

[23] M. H. Rosner and M. D. Okusa, "Acute kidney injury associated with cardiac surgery," *Clin J Am Soc Nephrol,* vol. 1, pp. 19-32, Jan 2006.

[24] K. A. Ramos and C. B. Dias, "Acute Kidney Injury after Cardiac Surgery in Patients Without Chronic Kidney Disease," *Braz J Cardiovasc Surg,* vol. 33, pp. 454-461, Sep-Oct 2018.

[25] Y. Lin, J. Zheng, and Y. He, "VICO: Ontology-based representation and integrative analysis of vaccination informed consent forms," *J Biomed Semantics,* vol. 7, p. 20, 2016.

[26] S. H. Brown, P. L. Elkin, B. A. Bauer, D. Wahner-Roedler, C. S. Husser, Z. Temesgen*, et al.*, "SNOMED CT: utility for a general medical evaluation template," *AMIA Annu Symp Proc,* pp. 101-5, 2006.

[27] C. H. Lin, N. Y. Wu, and D. M. Liou, "A multi-technique approach to bridge electronic case report form design and data standard adoption," *J Biomed Inform,* vol. 53, pp. 49-57, Feb 2015.

[28] H. Yu, J. Zheng, H. Wang, E. Ong, X. Ye, Z. Zhang*, et al.*, "ORDI: An integrative community-driven ontology to support standardized representation and data analysis for respiratory disease investigations " in *The 11th International Biocuration Conference (BioCuration-2018)*, Shanghai, China, April 8-11, 2018.