Published in final edited form as:

J Biomed Inform. 2015 June ; 55: 132–142. doi:10.1016/j.jbi.2015.03.008.

An Integrated, Ontology-Driven Approach to Constructing Observational Databases for Research

William Hsu, PhD[1], **Nestor R. Gonzalez, MD**[1,2], **Aichi Chien, PhD**[1], **J Pablo Villablanca, MD**[1], **Päivi Pajukanta, MD, PhD**[3], **Fernando V Vinuela, MD**[1], and **Alex AT Bui, PhD**[1]

[1]Department of Radiological Sciences, UCLA David Geffen School of Medicine, Los Angeles, CA

[2]Department of Neurosurgery, UCLA David Geffen School of Medicine, Los Angeles, CA

[3]Department of Human Genetics, UCLA David Geffen School of Medicine, Los Angeles, CA

Abstract

The electronic health record (EHR) contains a diverse set of clinical observations that are captured as part of routine care, but the incomplete, inconsistent, and sometimes incorrect nature of clinical data poses significant impediments for its secondary use in retrospective studies or comparative effectiveness research. In this work, we describe an ontology-driven approach for extracting and analyzing data from the patient record in a longitudinal and continuous manner. We demonstrate how the ontology helps enforce consistent data representation, integrates phenotypes generated through analyses of available clinical data sources, and facilitates subsequent studies to identify clinical predictors for an outcome of interest. Development and evaluation of our approach are described in the context of studying factors that influence intracranial aneurysm (ICA) growth and rupture. We report our experiences in capturing information on 78 individuals with a total of 120 aneurysms. Two example applications related to assessing the relationship between aneurysm size, growth, gene expression modules, and rupture are described. Our work highlights the challenges with respect to data quality, workflow, and analysis of data and its implications towards a learning health system paradigm.

Graphical abstract

Keywords

Data extraction; biomedical ontology; retrospective study; image analysis; database; intracranial aneurysm

Corresponding Author Information: William Hsu, PhD, UCLA Medical Imaging Informatics, 924 Westwood Blvd, Suite 420, Los Angeles, CA, 90024, Tel: (310) 794-3536, Fax: (310) 794-3546, willhsu@mii.ucla.edu.

1. INTRODUCTION

The promise of using electronic health records (EHRs) for secondary uses such as studying the natural evolution of diseases, characterizing recent trends in exposures, and improving the quality and delivery of care has been well-documented [1–3]. Recent reports from the Institute of Medicine on health information technology motivate the reuse of EHR data to achieve a learning healthcare system that permits real-time analysis of data collected about patients [5, 6]. Efforts to mine the patient record have demonstrated promising results in identifying patient cohorts for clinical trials [7], detecting adverse drug events [8], and providing clinical phenotypes for correlation with genetic findings [9]. Infrastructure tools such as i2b2 and SHARPn have made searching, summarizing, and retrieving data from cohorts captured by the EHR more feasible [10, 11]. These developments have supported a growing body of work that utilize EHR data in identifying patient cohorts with specific diseases and conducting large population studies to mine associations between gene variants and clinical phenotypes [12–14]. Nevertheless, fulfilling the promise of precision medicine necessitates not only the ability to aggregate and mine information from multiple clinical data sources, but also novel approaches to obtain detailed characterizations of observations that provide sufficient context for studying the evolution of a patient's condition. Moreover, the results of observational studies leveraging EHR data are influenced by inherent variability in reporting quality, which may result in biased, incomplete, and inconsistent information [15, 16]. Sample sizes that are larger than a single institution are frequently required to establish statistical significance, but EHR data are typically kept in silos that make pooling information across multiple populations difficult. Common data models such as the Clinical Element Model (CEM) [17] and the Observational Medical Outcomes Partnership (OMOP) [18] are beginning to address these data integration issues, but the implementation of these models remains limited in scope. Establishment of a computation framework and systematic workflow is needed to address a number of caveats in transforming data originally collected for clinical and billing purposes into data usable for research [19]

In this paper, an ontology-driven framework is presented for representing and validating observational clinical, imaging, and genomic findings from clinical records at our institution. The goal is to facilitate the systematic extraction and longitudinal representation of detailed observations in a standardized manner, allowing a large cohort of individuals to be identified from the EHR and subsequently used to address research aims. The ontology plays a central role in data integration, information extraction, quality assessment, and retrieval.

1.1 Related Work

Components of this work are similar to existing approaches, but several distinctions can be made. In [20], Min et al. demonstrated the application of an ontology to assist in the integration and querying of heterogeneous information across a prostate cancer database and tumor registry. A prostate cancer ontology was created and mapped to multiple custom relational databases using a mediator (D2RQ). The use of the ontology permitted queries to be posed using the SPARQL Protocol and RDF Query Language (SPARQL), allowing data

to be easily retrieved from multiple sources using a single query. In our work, the role of the ontology is similar, but we demonstrate how the ontology is integrated with the OMOP common data model, rather than formulating a custom data model and maintaining a mapping. REDCap [21] is an electronic data capture tool that permits users to easily design and deploy standardized forms with data validation functions. Observations and measurements associated with each instance of a disease are tracked independently over time in our approach, permitting investigators to study a particular instance (e.g., evolution of a single lesion) or the patient as a whole (e.g., total number of lesions in a patient), maintaining this distinction in REDCap is difficult using the current paradigm employed to organize its data. OPIC (Ontology-driven Patient Information Capture) is a data collection framework that utilizes the Epilepsy and Seizure ontology to standardize data entry and representation, proactively enforce data accuracy, and support logical skip patterns [22, 23]. While related in approach, our work focuses on using the ontology to extract and represent clinical data; we also show how an ontology-driven approach can be applied to another domain with differing information requirements. In summary, the contributions of our work are three-fold: 1) to present a framework that combines the OMOP common data model with an application ontology to integrate heterogeneous data across multiple sources; 2) to employ the ontology as the central source of domain knowledge to transform clinical data into a structured representation that characterizes clinical observations longitudinally; and 3) to discuss the strengths and limitations of utilizing an ontology-driven approach to improve the quality of data extracted from the EHR for secondary use.

1.2 Driving Example: Intracranial Aneurysms

As a running illustration, our efforts are described in the context of analyzing data from patients with intracranial aneurysms (ICA). ICAs are complex lesions that are only partially understood: their multifaceted nature has hampered efforts to explain its pathophysiology and thus the development of effective therapies. Ruptured ICAs result in subarachnoid hemorrhage (SAH) associated with a poor 30-day mortality rate of 17–35% [24]. Recent studies on ICA have focused on imaging studies [25], tissue samples [26], and genomic analyses [27, 28]; such studies have identified variables (e.g., wall shear stress, familial influence, environmental factors) that influence how ICAs form and evolve. However, these studies typically analyze facets of the disease in isolation: while predictive factors are reported, no study has attempted to comprehensively understand the relationship between pathophysiological and genetic factors with clinical observables. Hence, gaps persist in our knowledge of what is known about the disease, what areas need to be further understood, and how the knowledge gained can influence routine clinical decisions. One notable effort was the @neurIST project [29], which implemented a grid-based infrastructure and application ontology for standardizing and sharing data from multiple sources and analyzing the data to generate computational models of risk for patients. While their efforts resulted in a platform for aggregating and sharing data on a specific patient cohort across multiple institutions, the ability to generalize their work to other research environments has yet to be demonstrated. Two example applications based on our developed framework are described to illustrate how the ontology is helping yield new insights into the management of this patient population.

2. MATERIALS & METHODS

The overarching goal in utilizing an ontology-driven approach is to improve the quality and accessibility of clinical observations and relevant contextual information to answer questions related to rupture risk and treatment selection in individuals with ICA. Our local institutional review board approved a waiver of consent for retrospective review of past ICA cases and consent materials for prospective cases. The following sections describe the approaches used to formulate the application ontology, establish a data extraction and integration workflow, generate detailed contextual information from the clinical data, and monitor the collection process.

2.1 Intracranial Aneurysm Ontology

2.1.1 Scope—The Intracranial Aneurysm Ontology serves as a unifying representation for modeling relevant findings related to ICA. The ontology covers a broad range of information specific to aneurysm risk, morphology, and treatment that is reported in or derived from clinical data sources, as listed in Table 1. Information related to biological processes is not explicitly represented in the ontology; rather, existing ontologies such as Gene Ontology are referenced [30]. Examples of how the ontology influences data extraction and integration are depicted in Figure 1. Entity names and synonyms captured in the ontology are used to generate term lists that are used in named entity recognition (Fig. 1a). Annotations associated with each entity specifying data type and permissible values constrain how fields are presented to the user in the web-based form (Fig. 1b). Queries that exploit the ontology's graph structure and semantic relationships can be executed using SPARQL to retrieve related entities (Fig. 1c).

2.1.2 Approach—Using the Basic Formal Ontology (BFO) [31] as the overarching organization, the ontology was developed using a top-down, bottom-up approach to catalog relevant entities, relations, and attributes.

Top-down approach: A top-down approach was followed to identify candidate entities based on the input of clinical investigators and examination of current scientific literature. During the elicitation process, individuals with backgrounds in neurosurgery (NRG), interventional and diagnostic neuroradiology (FV, JPV), and hemodynamic analysis (FV, AC) were asked to enumerate all known variables with relevance to aneurysm growth and rupture. In addition, a review of published systematic reviews and reporting standards was conducted to identify relevant entities documented in the literature [32–39]. In total, the top-down approach yielded 398 unique entities.

Bottom-up approach: A bottom-up approach was used to characterize the entities that have been reported as part of standard of care and extractable from the patient record. The chart abstraction task consisted of two parts: 1) enumeration of variations (e.g., synonyms, abbreviations) identified during the top-down approach as expressed in practice; and 2) identification of new entities documented in the record that were not initially identified through the top-down approach. Thirty retrospective cases were identified, resulting in a total of 307 documents comprising neurology consults, radiology reports, inpatient

admission and discharge summaries, and referral letters. To reduce the burden on the human annotator, an information extraction pipeline was implemented using the Unstructured Information Management Architecture (UIMA) [40] to identify all noun phrases and generate mappings between these phrases to UMLS (Unified Medical Language System) via MetaMap [41]. Of the 1,341 unique noun phrases, 261 mapped to concepts in UMLS. Upon review by a human annotator, 85 unique entities were identified, relating to specific interventional devices (e.g., ICA treatment coil types such as Matrix2 360 or GDC-18 360) and documented findings (e.g., stent-induced hyperplasia, coil compaction) that were not brought up during the top-down elicitation process.

An initial set of semantic relationships was delineated based on ones defined by the OMOP common data model; these relationships were drawn from existing terminologies such as SNOMED CT and RxNORM. To represent semantics related to scientific investigation, additional relationships were drawn from our prior work on the Phenomenon-centric Data Model and the Ontology for Biomedical Investigations [42]. These relationships were reviewed by domain experts who pruned the existing set and added relationships they perceived as missing. As a result, 145 relationships (140 object properties and 5 data properties) were defined.

2.1.3 Implementation—Protégé 4.3 [43] was used to formulate the ontology, and webProtégé [44] was used as part of the review process, allowing two investigators to modify the ontology through a web-based interface. The investigators worked collaboratively, finalizing the list of entities based on their domain expertise. The resulting ontology is comprised of 589 logical axioms, 483 entities, 140 object properties, and 5 data properties and has been made available in Web Ontology Language (OWL) format[1]. A breakdown of the entities represented in the ontology and whether the entities are new or drawn from existing ontologies is provided in Table 2.

2.2 Data Representation

An important requirement of the underlying representation was to support concurrent retrospective and prospective data collection. Retrospective cases were identified based on a list compiled by clinical investigators and patient cohort searches using diagnostic codes such as ICD-9 codes 430 (Subarachnoid hemorrhage) and 437.3 (Cerebral aneurysm, nonruptured). Prospective cases were recruited from individuals who were diagnosed with an aneurysm and were scheduled for a consultation at our institution. Individuals who consented to participate in the study completed a standardized questionnaire and underwent a blood draw that provided specimen for subsequent genetic analysis.

The database that stored information about each patient was organized using the OMOP v5.0 common data model, which provided a standardized representation for capturing observational data for research. OMOP was chosen because of its ability to: 1) provide a flexible representation for capturing new information without having to change the structure of the underlying database for different domains; 2) leverage ontologies and existing

[1]Accessible at http://www.mii.ucla.edu/~willhsu/icaproject

controlled vocabularies to represent the content of observations; and 3) share collected data across multiple institutions by defining a common data representation. Of the 39 tables defined in the v5.0 specification, a subset of 26 tables related to standardized clinical data, metadata, and vocabularies were populated with data from clinical sources as part of this work. Tables related to provider details (e.g., standardized health system data), costs of healthcare delivery (e.g., economics), and information about clinical events generated from other tables (e.g., derived elements) were not populated given the initial scope of the research questions (but will be added as part of future work). A modification to the OMOP model was introduced to permit instances of observations, measurements, procedure occurrences, and drug exposure to be associated with a particular condition occurrence. The existing OMOP model currently only associates an observation with unique person and visit instances; observations cannot refer to specific condition instances. An example of this limitation is if a patient has two distinct aneurysms with measurements taken of each aneurysm. The current model does not support uniquely associating a measurement to a specific aneurysm. Hence, additional foreign keys were defined between observations and measurements allowing them to reference a specific condition occurrence.

OMOP was also selected because of the incorporation of standardized vocabularies to represent the content of observational data. The design of the data model is a hybrid between entity-relationship (ER) and entity-attribute-value (EAV) models. In OMOP, new information collected or derived from the patient record can be captured by adding entities to the vocabulary without changing the structure of the database. Researchers collect information about new variables by specifying a corresponding entity to the ontology, which in turn augments the possible entities and values that are captured by the database. This approach allows the database to adapt to changes in data collection by adding or modifying entities in the ontology through a web-based interface, reflecting these changes immediately in the data entry and querying components of the workflow. An ongoing challenge is ensuring the consistent usage of a given entity to represent clinical information across individuals or institutions; an established method has been to create data dictionaries at each institution that are then mapped and harmonized to standard terminology resources [45]. The ICA ontology developed at our institution is used to populate the standardized vocabulary tables in the OMOP common data model, dictating how values extracted from observational clinical data are encoded.

2.3 Analysis Workflow

The analysis workflow is illustrated in Figure 2. Each patient case is analyzed along four perspectives: clinical (patient demographics, presentation, treatment, follow-up), quantitative imaging (3D morphological analysis), hemodynamics (computational fluid dynamics), and whole genome sequencing (weighted gene co-expression network analysis) on individuals who provided blood samples. Once patient data is retrieved and entered into the research repository, investigators involved in each aspect of the analysis are notified of pending cases. Characterization of each data type is described in the following paragraphs.

2.3.1 Clinical—Initially, data entry was performed by two human annotators (a resident and a research associate) who reviewed each case, entering information into a web-based

form. A neurosurgery fellow provided oversight to ensure the validity of annotations being generated. Given the number of variables being collected (summarized in Table 1), a UIMA pipeline was developed to automate the information extraction process. Annotators targeted reported measurements along different axes (e.g., anteroposterior, cranio-caudal), characteristics about the aneurysm (e.g., shape), clinical presentation (e.g., subarachnoid hemorrhage, incidental aneurysm), anatomical location, and interventions (e.g., whether an individual underwent stent-assisted coiling). A combination of regular expressions, dictionary lookup, and rule-based methods were used to implement each annotator, depending on the complexity of information being targeted. Entities and their synonyms drawn from the ontology were incorporated into term lists that were utilized in the lookup process. For example, to identify mentions of aneurysm location, all child entities of "aneurysm location" and their synonyms were retrieved. During the development process, the UIMA pipeline performed a first-pass review of each case with the human annotator validating the results and committing them to the database.

2.3.2 Quantitative Image Analysis—Computed tomography angiography (CTA) and magnetic resonance angiography (contrast-enhanced MRI, time-of-flight MRA) provided three-dimensional, *in vivo* characterization of the aneurysm location and morphology. For this study, all MRA and CTA studies acquired were considered for analysis. For each imaging study, a trained specialist utilized the Vitrea Enterprise Suite (ViTAL Images, Minnetonka, MN), an image processing workstation, to generate volume-rendered reconstructions of the angiography data and perform measurements along three dimensions. Other findings of interest such as arteriovenous malformations (AVMs), calcifications, thrombi, and geometry were interpreted by the specialist and recorded as part of semantic labels assigned to each study. As depicted in Fig. 1, the ontology captured each imaging study as individual instances that were semantically related to other entities such as aneurysm location, shape, and morphology. Morphology was characterized by entities representing measurements related to the size and orientation of the aneurysm dome and neck. Measurements could then be obtained for individual aneurysms by querying the ontology to return all instances of aneurysm morphology related to the imaging history associated with that aneurysm.

2.3.3 Hemodynamics—CTA studies were also analyzed using a computational fluid dynamic simulation developed by collaborators (FV, AC) [46–48]. For each individual, the angiography data was used as the input to the simulation; based on the numerical output, a researcher (AC) assigned semantic labels to describe the blood flow characteristics (e.g., stability, pattern, division, impact, concentration). Numerical values reflected wall shear stresses at six different measurement points. As our institution was one of the first to report on these types of blood flow characteristics, our clinical investigators defined the controlled terms used to characterize the blood flow. Corresponding entities were added to the ontology, permitting the integration of hemodynamic simulation results.

2.3.4 Sequencing of Peripheral Blood—Samples taken from prospective study participants who consented to providing blood were stored in a −80°C freezer within 24 hours of extraction and sent to a sequencing core and sequenced using a 100-bp RNAseq

protocol on Illumina HiSeq 2000 equipment. The resulting data were aligned using the STAR RNA-seq sequence alignment program [49]. The aligned sequences were then used to generate a weighted gene co-expression network (WGCNA) following a protocol similar to [50] developed by our genetics collaborator (PP). Using demographic and family history information from the database, gene expression values were corrected for ancestry informative markers. The WGCNA identifies co-expression networks that capture the main patterns of variation. Using WGCNA, the initial set of 16,535 genes measured was reduced to 30 modules by clustering them functionally. These modules were then correlated with clinical phenotypes such as aneurysm size, growth, and rupture.

2.4 User Interface

2.4.1 Data Entry—A web-based application was developed to facilitate the multiple users inputting data. The interface consisted of a series of structured forms organized by patient, aneurysm, and observation. Each observation was associated with a patient, condition occurrence (e.g., each instance of an aneurysm), and visit, allowing collected values to be summarized at the patient-level (e.g., how many aneurysms does an individual have) or observation-level (e.g., how many measurements does a specific aneurysm have). In addition, information that overlapped across sources (e.g., radiology report versus image analysis) were maintained separately, allowing investigators to assess whether measurements taken of the same aneurysm using different methods varied significantly.

2.4.2 Research Dashboard—An administrator oversaw the entire collection effort through a web-based dashboard, as depicted in Figure 3a. The interface provided a summary about individual cases, including whether the analysis workflow had completed, whether the inputted information was validated, and whether biospecimens and sequencing data were available for that individual. Using relationships defined in the ontology (e.g., is-a, part-of), the dashboard grouped entities that were related to a given finding together, which allowed investigators to determine what information was missing or inconsistent. For example, at each imaging observation, measurements of aneurysm dome size along the three axes could be derived from the radiology report or volumetric image analysis. However, differences in measurements existed when comparing the values of these two sources depending on the perspective and tool selected by the individual performing the analysis. The dashboard highlighted these differences and missing values (e.g., if a dimension could not be measured because of an image artifact) to the user. Administrators had the ability to view all of the data on the entire population and were able to edit or delete information for any of the modules, as necessary. Other types of users who viewed the dashboard were presented with a tailored version of the interface with a limited set of functions. For example, researchers not involved in the collection of the data were presented with a read-only, de-identified view of the collected information and a basic set of filters for cohort identification; only aggregate statistics about the population, not individual information such as patient names and specific dates of procedures, were made available. For approved investigators, the research dashboard facilitated patient cohort identification by providing users with a graphical interface for formulating SPARQL queries using the entities and relationships defined in the ontology. The default filters (Figure 3b) allowed users to pose queries such as *retrieve all cases involving female individuals who have undergone an endovascular coiling and have*

imaging follow-up. The list of available filters and permissible values were directly drawn from entities in the ontology. Once the appropriate patient cohort had been found, the export functionality (Figure 3c) allowed users to specify the variables that were included and whether aggregate values at the patient level or observations at the aneurysm level were returned.

3. RESULTS

The following sections describe the benefits and challenges of developing an ontology-driven framework for representing observational data and two applications illustrating the role of the ontology in facilitating data analysis. Insights were derived from applying the described framework to the analysis of 78 patient cases (30 prospectively consented cases and 48 retrospective cases).

3.1 Role of the Ontology

3.1.1 Data Representation—The ontology played a central role in extracting and integrating observational clinical data for research. Entities in the ontology defined what information was collected (or extracted) from the patient record and provided relevant context (e.g., synonymous terms) for standardizing the collected data. The underlying data model was domain-agnostic, relying on the application ontology to provide tailored knowledge for executing specific queries. While manual modifications were necessary to revise data entry forms to accommodate new entities, future development will permit the web-based interface o dynamically adapt to reflect any changes to the ontology.

3.1.2 Data Quality—The collected observational data had inconsistencies in data quality and level of granularity. For instance, the exact location of each aneurysm was often not explicitly reported in clinical documents but was determined by reviewing the medical images. In only a small fraction of cases, the record included survival information for individuals who had died from aneurysm rupture, but most individuals were lost to follow-up. In addition, some desired information was not explicitly reported. Neurological scores (e.g., Hunt and Hess score [51]) for individuals who presented with a rupture were only reported in a small percentage of cases. Subsequent review of the data by a trained individual was required to derive estimates for these scores based on documented evidence. Nevertheless, without explicitly reporting such information, ambiguity was introduced into the information extraction process. The ontology assisted in data validation tasks in two ways. First, each entity was defined with a set of annotation properties, which specified attributes such as source (e.g., original vocabulary from which the entity is derived), valid values (e.g., enumerated responses that are permissible), synonyms (e.g., for linking terms with the same meaning), permissible data types (e.g., text, number, date) and definition (e.g., the formal meaning of the entity). As illustrated in Figure 1b, these annotations were reflected in the web-based interface where fields related to a specific entity were constrained to only accept predefined values (i.e., drawn from the list of valid values) or for input fields, the appropriate data type (i.e., as specified by the data property). Second, the ontology provided a basis for formulating structured reporting templates that can be used clinically. Clinical documents related to aneurysm treatment and follow-up were largely unstructured,

complicating efforts to track individual aneurysms over time. Feedback from the research dashboard that summarized the percentage of missing information for a given entity (e.g., what proportion of patients have smoking history reported) can be used to guide reporting improvement initiatives. As part of ongoing work, we are working collaboratively with the clinical investigators to (automatically) develop reporting templates for intracranial aneurysms based on the developed ontology.

3.2 Information Extraction

The performance of the information extraction pipeline was evaluated based on the identification of anatomical location, aneurysm shape, interventions, measurements, and clinical presentation from free-text reports. The evaluation was performed using a set of 916 sentences, randomly drawn from previously unseen prospective cases. Documents from which sentences were identified included radiology reports, operative reports, consultation notes, and admission/discharge summaries. For each sentence, the pipeline was used to highlight entities along with associated values (e.g., aneurysm dome size and numerical dimensions); sentences without any detected information were also presented. A single human annotator than manually reviewed each of the sentences to verify whether the highlighted entities were indeed correctly categorized or sentences without identified entities did not contain relevant information. Table 3 summarizes the system performance using metrics such as precision, recall, f-score, and accuracy. Overall, the annotators were able to consistently extract the targeted information with high precision. Extracting clinical presentation achieved the highest recall given the relatively constrained nature of the terminology used (e.g., incidental versus subarachnoid hemorrhage). Measurements had the lowest recall (0.90) and f-score (0.95) because several values were not captured by the employed regular expression pattern. Anatomic location had lower accuracy (0.97) given its high number of false negatives, demonstrating the difficulty of consistently detecting variations in location names. To improve recall rates, annotators for aneurysm location, interventional complications, and clinical assessment will be implemented using conditional random fields (using Mallet [52]) rather than relying solely on syntactic or lookup-based approaches based on the degree of variability in the way terms are expressed. While this approach may decrease overall precision, a human annotator is ultimately responsible for finalizing and curating cases before being entered into the database.

3.3 Example Applications

3.3.1 Aneurysm morphology and rupture—Several of the authors (JPV, NRG, FVV) recently published a study on the natural history of asymptomatic unruptured ICAs, which followed 165 individuals with 258 aneurysms over a period of 10 years. The study found that risk of aneurysm rupture was significantly increased with growth (2.4% per year) versus those that remained unchanged (0.2%) [29]. The findings underscored the importance of methodically capturing temporal information about individuals to elucidate factors that affect clinical outcomes of long-term diseases such as ICA. The original data collection and analysis was performed using a single spreadsheet that contained a series of columns representing measurements for each dimension. Maintaining and utilizing this spreadsheet for analysis proved to be a time-consuming process because new columns were added for each observation. The ontology-driven approach simplified the tracking and analysis of

individual aneurysm measurements. The ontology contained entities representing diameters across three measurement axes. Each dimension entity is related to the entities “dome size” and “neck size” through the relationship “has measurement”. The entity “dome size” is semantically related to “aneurysm morphology”, which is in turn associated with the finding “aneurysm”. Given these relationships, a SPARQL query can be used to retrieve all measurements for an aneurysm of interest. Using a combination of SPARQL queries, metrics could be automatically generated from the database. For example, the average follow-up time for these cases was 26.9 months (minimum: 1, maximum: 98). The ABC/2 method [53] was used to estimate aneurysm volume and assess for growth, accounting for measurement error when utilizing the image processing software. Examining the evolution of the 120 aneurysms in our dataset, 97 aneurysms remained unchanged while 23 enlarged in size. To date, two aneurysms that were followed had ruptured: one non-growing aneurysm and one growing aneurysm. As new cases and follow-up information are added to the database, rupture rate based on characteristics such as aneurysm shape, location, size, and growth can be assessed in real-time. Figure 4 demonstrates how linked observations can be represented as a bubble plot where the horizontal axis represents time and the visual appearance represent the number and sizes of aneurysms for a given patient. Additional events drawn from the patient record could be overlaid to provide context. In summary, this example application illustrated how the ontology assisted in the organization and semantic retrieval of information from the database, allowing researchers to easily track and retrieve longitudinal measurements for analysis.

3.3.2 Genes and pathways involved in aneurysm rupture—The second example utilizes the database to support exploratory analysis of human whole blood transcriptomes with clinical and imaging phenotypes to identify WGCNA modules associated with aneurysm size and rupture. An initial batch of 12 individuals with blood samples was identified from the database to perform a pilot study, isolating and sequencing their RNA. The goal was to search for novel genes and pathways involved in ICA progression using RNA sequencing data, genetic variant data, and clinical phenotypes. From the set of 16,535 genes, 30 functionally related gene modules were enriched through WGCNA analysis using annotations provided by Gene Ontology. The hierarchical structure of the ICA ontology facilitated the retrieval of clinical phenotypes for the correlation network analysis. For example, in the absence of the ontology, queries that specified particular columns in the database were needed to retrieve all of the relevant clinical values. With the ontology, parent-child relationships were exploited to identify relevant entities. For example, to identify values related to family history, a SPARQL query was written to return child entities of “Family History” yielding entities such as “Family History of Arteriovenous Malformation” and “Family History of Diabetes”. These results were used to identify matching rows in the observational table of the OMOP common data model. Further, because RNA expression is influenced by environmental factors, contextual information related to the environment and patient state at the time of blood collection needed to be captured. In the ontology, observations were annotated as being static or time-dependent. Unchanging entities such as patient demographics and medical history were semantically grouped as part of a static patient context. Time-dependent entities such as aneurysm size and rupture status were grouped as part of an aneurysm context. These semantic groupings

aided in the identification of entities that may have changed after the date of specimen collection. The list of gene modules and clinical phenotypes were then used as inputs into a Fisher's exact test to determine whether relationships between gene modules and clinical phenotypes had statistical significance. Down regulation of two transcript modules were correlated with ruptured ICA status ($r = -0.78$, $p = 0.08$ and $r = -0.77$, $p = 0.09$) while up-regulation of two cellular respiration and translation modules were associated with aneurysm size ($r = 0.86$, $p = 0.02$ and $r = 0.9$, $p < 0.001$) [54]. This example demonstrated how the structure of the ontology simplified retrieval of information from the database. Further study is necessary to validate the influence on these two transcript modules.

4. DISCUSSION

Our experience in deploying an ontology-driven approach for characterizing and analyzing data captured routinely in the EHR has demonstrated the promise of utilizing such information to generate a detailed set of clinical phenotypes for subsequent clinical-genomic analyses. Despite the inherent shortcomings of observational data, the ability to query the resulting repository and utilize the ontology to address certain limitations has been useful.

4.1 Strengths and Limitations

The described ontology-driven approach addresses issues related to the extraction, representation, and validation of clinical observational data for research. First, the ontology provides a centralized source for defining variables to be captured, how each variable is represented, what variables already exist in other controlled vocabularies, and explicit definitions of relationships between variables and permissible values. Second, using the domain knowledge encoded in the ontology, data quality issues that are inherent in clinical data can be identified, if not addressed. Table 4 summarizes how the ontology is being applied to handle various dimensions of data quality assessment [4]. Finally, the usage of ontologies facilitates the integration of multiple data sources by explicitly modeling the relationships between variables and providing a formal means to query across distributed sources using SPARQL. Several limitations should be noted. The creation and maintenance of an ontology is a significant undertaking, requiring a series of meetings among investigators to come to a consensus about the variables and properties to collect. These discussions yielded useful insights on defining the scope of the data collection and improving the consistency of how observational data was represented. The exercise of formulating the ICA ontology also helped identify variables that were important risk factors in aneurysm rupture as reported in current literature. A second limitation is that the presented approach remains an indirect way to improve clinical data capture, given that the ontology is not directly integrated with the EHR and does not influence the data entry interfaces that are directly utilized by healthcare providers. While the ontology and infrastructure can quickly identify inconsistencies in data entry and missing information, the ontology does not prevent such events from happening at the time of data input. Nevertheless, this process has been informative to clinicians who can measure the conformance rates to current reporting standards, providing feedback as to what information needs to be more consistently reported or sources of measurement error. Finally, not all data collection projects are ideally suited for this approach. Existing tools such as REDCap

require less upfront development and provide sufficient validation and reporting tools. However, projects that require the integration of multiple heterogeneous and unstructured clinical data sources and require access to detailed clinical observations will benefit from the consistency, preciseness, and domain-awareness permitted through this approach.

4.2 Challenges in Representing Longitudinal Observational Data

The described efforts underscore the need for more accurate methods for data capture and representation, particularly as studies move towards longitudinal rather than cross-sectional analysis. Combining retrospective and prospective cases as part of a single research study presents unique challenges. While the review of retrospective cases helped increase sample sizes for analysis and provided cases that had long-term follow-up, issues arose when integrating this information with prospective data without appropriate context. For example, measurements of aneurysm size using acquisition protocols and image analysis software from ten years ago differed from measurements taken of the same aneurysm using today's technology. These differences could incorrectly have indicated a significant change in aneurysm size when the variations were not significant in reality. Furthermore, reporting style, available diagnostic tests, and interventions all evolved over time. For instance, the influence of dome direction on the success of endovascular coiling was not well understood until the past five years. While older radiology reports omit this information, more recent reports documented dome direction as an indicator of treatment success. Care must be given in recording the context in which measurements are taken (e.g., what imaging modality was used, whether the measurement was taken pre/post treatment). The ontology encoded basic information related to data provenance that permitted investigators to determine the source of the measurements and methods used to derive specific values. For example, data properties such as "has lower bound" and "has upper bound" were defined to capture uncertainty related to aneurysm measurements made using a specific software tool based on specifications provided by the manufacturer. Ongoing work is addressing the need to capture additional layers of provenance information.

Moreover, being able to precisely track individual findings (e.g., aneurysms, lesions, nodules) over time is critical, particularly as the technology to identify and characterize precise locations and minute changes in the biology of these findings are translated and adopted clinically. While current clinical documentation has generally been sufficient to perform coreference resolution of findings over time, the process of tracking findings would be greatly facilitated by identifying unique characteristics at the outset that are easily identifiable by individuals reporting or analyzing the clinical data. In a small percentage of cases, we needed to manually review source data (e.g., imaging studies) to verify the location of each reported aneurysm (e.g., in cases where laterality of the aneurysm was not documented in the reports). Our blood collection is being expanded to obtain samples from individuals as part of their routine blood work to provide new insights into how their transcriptome changes as the aneurysm evolves over time. While results of this initial study are promising, we are continuing to collect prospective cases to increase our sample size to achieve more significant statistical power to validate the clinical significance of the identified modules.

Nevertheless, a longstanding difficulty of longitudinal data analysis is inferring what occurs between time points. Patients may have multiple healthcare providers who may not be part of the same health system. Data captured during an encounter with that provider (e.g., assessment of neurologic state) may not be recorded until the patient is seen again at our institution. A mechanism to routinely follow-up with treated patients would provide important insights into the environmental and lifestyle factors that influence aneurysm recurrence. Ultimately, a broader framework involving a national-level learning health system, as motivated in [55], would be beneficial towards improving the quality and coverage of clinical information about individuals. A strength of utilizing ontologies to standardize the representation of observational data is the ability to interoperate across geographically separated databases. SPARQL queries may be formulated to retrieve instances of entities from different knowledge sources. For instance, semantic queries can be executed against clinical data obtained at our institution and other resources such as the @neurIST project or public data repositories. The implementation of a federated database is left for future work but would be modeled after similar efforts [56].

4.3 Generalizability

While demonstrated in the domain of ICA, this approach can be generalized and scaled to other domains as well. The underlying OMOP common data model is domain-agnostic, allowing the schema to be reused across domains. Modifications to the application ontology will be needed to ensure that the standardized vocabulary reflects the domain being captured; this process presents the most labor intensive step in adapting and scaling the entire workflow. Still, given the modular architecture of the system, additional analyses can be incorporated by defining new entities in the ontology or linking to other external sources. The research dashboard automatically displays filters based on entities in the ontology and in the future will dynamically update the forms based on modifications to the ontology. Currently, only basic provenance information is addressed in this work. Expanding the capture of semantic provenance is critical, particularly in interpreting derived information. For instance, results of the information extraction process and quantitative image analysis need to be supplemented with the steps used to generate such information (e.g., what workflow pipeline was used to extract the observed value). We are exploring workflow management tools such as Taverna [57, 58] and semantic domain-aware provenance models like Janus [59] to capture intermediate results at each processing step.

4.4 Future Work

As part of future work, extraction of newer cases from the EHR will continue towards providing additional statistical power to validate the pilot imaging-genomic correlation study. Newer –omics techniques (e.g., DNA methylation studies) are also being pursued to provide new perspectives on a disease. Finally, this paradigm is being applied to other domains at our institution, namely in the context of lung cancer screening and prostate cancer diagnosis, utilizing the longitudinal data aggregation tools and workflow established to precisely characterize clinical phenotypes and image findings over time for individuals who are at risk for cancer.

5. CONCLUSION

We have presented an ontology-driven data integration and analysis workflow that facilitates secondary analysis of observational data collected during routine patient care. We have demonstrated two applications of how this approach assists in the consistent data capture and representation of patient information and facilitates data integration across sources to glean new insights into a complex disease.

Acknowledgments

The authors would like to acknowledge Lewellyn Andrada, Juan Anna Wu, Alex Juncosa, Edgar Rios, Patrick Langdon, Mark Connolly, and Joshua Dusick for their efforts in the data collection process. This work is funded by NIBIB R01 EB000362. Additional support is provided by NCI R01 CA157553 (AATB, WH), American Heart Association (NRG), and NIH R01 HL-095056 and NIH P01 HL-28481 (PP).

References

1. Safran C, Bloomrosen M, Hammond W, et al. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. J Am Med Inform Assoc. 2007; 14(1):1–9. [PubMed: 17077452]
2. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care. 2007; 13(6 Part 1):277–8. [PubMed: 17567224]
3. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med. 2009; 151(5):359–60. [PubMed: 19638404]
4. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013; 20(1):144–51. [PubMed: 22733976]
5. Levit, L.; Balogh, E.; Nass, S.; Ganz, PA. Delivering high-quality cancer care: Charting a new course for a system in crisis. National Academies Press; 2013.
6. Smith, M.; Saunders, R.; Stuckhardt, L.; McGinnis, JM. Best care at lower cost: the path to continuously learning health care in America. National Academies Press; 2013.
7. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care. 2007; 13(6 Part 1):281–8. [PubMed: 17567225]
8. Wang X, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. J Am Med Inform Assoc. 2009; 16(3):328–37. [PubMed: 19261932]
9. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011; 12(6):417–28. [PubMed: 21587298]
10. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010; 17(2):124–30. [PubMed: 20190053]
11. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform. 2012; 45(4):763–71. [PubMed: 22326800]
12. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011; 4(1):13. [PubMed: 21269473]
13. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L. Translational bioinformatics: Linking knowledge across biological and clinical realms. J Am Med Inform Assoc. 2011; 18(4):354–7. [PubMed: 21561873]

14. Luciano JS, Andersson B, Batchelor C, et al. The Translational Medicine Ontology and Knowledge Base: Driving personalized medicine by bridging the gap between bench and bedside. Journal of biomedical semantics. 2011; 2(Suppl 2):S1. [PubMed: 21624155]
15. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Medical care research and review. MCRR. 2010 Oct; 67(5):503–27. [PubMed: 20150441]
16. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Med Care. 2013; 51:S30–S7. [PubMed: 23774517]
17. Coyle, JF.; Heras, Y.; Oniki, T.; Huff, SM. Clinical Element Model Introduction & Data Types Reference Manual. 2008. [cited 2014 06–25]; Available from: http://informatics.mayo.edu/sharp/images/e/e2/CEM_Reference20081114.pdf
18. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: Rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010; 153(9): 600–6. [PubMed: 21041580]
19. Hersh WR, Cimino J, Payne PR, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2013; 1(1):14.
20. Min H, Manion FJ, Goralczyk E, Wong Y-N, Ross E, Beck JR. Integration of prostate cancer clinical data using an ontology. J Biomed Inform. 2009; 42(6):1035–45. [PubMed: 19497389]
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009; 42(2):377–81. [PubMed: 18929686]
22. Sahoo SS, Lhatoo SD, Gupta DK, et al. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. J Am Med Inform Assoc. 2014; 21(1):82–9. [PubMed: 23686934]
23. Sahoo SS, Zhao M, Luo L, et al. OPIC: ontology-driven patient information capturing system for epilepsy. AMIA Annu Symp Proc. 2012; 2012:799–808. [PubMed: 23304354]
24. Feigin VL, Lawes CMM, Bennett DA, Barker-Collo SL, Parag V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. Lancet Neurol. 2009; 8(4):355–69. [PubMed: 19233729]
25. Elsharkawy A, Lehecka M, Niemela M, et al. Anatomic Risk Factors for Middle Cerebral Artery Aneurysm Rupture: Computerized Tomographic Angiography Study of 1009 Consecutive Patients. Neurosurgery. 2013; 73(5):825–37. [PubMed: 24141397]
26. Vinters, HV.; Hammond, RR. Saccular (Berry) Aneurysms. In: Kalimo, H., editor. Pathology and Genetics: Cerebrovascular Diseases. Basel, Switzerland: ISN Neuropath Press; 2005. p. 104-11.
27. Yasuno K, Bilguvar K, Bijlenga P, et al. Genome-wide association study of intracranial aneurysm identifies three new risk loci. Nat Genet. 2010; 42(5):420–5. [PubMed: 20364137]
28. Ruigrok YM, Tan S, Medic J, Rinkel GJ, Wijmenga C. Genes involved in the transforming growth factor beta signalling pathway and the risk of intracranial aneurysms. Journal of neurology, neurosurgery, and psychiatry. 2008; 79(6):722–4.
29. Benkner S, Arbona A, Berti G, et al. @neurIST: Infrastructure for Advanced Disease Management Through Integration of Heterogeneous Data, Computing, and Complex Processing Services. IEEE Trans Inf Technol Biomed. 2010; 14(6):1365–77. [PubMed: 20435543]
30. Huang D, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009; 37(1):1–13. [PubMed: 19033363]
31. Arp R, Smith B. Function, role, and disposition in basic formal ontology. Nature Precedings. 2008; 1941(1):1–4.
32. Viñuela F, Duckwiler G, Mawad M. Guglielmi detachable coil embolization of acute intracranial aneurysm: perioperative anatomical and clinical outcome in 403 patients. J Neurosurg. 1997; 86(3):475–82. [PubMed: 9046305]

33. Jayaraman MV, Meyers PM, Derdeyn CP, et al. Reporting standards for angiographic evaluation and endovascular treatment of cerebral arteriovenous malformations. J Neurointerv Surg. 2011; 4(5):325–30. [PubMed: 22131440]
34. Ferns SP, Sprengers ME, van Rooij WJ, et al. Coiling of Intracranial Aneurysms A Systematic Review on Initial Occlusion and Reopening and Retreatment Rates. Stroke. 2009; 40(8):e523–e9. [PubMed: 19520984]
35. Meyers PM, Schumacher HC, Higashida RT, et al. Reporting standards for endovascular repair of saccular intracranial cerebral aneurysms. J Vasc Interv Radiol. 2009; 20(7):S435–S50. [PubMed: 19560031]
36. Currie S, Mankad K, Goddard A. Endovascular treatment of intracranial aneurysms: review of current practice. Postgrad Med J. 2011; 87(1023):41–50. [PubMed: 20937736]
37. Brown RD Jr, Broderick JP. Unruptured intracranial aneurysms: epidemiology, natural history, management options, and familial screening. Lancet Neurol. 2014; 13(4):393–404. [PubMed: 24646873]
38. Dhar S, Tremmel M, Mocco J, et al. Morphology parameters for intracranial aneurysm rupture risk assessment. Neurosurgery. 2008; 63(2):185. [PubMed: 18797347]
39. Schievink WI. Genetics of intracranial aneurysms. Neurosurgery. 1997; 40(4):651–63. [PubMed: 9092838]
40. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004; 10(3–4):327–48.
41. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010; 17(3):229–36. [PubMed: 20442139]
42. Brinkman RR, Courtot M, Derom D, et al. Modeling biomedical experimental processes with OBI. J Biomedical Semantics. 2010; 1(S-1):S7.
43. Tudorache T, Vendetti J, Noy NF. Web-Protege: A lightweight OWL ontology editor for the web. 5th OWL: Experiences and Directions (OWLED). 2008; 2008
44. WebProtege. [cited 2014-10-01]; Available from: http://webprotege.stanford.edu/
45. Pathak J, Wang J, Kashyap S, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc. 2011; 18(4):376–86. [PubMed: 21597104]
46. Chien A, Castro M, Tateshima S, Sayre J, Cebral J, Vinuela F. Quantitative hemodynamic analysis of brain aneurysms at different locations. AJNR Am J Neuroradiol. 2009; 30(8):1507–12. [PubMed: 19406766]
47. Chien A, Sayre J, Viñuela F. Quantitative comparison of the dynamic flow waveform changes in 12 ruptured and 29 unruptured ICA–ophthalmic artery aneurysms. Neuroradiology. 2013; 55(3): 313–20. [PubMed: 23443738]
48. Chien A, Tateshima S, Castro M, Sayre J, Cebral J, Viñuela F. Patient-specific flow analysis of brain aneurysms at a single location: comparison of hemodynamic characteristics in small aneurysms. Med Biol Eng Comput. 2008; 46(11):1113–20. [PubMed: 18931868]
49. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012 Oct 25. 2012.
50. Haas B, Horvath S, Pietilainen K, et al. Adipose Co-expression networks across Finns and Mexicans identify novel triglyceride-associated genes. BMC Med Genomics. 2012; 5(1):61. [PubMed: 23217153]
51. Hunt WE, Hess RM. Surgical risk as related to the time of intervention in the repair of intracranial aneurysms. Neurosurgical Classics II. 1967:342–6.
52. McCallum, A. MALLET: A Machine Learning for Language Toolkit. 2002. [cited 2014-10-01]; Available from: http://mallet.cs.umass.edu
53. Kothari RU, Brott T, Broderick JP, et al. The ABCs of measuring intracerebral hemorrhage volumes. Stroke. 1996; 27(8):1304–5. [PubMed: 8711791]
54. Haas B, Gonzalez NR, Nikkola E, et al. Abstract W P73: Feasibility and Preliminary Results of Whole Blood RNA-Sequencing Analysis in Patients With Intracranial Aneurysms. Stroke. 2014; 45(Suppl 1):AWP73.

55. Feeley TW, Sledge GW, Levit L, Ganz PA. Improving the quality of cancer care in America through health information technology. J Am Med Inform Assoc. 2013; 21(5):772–5. [PubMed: 24352553]

56. Gupta A, Bug W, Marenco L, et al. Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). Neuroinformatics. 2008; 6(3):205–17. [PubMed: 18958629]

57. Oinn T, Addis M, Ferris J, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004; 20(17):3045–54. [PubMed: 15201187]

58. Zhao J, Goble C, Stevens R, Turi D. Mining Taverna's semantic web of provenance. Concurrency and Computation: Practice and Experience. 2008; 20(5):463–72.

59. Missier, P.; Sahoo, SS.; Zhao, J.; Goble, C.; Sheth, A. Provenance and Annotation of Data and Processes. Springer; 2010. Janus: From workflows to semantic provenance and linked open data; p. 129-41.

Highlights

- Incomplete and inconsistent data limit the use of observational data in research.
- An ontology-driven framework is proposed for extracting and representing data.
- The ontology facilitates standardization, data extraction, and semantic retrieval.
- Two examples illustrate how the ontology supports the analytic workflow.

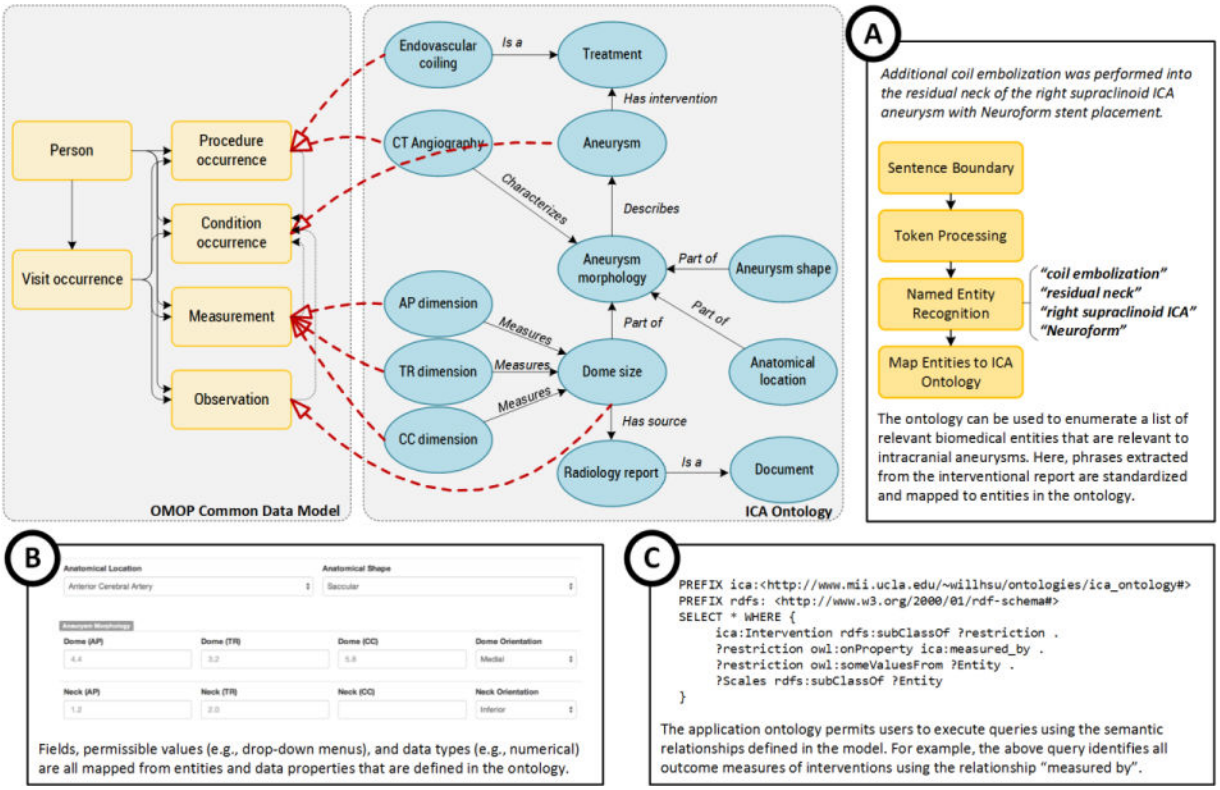

Figure 1.

A schematic illustrating how information in the ontology was utilized in different aspects of the data extraction and representation process. The ontology was leveraged in three ways: (a) As part of the information extraction task, entities in the ontology were used to provide terms and synonyms for the named entity recognition component; (b) the ontology also mapped to specific fields in the web-based data entry form and provided constraints for data validation; and (c) the contents of the ontology were queried using SPARQL to retrieve information from the underlying data model based on semantic relationships.

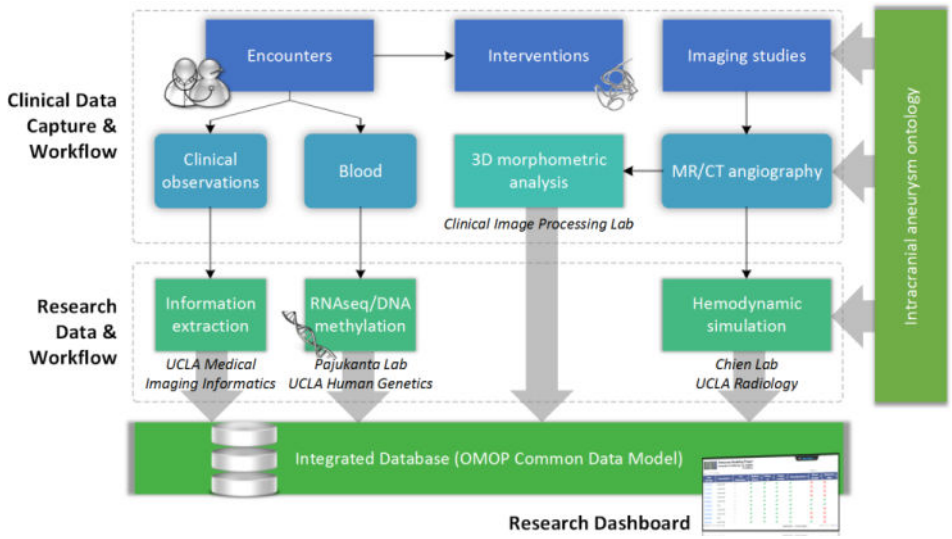

Figure 2.

The analysis workflow. Routinely generated clinical data were captured on retrospective and prospective cases. Each case was augmented with additional image processing and genetic analyses performed offline. The entire process was organized using a central ontology that linked with an OMOP-based database and a web-based dashboard.

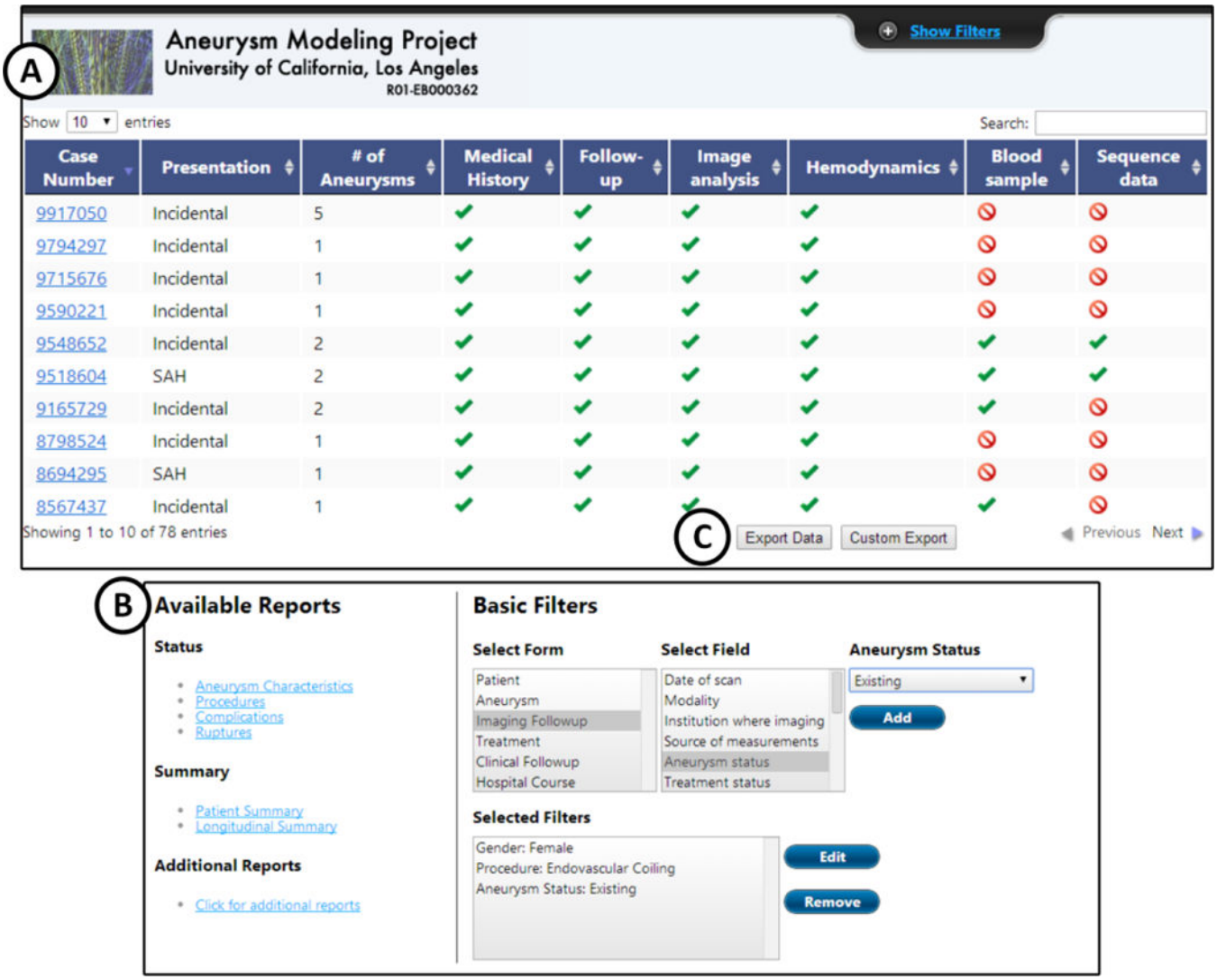

Figure 3.

A screen capture of the progress dashboard. (a) Users viewed the entire status of the analysis workflow and what information is available for each case from the dashboard. (b) Built-in filters permitted users to easily search for cases that have matching characteristics. A predefined set of reports was also provided, allowing users to generate visualizations such as the one shown in Figure 4 in real-time. (c) The export option allowed users to retrieve relevant information about matching patients in a format that was amenable for analysis in external programs such as R.

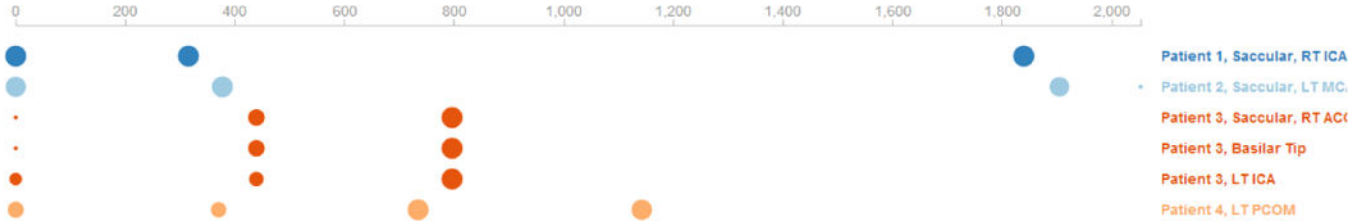

Figure 4.

Visualization that summarizes available observations of individual aneurysms for a subset of patients plotted by days since presentation. Observations are represented as circles, whose size is proportion to aneurysm diameter. Such visualization can be generated in real-time from the database to track changes in aneurysms among followed individuals.

Table 1

Summary of variables extracted or derived from observational clinical data, the data source from which the variable was derived, and any existing vocabularies to which the variable could be mapped.

	Parent concept	Concept	Data source	Source vocabulary
Clinical encounters and follow-up	Demographics, diagnosis, presentation	Age Clinical presentation Date of presentation Ethnicity Gender Signs and symptoms	Admission and discharge summaries; surgical reports, radiology consults (HIS/RIS)	SNOMED CT, NCI Thesaurus
	Family history, social history	Alcohol consumption Smoking ersonal/family history of chronic heart disease Personal/family history of diabetes Personal/family history of inherited disorders Personal/family history of other comorbidities	History & physical (HIS); patient self-reported surveys	Aneurist
	Neurological exam, follow-up and outcomes	Clinical assessment Date of follow-up Glasgow Coma Sore Modified Rankin Scale NIH Stroke Scale World Federation Scale	Neurology consults (HIS)	Aneurist
	Medications	Papaverine Triple-H	Computerized order entry (CPOE)	RxNORM
Imaging Studies	Image findings	Arterial incorporation Calcification Intraluminal thrombus Number of aneurysms Treatment status	CT/MR angiography; radiology reports (PACS, RIS)	Aneurist, RadLex
	Morphology	Anatomical location Date of scan Dome dimension Neck dimension Shape		Aneurist, Foundational Model of Anatomy
	Hemodynamics	Flow patterns Wall shear stress	Hemodynamic simulations from CT/MR angiography	
Genetics	Peripheral blood sequencing	DNA methylation RNA transcription factors	Biospecimen collection	Gene ontology
Interventions	Endovascular embolization	Balloon assistance Complications Hospital length of stay ICU length of stay Stent placement	Interventional neuroradiology reports (HIS)	SNOMED CT, Aneurist

	Parent concept	Concept	Data source	Source vocabulary
		Type of embolic material		
	Surgical clipping	Bypass type Complications Hospital length of stay ICU length of stay Temporary artery clipping Type of aneurysm clip Type of craniotomy		
	Outcome	Immediate anatomic result Immediate clinical result	Consult notes (HIS)	

Table 2

Number of entities in the ICA ontology broken down by source vocabulary.

Source Vocabulary	Concepts (%)
Aneurist	130 (26.9%)
SNOMED CT	106 (22%)
New entities that could not be mapped to other ontologies	96 (19.8%)
NCI Thesaurus	42 (8.7%)
Basic Formal Ontology	39 (8.1%)
RxNORM/MedDRA	21 (4.3%)
Foundational Model of Anatomy	20 (4.1%)
Other ontologies	15 (3.1%)
Logical Observation Identifiers Names and Codes (LOINC)	14 (3%)

Table 3

Information extraction performance for each category of information. TP, true positive; TN, true negative; FN, false negative.

Feature Type	TP	TN	FP	FN	Precision	Recall	F-score	Accuracy
Anatomical location	452	436	2	26	0.99	0.94	0.97	0.97
Morphology	40	872	0	4	1.0	0.91	0.95	0.99
Intervention	332	566	18	0	0.95	1.0	0.97	0.98
Measurement	114	790	0	12	1.0	0.98	0.99	0.99
Presentation	124	790	0	2	1.0	0.98	0.99	0.99

Table 4

The contribution of the approach in addressing different categories of data quality, as adapted from [4].

Domains	Description	Proposed approach
Correctness	Whether observations accurately reflect the true state of a patient	For each observation, the ontology captures not only the observed value but also annotations related to the unique definition of a given entity, permissible values and data types, and source of the information. Basic provenance information is also captured.
Plausibility	Whether an observation makes sense in the context of other observations	
Concordance	Whether independent observations are in agreement and reliable	Data from different clinical sources is semantically related based on patient, medical problem, states, findings, and observations. This organization allows values for a given entity to be linked and assessed to determine the level of concordance.
Currency	Whether an observation is a relevant representation of the patient at a given time point	Given that observations are grouped together temporally into states, the recency of observations can be evaluated. In addition, observations made during the time of biospecimen collection can be identified, providing the appropriate context for interpreting the results of gene expression analysis.
Completeness	Whether an observation is present or absent, as expected	Once mapped to the ontology, the level of missing or discordant information in any given data source can be measured, allowing the overall data quality to be characterized.