Journal of Biomedical Informatics 55 (2015) 132–142

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An integrated, ontology-driven approach to constructing observational databases for research

William Hsu [a,*], Nestor R. Gonzalez [a,b], Aichi Chien [a], J. Pablo Villablanca [a], Päivi Pajukanta [c], Fernando Viñuela [a], Alex A.T. Bui [a]

[a] Department of Radiological Sciences, UCLA David Geffen School of Medicine, Los Angeles, CA, United States
[b] Department of Neurosurgery, UCLA David Geffen School of Medicine, Los Angeles, CA, United States
[c] Department of Human Genetics, UCLA David Geffen School of Medicine, Los Angeles, CA, United States

ARTICLE INFO

Article history:
Received 21 March 2014
Revised 14 February 2015
Accepted 19 March 2015
Available online 26 March 2015

Keywords:
Data extraction
Biomedical ontology
Retrospective study
Image analysis
Database
Intracranial aneurysm

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 toward a learning health system paradigm.

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

* Corresponding author at: UCLA Medical Imaging Informatics, 924 Westwood Blvd, Suite 420, Los Angeles, CA 90024, United States. Tel./fax: +1 (310) 794 3536.
E-mail address: willhsu@mii.ucla.edu (W. Hsu).

http://dx.doi.org/10.1016/j.jbi.2015.03.008
1532-0464/

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 threefold: (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 Fig. 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.

2.1.2.1. 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.