Journal of Biomedical Informatics 57 (2015) 42–52

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Combining expert knowledge and knowledge automatically acquired from electronic data sources for continued ontology evaluation and improvement

Claire L. Gordon [a,b,c], Chunhua Weng [b,*]

[a] Department of Medicine, Columbia University Medical Center, 630 West 168th Street, New York, USA
[b] Department of Biomedical Informatics, Columbia University Medical Center, 622 West 168th Street, New York, NY 10032, USA
[c] Department of Medicine, University of Melbourne, Melbourne, VIC 3010, Australia

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 17 July 2015
Accepted 19 July 2015
Available online 23 July 2015

Keywords:
Infectious disease
Bacteria
Antibiotic
Ontology
Evaluation
Knowledge acquisition

ABSTRACT

Introduction: A common bottleneck during ontology evaluation is knowledge acquisition from domain experts for gold standard creation. This paper contributes a novel semi-automated method for evaluating the concept coverage and accuracy of biomedical ontologies by complementing expert knowledge with knowledge automatically extracted from clinical practice guidelines and electronic health records, which minimizes reliance on expensive domain expertise for gold standards generation.
Methods: We developed a bacterial clinical infectious diseases ontology (BCIDO) to assist clinical infectious disease treatment decision support. Using a semi-automated method we integrated diverse knowledge sources, including publically available infectious disease guidelines from international repositories, electronic health records, and expert-generated infectious disease case scenarios, to generate a compendium of infectious disease knowledge and use it to evaluate the accuracy and coverage of BCIDO.
Results: BCIDO has three classes (i.e., infectious disease, antibiotic, bacteria) containing 593 distinct concepts and 2345 distinct concept relationships. Our semi-automated method generated an ID knowledge compendium consisting of 637 concepts and 1554 concept relationships. Overall, BCIDO covered 79% (504/637) of the concepts and 89% (1378/1554) of the concept relationships in the ID compendium. BCIDO coverage of ID compendium concepts was 92% (121/131) for antibiotic, 80% (205/257) for infectious disease, and 72% (178/249) for bacteria. The low coverage of bacterial concepts in BCIDO was due to a difference in concept granularity between BCIDO and infectious disease guidelines. Guidelines and expert generated scenarios were the richest source of ID concepts and relationships while patient records provided relatively fewer concepts and relationships.
Conclusions: Our semi-automated method was cost-effective for generating a useful knowledge compendium with minimal reliance on domain experts. This method can be useful for continued development and evaluation of biomedical ontologies for better accuracy and coverage.

1. Introduction

Ontologies enable formal representation and sharing of domain knowledge [1] and can augment clinical decision support systems by providing a standard vocabulary for biomedical entities to help standardize and integrate heterogeneous data resources [2–4]. Ontologies are now pervasive in biomedicine and function to address multiple requirements including knowledge management, data integration, exchange and semantic interoperability, and decision support and reasoning [2]. However, ontology evaluation remains difficult [5]. Common methods for the evaluation of biomedical ontologies include conformance to a philosophical principle [6], application or task-based evaluation [7], user-based evaluation [8], data-driven evaluation [9] and gold standard-based evaluation [10]. Evaluation of a large clinical knowledge base often centers on example applications and involves comparing ontologies against pre-defined gold standards [11]. This can be problematic for domain-specific ontologies since there may be no available gold standard for comparison [11]. The development of a new gold standard requires extensive domain expertise through a process that can have poor cost-effectiveness and cause long time delays.

* Corresponding author at: Biomedical Informatics, Columbia University, 622 W 168th Street, PH-20 Rm 407, New York, NY 10032, USA. Tel.: +1 (212) 305 3317; fax: +1 (212) 305 5392.
E-mail address: cw2384@columbia.edu (C. Weng).

http://dx.doi.org/10.1016/j.jbi.2015.07.014
1532-0464/

Evaluation of the accuracy and comprehensiveness of a large domain-specific ontology typically relies on domain experts to manually develop a gold standard reference. This method faces several challenges. First, knowledge about a domain is constantly evolving, but knowledge acquisition directly from domain experts cannot happen as frequently as needed and often lags behind knowledge generation in any domain. Static gold standards can soon become outdated. Modern ontology design and evaluation requires an iterative and dynamic process so that newly emerging knowledge can be incorporated in frequent evaluations. Second, domain experts may not possess comprehensive knowledge about a domain all the time; therefore, relying on the single source of expert knowledge can lead to bias or limitations in the resulting gold standard.

In this paper, we presented a new semi-automated method for combining multiple knowledge sources to evaluate biomedical ontologies, which minimize the involvement of domain experts and augment them with knowledge automatically acquired from public electronic data sources, and applied it to evaluate a bacteria clinical infectious disease ontology. We explain how we utilized automated extraction of concepts and properties in conjunction with manual methods to integrate multiple diverse knowledge sources into a comprehensive compendium of infectious disease (ID) knowledge, and then compared BCIDO to this knowledge compendium. This method is superior to existing static methods for ontology development in that it can run anytime and multiple times so that emerging new domain knowledge can be incorporated in gold standard generation as often as preferred. On this basis, we discuss how this method can be used for the evaluation of other biomedical ontologies. Another contribution of this work is a validated bacteria clinical infectious disease ontology that provides comprehensive concept and concept relationships that are useful for portable decision support for antimicrobial prescription.

2. Materials and methods

2.1. Combining expert knowledge and electronic data for BCIDO ontology development

Antimicrobial resistance is an increasing problem worldwide and is often caused by inappropriate antimicrobial prescribing. Antibiotic resistance is now a major threat to public health and has the potential to affect anyone, of any age, in any country [12]. Incorporating an antibiotic decision support system (ADSS) into clinical decision-making has been shown to be effective at reducing inappropriate antibiotic prescribing and lowering local antimicrobial resistance [13–15]. However, despite their apparent benefits, ADSSs are infrequently used in the hospital in-patient setting [16]. The barriers to widespread adoption and implementation of successful ADSSs include standalone systems that are independent from the electronic health record (EHR) and require interruption of the clinical workflow to use [14,17], a single infectious disease focus (i.e., acute bronchitis) [18,19] or single clinical location (i.e., intensive care or primary care) [14,20–24], and ADSSs that use their own terminology and cannot be transferred to other EHR systems [15,18,25,26]. To improve the interoperability of future portable ADSSs, we developed and published a bacterial clinical infectious diseases ontology (BCIDO) [16].

BCIDO defines common concept definitions for clinical infectious diseases along with domain knowledge commonly used in the hospital in-patient setting for the diagnosis of these diseases. BCIDO encompasses concept definitions for common clinical presentations of infections, patient-specific factors that influence differential diagnoses and treatment options, the organisms themselves, and the antimicrobial agents used to treat infections. The design of BCIDO has been described previously [16]. In brief, the ontology covers factors relevant to making an antimicrobial decision in the hospital setting, including patient factors and microbiology results, such as gram stain and culture results. Specific antimicrobial treatment recommendations are not defined in BCIDO because they vary widely among clinicians, institutions and countries and are therefore not "universal truths". However, the factors required for making an antimicrobial treatment decision are included so that treatment decisions in an ADSS can be tailored to local preferences. BCIDO is limited to bacterial infections. However, it has been designed to be easily extended to include antimicrobial treatments for mycobacterial, viral, and fungal infections. The concept granularity of the ontology is often chosen to ensure a diagnosis or treatment recommendation can be made at this granularity level.

When designing BCIDO, the Infectious Disease Ontology (IDO) [27] (http://infectiousdiseaseontology.org/) was selected as the upper ontology. IDO is a suite of interoperable ontology modules that together aim to cover the entire infectious disease domain. The suite consists of the core IDO, covering terms and relations generally relevant to the infectious disease domain, and a set of domain-specific ontologies developed as extensions from the core [28]. To date, disease and pathogen specific extension ontologies have been developed for malaria [29], dengue fever [30], brucellosis [31], and *Staphylococcus aureus* [32,33]. The primary purpose of the core IDO is to maximize interoperability between IDO extensions as well as with ontologies outside the IDO suite. To accomplish this, IDO is developed within the framework of the OBO Foundry [28] (http://obofoundry.org/) and adheres to the Foundry's ontology development guidelines. BCIDO was developed using the core IDO as an upper ontology, and thus the Basic Formal Ontology and Ontology of General Medical Science, which serve as upper ontologies for the IDO suite. BCIDO adheres to the Foundry's ontology development guidelines and to Cimino's Desiderata for terminologies [1]. Together these include: (1) using Aristotelian definitions with a single mode of classification, (2) using single inheritance hierarchies, (3) using relations with formal, logical definitions based on a distinction between types and instances, and (4) writing definitions and ontology assertions as compositions of ontology terms and relations.

To help standardize and integrate data resources, clinical infectious disease concepts and antibiotics in BCIDO were mapped to the reference resource, the Unified Medical Language System (UMLS) [34] concept unique identifiers (CUIs), where possible. UMLS integrates many terminologies and coding standards. Mapping BCIDO to UMLS CUIs enables BCIDO to be linked to many other relevant biomedical resources such as SNOMED-CT and ICD version 9 or 10 [35]. Bacterial terms were imported from the National Center for Biotechnology Information Organismal Classification (NCBITaxon). Anatomical terms were imported from The Foundational Model of Anatomy (http://sig.biostr.washington.edu/projects/fm/index.html) (FMA) [11] and were used to define the location of infectious processes (i.e., osteomyelitis *located_in* some bone).

The ontology was represented in the OWL 2 EL Web Ontology Language (OWL) as a single hierarchical structure using the Protégé-OWL editor (http://protege.stanford.edu). The entire IDO core ontology was imported as the upper ontology (http://purl.obolibrary.org/obo/ido.owl). The Basic Formal Ontology was used to assist in designing the structure of our ontology and defining additional ontology classes and properties. Clinical infectious disease concepts and antibiotics were mapped to UMLS using the "identifier" annotation property, and synonyms or related terms were recorded using the "has_related_term", "has_exact_synonom" or "has_broad_synonom" annotation properties, as defined by the Dublin core. Bacterial terms were imported from the