OCRe: An Ontology of Clinical Research

Samson W. Tu, MS,[1] Simona Carini, MA,[2] Alan Rector, MD PhD,[3]
Peter Maccallum, PhD,[4] Igor Toujilov, PhD,[5] Steve Harris, PhD,[6] Ida Sim, MD PhD[2]

[1]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[2]Division of General Internal Medicine, UCSF, San Francisco, CA
[3]Department of Computer Science, University of Manchester, Manchester, UK
[4]Cancer Research UK Cambridge Research Institute, Cambridge, UK
[5]University College London Cancer Institute, London, UK
[6]Oxford University Computing Laboratory, Oxford, UK

ABSTRACT

Human studies are the most important source of evidence for advancing our understanding of health and disease. Yet there is no standard method for investigators to query for studies that are relevant to their scientific hypothesis. Querying data and meta-data across clinical trials and observational studies is difficult because of the lack of semantic and terminology standards for describing the design and methods of human studies, and because of the variety of clinical terminology standards used. The Ontology of Clinical Research (OCRe) is a formal ontology for describing human studies that provides methods for binding to external information standards (e.g. BRIDG) and clinical terminologies (e.g. SNOMED CT). It allows the indexing of research studies across multiple study designs, interventions/exposures, outcomes, and health conditions. With such indexing, investigators interested in the evidence pertaining to a particular question (e.g., what is the effect of A on B in people with C) will be able to locate relevant research studies more easily across disparate data sources. The ontology was developed using Protégé 4 beta.

INTRODUCTION

Human studies — encompassing both interventional and observational studies — are the most important source of evidence for advancing health science. These studies are expensive, logistically complex, and labor intensive to design, perform, and analyze. Important tasks for clinical and translational research include searching for studies that involve particular designs, interventions, or outcomes, and searching for study components or data about particular types of subjects. Such queries are currently difficult to execute because there is no standard terminology or information model for the design and methods of human studies and because clinical terms are not standardized across studies. To address these difficulties, we developed the Ontology of Clinical Research (OCRe), a formal ontology, developed using Protégé 4 beta, that represents the entities and relationships related to the design and analysis of human studies.

ONTOLOGY DESCRIPTION

We conceptualize a study as a real-world entity (like a person) that has associated processes (like a person's life) during which its properties and components parts evolve. At the design stage, the study authors formulate a set of documents (i.e., informational entities) that spell out the scientific hypothesis being studied, the design of the study, and planned activities of the study. At the execution phase, participants of a study carry out activities that are recorded and that result in a body of collected data. In the analysis phase, investigators transform the data and perform statistical analysis on them, resulting in publications and other artifacts (e.g., submissions to ClinicalTrials.gov).

We formalize the conceptualization of human studies as an OWL 1.1 ontology. Following Smith et al. [1], we view an ontology as a logical specification of the universals and defined classes in a specific domain and, at a minimum, of the subsumption relationships among the classes. The universals are types of entities that share some intrinsic characteristics. The domain may include informational entities, such

as the content of protocol documents and clinical statements. The portion of the ontology that specifies the structure of these informational entities constitutes a *model of information*, that is related to, but is not the same as conventional information models, such the Health Level Seven Reference Information Model (HL7 RIM)[2] and BRIDG[3], which is a domain analysis model adopted by FDA and HL7 for clinical trial applications and messages. These conventional information models are best seen as data structures that have some correspondence to entities in the ontology. An observation, recorded by a particular clinician, about a particular patient having a rash is a data structure that holds some information content: the rash that the patient is experiencing at a certain time. Given our goal of developing an expressive yet easy to apply ontology suitable for annotating human studies, it is imperative that we reuse, as much as possible, ontologies, information models, and terminologies, such as SNOMED CT and BRIDG that have already covered relevant domains in great detail. The use of ontologies and conventional information models requires careful mapping[4]. Some classes in an ontology of human studies, such as protocol and trial data, are informational entities that map directly to conventional information model classes in BRIDG or HL7 RIM. Others, such as `Organization` and `Person`, require a shift from an ontological view (e.g., seeing an instance of `Person` as denoting a real person) to an informational view (e.g., seeing the instance as an information record about a person).

OCRe is a set of modular components related by their import relationship The core modules are *clinical* (containing shared upper-level entities), *study design* (containing descriptors of study design and a categorization of studies by their design descriptors) and *research* (containing terms and relationships that characterize a study). The *study_protocol* module is an extension, based on the BRIDG model, from which we import terms that specify the temporal aggregates (e.g., epochs and arms in clinical trials) and sequencing relationships among protocol-driven activities. The *research*, *study design*, *study_protocol* and *clinical* modules are applicable to any clinical domain. They are designed to be used in conjunction with domain-specific models and terminologies of health conditions, interventions, and measurements. For the purpose of annotating the set of sample studies (the *test-trials* module) we use SNOMED CT terms that are spelled out in the *snomed_interface* module. Finally, the *bfo-mapping* module contains the connecting relationships that map OCRe entities to those of Basic Formal Ontology (BFO), an upper ontology shared by a number of biomedical ontologies[5].

OCRe focuses on entities of the design and analysis phase of studies. The entities in the ontology include `Study` and physical entities such as `Person` and `Material`. Aggregates of entities include `Population` and `Organization`. Entities and aggregates of entities can take up `Roles` such as `Study Subject` and `Healthcare Provider`. A `Study` includes the scientific hypothesis being tested, the study protocol (i.e., the study plan), investigators, subjects, data sets, attributes such as enrollment start date and end date, and may include one or more study sites. It can be characterized in terms of qualities, such as study purpose (e.g., prevention, diagnostic, treatment), study design features (e.g., prospective or retrospective), or study status (e.g., planned, enrolling, or completed).

A `Study Protocol` describes the activities planned to achieve the objectives of the study. The study protocol specifies, for example, characteristics of the subjects to be enrolled, activities to be performed, data to be collected, outcomes to be assessed, and method of data analysis. A study protocol is an informational entity that is ontologically distinct from events that occur as part of the study.

`Events` are occurrences that happen to study subjects (called `Clinical Events` in OCRe) or to studies as a whole (called `Management Events`, such as adding sites or stopping recruitment). The most important clinical events are the making of observations (including assessments, diagnosis, and adverse events) and the administration of interventions (e.g. procedures and pharmacological treatments). Additional research-related events include enrollment, treatment assignment, and sample collection.

DETAILED EXAMPLES

To illustrate the type of modeling OCRe encompasses, we describe two components of OCRe in more

detail.

(1) A study is characterized by a set of *study characteristics*, which we model as a hierarchy of study design types (Figure 1). The hierarchy constitutes a small terminology that allows the specification of necessary and sufficient definitions of study design types. Thus, for example, we can define a `Parallel_group_study` as a `Quantitative_human_study` that has the characteristics of `Investigator_assigned_intervention`, and `External_control_ group` (i.e., subjects do not serve as their own control).

(2) In OCRe, a planned outcome is characterized as either a `Subject_outcome` (outcome described at the level of a subject) or a `Study_outcome` (outcome aggregated from multiple subjects). A `Subject_outcome` specifies (1) some outcome phenomena being assessed (e.g., a clinical phenotype such as stroke), (2) the quantitative variable by which an outcome phenomenon is measured, (3) annotations on the qualities of those variables (e.g., continuous), and (4) the assessment method(s) and time(s). A study protocol has a set of `Outcome_analyses_specification` that specify the relevant outcome variables, the study groups being compared, the type of statistical analysis being performed and the statistical methods being used. Furthermore, for a given study design, we can constrain the types of statistical analysis that are appropriate. For example, we can define an outcome analysis specification that is part of an interventional study, and constrain its statistical analysis to those where the main predictor variable has the nominal data

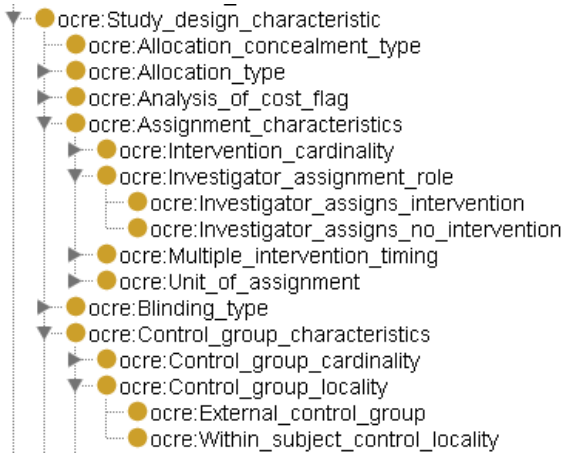

Figure 1. Part of the study design characteristics hierarchy

type, and thus further constraining the statistical methods that are appropriate.

DISCUSSION

OCRe is part of the Trial Bank Project[1] to create an interoperable federated database system of study design from all human studies, where data elements are standardized to controlled vocabularies and shared

ontologies to enable cross-study comparison and synthesis. OCRe is being deployed for this purpose via the CTSA Human Studies Database Project[2]. OCRe differs from other clinical research modeling efforts in that it focuses on providing a rich vocabulary and structures to characterize different types of human studies. Unlike BRIDG, which focuses on clinical trials, OCRe does not provide detailed data elements needed for interoperability of applications at the execution phase of trials. Furthermore, instead of being a data model, OCRe takes a more analytical approach, seeking to define the basic constituents of human

[1] http://rctbank.ucsf.edu/

[2] http://rctbank.ucsf.edu/home/hsdb.html

studies and uses them to build more complex concepts. Thus, for example, OCRe models the outcomes of a study in detail, linking them to the statistics used to analyze them.

Another feature of OCRe is that it is narrowly focused on concepts and relationships needed to characterize only studies involving individual human subjects, which is a distinct and highly valuable subset of all experimental studies. It provides terms and relationships for characterizing the essential design and structure of these studies, but relies on external vocabularies for terms related to the domain under study. The goal of OCRe in this context is to identify the common features of disparate research studies to allow their comparison over long time scales and varied experimental approaches. Domain specific scientific knowledge or techniques, which may vary over short timeframes time or between experts, is included only through bindings and the referencing of clinical thesauri and coding systems. the Ontology for Biomedical Investigations (OBI)[3], in contrast, contains detailed domain information, such as "*immortalized cell line derived from some macroscopic part of multicellular organism or organism*" that would not be part of OCRe.

OCRe has been subjected to initial formative evaluation, in which we annotated published clinical studies with OCRe terms and verified that we can query the repository of studies to select for studies that satisfy specific criteria. This has included cancer clinical trials annotated as part of the UK Medical Research Council funded CancerGrid[4] project. Within the US National Institutes of Health CTSA Human Studies Database Project, it will be evaluated and further developed as a key component of a federated multi-centre database of human studies.

ACKNOWLEDGMENTS

The work on OCRe was supported in part by R01-LM06780 and MRC-G0100852.

REFERENCES

[1] Smith B, Kusnierczyk W, Schober D, Ceusters W, editors. Toward a Reference Terminology for Ontology Research Development in the Biomedical Domain. KR-MED 2006; 2006.

[2] Health Level Seven. HL 7 Reference Information Model. http://www.hl7.org/library/data-model/RIM/modelpage_non.htm2006 [cited 2008]; Available from: http://www.hl7.org/library/data-model/RIM/modelpage_mem.htm.

[3] Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG Project: A Technical Report. J Am Med Inform Assoc2008 March-April;15(2):130-7.

[4] Rector A, Qamar R, Marley T, editors. Binding Ontologies & Coding Systems to Electronic Health Records and Messages. Proc of the Second International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2006); 2006.

[5] Grenon P, Smith B, Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli DM, editor. Ontologies in Medicine. Amsterdam: IOS Press; 2004. p. 20-38.

[3] http://obi-ontology.org/

[4] http://www.cancergrid.org