The Human Studies Database Project: Federating Human Studies Design Data Using the Ontology of Clinical Research

Ida Sim, MD, PhD[1], Simona Carini, MA[1], Samson Tu, MS[2], Rob Wynden[1], Brad H. Pollock, PhD[3], Shamim A. Mollah, MA[4], Davera Gabriel, RN[5], Herbert K. Hagler, PhD[6], Richard H. Scheuermann, PhD[6], Harold P. Lehmann, MD, PhD[7], Knut M. Wittkowski, PhD[4], Meredith Nahm, MS[8], Suzanne Bakken, RN, DNSc[9]

[1]Univ. of California San Francisco, CA; [2]Stanford University, Stanford, CA; [3]Univ. of Texas Health Science Center, San Antonio, TX; [4]The Rockefeller University, New York, NY; [5]Univ. of California Davis, CA; [6]UT Southwestern Medical Center, TX; [7]Johns Hopkins University, Baltimore, MD; [8]Duke University, Durham, NC; [9]Columbia University, New York, NY

Abstract

Human studies, encompassing interventional and observational studies, are the most important source of evidence for advancing our understanding of health, disease, and treatment options. To promote discovery, the design and results of these studies should be made machine-readable for large-scale data mining, synthesis, and re-analysis. The Human Studies Database Project aims to define and implement an informatics infrastructure for institutions to share the design of their human studies. We have developed the Ontology of Clinical Research (OCRe) to model study features such as design type, interventions, and outcomes to support scientific query and analysis. We are using OCRe as the reference semantics for federated data sharing of human studies over caGrid, and are piloting this implementation with several Clinical and Translational Science Award (CTSA) institutions.

Introduction

Human studies are one of the most central and valuable activities in biomedical research. Study designs and results should be made machine-readable to facilitate large-scale data mining and synthesis.

The Human Studies Database (HSDB) Project is a consortium of research institutions that is developing semantic and data sharing technologies to federate descriptions of human studies design over caGrid. In this paper, we describe 1) our use cases; 2) the Ontology of Clinical Research (OCRe), a rich model of human study designs; 3) our HSDBgrid data sharing architecture incorporating i2b2 and caGrid technologies; and 4) early results on using OCRe as the semantic standard for sharing over HSDBgrid.

Overview

There is a growing interest in sharing raw clinical research data to facilitate science and to promote transparency and accountability (1, 2). Because of competing regulatory and intellectual property concerns, it is unlikely that such sharing will be accomplished by aggregating all data into a single database. Instead, the most feasible data sharing approach is to "federate" queries over locally controlled databases that are standardized to a common model of clinical research.

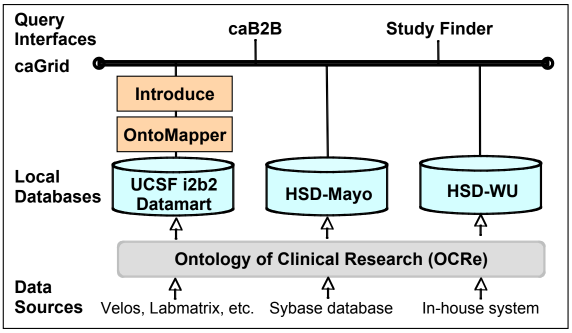

Figure 1. HSDBgrid Data Federation Architecture. HSD = Human Studies Database Service using caCORE SDK.

Figure 1 illustrates the HSDBgrid architecture for federating human studies databases. OCRe serves as the common semantic model, which defines the concepts that can be queried over the individual databases. The content and richness of HSDBgrid queries is therefore critically dependent on OCRe. For example, if OCRe does not include the concept of primary outcome, then HSDBgrid cannot support queries about primary outcomes.

It is especially important that OCRe be a rich model of human study designs, because the use and interpretation of study data depends critically on the context in which those data were collected. For example, data from a trial enrolling only patients with advanced breast cancer will not be representative of breast cancer patients in general. Similarly, a diabetes study that excludes patients with heart disease is non-representative of diabetes patients in general. Studies may even include subjects who do not have the condition of interest: e.g., a study with a non-specific case definition, or a study with healthy volunteers.

Effective sharing of clinical research data therefore requires sharing study design metadata as well as results data. OCRe is a model of human studies design and results data that can serve as a common semantic for data sharing. The HSDB Project's initial goal is to share study design information (e.g., study design type, eligibility criteria, outcomes) among CTSA and other institutions as a prelude to the more complicated task of sharing results data.

Use Cases for Sharing Human Studies Data

Shared human studies data have two broad uses: 1) for researchers, to inform the design of new studies and to aggregate and analyze existing data for new findings; and 2) for research administrators, to inform the optimization of research oversight and processes. We canvassed researchers and administrators from six CTSA institutions to describe and prioritize their needs for shared human studies data (3). The top three priority needs were 1) research characterization (e.g., of population characteristics, outcome variables); 2) registration of studies into ClinicalTrials.gov; and 3) facilitating research collaborations. We broadened research characterization to cover scientific query and analysis in general, and adopted that as our target need for HSDB. For example, Dr. A, a researcher, seeks data on the prevalence of asthma in school-aged children to inform the design of a new study. These needs would not be met by searching PubMed or ClinicalTrials.gov, because studies may have collected relevant data without that collection being mentioned. In contrast, HSDB would meet Dr. A's needs by supporting queries of key study features standardized across large numbers of human studies of varied design. But what kind of queries would Dr. A submit to HSDB? And what modeling is needed in OCRe to support those queries?

To determine whether a study's asthma prevalence data is relevant to her needs, Dr. A must first identify studies enrolling school-aged children. Next, she needs to select studies whose design types are suitable for assessing prevalence (e.g., observational cohort and cross-sectional studies). She then needs to examine how individual studies specified the phenomenon of asthma (e.g., extrinsic asthma, status asthmaticus), how the phenomenon was represented as study variables (e.g., peak flow, billing code), how and when these variables were measured, and whether any study interventions might have increased or decreased the reported prevalence of asthma. She will also want to adjust for clinically relevant covariates (e.g., air quality) and will want to know if they were measured. This use case illustrates the depth to which OCRe must model eligibility criteria, study design types, study outcomes and variables, and study exposures. This scientific depth of modeling is not present in existing clinical research models (e.g., BRIDG, CDISC SDTM) that serve primarily operational and administrative needs.

Ontology of Clinical Research

OCRe is an OWL 1.1 ontology that focuses on the design and analysis of human studies. Its scope includes human investigations of any design type (e.g., interventional, observational) for any intent (e.g., therapeutic, diagnostic, preventive) in any clinical domain on any type of data (e.g., clinical, imaging, genomics). OCRe includes 1) a representation of the structure of human studies and associated entities, 2) informational entities (e.g., study protocols), 3) terms for describing study characteristics, and 4) bindings to standard terminologies (e.g., SNOMED CT).

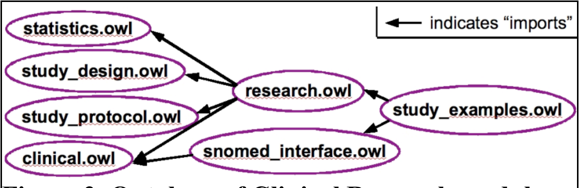

Figure 2. Ontology of Clinical Research modules

OCRe is organized as a set of modular components related by their import relationship (Figure 2). The *research* module imports the *clinical*, *study_design*, *statistics*, and *study_protocol* modules to describe a study. The *study_protocol* module imports from the BRIDG model (4) terms that specify temporal aggregates (e.g., epochs and arms) and sequencing relationships among protocol-driven activities.

OCRe modules are independent of any clinical domain because the clinical content is expressed through external ontologies and terminologies such as NCI Thesaurus or SNOMED-CT. OCRe interfaces to these terminologies by relating OCRe entities (e.g., outcome phenomenon) to these external concepts (e.g., acute myocardial infarction) and their associated terminology codes (e.g., SNOMED-CT code for acute myocardial infarction).

In the next sections, we discuss OCRe's modeling of several key domains of clinical research.

Study Design Typology

We postulated that there exist a small number of high-level study design types that represent distinct approaches to human investigations, and that we could reliably classify all human studies into these design types. Since each study type is subject to a distinct set of biases and interpretive pitfalls, a study's design type would strongly inform the interpretation and reuse of its data and biosamples.

Through iterative consultation with statisticians and epidemiologists, we defined a typology of study designs based on discriminating factors that define mutually exclusive and exhaustive study types (hybrid studies can be of more than one type). We use these factors as questions in a web-based classification tool (5). Our tool first classifies studies into human and non-human studies (Does the study use or collect measurements, assessments or observations about individual humans?). It then classifies human studies into qualitative or quantitative studies, and subsequently classifies quantitative studies into four interventional or four observational high-level design types (in red in Figures 3 and 4).

For interventional studies (Figure 3), discriminating factors include whether the investigator has a choice of interventions to which s/he can assign participants, whether the main comparison is within or across participants, and whether intervention assignment and data analysis are only within a single participant. Additional descriptors elaborate on secondary design features (e.g., randomization, blinding) that introduce or mitigate additional interpretive concerns.

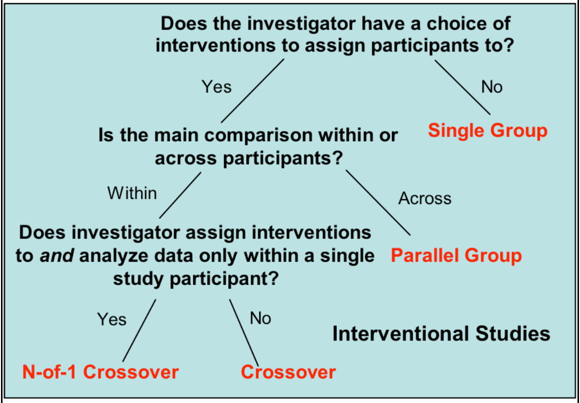
Does the investigator have a choice of interventions to assign participants to?
Yes
No
Is the main comparison within or across participants?
Single Group
Within
Across
Does investigator assign interventions to *and* analyze data only within a single study participant?
Parallel Group
Yes
No
Interventional Studies
N-of-1 Crossover
Crossover

Figure 3. Interventional studies

For observational studies (Figure 4), the four design types are based on whether the main control group is defined by case (outcome) or exposure (predictor) status, whether the case and control are in the same person, and whether outcomes are measured at the same time as predictors or after. Additional descriptors other than the ones for interventional studies apply to these observational study types (e.g., retrospective or prospective). The design typology is formalized in OCRe as an OWL hierarchy.

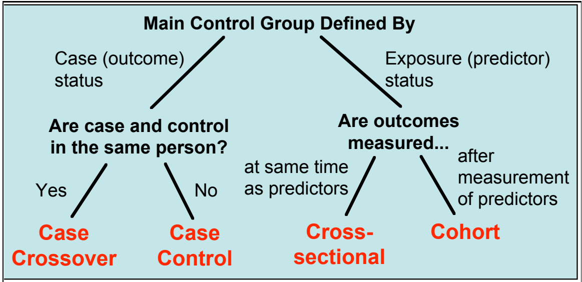
Main Control Group Defined By
Case (outcome) status
Exposure (predictor) status
Are case and control in the same person?
Are outcomes measured...
Yes
No
at same time as predictors
after measurement of predictors
Case Crossover
Case Control
Cross-sectional
Cohort

Figure 4. Observational studies

Eligibility Criteria

OCRe uses Eligibility Rule Grammar and Ontology (ERGO) Annotation (6) to capture the clinical content of eligibility criteria in machine-readable form. ERGO Annotation is a declarative representation of eligibility criteria that is informed by both the complexity of natural language and the requirements for computability. ERGO Annotation models three statement types: 1) simple statements making single assertions, 2) statements about quantitative comparisons, and 3) complex statements, which are simple and/or comparison statements joined by Boolean connectives or semantic connectors (e.g., evidenced_by).

Study Outcomes and Analyses

In OCRe, the study protocol specifies the study activities to achieve the study's scientific objectives, such as the collection and analysis of study data. Figure 5 shows our conceptualization of the entities related to outcomes and analyses in human research. We first define a study phenomenon as "a fact or event of interest susceptible to description and explanation." Study phenomena are represented by one or more specific study variables that may be derived from other variables. For example, the study phenomenon of cardiovascular morbidity may be represented as a composite variable derived from cardiovascular death, myocardial infarction (MI), and stroke variables. Each variable can be further described by its type (e.g., dichotomous), coding (e.g., death or not), timepoints of assessment (e.g., 6 months after index MI), and assessment method (e.g., death certificate). All variables are associated with participant-level and study-level observations (observations aggregated across subjects).

A study protocol may specify several analyses, each having dependent and independent variables that represent various study phenomena. Variables may play the role of dependent or independent variables in different analyses. If the study protocol designates a primary analysis, the dependent variable of that analysis represents what is conventionally known as the primary outcome of the study. To our knowledge, OCRe is the first model to disambiguate study phenomena of interest from the variables that code observations of those phenomena, and from the use of those variables in study analyses. This clarity of modeling should provide a strong ontological foundation for scientific query and analysis in HSDB.

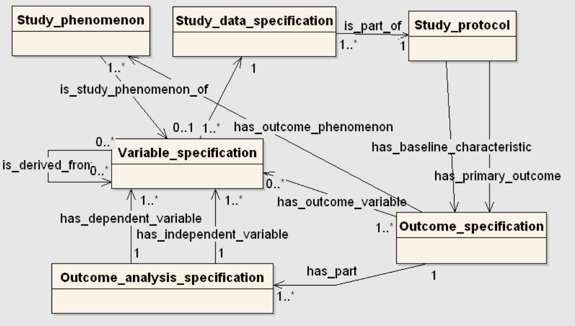

Figure 5. **Study Outcomes and Analyses**

Data Federation Architecture

Figure 1 shows the HSDBgrid data federation architecture. We will describe how we use OCRe as the semantic standard for interoperating human studies data, and how participating institutions are using i2b2 and/or caBIG technology to implement HSDBgrid federation of human studies databases.

Using OCRe as the semantic standard in HSDBgrid

As evidenced by our discussions above, OCRe is more than a subsumption hierarchy of terms. To bring the rich OWL-based modeling of OCRe into caGrid, we are uploading it to LexEVS (7), an ontology/terminology server for caGrid.

Next, we need to define and standardize the meaning of OCRe entities within the caGrid environment. This can be done by finding existing Common Data Elements (CDEs) in a caGrid data standards repository (e.g., caDSR (8) or openMDR (9)) that correspond to OCRe entities (e.g., outcome_ specification). We then annotate the OCRe entity with the corresponding CDE's unique ID, thereby defining and exposing the OCRe entity in a standard way to all caGrid services. If no corresponding existing CDE is found, we define and check in a new CDE into a data standards repository.

Implementing sharable human studies repositories

Local human studies data repositories will need a database model, e.g., a Unified Modeling Language (UML) model. The UML classes should be annotated with the unique IDs of the appropriate HSDB CDEs (e.g. outcome_specification) from either caDSR or openMDR. We have used the OntoMapper tool to perform this annotation. It appears that because the rich modeling of OCRe is available on caGrid via LexEVS, the class structure of the UML model does not need to fully replicate OCRe's semantics, but this is not yet clear. HSDB is one of the most semantically demanding data sharing projects on caGrid and our findings should serve as a template for other data sharing projects in biomedicine.

These local repositories whose database models are annotated to HSDB CDEs then need to be exposed on caGrid. Among CTSA institutions, the technology platforms used for repositories include caBIG and i2b2. HSDBgrid accommodates both platforms (Figure 1). caGrid databases that are built using the caCORE SDK are directly grid-accessible. i2b2 databases can be exposed on caGrid using the Introduce Toolkit (10). We are currently testing a virtual machine for exposing i2b2 repositories that are annotated with CDEs onto caGrid.

Results

Evaluation of OCRe

For the study design typology, we performed a pilot masked evaluation of rater agreement on active research protocols from four institutions (11). This pilot showed that an early version of our typology achieved a moderately high classification agreement (Fleiss' kappa = 0.442) across a broad range of studies, and a higher agreement (Fleiss' kappa = 0. 463) on quantitative studies only. We refined our typology based on these results and are now performing a larger scale evaluation.

In separate work, we showed how eligibility criteria can be formulated as ERGO Annotation statements that are precise description-logic expressions involving terms from standard terminologies. Moreover, for 60 free-text eligibility criteria drawn from four trials in ClinicalTrials.gov, we showed that a semi-automated natural language processing process achieved a 70% full or partial match to hand-coded ERGO Annotation statements (6). Other parts of OCRe (e.g., study outcomes) have not yet been evaluated.

Pilot Data Federation over caGrid

We piloted the OCRe-based caGrid federation of 10 data elements from 5 randomized trials from an i2b2 database at UCSF. We mapped four OCRe entities to CDEs in caDSR and created new CDEs for the other entities in openMDR. We used OntoMapper to standardize data elements from the i2b2 data model to the CDEs, and used Introduce to expose our i2b2 datamart on caGrid. We were able to issue CQL queries over caGrid to successfully retrieve data from the UCSF datamart. We have thus demonstrated the first end-to-end use of ontologies to share semantically standardized data over caGrid.

Discussion

The HSDB project is a multi-institutional collaboration that has made substantial technical progress towards integrating human studies design data to address high priority scientific query and analysis needs. Our approach uses OCRe, a semantic model of human studies design and analysis, as the common semantics for interoperating local caBIG and i2b2 human studies databases over caGrid.

There are several challenges facing this project and caGrid data sharing in general. One challenge is to more fully understand the respective roles of OWL ontologies and UML models in federated data sharing, and to use or develop appropriate caGrid and other technologies consonant with these roles. We are exploring using SPARQL views of OCRe to generate reproducible mappings between OCRe and HSDB UML models. This mechanism will be important for propagating OCRe updates to UML models.

A second challenge is data acquisition. How will disparate human studies design data be gathered and aggregated from study protocol documents, ethics board applications, and clinical research management systems (CRMSs) into HSDB repositories? Presently, the process is entirely manual, but increasing automation will be possible if ethics boards and CRMSs begin to adopt OCRe's conceptualization and definitions of human research. We also need to continue our early efforts at harmonizing OCRe with BRIDG (and thus to HL7's Clinical Trials Registry and Results Project), to Open Biomedical Ontologies, and to other data sharing and reporting initiatives (e.g., FDA Amendments Act of 2007, NCI's Clinical Trials Reporting Program, NIH's Data Sharing Initiative). A multi-institutional international human studies database will be an incomparably rich resource for clinical and translational research.

Acknowledgements

Funded in part by R01-LM-06780 and by UL1RR024156, UL1RR024128, UL1RR02500, UL1RR024143, UL1RR024146, UL1RR024131, UL1RR025767, and UL1RR024982 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. We thank Shannon Hastings and Rakesh Dhaval from OSU for their technical contributions.

References

1. Kaiser J. Making clinical data widely available. Science 2008; 322(5899):217-8.
2. Krleza-Jeric K, Chan AW, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (Part 1). BMJ 2005;330(7497):956-8.
3. Data Element Use Cases - Informatics - CTSA Wiki. [cited 2009 October 29]; Available from: https://www.ctsawiki.org/wiki/display/INF/Data+Element+Use+Cases
4. BRIDG. 2009 [cited 2009 October 10]; Available from: http://bridgmodel.org
5. HSDB Project Study Design Categorizer II. [cited 2009 October 15]; Available from: https://www.surveymonkey.com/s.aspx?sm=FeOpUXvNrlAYTrSC%2fsKQ8w%3d%3d
6. Tu S, Peleg M, Carini S, Bobak M, Rubin D, Sim I. A Practical Method for Transforming Free-Text Eligibility Criteria into Computable Criteria. In: AMIA Annual Symposium; San Francisco, CA; 2009, in press.
7. LexEVS (Version 5.0). [cited 2009 October 20]; Available from: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_Version_5.0
8. NCICB: Cancer Data Standards Registry and Repository (caDSR). [cited 2009 October 31]; Available from: http://ncicb.nci.nih.gov/infrastructure/cacore_overview/cadsr
9. Overview -- Metadata Repository -- caGrid.org. [cited 2009 October 20]; Available from: http://cagrid.org/display/MDR/Overview
10. Hastings S, Oster S, Langella S, Ervin D, Kurc T, Saltz J. Introduce: An Open Source Toolkit for Rapid Development of Strongly Typed Grid Services. J of Grid Computing 2007;5(4):407-427.
11. Carini S, Pollock BH, Lehmann HP, Bakken S, Barbour EM, Gabriel D, et al. Development and Evaluation of a Study Design Typology for Human Subjects Research. In: AMIA Annual Symposium; San Francisco, CA; 2009, in press.