Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups

Guoqian Jiang, Harold R Solbrig, Christopher G Chute

Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, Minnesota, USA

Correspondence to
Dr Guoqian Jiang, Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 First Street, SW, Rochester, MN 55905, USA; jiang.guoqian@mayo.edu

Received 30 November 2011
Accepted 18 March 2012
Published Online First
17 April 2012

ABSTRACT

Objective The objective of this study is to develop an approach to evaluate the quality of terminological annotations on the value set (ie, enumerated value domain) components of the common data elements (CDEs) in the context of clinical research using both unified medical language system (UMLS) semantic types and groups.

Materials and methods The CDEs of the National Cancer Institute (NCI) Cancer Data Standards Repository, the NCI Thesaurus (NCIt) concepts and the UMLS semantic network were integrated using a semantic web-based framework for a SPARQL-enabled evaluation. First, the set of CDE-permissible values with corresponding meanings in external controlled terminologies were isolated. The corresponding value meanings were then evaluated against their NCI- or UMLS-generated semantic network mapping to determine whether all of the meanings fell within the same semantic group.

Results Of the enumerated CDEs in the Cancer Data Standards Repository, 3093 (26.2%) had elements drawn from more than one UMLS semantic group. A random sample (n=100) of this set of elements indicated that 17% of them were likely to have been misclassified.

Discussion The use of existing semantic web tools can support a high-throughput mechanism for evaluating the quality of large CDE collections. This study demonstrates that the involvement of multiple semantic groups in an enumerated value domain of a CDE is an effective anchor to trigger an auditing point for quality evaluation activities.

Conclusion This approach produces a useful quality assurance mechanism for a clinical study CDE repository.

INTRODUCTION

With the importance of value sets gradually being recognized by the clinical research community,[1] standardization of value sets is becoming imperative, as it can enable comparison across disparate datasets and facilitate reuse of well-defined value sets to advance clinical research studies. This standardization involves two components—access and content.

Access to standardized value sets is a component of standardized terminological services. Notably, Health Level Seven International (HL7) has been developing effective and deployable solutions for value sets for many years, and is presently working on extending its standard specification Common Terminology Services 2 (CTS2) to support formal subsets and value set definition functionality.[2] The Centers for Disease Control and Prevention (CDC) has been publishing value sets through its Public Health Information Network Vocabulary Access and Distribution System (PHIN VADS).[3] In the context of terminology/ontology services, a value set is a uniquely identifiable set of discrete meanings or 'concepts' that can be resolved as a unique collection of identifiers.[4] Value sets can be derived from existing coding schemes such as SNOMED CT, either by enumerated lists or by constraints based on logical expressions (eg, all sub-codes of the code 'breast cancer').

Standardized value set content can be realized through the perspective of meta-data management and registry. One such effort is the ISO/IEC 11179 standard, which specifies a meta-data model for representing the common data elements (CDEs).[5 6] In the ISO/IEC 11179 standard, a 'data element' is defined as 'a unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes'.[6] ISO/IEC 11179 refers to the set of permissible values for a CDE as a 'value domain'. These value domains come in two 'flavors'—'described value domains' and 'enumerated value domains'. Described value domains represent free text and numeric fields, while enumerated value domains represent coded information—a finite set of discrete values, each of which represents a predefined meaning. Typical examples of a described value domain might include a field that represents a person's name or age, while the person's gender would usually be represented as an enumerated domain. Enumerated value domains are composed of one or more 'permissible values', each of which represents a valid value for the field. Each of these values, in turn, is tied to a corresponding 'value meaning', which represents the intended meaning of the permissible value in the context of the value domain. As an example, a patient gender value domain might have the permissible values '0', '1' and '9', where '0' represents the meaning 'male', '1' 'female' and '9' 'indeterminate'. Sets of value meanings are grouped into 'enumerated conceptual domains'. The solid lines in figure 1 illustrate linkages among the components of the standard.

Part of value of the ISO/IEC 11179 standard comes from the ability to align fields from different databases by connecting local value domains with common conceptual domains. A field that represents a person's age in weeks can be aligned with a field in another database that represents it in years. Similarly, a field that represents patient

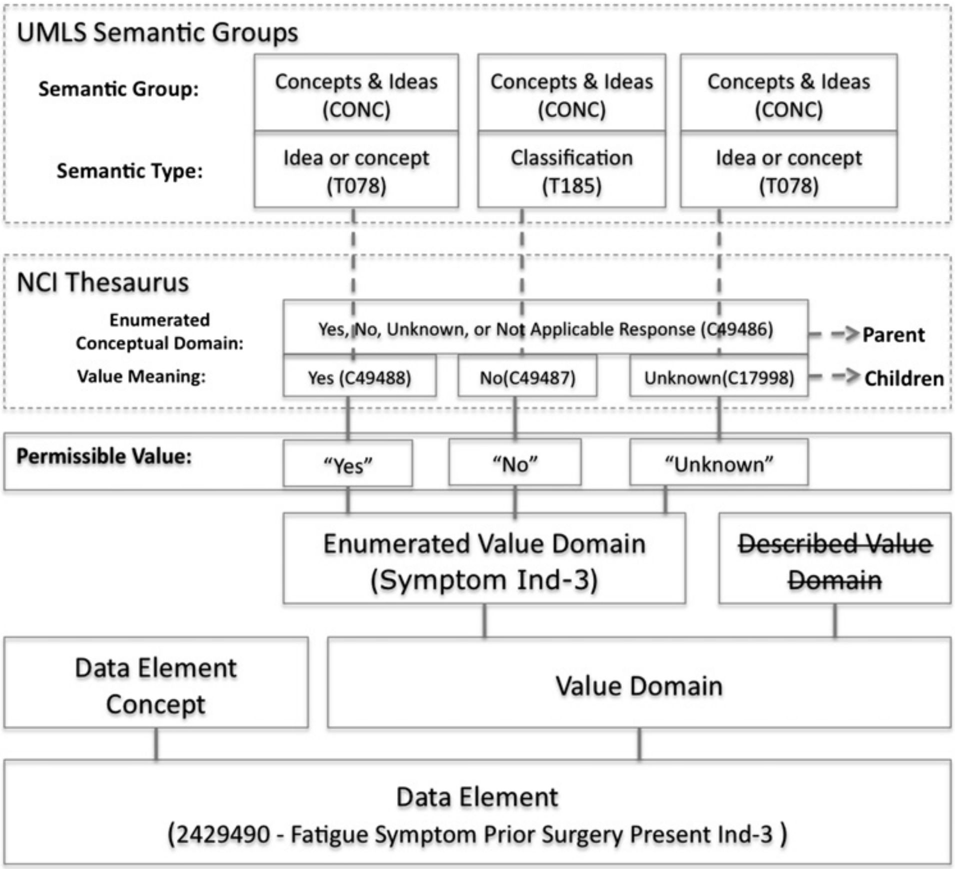


Figure 1 The linkage between the value set constructs (ie, enumerated value domain, permissible value, value meaning, and enumerated conceptual domain) of a data element and the UMLS semantic groups (with an example data element, 2429490—Fatigue Symptom Prior Surgery Present Ind-3). Note that the 'Described Value Domain' is crossed out, as it is not the focus of this paper. The solid lines between the components denote the linkage specified in the ISO/IEC 11179 standard, and the dotted lines denote the linkage between the terminological annotations and their corresponding semantic types and groups.

gender using the codes 'M' and 'F' can be aligned with another one using the values '0', '1' and '9' by connecting them to the same enumerated conceptual domain.

The National Cancer Institute (NCI) has implemented the ISO/IEC 11179 standard in the Cancer Data Standards Repository (caDSR) for cancer studies.[7] However, tools for validating the quality of the NCI registry meta-data are still somewhat limited.[8 9] In this context, profiling the terminological annotations using the unified medical language system (UMLS) semantic network (SN) has previously been demonstrated as an effective approach for quality evaluation of the standard structure of cancer study CDEs.[9] This study focused on the data element concept components—object class and property. It was clear that a principled auditing mechanism would be crucial for ensuring the quality of semantic annotations of the CDE structures in a CDE repository, thus potentially enabling data integration and interoperability across disparate clinical research datasets.

The objective of the present study is to extend our previous approach,[9] to evaluate the quality of terminological annotations on the value set (ie, enumerated value domain) components of the CDEs in the context of clinical research using both the UMLS semantic types and groups. We integrate the CDEs of the NCI caDSR, the NCI Thesaurus (NCIt) concepts and the UMLS semantic groups using a semantic web framework for a SPARQL-enabled evaluation. We demonstrate that the involvement of multiple semantic groups for asserted terminological annotations of permissible values in an enumerated value domain of a CDE is an effective anchor to trigger an auditing point for quality evaluation activities. We perform a preliminary evaluation of the usefulness of our approach.

BACKGROUND

ISO/IEC 11179 standard

'The ISO/IEC 11179 standard, formally known as the ISO/IEC 11179 Metadata Registry (MDR) standard, is an international standard for representing metadata for an organization in a metadata registry'.[5] 'The six part standard focuses on the data element as one of the foundational concepts. The purpose of the standard is to maintain a semantically precise structure of data elements'.[10] 'Each data element in an ISO/IEC 11179 metadata registry: (1) should be registered according to the Registration guidelines; (2) will be uniquely identified within the register; (3) should be named according to Naming and Identification Principles; (4) should be defined by the Formulation of Data Definitions rules; and (5) may be classified in a Classification Scheme.'[10]

National Cancer Institute caDSR CDEs

The caDSR is part of the NCI Cancer Common Ontologic Representation Environment (caCORE) infrastructure, and uses caCORE resources to support data standardization in clinical research cancer studies.[7] 'The system includes an administrator web interface for overall system and CDE management activities. In addition, a suite of specialized end user tools simplify the development, management, and deployment of ISO/IEC 11179 compliant CDEs. The CDE Browser provides public access to caDSR contents for searching, viewing and downloading CDEs in Excel or XML format. Integrated with caCORE Enterprise Vocabulary Services (EVS), the CDE Curation Tool aids developers in consumption of NCI controlled vocabulary and standard terminologies for naming and defining CDEs.'[11]

NCI Thesaurus

The NCIt is a reference biomedical ontology used by the NCI and a growing number of other systems.[12] 'NCIt covers vocabulary for clinical care, translational and basic research, and public information and administrative activities. The content is focused on cancer, but contains an increasing amount of terminology that is not specific to cancer as the number of non-cancer users and partners increases. NCIt is a concept-based terminology, with more than 70 000 concepts hierarchically organized in 19 distinct domains. It provides terminological information—definitions, synonyms, and other concept properties. The NCIt is the main source used for annotating the data elements in the caDSR; further NCIt entries reference corresponding nodes in the UMLS SN.'[9 12]

UMLS semantic network and groups

'The UMLS SN provides information about the set of basic semantic types, or categories, which may be assigned to NCIt concepts; it also defines the set of relationships that may hold between semantic types.'[9 13] 'The SN contains 133 semantic types and 54 relationships. The SN serves as an authority for the semantic types that are assigned to concepts in the Metathesaurus. The SN defines these types, both with textual descriptions and by means of the information inherent in its hierarchies. There are major groupings of semantic types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas.'[9 14 15]

Semantic web technologies

The world wide web consortium (W3C) is the main standards body for the world wide web. The goal of the W3C is to develop interoperable technologies and tools as well as specifications and guidelines to lead the web to its full potential. W3C recommendations have several maturity levels: working draft, candidate recommendation, proposed recommendation, and W3C recommendation. The resource description framework (RDF),[16] web ontology language (OWL),[17] and SPARQL[18] specifications have all achieved the level of W3C recommendations, and are becoming generally accepted and widely used. OWL is a standard ontology language for the semantic web, and is used for ontology modeling by building hierarchies of classes describing concepts in a domain and relating the classes to each other using properties. RDF is a model of directed, labeled graphs that is used to represent information in the web.[16] SPARQL is a query language for RDF graphs. SPARQL queries are expressed as constraints on graphs, and return RDF graphs or sets as results. A 'triple store' is a database for the storage and retrieval of RDF meta-data, ideally through standard SPARQL query language. Some triple stores can store billions of triples. In addition, the XML Semantics Reuse methodology is a typical semantic web technology, which focuses on moving meta-data from the XML domain to the semantic web.[19]

MATERIALS AND METHODS

We examined three resources—the caDSR CDEs, the NCIt and the UMLS semantic types and groups. Figure 1 shows the linkage between the data element, value domain, enumerated value domain, permissible values, NCIt-based concept annotations, and UMLS semantic types and groups, as well as a typical example of each element. Note that the value meaning of 'No (C49487)' is coupled with a different semantic type than 'Yes (C49488)'. Although these still fall within the same semantic group 'CONC', and would not be caught by this study, this juxtaposition is still curious and should probably be examined further.

Materials

First, we accessed the caDSR CDE browser[11] and downloaded all non-retired production CDEs (ie, the CDEs that do not have Workflow status = RETIRED) as of April 14, 2011. These CDEs were rendered in XML format.

Second, we downloaded the latest version 11.04d of NCIt from the NCI EVS Downloads website.[20] We used a copy of the inferred OWL file that contains the terminology from the NCIt that includes both stated and inferred relationships and excludes retired concepts. Note that excluding the retired concepts may affect currently generated result sets, as the terminological annotations asserted in the caDSR may be based on an old version of NCIt (see the Discussion section).

Third, we downloaded a copy of the UMLS semantic groups file that provides the mappings between 15 semantic groups and 133 semantic types.[13]

Methods

Figure 2 shows the system architecture of the semantic web infrastructure implemented for quality evaluation of CDE value set components. As illustrated in this diagram, the system comprises three layers: data integration, data access and quality evaluation.

Data integration using a semantic web framework

We converted the caDSR CDE data rendered in XML format into the RDF format using the XML2RDF web service API implemented in the ReDeFer project.[19] We used 4store, a scalable open source RDF database developed at Garlik (Nottingham, Nottinghamshire, UK),[21] as the data persistence layer, and accessed it via third-party Java client wrapper. Figure 3 shows a CDE example in the transformed RDF format. Once the CDE XML data were transformed into the RDF format, they were loaded into the RDF store by an import script.

The NCIt was already rendered in RDF/OWL, so its contents were loaded directly into the RDF store using a built-in script. We used a third-part open source script[22] to convert the UMLS SN into RDF format and loaded the contents into the RDF store.

Data access using SPARQL queries

With the RDF store loaded with the caDSR CDEs, the NCIt concepts and their semantic types, and the UMLS SN, we could

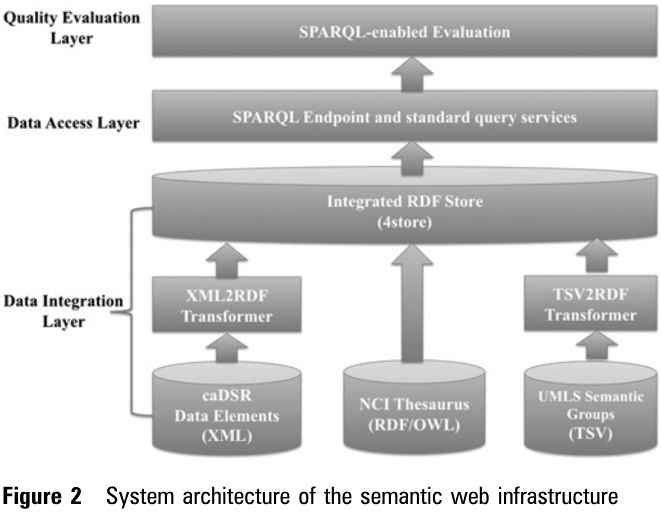
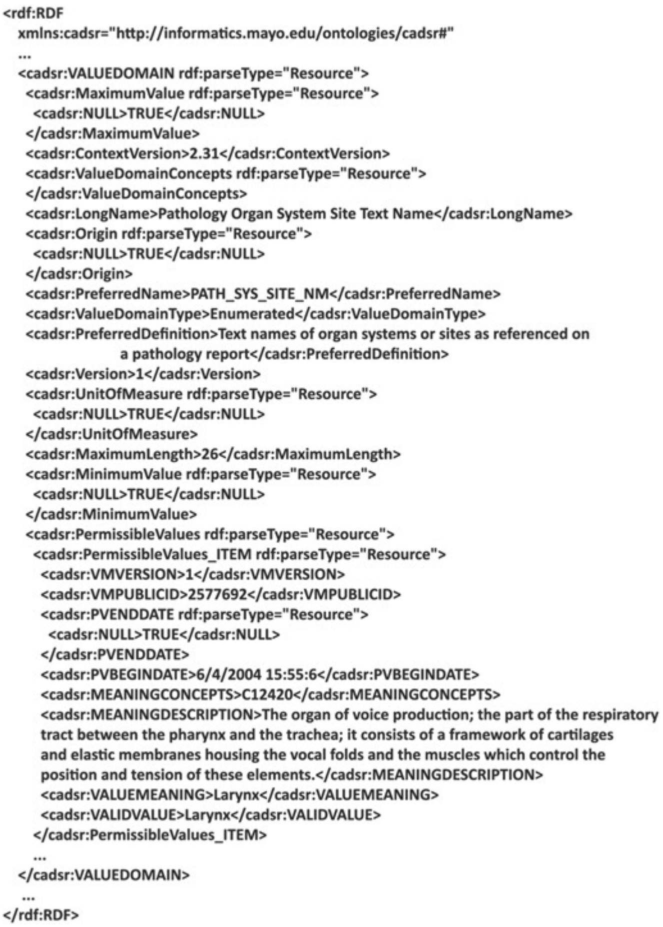

Quality Evaluation Layer
SPARQL-enabled Evaluation
Data Access Layer
SPARQL Endpoint and standard query services
Integrated RDF Store (4store)
Data Integration Layer
XML2RDF Transformer
TSV2RDF Transformer
caDSR Data Elements (XML)
NCI Thesaurus (RDF/OWL)
UMLS Semantic Groups (TSV)

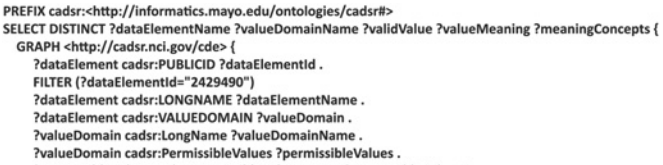
Figure 2 System architecture of the semantic web infrastructure implemented for quality evaluation of common data element (CDE) value set components. caDSR, Cancer Data Standards Repository; OWL, web ontology language; RDF, resource description framework; TSV, tab separated values.



```
<rdf:RDF
  xmlns:cadsr="http://informatics.mayo.edu/ontologies/cadsr#"
  ...
  <cadsr:VALUEDOMAIN rdf:parseType="Resource">
   <cadsr:MaximumValue rdf:parseType="Resource">
    <cadsr:NULL>TRUE</cadsr:NULL>
   </cadsr:MaximumValue>
   <cadsr:ContextVersion>2.31</cadsr:ContextVersion>
   <cadsr:ValueDomainConcepts rdf:parseType="Resource">
   </cadsr:ValueDomainConcepts>
   <cadsr:LongName>Pathology Organ System Site Text Name</cadsr:LongName>
   <cadsr:Origin rdf:parseType="Resource">
    <cadsr:NULL>TRUE</cadsr:NULL>
   </cadsr:Origin>
   <cadsr:PreferredName>PATH_SYS_SITE_NM</cadsr:PreferredName>
   <cadsr:ValueDomainType>Enumerated</cadsr:ValueDomainType>
   <cadsr:PreferredDefinition>Text names of organ systems or sites as referenced on
                 a pathology report</cadsr:PreferredDefinition>
   <cadsr:Version>1</cadsr:Version>
   <cadsr:UnitOfMeasure rdf:parseType="Resource">
    <cadsr:NULL>TRUE</cadsr:NULL>
   </cadsr:UnitOfMeasure>
   <cadsr:MaximumLength>26</cadsr:MaximumLength>
   <cadsr:MinimumValue rdf:parseType="Resource">
    <cadsr:NULL>TRUE</cadsr:NULL>
   </cadsr:MinimumValue>
   <cadsr:PermissibleValues rdf:parseType="Resource">
    <cadsr:PermissibleValues_ITEM rdf:parseType="Resource">
     <cadsr:VMVERSION>1</cadsr:VMVERSION>
     <cadsr:VMPUBLICID>2577692</cadsr:VMPUBLICID>
     <cadsr:PVENDDATE rdf:parseType="Resource">
      <cadsr:NULL>TRUE</cadsr:NULL>
     </cadsr:PVENDDATE>
     <cadsr:PVBEGINDATE>6/4/2004 15:55:6</cadsr:PVBEGINDATE>
     <cadsr:MEANINGCONCEPTS>C12420</cadsr:MEANINGCONCEPTS>
     <cadsr:MEANINGDESCRIPTION>The organ of voice production; the part of the respiratory
     tract between the pharynx and the trachea; it consists of a framework of cartilages
     and elastic membranes housing the vocal folds and the muscles which control the
     position and tension of these elements.</cadsr:MEANINGDESCRIPTION>
     <cadsr:VALUEMEANING>Larynx</cadsr:VALUEMEANING>
     <cadsr:VALIDVALUE>Larynx</cadsr:VALIDVALUE>
    </cadsr:PermissibleValues_ITEM>
    ...
  </cadsr:VALUEDOMAIN>
   ...
</rdf:RDF>
```

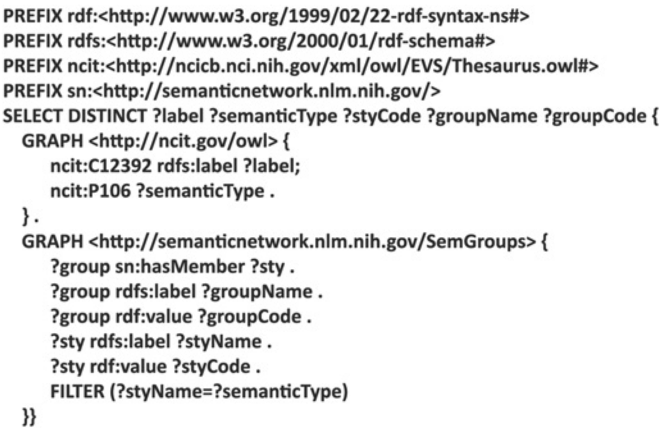
Figure 3 A portion of a common data element (CDE) example in the transformed resource description framework (RDF) format.

access all integrated data resources through standard query services of a built-in SPARQL end point. We then defined SPARQL queries that: (1) extracted the IDs and names of all the data elements; (2) extracted the IDs and names of all of the enumerated value domains; (3) extracted all the NCIt concept codes linked to the permissible values for the enumerated value domain; and (4) extracted the semantic type and group information for each of the NCIt concept codes isolated in step 3. Figure 4 shows a sample query that selects the permissible values and corresponding concept codes for a specific data element. Figure 5 shows a query example that extracts the semantic type and group information for a specific NCIt concept code.



```
PREFIX cadsr:<http://informatics.mayo.edu/ontologies/cadsr#>
SELECT DISTINCT ?dataElementName ?valueDomainName ?validValue ?valueMeaning ?meaningConcepts {
  GRAPH <http://cadsr.nci.gov/cde> {
     ?dataElement cadsr:PUBLICID ?dataElementId .
     FILTER (?dataElementId="2429490")
     ?dataElement cadsr:LONGNAME ?dataElementName .
     ?dataElement cadsr:VALUEDOMAIN ?valueDomain .
     ?valueDomain cadsr:LongName ?valueDomainName .
     ?valueDomain cadsr:PermissibleValues ?permissibleValues .
     ?permissibleValues cadsr:PermissibleValues_ITEM ?permissibleValuesItem .
     ?permissibleValuesItem cadsr:VALIDVALUE ?validValue .
     ?permissibleValuesItem cadsr:VALUEMEANING ?valueMeaning .
     ?permissibleValuesItem cadsr:MEANINGCONCEPTS ?meaningConcepts .
  }}
```

Figure 4 A SPARQL query example that extracts data element name, value domain name, valid value, value meaning, and meaning concepts (ie, terminological annotations by National Cancer Institute (NCI) concept codes) for a specific data element ID (ie, 2429490).



```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ncit:<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX sn:<http://semanticnetwork.nlm.nih.gov/>
SELECT DISTINCT ?label ?semanticType ?styCode ?groupName ?groupCode {
  GRAPH <http://ncit.gov/owl> {
     ncit:C12392 rdfs:label ?label;
     ncit:P106 ?semanticType .
  } .
  GRAPH <http://semanticnetwork.nlm.nih.gov/SemGroups> {
     ?group sn:hasMember ?sty .
     ?group rdfs:label ?groupName .
     ?group rdf:value ?groupCode .
     ?sty rdfs:label ?styName .
     ?sty rdf:value ?styCode .
     FILTER (?styName=?semanticType)
  }}
```

Figure 5 A SPARQL query example that extracts the semantic type and group information for a specific National Cancer Institute Thesaurus (NCIt) concept code (ie, C12392|liver). The code P106 denotes semantic type in the NCIt.

Quality evaluation of CDE value set components

The goal of the evaluation is to develop a mechanism using UMLS semantic types to audit the semantic consistency of the terminological annotations on enumerated CDE value set components. We started by isolating the set of CDE-permissible values that have corresponding meanings in external controlled terminologies. Next, we linked the corresponding value meanings against their NCI- or UMLS- generated SN mapping to determine whether all of the meanings fell within the same semantic group. We posited that if the meaning concepts fell into more than one group, this might be an indication of a curation problem and could be used as a trigger for an auditing process. To test this hypothesis, we randomly selected a subset of failing CDEs and manually examined their permissible values and the corresponding asserted terminological annotations.

RESULTS

We converted and loaded three data sources, comprising the caDSR CDEs, the NCIt concepts, and UMLS semantic groups, into the RDF store. For the caDSR source, there were 47 312 CDEs as of April 14, 2011. The XML2RDF transformation of the caDSR CDE source produced about 11.55 million RDF triples. For the NCIt source, 97 354 concepts were represented as approximately 2.27 million RDF triples. For the UMLS SN, 15 semantic groups were aggregated from 116 semantic types. The system produced 562 RDF triples for the data source. Of the 47 312 CDEs from the caDSR source, 16 888 (35.7%) were of the type 'enumerated', which were associated with 6534 distinct enumerated value domains comprising 153 435 distinct permissible values, of which 74 686 (48.7%) had associated value meanings. These 74 686 permissible values with asserted NCIt concept codes corresponded to 11 883 distinct NCI concept code groups. Of these 11 883 distinct groups, 8859 contained only a single NCIt concept code, whereas 3024 contained multiple NCIt concept codes.

Using SPARQL, we mapped each asserted NCIt concept code with its semantic type(s) and group(s). Of the 16 888 CDEs, 11 798 (69.9%) had at least two asserted NCI concept codes mapped to their semantic types and groups, 1587 had only one asserted NCI concept code mapped, and 3503 had no asserted NCIt concept codes mapped. Then we focused on the 11 798 CDEs and screened them using the single semantic group criteria described above. We found that 8705 (73.8%) CDEs had a unique

Table 1 Permissible values of a data element 'Lesion Measurable Evaluation Anatomic Site' from the selected data element samples for human review

Valid value	Value meaning	NCIt code (meaning concept)	NCIt concept label	Semantic type	Type code	Semantic group	Group code
Liver	Liver	C12392	Liver	Body Part, Organ, or Organ Component	T023	Anatomy	ANAT
Brain	Brain	C12439	Brain	Body Part, Organ, or Organ Component	T023	Anatomy	ANAT
Lung	Lung	C12468	Lung	Body Part, Organ, or Organ Component	T023	Anatomy	ANAT
Bone	Bone	C12366	Bone	Tissue	T024	Anatomy	ANAT
Skin	Skin	C12470	Skin	Anatomical Structure	T017	Anatomy	ANAT
Breast	Breast	C12971	Breast	Body Part, Organ, or Organ Component	T023	Anatomy	ANAT
Other	Other	C17649	Other	Qualitative Concept	T080	Concepts & Ideas	CONC

Each row represents a permissible value, its asserted NCIt concept code, the linkage to the semantic type and code, and the linkage to its corresponding semantic group name and code. Data element: Lesion Measurable Evaluation Anatomic Site (2003735). Value domain: Lesion Anatomic Site (2017124).
NCIt, National Cancer Institute Thesaurus.

UMLS semantic group for their asserted NCI concept codes, covering 1350 distinct enumerated value domains, whereas 3093 (26.2%) CDEs had multiple UMLS semantic groups mapped to their asserted NCI concept codes, covering 1848 distinct enumerated value domains.

We randomly selected 100 CDEs out of the 3093 CDEs that failed the single semantic group criteria and performed a preliminary evaluation on the permissible values of their enumerated value domains through a manual review. In these 100 CDEs, there were 1415 permissible values from 92 distinct enumerated value domains. The number of permissible values in each enumerated value domain ranged from 2 to 156. For each CDE, we extracted its public ID, long name, the public ID and long name of its value domain, and the data on each permissible value (comprising valid value and value meaning). For each asserted NCIt concept code, we extracted its label name, its semantic type and code, and its corresponding semantic group name and code. Table 1 shows the permissible values of a data element 'Lesion Measurable Evaluation Anatomic Site' from the selected data elements.

Anchored by the semantic group, the permissible values of all 100 CDEs were reviewed. All three of us have expertise and experience with data collection standards for clinical research, and two (GJ and CGC) have backgrounds in clinical medicine. We identified 17 CDEs that had inconsistency issues associated with their permissible values and corresponding asserted NCIt concept codes. Table 2 shows the summary of the evaluation results for the 17 CDEs, including the dominant semantic group of permissible values and inconsistent code examples for each CDE. As a case study, we chose two example CDEs to explain how we identified the inconsistency for their permissible values and corresponding asserted NCIt concept codes.

Table 2 Summary of evaluation results for 17 data elements including the dominant semantic group of permissible values and inconsistent code examples for each CDE

Data element ID	Data element name	Number of asserted NCIt codes	Dominant semantic group	Other semantic group	Inconsistent asserted NCIt code examples
3179024	Malignant Neoplasm Measurable Disease Evaluation Method Clinical Trial Eligibility Criteria Type	6	PROC	PHEN	C17262|X-Ray
2663176	Template (Object Class) Name Prefix Code	13	CONC	LIVB	C25174|Father
2188290	Disease Response Site Status	9	CONC	DISO	C35571|Progressive Disease
2672955	Preparative Regimen Other Finished Pharmaceutical Product Planned Administered Dose Unit of Measure Name	49	CONC	PHYS, DISO, ACTI	C25379|Course, C28245|Inhalation, C67447|Session
2673966	Glioblastoma Pathology Or Primary Neoplasm Metastatic Neoplasm Status Tumor Status Text Name	8	CONC	DISO	C14174|Metastatic
62339	Gynecologic Malignant Neoplasm Progression Anatomic Site	13	ANAT	DISO	C3331|Pleural Effusion, C2885|Ascites
2785775	Hematopoietic Stem Cell Graft Arrival Facility Shipping Environment Type	10	CONC	ANAT, DEVI, CHEM. OBJC, ACTI	Multiple codes
2783910	New Specimen Order Object Specimen Type Specimen Type Collection Text Type	64	ANAT	CONC, OBJC, ACTI, CHEM, PHYS,	Multiple codes
3109755	National Surgical Adjuvant Breast and Bowel Project Laboratory Procedure Outcome Type	4	CHEM	PROC, ANAT	C51951|Platelet Count
2963645	Bone Sarcoma Or Soft Tissue Sarcoma Disease or Disorder Primary Occurrence Anatomic Site Type	47	ANAT	CONC	C63921|Radius, C25253|Multifocal
3197150	Participant Personal Medical History Cardiac Surgery Type	4	PROC	OCCU, ANAT	C17173|Surgery
2784056	Kidney Biopsy Pathology Surgical Procedure Specimen Procedure	53	PROC	OCCU, CONC, ANAT, ACTI, OBJC	C17173|Surgery, C25436|Block
2919627	Molecular Specimen Class Specimen Class Text Type	4	ANAT	CONC, OBJC	C25574|Molecular, C25278|Fluid
3162656	Study Agent Dosage Unit of Measure Code	22	CONC	CHEM, ACTI	C25158|Capsule, C25397|Application
2695092	Laboratory Procedure IgG Or Total Cytomegalovirus Antibody Laboratory Finding Result	38	CONC	DISO, PHEN, PROC, LIVB, ACTI	C38757|Negatvie Finding, C38758|Positive Finding
3012793	First Electrocardiogram Right Bundle Branch Block Status	5	DISO	CONC, LIVB	C17734|At-Risk Population
2755993	Composition Element Type Composing Element Type	16	CHEM	OBJC, LIVB, CONC	Multiple codes

CDE, common data element; NCIt, National Cancer Institute Thesaurus.

Table 3 Permissible values of the first example from the data elements identified with inconsistent codes (highlighted in bold and italic)

Valid value	Value meaning	NCIt code (meaning concept)	NCIt concept label	Semantic type	Type code	Semantic group	Group code
Palpatation	Palpation	C16950	Palpation	Diagnostic Procedure	T060	Procedures	PROC
CT	Computed Tomography	C17204	Computed Tomography	Diagnostic Procedure	T060	Procedures	PROC
Chest x-ray	Chest Radiography	C38103	Chest Radiography	Diagnostic Procedure	T060	Procedures	PROC
Spiral CT	Spiral CT	C20645	Spiral CT	Diagnostic Procedure	T060	Procedures	PROC
Plain x-ray	***X-Ray***	**C17262**	***X-Ray***	***Natural Phenomenon or Process***	***T070***	***Phenomena***	***PHEN***
MRI	Magnetic Resonance Imaging	C16809	Magnetic Resonance Imaging	Diagnostic Procedure	T060	Procedures	PROC

Data element: Malignant Neoplasm Measurable Disease Evaluation Method Clinical Trial Eligibility Criteria Type (3179024). Value domain: Malignant Neoplasm Measurable Disease Evaluation Method Type (3179022).
NCIt, National Cancer Institute Thesaurus.

Table 3 shows the first example CDE and permissible values. Using the semantic groups linked to each permissible value, we could see that the dominant semantic group of the permissible values is PROC (Procedures), whereas there is an outlier, the semantic group of which is PHEN (Phenomena). Although the label of the asserted NCIt concept code C17262|X-Ray of the outlier matched exactly the strings of the valid value and value meaning, this is not a correct code for the case. The value domain of the CDE actually references the values related to Procedure. The valid value Plain X-Ray here corresponds to the concept C38101|Radiography, the definition of which is 'A radiographic procedure using the emission of X-Rays to form an image of the structure penetrated by the radiation'. The term 'X-Ray' has been listed as a synonym of the concept in the NCIt.

Table 4 shows the second example CDE and permissible values. Using the semantic groups linked to each permissible value, it is clear that the dominant semantic group of the permissible values is CONC (Concepts & Ideas), whereas there is an outlier, the semantic group of which is DISO (Disorders). The asserted NCIt concept code of the outlier is C14714|Metastatic, the semantic type of which in the NCIt is Finding. We argue that the valid value Metastatic in the context of this CDE is used to indicate the tumor status and should be a code with the semantic type Qualitative Concept or Functional Concept as those other asserted codes. More interestingly, we found that the corresponding NCI Meta-thesaurus code of the NCIt code C14714|Metastatic is C0036525, the semantic type of which is classified as Functional Concept. This code may be more appropriate for the annotation.

DISCUSSION

Our results demonstrate that the semantic group is a potential target for detecting inconsistent permissible values for a CDE. We hypothesized that all the permissible values of the enumerated value domain of a CDE should come from a single semantic group. A total of 8705 (73.8%) enumerated CDEs met the criteria, while the remaining 3093 (26.2%) fell into multiple groups. A manual evaluation on 100 of the 3093 CDEs identified 17 that had inconsistency issues. The majority of the remaining 83 exhibited a dominant + residual/contextual pattern, where the permissible values of a CDE had a dominant semantic group (ie, the semantic group for the majority of permissible values), in addition to one or more residual/contextual semantic group(s).

A typical example is illustrated in table 1. For the permissible values of the value domain Lesion Anatomical Site, the dominant semantic group is ANAT (Anatomy), and we can also see the other semantic group CONC (Concepts & Ideas), which is linked to a residual value, C17649|Other. From other examples, in addition to a dominant semantic group, we could see the contextual values such as C17998|Unknown, C41132|None, C48660|Not Applicable, etc, which mostly belong to the semantic group, CONC (Concepts & Ideas). The residual category issue is a variant of the categorization NEC (not elsewhere classified) issue,[23] where the code is intended to represent a member of the base class not included in the set itself (eg, other 'anatomical site', other 'procedure', etc). The contextual values represent a mixture of negation ('NO' anatomical site) and meta-data about the data element itself ('it doesn't apply', 'we weren't able to fill it out because …'). Rules could be put into place to detect these patterns that could improve the sensitivity of the inconsistency auditing, but it should be noted that they represent another potential barrier to interoperability and will need to be addressed.

Within the 17 CDEs with identified inconsistency issues, we first identified situations where permissible values were used to create a single, composite meaning. As an example, the value Hormone Therapeutic Procedure For Breast Carcinoma was associated with four NCIt concept codes: C2315|Hormone, C49236|Therapeutic Procedure, C64956|For and C4872|Breast Carcinoma. This posed a challenge to our approach, as we had to determine a primary code to represent the semantic group of the permissible value. In our manual review, we heuristically

Table 4 Permissible values of the second example from the data elements identified with inconsistent code (highlighted in bold and italic)

Valid value	Value meaning	NCIt code (meaning concept)	NCIt concept label	Semantic type	Type code	Semantic group	Group code
Residual	Residual	C37895	Residual	Qualitative Concept	T080	Concepts & Ideas	CONC
Recurrent	RECURRENT	C14173	Recurrent	Qualitative Concept	T080	Concepts & Ideas	CONC
Invasive	Invasive	C14159	Invasive	Qualitative Concept	T080	Concepts & Ideas	CONC
Primary	PRIMARY	C25251	Primary	Qualitative Concept	T080	Concepts & Ideas	CONC
Metastatic	***METASTATIC***	**C14174**	***Metastatic***	***Finding***	***T033***	***Disorders***	***DISO***
Malignant	MALIGNANT	C14143	Malignant	Qualitative Concept	T080	Concepts & Ideas	CONC
N/A	Not Applicable	C48660	Not Applicable	Qualitative Concept	T080	Concepts & Ideas	CONC
Benign	BENIGN	C14172	Benign	Qualitative Concept	T080	Concepts & Ideas	CONC

Data element: Glioblastoma Pathology Or Primary Neoplasm Metastatic Neoplasm Status Tumor Status Text Name (2673966). Value domain: Tumor Status Text Name (2231026).
NCIt, National Cancer Institute Thesaurus.

determined that, if the semantic group of one of the asserted codes were consistent with the dominant semantic group of the CDE, then the permissible value would pass the check. There are some examples in table 2 that did not pass the check and were marked as 'multiple codes' to indicate their inconsistency. For the longer run, we would advocate that a formal post-coordination model such as that described in the SNOMED CT *Technical Implementation Guide*[24] should be adopted.[25] The structure described in this model identifies the focus concept(s) of the expression, and only the semantic groups that correspond to those concepts would be used for the CDE consistency checking. That said, the generation and comparison of post-coordinated concepts is a sufficiently complex task that the added contribution to interoperability may still be limited.

The remaining inconsistencies consisted of the errors in the assignment of value meanings to permissible values, where one of the terms associated with the assigned concept is a homonym for the real target term (eg, 'X-Ray', the physical phenomenon, vs 'X-Ray', the diagnostic procedure). As mentioned above, the caDSR is part of the NCI caCORE infrastructure, and uses caCORE resources to support data standardization in clinical research cancer studies. Specifically, the CDE Curation Tool in the caDSR used the NCI EVS vocabulary services for annotating the permissible values of a CDE. The regular term search based on the vocabulary services is usually used to help assign the NCI concept codes (ie, the value meaning) to the permissible values.[26] Errors are prone to occur if curators just see the surface meaning based on term strings. Although careful reading of the textual definition (if available) of a concept code would be helpful for determining a correct concept code, an automatic alerting functionality would be more desirable. For example, once an incorrect code is assigned, the curation system will raise an alerting flag for quality checking. Note that we are aware that some of the errors may be caused by other factors. These factors include SN classification error on the part of the NCIt (eg, the semantic type of T185 on NO described in figure 1).

We also uncovered a small portion of terminological annotations by the NCIt concept codes that were no longer available, as they had been retired or merged in the latest version of NCIt.[27] From the perspective of quality control, we consider that a versioning control mechanism for the terminological annotations is an important aspect to ensure the quality of value set components of the CDEs. We also note that about 51.3% of permissible values in the caDSR still do not have any terminological annotations asserted at all. This reflects a very important aspect of challenges for the meta-data standard (such as ISO/IEC 11179)—that is, it is not an easy task to author the terminological annotations for a large set of permissible values. Without having the terminological annotations asserted, it would be hard to ensure the quality of the value set components of a CDE for its downstream use. Although the automatic approach using the natural language processing-based ontology annotation tools may partially help resolve the problem, we consider the recent advance in value set definition services (eg, one of the CTS2 services[2]) in the community of terminological services would also be useful for dealing with the challenge.

While the potential causes of the semantic inconsistencies discovered in our study varied, we would assert that *all* of them could have a negative impact on data integration and interoperability across disparate clinical research datasets. Residual context codes have already been well documented by Cimino[23] when it comes to meaning drift, but the circumstances in the CDEs render it even more problematic in that the code by itself does not indicate 'a process/disease/organism/etc not elsewhere classified'—instead providing one big category stating that 'something' was reported that we were unable to name. The juxtaposition of multiple codes in an attempt to create a post-coordinated grammar presents many problems, including isolating the actual head code and mapping between the expressions and local data. Errors due to homonyms and false similarity have the potential to result in incorrect information (eg, classifying a particular surgical procedure as a profession), confusion, and wholesale errors in the content of data itself. We should also note that the loss of information due to reorganization of ontology content has the potential to render dependent information useless.

We also observe in conclusion that the semantic web framework we used in this study has been demonstrated to be very useful for achieving our goal. Similarly, semantic web technologies have been leveraged to audit the large-scale biomedical terminologies such as NCIt and SNOMED CT for the purpose of quality assurance.[28 29] In this study, we utilized the XML Semantics Reuse technology—a XML2RDF converter, the RDF store and the standard SPARQL query services. With the capacity of the RDF store, we were able to integrate multiple, large-scale, heterogeneous CDE data and ontology resources easily in an agile manner. The underlying RDF model encoding of knowledge in the form of triples plays a key role in this, as the RDF can be used as a schema-less data representation format. This ensures the flexibility and scalability of our system and enables a high-throughput mechanism for quality evaluation. Using the powerful SPARQL query language, we were able to map the CDE-permissible values, their annotations, semantic types and groups across different graph models for the purpose of designed evaluation. In addition, we will be able to build a community-based collaborative framework for quality assurance of clinical study CDE authoring, as we have explored in our previous studies.[30 31]

Finally, there are several limitations in this study. First, the evaluation was not focused on the clinical validity of the value sets, although we believe that the original value set creation process may possibly be a contributing factor to the semantic inconsistency of the terminological annotations downstream. In addition, it would be interesting to explore in future whether there are clinical research cases where a value set for a CDE would have multiple concepts from multiple semantic groups and that might be ideal for that CDE. Second, in this study, we only used the NCI caDSR CDE repository as an example to develop a quality evaluation mechanism. We believe that the proposed approaches could be generalized to audit the CDE repository developed in individual institutes.

CONCLUSION

In conclusion, we have developed a novel SPARQL-enabled approach for quality evaluation of terminological annotations on the value set components (ie, the permissible values of the enumerated value domain) of a CDE using UMLS semantic types and groups. We believe that our approach produces a useful quality assurance mechanism for a clinical study CDE repository.

Acknowledgments The authors thank Ms Ann Ihrke, a lead copy editor/proofreader at Mayo Clinic, who provided copy editing services.

Contributors GJ and HRS conceived and designed the evaluation; GJ analyzed the data; GJ, HRS and CGC performed the evaluation; GJ and HRS wrote the paper; CGC reviewed the paper.

Funding This study is supported in part by NCI caBIG Vocabulary Knowledge Center Solicitation S07-03 (Contract Number 28XS192).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Wynden R,** Solbrig HR, Tu S, *et al*. The value of value sets. *AMIA Summits Transl Sci Proc* 2011:162—3.
2. **The CTS2 wiki.** http://informatics.mayo.edu/cts2 (accessed 23 Jul 2011).
3. **PHIN Vocabulary Access and Distribution System.** http://www.cdc.gov/phin/tools/PHINvads/index.html (accessed 22 Dec 2011).
4. **Pathak J,** Jiang G, Dwarkanath SO, *et al*. LexValueSets: an approach for context-driven value sets extraction. *AMIA Annu Symp Proc* 2008:556—60.
5. **The ISO/IEC 11179 URL.** http://metadata-standards.org/11179/ (accessed 23 Jul 2011).
6. **The ISO 11179 Term Definition.** https://wiki.nci.nih.gov/display/caDSR/ISO+11179+Term+Definitions (accessed 23 Jul 2011).
7. **Komatsoulis GA,** Warzel DB, Hartel FW, *et al*. caCORE version 3: implementation of a model driven, service-oriented architecture for semantic interoperability. *J Biomed Inform* 2008;**41**:106—23.
8. **de Coronado S,** Wright LW, Fragoso G, *et al*. The NCI Thesaurus quality assurance life cycle. *J Biomed Inform* 2009;**42**:530—9.
9. **Jiang G,** Solbrig HR, Chute CG. Quality evaluation of cancer study common data elements using the UMLS Semantic Network. *J Biomed Inform* 2011;**44**(Suppl 1):S78—85.
10. **The ISO/IEC 11179 Standard article in WikiPedia.** http://en.wikipedia.org/wiki/ISO/IEC_11179 (accessed 23 Jul 2011).
11. **The caDSR CDE Browser.** https://cdebrowser.nci.nih.gov/CDEBrowser/ (accessed 23 Jul 2011).
12. **de Coronado S,** Tuttle MS, Solbrig HR. Using the UMLS Semantic Network to validate NCI Thesaurus structure and analyze its alignment with the OBO relations ontology. *AMIA Annu Symp Proc* 2007:165—70.
13. **The UMLS Semantic Network Documentation.** http://www.nlm.nih.gov/research/umls/meta3.html (accessed 23 Jul 2011).
14. **The UMLS Semantic Groups File.** http://semanticnetwork.nlm.nih.gov/SemGroups/ (accessed 23 Jul 2011).
15. **McCray AT,** Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform* 2001;**10**:216—20.
16. **The RDF.** http://www.w3.org/RDF/ (accessed 23 Jul 2011).
17. **OWL Overview.** http://www.w3.org/TR/owl-features/ (accessed 23 Jul 2011).
18. **The SPARQL.** http://www.w3.org/TR/rdf-sparql-query/ (accessed 23 Jul 2011).
19. **The XML2RDF.** http://rhizomik.net/html/redefer/xml2rdf/ (accessed 23 Jul 2011).
20. **The NCI Thesaurus Downloads.** http://ncicb.nci.nih.gov/download/evsportal.jsp (accessed 23 Jul 2011).
21. **The 4store URL.** http://4store.org/ (accessed 23 Jul 2011).
22. **RDF extension for Google Refine.** http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/ (accessed 23 Jul 2011).
23. **Cimino JJ.** Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med* 1998;**37**:394—403.
24. The IHTSDO. SNOMED CT® Technical Implementation Guide, January 2012. Available at http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_TechnicalImplementationGuide_Current-en-US_INT_20120131.pdf (accessed 30 Mar 2012).
25. **Jiang G,** Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. *J Am Med Inform Assoc* 2009;**16**:89—102.
26. **The CDE Curation Tool Help Document.** https://cdecurate.nci.nih.gov/help/wwhelp/wwhimpl/js/html/wwhelp.htm (accessed 17 Feb 2012).
27. **NCIt Retired and Merged Concepts.** https://wiki.nci.nih.gov/download/attachments/8168350/Retired+and+Merged+Concepts.xls (accessed 23 Jul 2011).
28. **Mougin F,** Bodenreider O. Auditing the NCI thesaurus with semantic web technologies. *AMIA Annu Symp Proc* 2008:500—4.
29. **Zhang GQ,** Bodenreider O. Large-scale, Exhaustive Lattice-based Structural Auditing of SNOMED CT. *AMIA Annu Symp Proc* 2010;**2010**:922—6.
30. **Jiang G,** Solbrig HR, Iberson-Hurst D, *et al*. A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki. *AMIA Summits Transl Sci Proc* 2010;**2010**:11—15.
31. **Jiang G,** Solbrig HR, Chute CG. ADEpedia: a scalable and standardized knowledge base of adverse drug events using semantic web technology. *AMIA Annu Symp Proc* 2011;**2011**:607—16.