Harmonization of Detailed Clinical Models with Clinical Study Data Standards

G. Jiang; J. Evans; T. A. Oniki; J. F. Coyle; L. Bain; S. M. Huff; R. D. Kush; C. G. Chute

1Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA;
2Clinical Data Interchange Standards Consortium (CDISC), Austin, Texas, USA;
3Intermountain Medical Center, Intermountain Healthcare, Murray, Utah, USA

Keywords
Detailed Clinical Model (DCM), data standards, Clinical Information Modeling Initiative (CIMI), Clinical Data Interchange Standards Consortium (CDISC), clinical study, electronic health records (EHR)

Summary
Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on “Managing Interoperability and Complexity in Health Systems”.

Background: Data sharing and integration between the clinical research data management system and the electronic health record system remains a challenging issue. To approach the issue, there is an emerging interest in utilizing the Detailed Clinical Model (DCM) approach across a variety of contexts. The Intermountain Healthcare Clinical Element Models (CEMs) have been adopted by the Office of the National Coordinator awarded Strategic Health IT Advanced Research Projects for normalization (SHARPn) project for normalizing patient data from the electronic health records (EHR).

Objective: The objective of the present study is to describe our preliminary efforts toward harmonization of the SHARPn CEMs with CDISC (Clinical Data Interchange Standards Consortium) clinical study data standards.

Methods: We were focused on three generic domains: demographics, lab tests, and medications. We performed a panel review on each data element extracted from the CDISC templates and SHARPn CEMs.

Results: We have identified a set of data elements that are common to the context of both clinical study and broad secondary use of EHR data and discussed outstanding harmonization issues.

Conclusions: We consider that the outcomes would be useful for defining new requirements for the DCM modeling community and ultimately facilitating the semantic interoperability between systems for both clinical study and broad secondary use domains.

Correspondence to:
Guoqian Jiang, MD, PhD
Department of Health Sciences Research
Mayo Clinic
200 First St SW
Rochester, MN 55905
USA
E-mail: jiang.guoqian@mayo.edu

Methods Inf Med 2015; 54: 65–74
http://dx.doi.org/10.3414/ME13-02-0019
received: June 7, 2013
accepted: April 23, 2014
Epub ahead of print: November 26, 2014

1. Introduction

Data sharing and integration between clinical research data management systems (CDMS) and electronic health record (EHR) systems remains a challenging issue. To achieve interoperability, the systems must adhere to shared data interchange standards. However, considerable efforts for standardization and communication of medical information in the areas of health care and clinical research have been carried out in independent ways [1]. Realizing that collaboration can save the standards-developing community resources and improve the quality of the standards they develop, the standards-developing organizations for both health care and clinical research tend to collaborate in the production of standards [2]. Notably, the collaboration between CDISC (Clinical Data Interchange Standards Consortium) [3] and Health Level 7 [4] (HL7) provides such an example on how the standards-developing organizations can work together. In 2004, CDISC initiated the Biomedical Research Integrated Domain Group (BRIDG) in collaboration with HL7 and National Cancer Institute. The collaboration effort developed a domain analysis model that is a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artifacts [5]. The BRIDG model was designed as a step outlined in the HL7 Development Framework. In addition, a HL7 Reference Information Model-based layer has been introduced into the model starting with the BRIDG model Release 3.0.

Detailed Clinical Models (DCMs) have been regarded as the basis for retaining computable meaning when data are exchanged between heterogeneous computer systems [6]. In HL7, a DCM is defined as an information model of a discrete set of precise clinical knowledge that can be used in a variety of contexts [7]. Besides the HL7 efforts on DCMs, a number of similar initiatives have been going on independently. Goossen et al. [8] reviewed six such initiatives of DCMs, including ISO/CEN EN13606/Open-EHR Archetypes [9, 10] and Intermountain Healthcare Clinical Element Models (CEMs) [11]. The authors found that 1) there are many commonalities and differences between initiatives; 2) important differences include the use of...
or lack of a reference model and expressiveness of models; and 3) applying clinical data element standards facilitates the use of conceptual DCM models in different technical representations. To provide guidance on representation format and processes to improve the quality of modular data specifications for clinical information (i.e., DCMs), an ISO 13972 standard of New Work Item Proposal in quality criteria and methodology for DCMs was accepted by ISO member states in July 2009 [12] and are currently under development [13]. To synergize all these independent national or international efforts on standardization of DCMs, fortunately, the community has recently initiated an international collaboration effort known as the Clinical Information Modeling Initiative (CIMI) [14]. In its public statement [15], the CIMI is described as “an international collaboration that is dedicated to providing a common format for detailed specifications for the representation of health information content so that semantically interoperable information may be created and shared in health records, messages and documents”.

While these DCM modeling efforts are mainly focused on the context of improving the interoperability between EHR systems, there are emerging interests in the use of DCMs in the context of both clinical research and broad secondary use of EHR data. A notable example for the context of secondary use is the Office of the National Coordinator awarded Strategic Health IT Advanced Research Projects Area 4 (SHARPN) [16], in which the Intermountain Healthcare CEMs have been adopted by the SHARPN project for normalizing patient data for the purpose of secondary use. The secondary use of EHR sourced data is a broad domain. As defined by the SHARPN consortium, “it includes patient safety and clinical quality metrics and development programs as the most obvious, but other clinical applications range from clinical decision support to practice variation monitoring. The entire categories of clinical and translational research are fundamentally dependent on effective secondary use of clinical information, including clinical trials, observational cohorts, outcomes research, comparative effectiveness, and best evidence discovery” [16]. In the context of clinical research, for another example, CDISC has initiated the SHARE project [17]. CDISC SHARE is “a global, accessible electronic library, which enables standardized data element definitions and richer metadata to improve biomedical research and its link with healthcare” [17, 18]. CDISC SHARE also intends to build reusable domain-specific templates and this is analogous to the DCM modeling efforts as described above.

As part of the SHARPN data normalization project, we consider that it is necessary to build interoperability between SHARPN CEMs and existing data standards (e.g., CDISC standards [19] and ISO 11179 standards [20]) as these efforts would maximize the reusability of the CEMs in a variety of use cases across both clinical study and secondary use. The objective of the present study is to describe our preliminary efforts on harmonization of the SHARPN CEMs with CDISC SHARE clinical study data standards. As our starting point, we have focused on three generic domains: demographics, lab tests, and medications. We formed a CSHARE CEMs Harmonization Working Group with representatives from CDISC, Intermountain Healthcare, and Mayo Clinic. We performed a panel review on each data element extracted from the CDISC templates and SHARPN CEMs. We identified the common data elements (CDEs) across the context of clinical study and broad secondary use of EHR data and discussed the outstanding issues out of the harmonization process.

2. Background

2.1 Clinical Element Models

As we mentioned in the previous section, the SHARPN project has adopted the Intermountain Healthcare CEMs for data normalization [16]. The CEM presents a model for describing and representing detailed clinical information through defining a standard data structure. The CEM is Intermountain Healthcare’s strategy for detailed clinical models. Presently, over 4,000 CEMs, such as blood pressure measurement or specific laboratory tests, are defined [16, 21]. Figure 1 shows a diagram illustrating the data structure of the Blood Pressure Panel CEM.

The Blood Pressure Panel CEM groups two measures, systolic blood pressure and diastolic blood pressure, together. The key word "key" is a constraint that declares a terminology concept that defines the semantic meaning of the model in which it appears. CEMs are authored using computable formalisms: the Constraint Definition Language is currently being developed in the Intermountain Healthcare as the formalism for authoring and representing CEM models [22], whereas historically the Abstract Syntax Notation One [23] and Clinical Element Markup Language [6] had been used. In the SHARPN project, CEMs are authored in Constraint Definition Language and then are converted to XML schema renderings that are used in data normalization pipeline.

2.2 CDISC Standards

The mission of CDISC is “to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare” [3]. Over the past decade, CDISC has fulfilled its mission by publishing and supporting a suite of standards that enable the electronic interchange of data throughout the lifecycle of a clinical research study [18]. Figure 2 shows a diagram illustrating the CDISC

Figure 1  A diagram illustrating the data structure of BloodPressurePanel Clinical Element Model. BP indicates blood pressure.

Methods Inf Med 1/2015 © Schattauer 2015

Downloaded from www.methods-online.com on 2018-01-14 | IP: 54.70.40.11
For personal or educational use only. No other uses without permission. All rights reserved.
standards across the lifecycle of clinical research studies.

Specifically, CDISC has developed standards for use across the various points in the research study lifecycle:

- **Planning**: Protocol Representation Model Version 1, which includes study design, eligibility criteria, and clinical trial registration
- **Data collection**:
  - Clinical Data Acquisition Standards Harmonization for the collection of data through case report forms
  - Operational Data Model for the collection of operational data through electronic data exchange
  - Laboratory Model for the collection of clinical laboratory data through electronic data exchange
- **Data tabulations**:
  - Study Data Tabulation Model for submission of human subject data to regulatory agencies
  - Standard for the Exchange of Non-clinical Data for submission of non-human subject data to regulatory agencies
- **Statistical analysis**: Analysis Data Model for submission of statistical analysis data to regulatory agencies.

As adopters have realized the benefits of these standards, it has become apparent that there is a need for a foundational standard to support *computable semantic interoperability* – the predictable exchange of meaning between two or more systems – across multiple standards including, but not limited to, those developed by CDISC.

### 2.3 SHARP Area 4 Project

The focus of SHARPn is on building a scalable and standards-driven infrastructure for secondary use of EHR data. [http://sharpn.org] [16]. A collaboration of 16 academic and industry partners participate in SHARPn, including Mayo Clinic, University of Utah, Intermountain Healthcare, CDISC. The vision of the project is to develop and foster a federated informatics research community committed to open-source resources that can industrially scale to address barriers to the broad-based, facile and ethical use of EHR data for secondary purpose. SHARPn is comprised of six strongly intertwined projects: 1) Clinical Data Normalization; 2) Natural Language Processing; 3) Phenotyping Applications; 4) Performance Optimizations and Scalability; 5) Data Quality Metrics; and 6) Evaluation Framework. Notably, SHARPn has adopted CEMs, developed by Intermountain Healthcare, as the canonical representation for both syntactic and semantic normalization of clinical data.

### 3. Materials and Methods

#### 3.1 Materials

- **CDISC SHARE Domain Templates**

CDISC SHARE contributed domain-specific templates in the following three domains: 1) demographics; 2) laboratory tests; and 3) concomitant medication. The templates are provided in the format of Microsoft Excel spreadsheets. In the template for demographics, there are 13 data elements (i.e., variables). Of these data elements, 5 have value sets (i.e., code list) defined. Table 1 shows the variables and their definitions listed in the template for the demographics domain. Table 2 shows the description of the code list C66731 for the variable SEX listed in the

| Table 1 | Table 2 | Figure 2 | A diagram illustrating CDISC standards defined across the lifecycle of clinical research studies |
Table 3 shows the permissible values in the code list C66731 for the variable SEX in the template for the demographics domain. In the template for laboratory tests, there are 53 data elements. Fifteen of the 53 data elements have value sets defined. In the template for concomitant medications, there are 61 data elements and 16 of them have value sets defined.

### 3.1.2 SHARPn Clinical Element Models

The SHARPn project provided three CEM models: SecondaryUsePatient, SecondaryUseLabObs and SecondaryUseNotedDrug in XML Schema format. Three hundred seventy-one data elements (with 27 root elements) from SecondaryUsePatient Model, 478 data elements (with 22 root elements) from SecondaryUseLabObs model, and 281 data elements (with 29 root elements) from SecondaryUseNotedDrug model are extracted and represented in Excel Spreadsheet for the harmonization.

#### 3.1.2.1 Figure 3

Shows a compiled tree representation of root elements in the SecondaryUsePatient CEM. Each root element has a set of sub-elements. For example, for the root element “citizenship” in the SecondaryUsePatient CEM as shown in Figure 3, it has eight sub-elements, in which CD (ie, Concept Descriptor) and TS (ie, Time Stamp) denote data types of corresponding data elements.

**citizenship**:  
- citizenship.key
- citizenship.CD
- citizenship.startTime
- citizenship.startTime.key
- citizenship.startTime.TS
- citizenship.endTime
- citizenship.endTime.key
- citizenship.endTime.TS

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Variable Label</th>
<th>SHARE Generic Definition Variable Name</th>
<th>Codelist Master*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOMAIN</td>
<td>Domain</td>
<td>A domain is a grouping of related information</td>
<td>C66734</td>
</tr>
<tr>
<td>--DTC</td>
<td>Date/Time of Collection</td>
<td>The date and time demographic data was collected</td>
<td></td>
</tr>
<tr>
<td>COUNTRY</td>
<td>Country</td>
<td>Country of the investigational site in which the subject participated in the trial. Country is a collective generic term that refers here to a wide variety of dependencies, areas of special sovereignty, uninhabited islands, and other entities in addition to the traditional countries or independent states.</td>
<td>C66786</td>
</tr>
<tr>
<td>BRTHDTC</td>
<td>Date/Time of Birth</td>
<td>Date/time of birth</td>
<td></td>
</tr>
<tr>
<td>SEX</td>
<td>Sex</td>
<td>The assemblage of physical properties or qualities by which male is distinguished from female; the physical difference between male and female; the distinguishing peculiarity of male or female.</td>
<td>C66731</td>
</tr>
<tr>
<td>RACE</td>
<td>Race</td>
<td>An arbitrary classification of a taxonomic group that is a division of a species. It usually arises as a consequence of geographical isolation within a species and is characterized by shared heredity, physical attributes and behavior, and in the case of humans, by common history, nationality, or geographic distribution.</td>
<td>C74457</td>
</tr>
<tr>
<td>ETHNIC</td>
<td>Ethnicity</td>
<td>A social group characterized by a distinctive social and cultural tradition that is maintained from generation to generation. They share a common history and origin and a sense of identification with the group. Members of the group have similar and distinctive features in their lifestyle habits and shared experiences. They often have a common genetic heritage which may be reflected in their experience of health and disease.</td>
<td>C66790</td>
</tr>
<tr>
<td>INVID</td>
<td>Investigator Identifier</td>
<td>Unique identifier for an investigator within a study</td>
<td></td>
</tr>
<tr>
<td>INVNAM</td>
<td>Investigator Name</td>
<td>The investigator’s name</td>
<td></td>
</tr>
<tr>
<td>STUDYID</td>
<td>Study Identifier</td>
<td>A sequence of characters used to identify the study</td>
<td></td>
</tr>
<tr>
<td>USUBJID</td>
<td>Unique Subject Identifier</td>
<td>An identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product</td>
<td></td>
</tr>
<tr>
<td>SUBJID</td>
<td>Subject Identifier for the Study</td>
<td>Subject identifier for the study which must be unique within a study</td>
<td></td>
</tr>
<tr>
<td>SITEID</td>
<td>Study Site Identifier</td>
<td>Unique identifier for a site within a study</td>
<td></td>
</tr>
</tbody>
</table>

*Codes for those variables with coded values
Table 2  The description of the Codelist C66731 for the variable SEX listed in the template of CDISC SHARE demographics domain

<table>
<thead>
<tr>
<th>Code</th>
<th>CDISC Submission Value</th>
<th>CDISC Synonym(s)</th>
<th>CDISC Definition</th>
<th>NCI Preferred Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>C66731</td>
<td>SEX</td>
<td>Sex</td>
<td>The assemblage of physical properties or qualities by which male is distinguished from female; the physical difference between male and female; the distinguishing peculiarity of male or female. (NCI)</td>
<td>CDISC SDTM Sex of Individual Terminology</td>
</tr>
</tbody>
</table>

Figure 3  A compiled tree representation of root elements in the SecondaryUsePatient Clinical Element Model. The root element “Citizenship” is expanded to show its sub-elements. The CEM Browser is accessible at http://www.clinicalelement.com.
3.2 Methods

We formed a CSHARE CEMs Harmonization Working Group with representatives from CDISC, Intermountain Healthcare and Mayo Clinic in October 2011. We performed a panel review on each data element extracted from the CDISC templates and SHARPn CEMs.

As an initial plan, the harmonization work is divided into two phases. The goal for Phase I is to identify CDEs across two contexts. When a consensus is achieved, a data element is classified into one of the following three context categories: Common, CDISC-Template-Specific or SHARPn-CEM-Specific. The goal for Phase II is to harmonize the data types of the CDEs identified and to harmonize the contexts. When a consensus is achieved, for Phase I is to identify CDEs across two contexts. When a consensus is achieved, for Phase II is to harmonize the data types of the CDEs identified and to harmonize the contexts. When a consensus is achieved, for Phase III is to harmonize the data types of the CDEs identified and to harmonize the contexts. When a consensus is achieved, for Phase IV is to harmonize the data types of the CDEs identified and to harmonize the contexts. When a consensus is achieved, for Phase V is to harmonize the data types of the CDEs identified and to harmonize the contexts.

3.3 Results

In total, we reviewed 127 data elements from the CDISC SHARE templates and 1,130 data elements extracted from the SHARPn CEMs. We classified the data elements into the context categories as described in the Methods section.

4. Results

In total, we reviewed 127 data elements from the CDISC SHARE templates and 1,130 data elements extracted from the SHARPn CEMs. We classified the data elements into the context categories as described in the Methods section.

4.1 Demographics Domain

We identified four CDEs from Demographics domain, accounting for 30.8% (4 out of 13) of all data elements in CDISC template and 14.8% (4 out of 27) of all root elements from SecondaryUsePatient CEM. This result indicates that only a small portion of data elements are common across the two contexts. ▶ Table 4 shows the CDEs identified from Demographics domain.

4.2 Lab Tests Domain

We identified 20 CDEs from Lab Tests domain, accounting for 37.8% (20 out of 53) of all data elements from CDISC Lab Tests template. The CDEs cover 9 root elements from SecondaryUseLabObs, accounting for 40.9% (9 out of 22) of all root elements. ▶ Table 5 shows the CDEs identified from Lab Tests domain.

4.3 Medications Domain

We identified 15 CDEs from the Medications domain, accounting for 24.6% (15 out of 61) of all data elements from CDISC Medications template. The CDEs cover 13 root elements from SecondaryUseNotedDrug CEM, accounting for 44.8% (13 out of 29) of all root elements. ▶ Table 6 shows the CDEs identified from Medications domain.

4.4 Outstanding Issues

We also identified a number of outstanding issues during the harmonization. The outstanding issues include 1) differences in implementation; 2) common potential; 3) data type harmonization issues; and 4) value set harmonization issue. We will discuss the issues in detail in the following section.

5. Discussion

5.1 Significance of the Study

It is well recognized that there is proliferation of standards rather than convergence so harmonization efforts become a crucial for alignment of the standards [24]. The DCM modeling community faces exactly the same issue [8]. However, harmonization has inherent challenges when more standards are to be aligned – the mappings just become more complex at the exponential scale as the number of standards in-
increases. Therefore, initiatives like CIMI may be our only hope for long-term global semantic interoperability. The common formalisms used to model DCMs are actively being developed in the CIMI community and the public wiki of the CIMI community provides rich information about the CIMI activities and roadmap including the milestones of the CIMI [14].

Secondary use of EHR data, in general, refers to uses of identifiable data in an EHR system outside of patient care delivery and beyond the purpose for which data were obtained. This includes clinical research among other domains. By identifying the CDEs, we actually specified the mappings between the data elements of the two standard specifications (ie, CSHARE templates vs. SHARPn CEMs). We show that the mappings produced in this study can be used to facilitate the meaningful use of EHR data for clinical studies. A typical scenario is described as follows. Using a form generator in a CDMS, a clinical researcher generates a case report form for data collection. The form generator utilizes the CDISC SHARE domain templates. After the clinical researcher selects the data elements from the domains Demographics, Lab Tests and Medications, a case report form is automatically generated. The CDMS is connected with a centralized EHR data repository in which all de-identified patient data have been normalized by utilizing a CEM-based normalization pipeline. When an eligible patient is identified, the CDMS prepopulates the case report form with the patient data out of the repository. This functionality utilizes the mappings between CDISC SHARE data elements and CEM data elements.

### 5.2 Outstanding Harmonization Issues

First, we found that a number of data elements are semantically common but the implementation of their definition is different between the two worlds. For example, a data element "Dose Form (--)DOSFRM)" from CDISC SHARE is defined as "the dosage form for the treatment administered". Its corresponding equivalent data element in the SecondaryUseNotedDrug CEM is identified as "formulation.data.code". The root element "formulation" itself is a data structure that not only captures the dosage form data through the data element "formulation.data.code," but also captures the usual route by the data element "formulation.usualRoute." We consider that the differences in the implementation actually reflect the importance of the mappings produced by the harmonization work when the patient data normalized using the CEM models are considered for clinical study use. Note that the approach used in defining the data structure in a CEM model is similar to the dissection approach that is a common practice used in the terminology space for development of re-usable terminologies. The dissection approach was originally used by the GALEN and GALEN-IN-USE projects [25, 26]. This kind of post-coordination-based representation would potentially provide the linkage between the data elements defined.

### Table 5  Common data elements identified from lab tests domain

<table>
<thead>
<tr>
<th>SHARE Variable Name</th>
<th>SHARE Variable Label</th>
<th>CEM Elements</th>
<th>CEM Datatype</th>
</tr>
</thead>
<tbody>
<tr>
<td>--ENDTC</td>
<td>End Date/Time of Observation</td>
<td>specimenCollected.endTime</td>
<td>TS</td>
</tr>
<tr>
<td>--LOINC</td>
<td>LOINC Code</td>
<td>key.code</td>
<td>CD</td>
</tr>
<tr>
<td>--METHOD</td>
<td>Method</td>
<td>method</td>
<td>CD</td>
</tr>
<tr>
<td>--NRIND</td>
<td>Normal/Reference Range Indicator</td>
<td>labInterpretation</td>
<td>CD</td>
</tr>
<tr>
<td>--ORNHI</td>
<td>Normal Range Upper Limit-Original Units</td>
<td>referenceRangeXXX.upperBoundXXX.data.XXX.unitOriginalText</td>
<td>string</td>
</tr>
<tr>
<td>--ORNRL 0</td>
<td>Normal Range Lower Limit-Original Units</td>
<td>referenceRangeXXX.lowerBoundXXX.data.XXX.unitOriginalText</td>
<td>string</td>
</tr>
<tr>
<td>--ORRES</td>
<td>Result or Finding in Original Units</td>
<td>PQ.value</td>
<td>double</td>
</tr>
<tr>
<td>--ORRESU</td>
<td>Original Units</td>
<td>PQ.unitOriginalText</td>
<td>string</td>
</tr>
<tr>
<td>--STDTC</td>
<td>Start Date/Time of Observation</td>
<td>specimenCollected.startTime</td>
<td>TS</td>
</tr>
<tr>
<td>--STNRRC</td>
<td>Normal Range for Character Results</td>
<td>referenceRangeNar</td>
<td>ST</td>
</tr>
<tr>
<td>--STNRHI</td>
<td>Normal Range Upper Limit-Standard Units</td>
<td>referenceRangeXXX.upperBoundXXX.data.XXX.value</td>
<td>string</td>
</tr>
<tr>
<td>--STNRL 0</td>
<td>Normal Range Lower Limit-Standard Units</td>
<td>referenceRangeXXX.lowerBoundXXX.data.XXX.value</td>
<td>string</td>
</tr>
<tr>
<td>--STRESCE</td>
<td>Character Result/Finding in Std Format</td>
<td>data</td>
<td>CD</td>
</tr>
<tr>
<td>--STRESN</td>
<td>Numeric Result/Finding in Standard Units</td>
<td>data</td>
<td>CD</td>
</tr>
<tr>
<td>--STRESU</td>
<td>Standard Units</td>
<td>data.unit.normal</td>
<td>string</td>
</tr>
<tr>
<td>--TEST</td>
<td>Test Name</td>
<td>key.OriginalText</td>
<td>string</td>
</tr>
<tr>
<td>--TESTCD</td>
<td>Test Short Name</td>
<td>key.code</td>
<td>CD</td>
</tr>
<tr>
<td>COVAL</td>
<td>Comments</td>
<td>comment</td>
<td>ST</td>
</tr>
<tr>
<td>LBNAM</td>
<td>LBNAM</td>
<td>performingLaboratory.laboratoryId</td>
<td>ST</td>
</tr>
<tr>
<td>LBNAM</td>
<td>LBNAM</td>
<td>performingLaboratory.laboratoryName</td>
<td>ST</td>
</tr>
</tbody>
</table>
in the CEM models and the terminological codes defined in controlled vocabularies.

Second, some data elements are identified as “Common Potential” because these data elements are specified in CDISC templates and do not have corresponding data elements in the CEMs for secondary use while we consider that they have potential to be used in EHR use case. For example, a data element “Reason for Dose Adjustment (---ADJ)” in CDISC SHARE template is defined as “Reason why the treatment dose was adjusted.” There is no corresponding data element in the SecondaryUseNoteDrug CEM for secondary use, but we consider that this data element may be applicable in EHR use case and should be reviewed further in future. Note that the rest of the data elements are left as either CDISC-Template-Specific or SHARPn-CEM-Specific. These model-specific elements are semantically unrelated and generally use-case specific. For example, the data elements INVID, INVNAV, STUDYID, USUBID, SUBID, SITEID in the template of CDISC SHARE Demographics domain (Table 1) are designed for recording the information about investigators, study subjects and sites in the context of clinical study, which apparently are not the focus of the SHARPn CEMs.

Third, data types are one of the major constraints commonly used to standardize the data elements in meta-data management. Though we are in the process of harmonizing the data types for the data elements from the two worlds, we have realized the CEMs use HL7 data types with a few of extensions whereas CDISC templates use basic data types (e.g., Char, Num, etc.) with the mappings to ISO 21090 data types [27]. As the HL7 extensively uses ISO 21090 data types, we believe it would not be very hard to harmonize the data types between the two worlds.

Fourth, value sets are also a major constraint for meta-data standardization [28]. CDISC defines standard terminology used to represent codelists (i.e., value set) for those data elements with coded data types. For example, Table 2 and Table 3 illustrate such a value set (i.e., the codelist C66731 for the variable SEX) and its permissible values. The CEMs provide an external value set reference ID while a set of extensions whereas CDISC templates use basic data types (e.g., Char, Num, etc.) with the mappings to ISO 21090 data types. The CEMs have realized the CEMs use HL7 data types with a few of extensions whereas CDISC templates use basic data types (e.g., Char, Num, etc.) with the mappings to ISO 21090 data types.

Table 6 Common data elements identified from medications domain

<table>
<thead>
<tr>
<th>SHARE Variable Name</th>
<th>SHARE Variable Label</th>
<th>CEM Element</th>
<th>CEM Datatype</th>
</tr>
</thead>
<tbody>
<tr>
<td>--DECOD</td>
<td>Standardized Treatment Name</td>
<td>data.value = CD/value</td>
<td>string</td>
</tr>
<tr>
<td>--DOSE</td>
<td>Dose</td>
<td>minDosePerAdministration.data.value</td>
<td>string</td>
</tr>
<tr>
<td>--DOSFRM</td>
<td>Dose Form</td>
<td>formulation.data.code</td>
<td>CD</td>
</tr>
<tr>
<td>--DOSFRQ</td>
<td>Dosing Frequency per Interval</td>
<td>codedFrequency</td>
<td>CD</td>
</tr>
<tr>
<td>--DOSTOT</td>
<td>Total Daily Dose</td>
<td>avgDailyDose.data.value</td>
<td>string</td>
</tr>
<tr>
<td>--DOSTXT</td>
<td>Dose Description</td>
<td>takenDoseLowerLimit</td>
<td>PQ</td>
</tr>
<tr>
<td>--DOSTXT</td>
<td>Dose Description</td>
<td>takenDoseUpperLimit</td>
<td>PQ</td>
</tr>
<tr>
<td>--DOSTXT</td>
<td>Dose Description</td>
<td>maxDosePerAdministration</td>
<td>PQ</td>
</tr>
<tr>
<td>--DOSU</td>
<td>Route of Administration</td>
<td>routeMethodDevice</td>
<td>CD</td>
</tr>
<tr>
<td>--DUR</td>
<td>Duration</td>
<td>duration</td>
<td>PQ</td>
</tr>
<tr>
<td>--ENDTC</td>
<td>End Date/Time of Observation</td>
<td>endTime</td>
<td>TS</td>
</tr>
<tr>
<td>--INGRD</td>
<td>Substance Active Ingredient</td>
<td>activeIngredient.data.code</td>
<td>CD</td>
</tr>
<tr>
<td>--TRT</td>
<td>Name of Treatment</td>
<td>data.originalText = CD/originalText</td>
<td>string</td>
</tr>
<tr>
<td>--STIM</td>
<td>Start Date/Time of Observation</td>
<td>startTime</td>
<td>TS</td>
</tr>
<tr>
<td>--DOSU</td>
<td>Dose Units</td>
<td>minDosePerAdministration.data.unit</td>
<td>CD</td>
</tr>
</tbody>
</table>

5.3 Implications to a Collaborative Harmonization Platform

In this study, we started from three generic domains and formed a Harmonization Working Group to discuss the data elements one by one in a panel review. To support harmonization, all data elements and their related meta-data are prepared in the format of Excel spreadsheets, which worked well in a small-scale project as we did in this study. However, we believe a collaborative harmonization platform should be developed to enable more scalable and efficient collaboration. In a previous study, we developed a CDISC SHARE prototype for supporting collaborative harmonization of clinical study data elements leveraging Semantic Wiki technology [18]. To meet the requirements for the harmonization between clinical study data standards and CEMs, we suggest an integrative semantic repository could be very useful for lowering the barriers to the access of both data standards and CEMs. We are in the process developing such a repository prototype using Semantic Web technology [30]. We are also working on soliciting the harmonization requirements and workflow process from the community, aiming at building a collaborative harmonization environment.

5.4 Implications to CEM Modeling

In this study, we have found CDEs identified from three target domains only cover a portion (less than 50%) of all data elements for each of two contexts: clinical study and secondary use. In addition, as one of outstanding issues as discussed above, a number of data elements have very different implementation though they are semantically common. Arguably, the CDEs identified from the study are the core of a CEM model that is reusable across different use cases. This actually poses a challenge to the CEM modeling if we would like a CEM model to be reused across a variety of contexts. In current CEM model-
ing approach, each CEM model starts from a core model where a comprehensive list of attributes (e.g., qualifiers, modifiers, attributes) are specified. When a submodel is defined by inheriting the core model for a specific use case, those attributes that are not applicable to the use case will be constrained out. To support clinical study use case, for example, those data elements in each specific domain that are classified into the context category “Clinical Study” will need to be added into a core model in each corresponding domain. The submodels produced for clinical study use case will be able to inherit these attributes in addition to core attributes while the models will be able to constrain out those attributes marked for other use cases. Meanwhile, those newly added attributes in a core model should be constrained out for their submodels defined for other use cases.

For the CEM modeling perspective, there are two implications in terms of modeling requirements. The first implication is that there should be a mechanism to record the context category information for the attributes in a core model. The second implication is that there should be a CEM versioning mechanism that can deal with the situation when new attributes are added in a core model. Note that ISO 13972 specification has already had sections with normative statements about the versioning control mechanism of DCMs [13], which is a good reference for the clinical modeling community. We consider that the two implications would be a useful contribution to the ongoing DCM modeling discussion in the context of CIMI community.

4. Conclusions

In conclusion, we have identified a set of data elements that are common to both CDISC templates and SHARPn CEMs while we identified a number of outstanding issues from the harmonization process. We consider that the outcomes produced by this working group would be useful for defining new requirements for the CEM modeling community, and ultimately facilitating the semantic interoperability between the systems used for both clinical study and broad secondary use domains as well. As mentioned in previous sections, we are at the Phase II harmonization process, aiming at data types and value sets harmonization. We will continue to discuss and analyze the outstanding issues identified in this study. We plan to expand our harmonization efforts to more other domains such as Diseases/Disorders, Diagnoses, Procedures and Pharmacogenomics, etc. We will continue to solicit requirements on building a collaborative platform for supporting efficient harmonization.

Note that as of April 2014, CIMI has released its reference model specification version 1.0.0 that consists of a core reference model, a data value type model and a party model [31]. In a recent study, we developed and evaluated a Semantic Web-based approach that enables a domain template generation mechanism for supporting clinical study meta-data standards development [32]. We also developed a RESTful web service for access to the generated domain templates in a Clinical Information Modeling Initiative (CIMI)-compliant format. We will continue to explore to author the CDISC clinical study data models using the formalisms developed in CIMI community.

Acknowledgments

This study is supported in part by the SHARP Area 4: Secondary Use of EHR Data (90TR000201). This paper is expanded from a 300-word abstract that was submitted and presented at 2012 SHARPn Submit “Secondary Use” June 11–12, 2012.

References