의학박사 학위논문

Ensuring semantic interloperability in the course of clinical document exchange using metadata registry related technologies

메타데이터 저장소 국제표준을 활용한 임상문서 정보교류의 의미론적 상호운용성 확보

2016 년 8 월

서울대학교 대학원

의과학과 의과학전공

김 혜 현

A thesis of the Degree of Doctor of Philosophy

메타데이터 저장소 국제표준을 활용한 임상문서 정보교류의 의미론적 상호운용성 확보

Ensuring semantic interoperability in the course of clinical document exchange using metadata registry related technologies

August 2016

The Department of Biomedical Science,
Seoul National University
College of Medicine
Hye Hyeon Kim

Ensuring semantic interoperability in the course of clinical document exchange using metadata registry related technologies

by
Hye Hyeon Kim

A thesis submitted to the Department of Biomedical Science in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Biomedical Science at Seoul National University College of Medicine

August 2016

Approved by Thesis Committee:

Professor _____Chairman
Professor _____Vice chairman
Professor _____
Professor _____
Professor _____

ABSTRACT

Introduction: Data standardization is crucial to facilitate understanding and sharing data across diverse translational studies. Common data elements (CDEs) based on the ISO/IEC 11179 Metadata Registry (MDR) standard provide well-defined and structured data that are feasible to incorporate in clinical documentation in such a way that supports semantic interoperability. However, structural limitations of MDR have been an obstacle for either composing CDEs in clinical forms or interpreting them from clinical forms. Though we developed simple extended relationships, we found it only covered simple relationships. The additional semantic relationships are needed.

Meanwhile, a clinical document is an essential tool to collect clinical information related to individual health. For comprehensive semantic representation and clear definition of clinical data, ISO/IEC 11179 standard based metadata, including CDEs has been used to compose clinical documents. When the decision is made to share clinical data through clinical documents, the data should be checked first for completeness and to ensure that no errors were introduced during the sharing process. The process of data validation significantly adds to the complexity of the data sharing process, but is critical to maintain the integrity of clinical data and to ensure high-quality data. Finding proper data elements from numerous data elements is essential to use them in metadata implemented clinical documents for effective semantic data exchange. It is required to develop an ontology to classify and search data elements.

Methods: We reviewed the CDEs currently being use in a clinical setting to understand the inter-related data elements. We then developed use case scenarios to describe common representational challenges in the inter-related CDEs, and extended the existing composite CDEs to address the identified challenges in data presentation, and data transformation. For developing validation process, we first defined what complex clinical document is as applying the developed several semantic relationships of data elements. As

considering these semantic relationships of data elements in clinical documents, we developed the process of syntactic and semantic validation of clinical documents, and specified the list of validation attributes.

Meanwhile, for developing Clinical Metadata Ontology (CMO), we adopted the General Formal Ontology method with a manual iterative process comprising five steps; (1) defining the scope of each ontology, (2) identifying concepts, (3) assigning hierarchical relationships among concepts, (4) development of properties (e.g., synonyms, preferred term, and definitions) for each concept, and (5) evaluating developed ontologies.

Results: We developed three types of extension to composite data element such as *Repeated composite data element* to resolve observational clinical data presentation challenges, and *Dictionary* and *Template composite data elements* to support knowledge data presentation and data model transformation respectively. In doing so, we defined four new constraints of CDEs in composite data element such as *Dependent, Operated, Ordered,* and *Required*. We also defined new types of the CDE such as *Hybrid relationship*. Base on the extension of semantic relationships of CDEs, we also developed the process of syntactic and semantic validation of clinical documents, and specified the list of validation attributes. We demonstrated and evaluated the feasibility of composite relationships as presenting a practical use case.

Tree structure based CMO was developed with 200 concepts under the four first-level terms including Description, Event, Finding and Procedure. CMO has 1060 synonyms for 151 (76%) CMO concepts, and 400 definitions for 137 (69%) CMO concepts. The Web-based CMO Browser and the CMO matched BMeSH DE Browser provide convenient access to CMO and help to understand how CMO concepts are matched to DEs in the practical clinical documents (http://www.snubi.org/software/cmo). CMO is the ontology as (1) a classification scheme for data elements for clinical documents, (2) an integration tool for data elements from a diversity of clinical documents, (3) a proper clinical data-organization scheme for data elements for developing clinical information systems including PHRs, and (4) a component ontology expendably connected to other healthcare data domains such as personal

lifelog data, which is supported by MELLO, and personal genomic data, which is supported by Health Avatar Project.

Conclusions: This paper investigated the feasibility of representing the complex clinical data in clinical forms with the extended MDR based extended semantic composite relationships and constraints. Our results indicates that our approach is able to comprehensively represent the CDEs in two perspectives; 1) the form-level data is represented by data item-level data without loss of the contextual semantic relationships between data elements and forms, 2) data integration and transformation across different standardized dictionaries or data models. The preliminary results of the present research can be used as a reference for future development of extended semantic composite relationships-applied system. It also emphasizes the extended MDR based value validation can help data error handling and provide clear error limits on data sharing. CMO is the ontology for classification of data elements. We can expect to search appropriate data elements and use them effectively in clinical documents.

Keywords: Common data element, Clinical docment exchange, Data interoperability, Metadata Registry

Student number: 2010-30607

CONTENTS

LIST OF TABLES AND FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF ABBREVIATION

AMC: Academic Medical Centers

A.P: Arterial Pressure

aCDE: Atomic Common Data Element

BMeSH: Biomedical Metadata Standard for Health

BMI: Body Mass Index

caBIG: Cancer Biomedical Informatics Grid

caDSR: Cancer Data Standards Registry and Repository

cCDE: Composite Common Data Element

CDISC: Clinical Data Interchange Standards Consortium

CDE: Common Data Element

CMO: Clinical Metadata Ontology

CRF: Case Report Form

CS: Classification Scheme

DBF: Diastolic Blood Pressure

eMERGE: electronic MEdical Records and GEnomics

EMR: Electronic Medical Records

IFCC: International Federation of Clinical Chemistry

MDR: Metadata Registry

NCI: National Cancer Institute

NGSP: National Glycohemoglobin Standardization Program

NIH: National Institute of Health

NINDS: National Institute of Neurological Disorders and Stroke

ODM: Operational Data Model

PHR: Personal Health Record

RIM: Reference Information Model

SBP: Systolic Blood Pressure

SNUH: Seoul National University Hospital

VDRL: venereal disease research laboratory

V.P: Venous Pressure

UMLS: Unified Medical Language System

GENERAL INTRODUCTION

The wider adoption of informatics systems and the rapidly development of new research area have resulted in the exponential growth of biological and clinical data and high requirement of exchanging those data among multiple institutions or patients and medical institutions. Several large scale of collaborative researches have been underway. We can expect that effective and seamless clinical data exchange between medical institutions can improve the quality of clinical care and reduce medical costs.

The use of data standard is a critical requirement for such harmonization. There are several efforts trying to address the standard data collection. One major approach is to build a common data model; it can be classified as a top-down approach where a top-level knowledge model agreement is forced for the underlying data models of the interoperating parties for successful data exchange. The research behind HL7 standards, OpenEHR, CDISC Standards, Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and i2b2 are among some of efforts that adopt this top-down strategy. This approach provides high comprehensiveness through well-structured model and specific description of each items. However, in contrast, it has low practicality as that it takes a long time to revise the model and reflect the recent advances.

Another major approach is to use controlled terminology to represent precise meaning of concepts; it can be classified as bottom-up approach. However,

each controlled terminology is only focused on each concept and concept relationships and is limited to represent superordinate concepts or complexed concepts that can be expressed as a combination of various concepts. For instance, for representing one diagnosis, 'Suspected heart failure caused by ischemic heart disease', it can be annotated by a SNOMED-CT single code 'Heart failure caused by ischemic heart disease' or by combination of concepts code '84114007 | heart failure |: 408729009 | finding context |: 415684004 | suspected |: 42752001 | due to |: 414545008 | ischaemic heart disease |'. Either way, it's fine to represent meaning of the phrase. However, it can give more confusion as high heterogeneity without standardized post-coordination syntax.

Metadata technology based on Metadata Registry (MDR) using international standard named ISO/IEC 11179 offers great implications for these limitations as providing an approach is to unify the top-down and bottom-up approaches. The ISO/IEC 11179 MDR standard describes the method of standardizing and registering data elements to make them understandable and sharable between organizations or systems. Its key concept is the data element, a unit of data for which definition, identification, representation, classification and permissible values are specified by means of a set of attributes. The data elements which reposted in MDR are designed whenever researchers need (bottom-up). These data elements are aggregated and integrated by research community (top-down). These well-defined data elements can be collected and reused as content standard.

The anticipated benefits of facilitating ISO/IEC 11179 based standardized data are multiple, including 1) effective and rapid data collection as reducing burden on investigators to facilitate their participation in clinical research, 2) enriched data sharing and data aggregation by employing common forms, and standard definitions, and 3) improved data quality by providing unified data and its descriptions. Moreover, as the National Institute of Health (NIH) encourages the use of the data elements, the data elements have been deployed in case report forms or clinical documents, and it has been proved the high effectiveness and usability of data elements.

However, structural limitations of MDR have been an issue as an obstacle for either composing data elements in clinical forms or interpreting them from clinical forms because data elements do not have a workable means to describe constraints or relationships that hold among different data elements. For instance, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) can be easily defined as two separate data elements annotated with standardized metadata conforming to ISO/IEC 11179. A constraint between the two data elements such as ‘SBP must be greater than DBP’ is needed described inside the data elements as there is no straightforward way to make data elements carry that information.

Another issue was that the method for syntactic and semantic validation for the developed clinical documents, containing ISO/IEC 11179 MDR standard based data elements, has not been established. This process is required during

exchanging or sharing clinical documents containing data elements to ensure integrity of clinical documents with metadata.

Meanwhile, we also needed a technique to classify and search data elements, which are rapidly developed from many large scale clinical studies to find and use specific data elements from numerous data elements from variable MDRs. Existed classification technique such as classification scheme structure in ISO/IEC 11179 is used limitedly or not used at all. Moreover, most data element browsers that developed by each research project remain to adopt simple keyword search engine.

In this study, I focused on developing MDR related technologies for ensuring semantic interoperability in the course of exchanging clinical document. First, we developed extended semantic relationships among data elements to represent complex clinical documents containing constraints and form-level based context information. Second, we developed the method for syntactic and semantic validation of clinical documents containing complex common data elements as identifying the scope of validation, and developing validation process in order to ensure the completeness and integrity of clinical document with metadata. Third, we also developed clinical metadata ontology, called CMO to classify numerous data elements and to search and use them in clinical documents.

CHAPTER 1

Composite common data elements: modeling composite relationships between common data elements for representing complex clinical metadata

INTRODUCTION

One of the essential works to facilitate understanding, sharing, and reusing of data across diverse medical institutions is to standardize the representation of clinical data [1]. Standardized representation of data does not only refer to using controlled biomedical vocabularies. Although using controlled biomedical vocabularies provide an effective way for supporting the conceptual level of consistency with comprehensive definition of concepts in semantic level of data, it has a limitation that it does not provide a consensual combined coding rule for the expression of the parent or a complex concept of data, which should be expressed as a combination of several concepts [2].

Common Data Elements (CDEs) can provide a means to incorporate these needs. In the perspective of metadata management and Metadata Registry (MDR), the ISO/IEC 11179 specifies a metadata model for representing the atomic CDE that is a logical data unit that consists of a data element concept having discrete set of precise clinical knowledge and a value domain including discrete data type, representation types, and unit of measures. It also includes additional information such as definitions of data, including an identifier [3-5].

The anticipated benefits of using ISO/IEC 11179 based standardized data in clinical research are 1) rapid and efficient study start-up as reducing burden on researchers by providing well-defined CDEs, 2) ease of sharing and aggregating data using standard definitions and forms, and 3) improved data quality by providing unified data and its description [6]. As the National

Institute of Health (NIH) encourages the use of the CDEs [7], CDEs have been deployed in case report forms (CRFs) and clinical documents [11-13], and it has been proved the high effectiveness and usability of CDEs.

In the view of content standard, the numerous large scale of clinical studies have been developed standardized CDEs by researchers and domain exports for providing unified data collection, and facilitating data sharing, for example, National Institute of Neurological Disorders and Stroke (NINDS) [6-10], National Cancer Institute (NCI)'s cancer Data Standards Registry and Repository (caDSR) [14], Parkinson's Disease Biomarker Program [15], and as well as a number of other clinical CDEs for a variety of different purposes [3, 16-18] have developed and used in their respective studies. Additionally, for facilitating Electronic Medical Records (EMR)-derived genomics studies, the electronic MEdical Records and GEnomics (eMERGE) study standardized representation of the phenotype data using CDEs through the mapping process [19].

However, structural limitations of MDR have been an issue as an obstacle for either composing CDEs in clinical forms or interpreting them from clinical forms [11, 20-21] because CDEs do not have a workable means to describe constraints or relationships that hold among different CDEs. For instance, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) can be easily defined as two separate CDEs annotated with standardized metadata conforming to ISO/IEC 11179. However, a constraint between the two CDEs

such as ‘SBP must be greater than DBP’ is usually described outside the CDEs as there is no straightforward way to make CDEs carry that information. To address this challenge, we identified three types of semantic relationships that represent the constraints or rules that hold among CDEs (i.e., *composite, dependent,* and *variable relationships*) in a prior study [19]. Composite relationship of CDEs mean that several CDEs may be grouped or tied together for giving more semantic meaning, for instance, as taking patient’s medical history, it is important which body system is related. We defined that these two CDEs, ‘DE: Body System for Medical History’, and ‘DE: Medical History Specify’ are in a composite relationship. Dependent relationship indicates that some CDEs are in a dependency relation, for instance one CDE may be activated or deactivated depending on the response of another CDE. Variable relationship indicates one representative CDE is developed and derived by concepts from a certain dictionary so that similar CDEs with different concepts can be covered by the representative CDE as having variable relationship. For instance, similar CDEs such as ‘normal value range of lab test Albumin’ and ‘normal value range of lab test Homocysteine’ can be covered by one variable data element, ‘normal value range of lab test X’.

However, it is our conclusion that our previous work supports relatively simple semantic relationships among CDEs and is not robust enough to cover many other specific challenges associated with CDE use in clinical forms. For example, some clinical items in clinical assessment forms are structured in a tabular form to make it easier to take values through repeated assessments

and/or observations. This value property information – i.e., *repeat* – needs to be made available to ensure that the values that belong to the same CDE are identified as such. We will describe other challenges by describing challenging scenarios in method.

We propose to develop one new semantic relationships to address the remaining challenges in representing the semantic relationships among the inter-related CDEs. We also developed constraints among atomic CDEs or sub-aCDEs in cCDEs. We demonstrated the newly defined semantic relationships using the data collection use cases from (1) National Institute of Neurological Disorders and Stroke (NINDS) [21] and (2) DialysisNet [22].

MATERIALS AND METHODS

1. Data resource: two CDE projects

NINDS CDE project is an ongoing effort that develops data standards for clinical research in neuroscience. It was initiated in 2006 to develop standardize data collection across neurological disorder related clinical studies funded by NINDS. As of today, NINDS CDE project includes 18 studies with 9,839 distinct CDEs. Those CDEs are not fully compliant with ISO/IEC 11179, as providing only simple data element and its definition. However, a part of NIND CDEs, registered in NCI caDSR and reviewed by the NCI's cancer Biomedical Informatics Grid (caBIG) project manager are conformed to ISO/IEC 11179 model fully. We found 308 (3.1%) Stroke and General CDEs of NINDS in caDSR and used those CDEs and their related CRFs to evaluate our approach.

DialysisNet is an iPad based-application developed through the Health Avatar Beans project to help clinicians manage their renal patients seamlessly and effectively. Health Avatar Beans is a project started in 2013 to establish data standards for managing and harmonizing hemodialysis data across multiple medical institutions. Health Avatar Beans aims to improve the management of chronic kidney disease and end-stage renal disease with an integrated mobile environment for data collection and documentation. DialysisNet was built upon the 122 distinct CDEs that were created based on the data collection forms used at the renal clinics of the four participating Academic Medical Centers (AMC).

2. Challenging cases

As examining CDEs from two different CDE projects, we found patterns that can be characterized as a new semantic relationship type of CDEs, and described these types with examples.

2.1 Data entries with multiple data types

A data type determines what kind of data can be stored and each data item is normally declared with one data type. However, unstructured free text based data in many clinical forms of EMRs might be allowed multiple data types of data. When the multiple data types of data are mixed, it can be a problem to harmonize data and to integrate data. For instance, the venereal disease research laboratory test is a blood test for syphilis that was developed by the eponymous lab. Normally, it has result data with numeric data type such as 0.8, but it also can be represented as string data type such as 'Negative'.

Figure 1-1 shows another example that a time related CDE 'DE50 Hemodialysis time hybrid' allows to have two data types; 1) hemodialysis observing time intervals related time data type as 'DE60 Hemodialysis time interval', 2) time stamp related enumerated string data type, such as *Finish, Start* as 'DE61 Hemodialysis time stamp'. This value property information – i.e., *hybrid* – needs to be made available to ensure that multiple data types are available in the CDE.

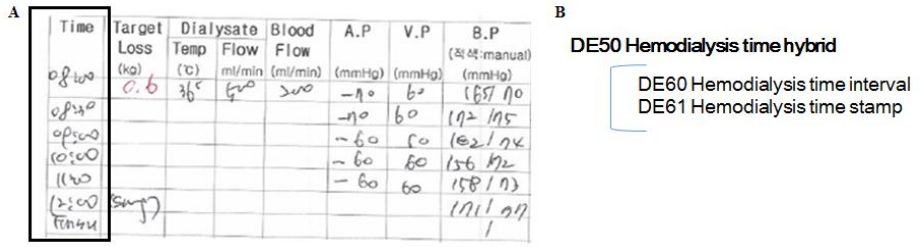

Figure 1-1. Example data from the part of hemodialysis report form containing one CDE with multiple data types. (A) Hemodialysis time has two different data types, (B) atomic CDE, 'DE50' has two different CDEs 'DE60', 'DE60' and either atomic CDEs are used in hybrid aCDE.

2.2 Transformation to different data models

Different types of stakeholders are involved in clinical studies. They usually have different needs for data representation, which might require the data being presented with different data models [24]. For instance, data managers may want to adopt Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) to manage the data in a way conforms to regulatory requirements, but study participants to see their clinical data collected during the study will require representing the data with a model that supports populating the data in, for example, personal health records (PHR) such as HL7 Continuity of Care Document (CCD) or ASTM Continuity of Care Record (CCR). Figure 1-2 shows how medication data related attributes are composed differently in different standard data models. Each attribute can be defined as one CDE, and each data model can be defined as one composite data element as including defined attributes related CDEs.

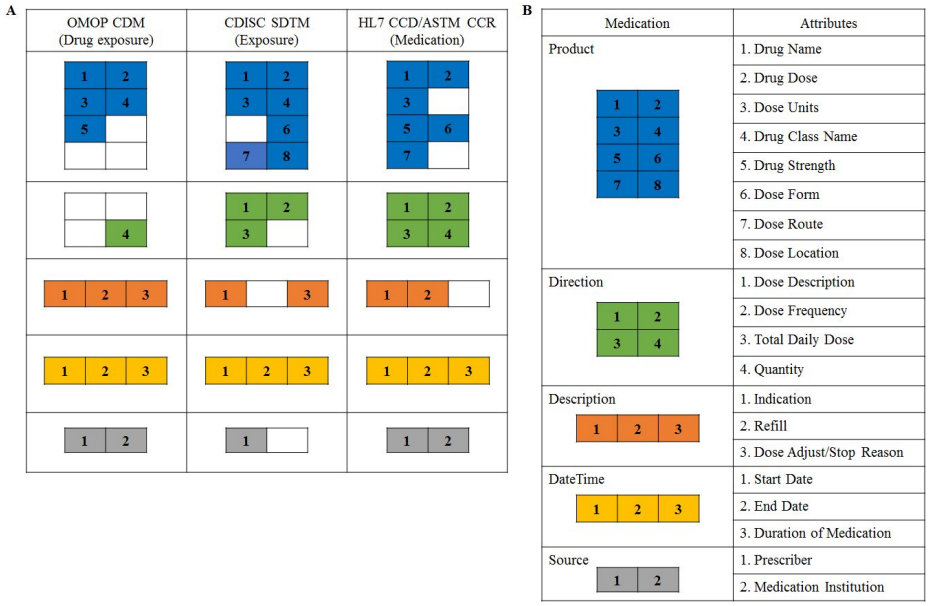

Figure 1-2. Different medication data composition according to data model. (A) Medication data composition from three different data models- Drug exposure table from OMOP CDM, Exposure domain from CDISC SDTM, and Medication section from CCD/CCR, (B) Reference data list of medication related attributes

2.3 Tabular data entries

Data in a tabular format can be found easily in clinical documentations. Figure 1-1(A) shows a typical set of hemodialysis report data in a tabular format. When hemodialysis was performed for chronic hemodialysis patients, the following constituents are recorded prior to the dialysis session and every 30 minutes during the session: interdialytic weight loss, dialysate temperature, dialysate flow rates, blood flow rates, blood pressure as arterial pressure (A.P)/venous pressure (V.P), blood pressure as systolic/diastolic measurement, pulse rates, and patient body temperature. For the case of receiving

hemodialysis several times in a day, it is possible to represent a set of hemodialysis related data in tabular format. This value property information – i.e., *repeat* – needs to be made available to ensure that the values that belong to the same CDE are identified as such.

2.4 Dictionary data entries

Data can be referring particular controlled biomedical vocabulary for several reasons including validation of value sets, and representing value sets with standardized terms for comprehensive understanding. This referencing information is needed to ensure that the value of the data element are connected to the standard terminology.

As we defined this referencing data as variable data element in our previous study [19], when a variable data element is included in a set of data, it can be distinguished from other general sets of data. For instance, figure 1-3(A) shows the typical set of lab test related data in a tabular format, including conducted lab tests, their results with unit of measures, value of indicator whether test result is abnormal, another value of indicator and whether it is clinically significant lab test when the test result is abnormal. The lab test related attribute among sub-attributes in a tabular format can be a key attribute to connect the lab dictionary through the attribute of *LabTest* (see Figure 1-3(B)).

A

Lab Class	Test	Result	Units for Result	Was test result abnormal?	If abnormal, Clinically Significant?
Electrolyte Laboratory Tests	Sodium (Na+)	138	mEq/L	☑ Normal ☐ Abnormal ☐ Unknown	☐ Clinically significant ☐ Not clinically significant

B

Panel	LabTest	Unit	Condition	RangeSt	RangeEd
CBC	Eosinophil	thousand/microL		0.0	0.5
CBC	HCT	%	Male	40	52
CBC	HCT	%	Female	31	43
	....		...		

Figure 1-3. Example for a set of lab test data connecting a lab dictionary. (A) Lab test related data set in a tabular format extracted from the part of a laboratory test form from NINDS CDE project, (B) Part of the laboratory dictionary from NINDS CDE project.

2.5 Derived data

Among atomic CDEs or some sub-CDEs in a composite relationship can be more intimately related each other. The value of the one CDE can be determined by calculating the values of the other related CDEs. For instance, the values of the questions of the same section in the questionnaire might be added for the total value for the same category. Several possible relationships or constraints are needed to be specified. For the formulation of the derived data, we used the method of prefix notation, a symbolic logic to describe the order of operations. For instance, the notation for the expression 3(4+5) could be expressed as *3+45.

3. Evaluation scheme

To evaluate the semantic relationships, we applied the extended relationships on the metadata from the practical clinical documents. We used 25 clinical documents, which are representative five clinical documents such as admission note, initial medical examination note, discharge note, emergency note, and operation note from five major Academic Medical Centers (AMCs) including Seoul National University Hospital, Ajou University Medical Center, Pusan National University Hospital, Gachon University Gil Hospital, and Chonnam National University Hospital.

The evaluation process consisted of the following three steps; atomic CDE extraction, atomic CDE integration as eliminating duplicated atomic CDEs, and construction of semantic relationships among the extracted atomic CDEs. We counted the number of CDEs and their relationships generated during each step as measuring of structural efficiency.

RESULTS

To address the specified challenging cases in the method, we developed new types of semantic relationships and constraints. Before defining new concepts, we first classified the three semantic relationships defined in our previous study [19] into three different types of data elements.

First, the variable relationship is considered one kind of atomic CDE (aCDE), as variable aCDE. Because it only has the characteristic to have reference information for certain controlled terminology. Additionally, we developed one new semantic relationship such as a *hybrid relationship*, as another kind of aCDE, hybrid aCDE.

Second, the composite relationship is considered high level of data elements, as composite CDE (cCDE). Because it includes several aCDEs and grouped them into one CDE for certain reasons. Not like aCDE is represented single aCDE, the cCDE includes inter-related aCDEs. There are three reasons for we distinctly defined the cCDE: 1) separate identification is needed since a set of CDEs in a composite relationship are reused across studies; 2) the clear purpose to be grouped is needed to be defined independently in the description of the cCDE; 3) several inter-relationships or constraints among sub-aCDEs in the cCDE are existed and needed to be defined in the description of cCDE. We also upgraded the cCDE as specifying three subtypes of cCDE such as *dictionary, template,* and *repeated cCDEs*.

Third, the dependent relationship is considered one kind of constraints, as dependent constraint. Because it affects the value among aCDEs in dependent

relationship. We also added other kinds of constraints such as *Ordered, Operated,* and *Required,* and defined them as constraints among aCDEs or sub-aCDEs in a cCDE. Figure 1-4 shows the overview of all of semantic relationships for several types of aCDEs, cCDEs, and constraints.

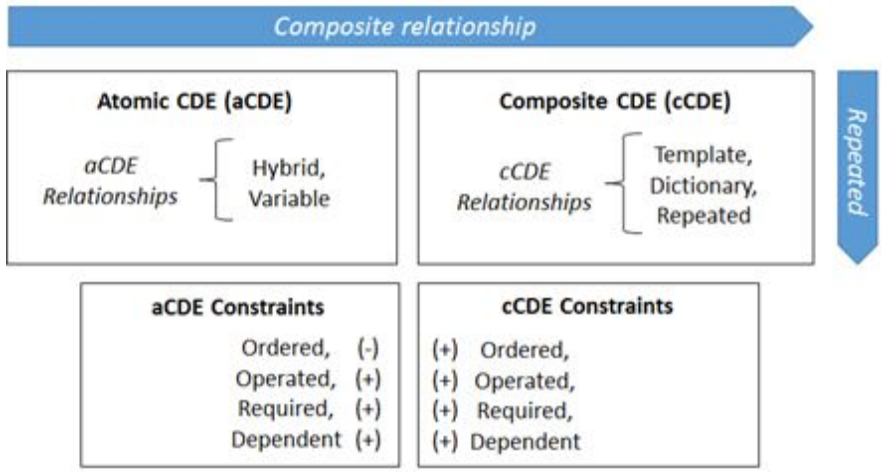

Figure 1-4. Overview of all of existed semantic CDE relationships including various composite relationships and constraint

Data entries with multiple data types: *Hybrid Relationship*

For a particular CDE, when it allows multiple data types, then it is in a *hybrid relationship* (see the top left of Figure 1-4). For instance, the result of HbA1c can be represented variously such as the percentage way of reporting HbA1c values, known as the National Glycohemoglobin Standardization Program (NGSP), and the numeric value way of reporting with unit of 'mmols/mol', known as the IFCC (International Federation of Clinical Chemistry) units. For some institutions, 'NGSP', the string value is included in the lab test result in

order to distinguish between NGSP and IFCC such as '5 (NGSP)'. When the result values of HbA1c are mixed, it can be a problem to measure the value distribution. As supporting conditional clause in hybrid relationship for handling different data types, possible errors can be prevent.

Derived data: *Constraints among aCDEs*

To support generation of robust data forms with individual aCDEs, it is necessary to record information [20] how aCDEs are inter-related. When the constraint is presented with prefix notation, several alphabets and operators are reserved as special key words and used as functional programming in a description (see Table 1-1).

Table 1-1 Functions for representing constraints.

Constraints		Functions
Order		Order
Required		Required
Operated	Assignment operator	=
	Arithmetic operator	+, -, *, /, %
	Logical operator	&, |
	Relational operator	<, <=, >, >=, ==, !=
Dependent		Dependent

1) **Operated**. Formulas for computation of certain aCDEs based on the values of other aCDEs that precede the computed aCDE in the form are defined as equation property. For example, Body Mass Index (BMI) related output aCDE is computed as a function of input aCDEs of body height and body weight by the formula, BMI in kg/m2 = weight / (height * height). Figure 1-5(A) shows BMI formula with aCDE that DE32 is output aCDE, and DE30 and DE31 are input aCDEs. It may provide a function for checking whether the value of output aCDE for the values of input aCDEs is correct. The prefix notation, 'DE32 = (/ (/ DE31 DE30) DE30)' means that 'DE32= DE31/DE30/DE30'. Meanwhile, we can see units of measure are different between DE31 and height value from the formula of DE32 such as 'cm' and 'm'. To handle this difference, we predefined conditional statements to change the unit for BMI formula.

2) **Required**. Required function in which certain aCDEs must be have values, and it does not allow to have null value of them. Figure 1-5(B) shows the part of demography information and asterisk marks on Patient Age, and Gender, which are DE40 and DE41 represent that those are required. The prefix notation, 'Required DE40 DE41' means that the response values of DE40 and DE41 are required.

3) **Dependent**. Dynamic enabling or disabling of certain aCDEs can be defined according to responses to preceding data elements (skip or exclusive logic). It works like a role as a conditional clause. For example, a set of aCDEs regarding hypertension is inapplicable if the patient does not have this condition. Figure 1-5(C) is another

example that for checking whether a patient is current or past smoker (DE20, DE21), if the patient have never smoked, then he or she can skip the DE22. The prefix notation, 'IIF((& [== DE20 'No'] [== DE21 'No']) DE22=null DE22)' means that 'if((DE20 == 'No') & (DE21 == 'No')) DE22=null, else DE22.

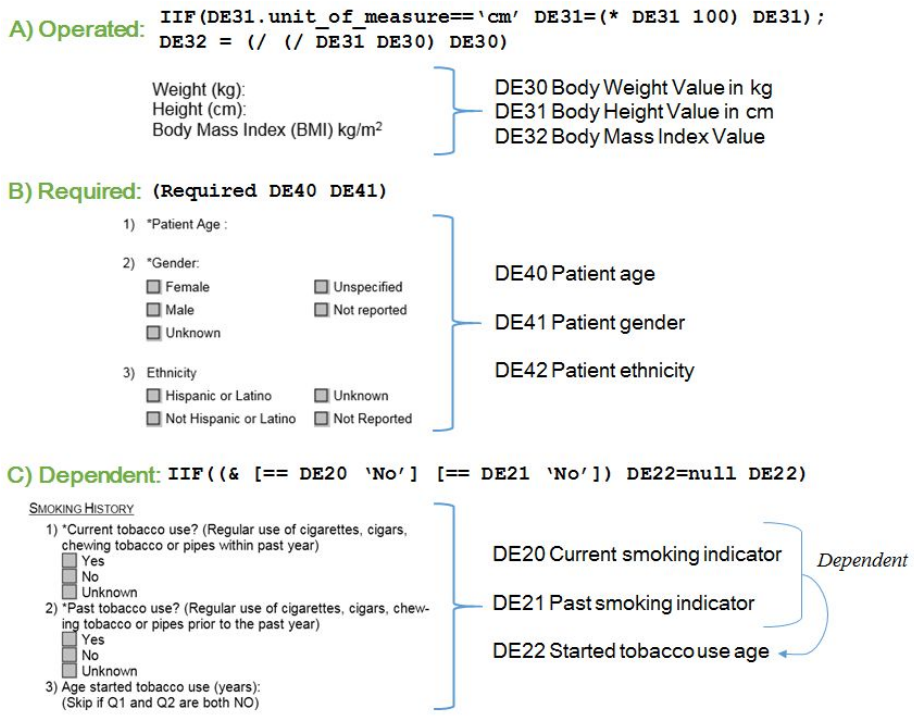

Figure 1-5. Three constraints with examples as representing them with prefix notation.

Transformation to different data models: *Template cCDE*

In biomedical data modeling, each data model from standard data models to proprietary data models has unique purpose to be organized for specifying how data items are related and which properties are needed to be represented.

When data items in a certain data model are defined as ISO/IEC 11179 based CDEs and they can be composed as one of cCDE according to the data model specification, we called that it is a *template* cCDE.

Main usage of the *template* cCDE is that it can help model converting function and model transforming function among the similar data items across different data models through metadata mapping process.

Figure 1-6 shows the example how the template cCDE is composed for certain data model, which is based on the part of comparison information among medication data composition from three different data models in figure 1-2. DE115 in figure 1-6(A) is a template cCDE for medication data following a CDISC SDTM model with having two cCDEs (DE15, DE16). DE15 is another cCDE for medication product information to have two sub-aCDEs (DE25, DE26). DE25 is about a drug name, indicating SDTM attribute, CMTRT. The CMTRT captures the name of the Concomitant Medications/Therapy and it is the topic variable in CDISC SDTM model. Like CMTRT, variable names in italic format are specific attributes of each data model. This mapping information among attributes of the data model and aCDEs are stored in DE115 like 'DE25 == CDISC.CMTRT' in figure 1-6(B). Meanwhile, DE215 in figure 1-6(C) is another template cCDE for medication data, following an OMOP CDM model. We can see the same aCDEs from DE1115 such as DE25, DE39, and DE41 are used as sub-aCDEs in DE215. This information can be used to transform among different data models.

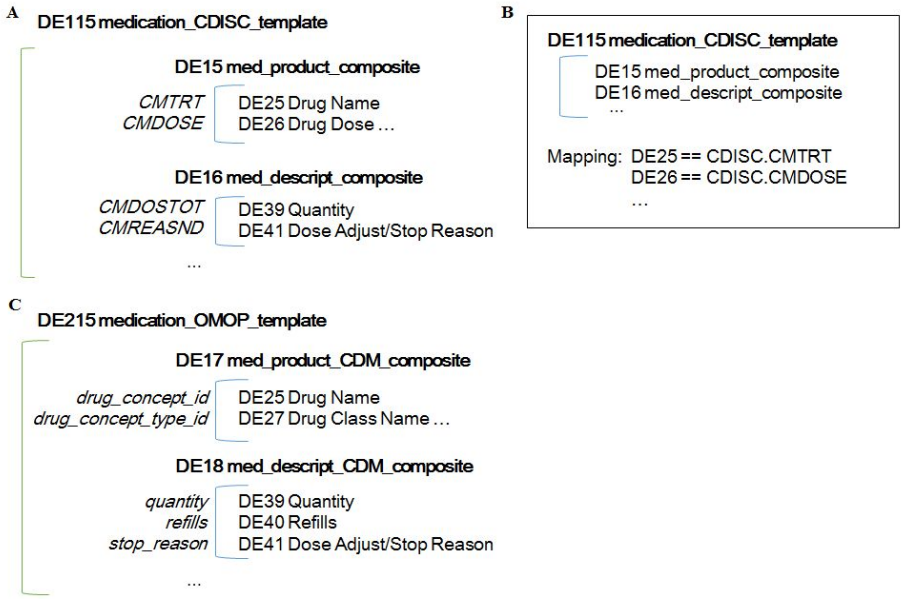

Figure 1-6. Part of different medication data composition in two different data models. (A) Template cCDE for medication data in CDISC SDTM, (B) mapping rule among attributes of the data model and aCDEs, (C) template cCDE for medication data in OMOP CDM.

Dictionary data entries: *Dictionary cCDE*

The cCDE is composed with several sub-aCDEs. When a particular sub-aCDE is in a variable relationship, which is referenced by external sources such as standard terminology, then it is called as *dictionary* cCDE. It may provide the way to connect various dictionaries from standardized biomedical controlled vocabulary to proprietary or self-defined dictionary from each organization, and may help to bring the knowledge information from them. As using the dictionary information, which is connecting to variable aCDE in the dictionary cCDE, the value-sets of sub-aCDEs can be validated. For instance,

four attributes such as *Lab Class, Test, Units for Result,* and *Was test result abnormal* from figure 1-3(A) are matched to a lab test related dictionary from figure 1-3(B) as *Lab Class* is matched to *Panel* and *Was test result abnormal* is matched to *RangeSt,* and *RangeEd.*

Tabular data entries: *Repeated cCDE*

The cCDE may be represented multiple times repeatedly in a tabular format, and is called it is *repeated* cCDE. It may prevent the unnecessary generation of redundant cCDEs. It can also show how values have been changed in a certain time interval or how related values are occurred in a certain time. For instance, figure 1-7 shows typical medical history items in tabular format, having six sub-attributes.

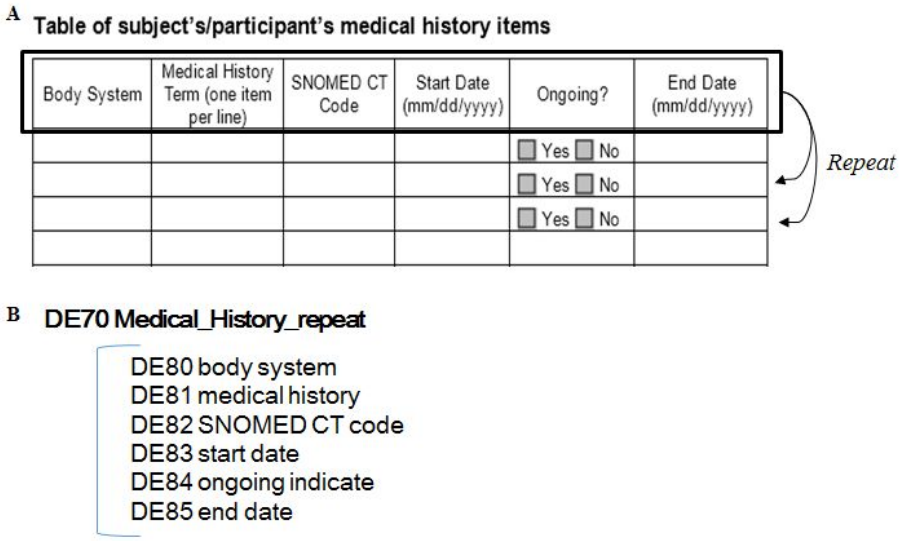

Figure 1-7. Repeated cCDE for medical history information. (A) Typical medical history items in tabular format, (B) defined repeated cCDE and its sub-aCDEs.

Derived data: ***Constraints among sub-aCDEs in cCDE***

We adopted the three constraints, which are defined among independent aCDEs such as *Operated, Required* and *Dependent* for completing constraints among sub-aCDEs in cCDE, and added one other constraint, *Ordered*. The order constraint restricts that certain sub-aCDE should be presented first then other sub-aCDEs in order. Changing the order of aCDEs may change the meaning or may be awkward in the context. For instance, end date related aCDE shouldn't be first than start date related aCDE such as the order of DE83, DE84, and DE85 in figure 1-7(B).

Evaluation

To show the efficiency of using semantic relationships among aCDEs and cCDEs, we counted the number of CDEs for each step from extracting CDEs to developing semantic relationships.

Table 1-2 shows the evaluation results that the number of aCDEs, cCDEs, and their relationships extracted from five document types from five hospitals and we see that how CDEs based semantic relationships are effective to reduce duplicate data. The each first row of the AMC shows the total numbers of aCDEs extracted from each document. The each second row of the AMC shows the number of the shared aCDEs among five documents from aCDEs shared by five document to aCDEs specific in one document. For instance, for the case of admission note from the hospital A, 6 aCDEs are shared in five documents, 3 aCDEs are shared in four documents, 10 aCDEs are shared in

three documents, 24 aCDEs are shared in two documents, and 38 aCDEs are specific for the admission note of the hospital A. The each third row of the AMC shows the number of cCDE and the number of sub-aCDEs of cCDEs are specified within brackets. The each fourth row of the AMC shows the number of the specific aCDEs and cCDEs, which are included sub-aCDEs. For instance, for the case of admission note from the hospital A, 43 sub-aCDEs are reduced from total 81 aCDEs and 10 cCDEs are additionally counted. In other word, the specific aCDEs and cCDEs are counted as 48, which is derived from the formula 81-43+10=48.

For three steps of evaluation, the first and second rows are represented the CDE extraction step, and the CDE integration step respectively. And the third and fourth rows are represented the steps for construction of semantic relationships of the extracted CDEs.

For all hospitals, the number of integrated CDEs were reduced by the lowest rate 25.2% to the highest rate 52.2% compared to the number of extracted CDEs. After applying composite relationships for integrated aCDEs, the number of unique aCDEs and cCDEs were reduced by the lowest rate 34% to the highest rate 80.9% compared to the number of extracted CDEs.

We descripted the summary of the number of aCDEs, and reducing rate by integrated aCDE and cCDE for each hospital in figure 1-8. All hospital get results from metadata based reducing effect. Hospital P has largest number of aCDEs compared to other hospitals. The clinical documents from the hospital S are relatively simple as it doesn't have any cCDEs.

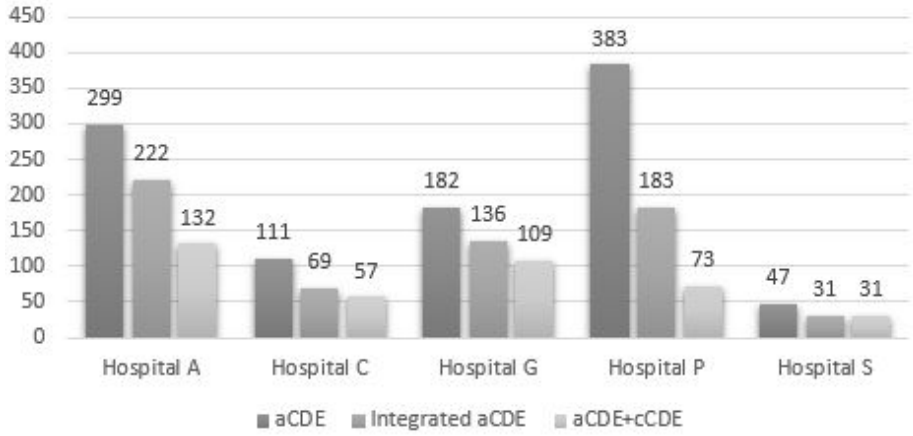

Figure 1-8. Summary of the number of aCDEs, integrated aCDEs and cCDEs for five documents from each hospital.

Meanwhile, we also observed how similar between hospitals. Table 1-3 shows another evaluation result as representing number of common aCDEs and cCDEs between hospitals. It shows the possibilities how different hospitals can share their data by using MDR based metadata. Through shared common metadata, it has effectiveness to reduce the duplicated data. The hospital G has most common data items to be shared then other hospitals as having high reducing rate of 28.1%, and 34.4% with hospital A, hospital P, respectively.

Table 1-2 Number of aCDEs and cCDEs extracted from five document types from five hospitals.

		Admission note	Initial medical report	Discharge note	Emergency note	Operation note	Total	Reducing rate
Hospital A	[a] |aCDE|	81	45	62	79	32	299	25.7% (222/299)
	[b] |Integrated aCDEs|	6/3/10/24/38	6/3/9/22/5	6/3/7/0/46	6/3/3/0/67	6/0/1/2/23	222	
	[c] |cCDE(sub-aCDE)|	10(43)	7(33)	3(14)	4(64)	3(12)	14(104) 27(166)	17.6% (104/166)
	[d] |aCDE+cCDE|	48	19	51	19	23	132 160	55.8% (132/299)
Hospital C	|aCDE|	24	28	14	23	22	111	37.8% (69/111)
	|Integrated aCDEs|	4/4/2/11/3	4/2/1/0/21	4/4/1/0/5	4/2/1/11/5	4/4/0/0/14	69	
	|cCDE(sub-aCDE)|	1(4)	3(14)	1(4)	1(9)	2(11)	3(15) 8(42)	62.5% (15/42)
	|aCDE+cCDE|	21	17	11	15	13	57 77	48.6% (57/111)
Hospital G	|aCDE|	58	26	34	53	11	182	25.2% (136/182)
	|Integrated aCDEs|	0/16/34/74/9	0/16/29/71/8	0/16/15/3/14	0/16/34/0/16	0/0/0/0/11	136	
	|cCDE(sub-aCDE)|	1(6)	1(6)	3(17)	3(23)	1(5)	4(31) 9(57)	73.3% (31/57)
	|aCDE+cCDE|	53	21	20	33	7	109 134	40.1% (109/182)

Hospital P	|aCDE|	133	124	48	66	12	383	52.2% (183/383)
	|Integrated aCDEs|	0/16/34/74/9	0/16/29/71/8	0/16/15/3/14	0/16/34/0/16	0/0/0/0/12	183	
	|cCDE(sub-aCDE)|	3(108)	2(83)	2(32)	2(32)	0	5(115)	44.4% (115/255)
							9(255)	
	|aCDE+cCDE|	28	43	18	36	12	73	80.9% (73/383)
							137	
Hospital S	|aCDE|	12	6	10	10	9	47	34% (37/47)
	|Integrated aCDEs|	0/1/3/6/1	0/1/3/0/2	0/1/0/0/9	0/1/3/6/0	0/0/0/0/9	31	
	|cCDE(sub-aCDE)|	0	0	0	0	0	0	-
	|aCDE+cCDE|	12	6	10	10	9	31	34% (37/47)

[a] Number of aCDEs,

[b] Number of common aCDEs for between hospitals,

[c] Unique number of aCDEs [reducing rate for common aCDEs, total number of aCDEs from five documents]

[d] Number of cCDEs (Number of sub-aCDEs in cCDEs).

Table 1-3 Number of common aCDEs and cCDEs between hospitals.

		Hospital A	Hospital C	Hospital G	Hospital P	Hospital S
Hospital A	|aCDE|	132	24	53	21	15
	|cCDE(sub-aCDE)|	12(104)	4(5)	2(10)	4(23)	2(6)
	|aCDE+cCDE|	222	25	61	40	19
	Mapping rate	100%	8%	28.1%	11.4%	9.2%
Hospital C	|aCDE|		57	17	14	8
	|cCDE(sub-aCDE)|		3(15)	1(4)	2(10)	3(14)
	|aCDE+cCDE|		67	20	22	19
	Mapping rate		100%	11.4%	12%	19.3%
Hospital G	|aCDE|			109	45	8
	|cCDE(sub-aCDE)|			4(31)	2(9)	3(14)
	|aCDE+cCDE|			136	52	19
	Mapping rate			100%	34.4%	19.3%
Hospital P	|aCDE|				73	9
	|cCDE(sub-aCDE)|				5(115)	3(14)
	|aCDE+cCDE|				183	20
	Mapping rate				100%	9.3%
Hospital S	|aCDE|					31
	|cCDE(sub-aCDE)|					0
	|aCDE+cCDE|					31
	Mapping rate					100%

DISCUSSION

Comparison with related studies

Standardized data through ISO/IEC 11179 based CDEs is one of effective ways to harmonize data collected from various clinical studies with following advantages; 1) providing consistent data collection tool, 2) improving study quality and reducing cost of data entry, cleansing by having uniform data. However, due to the limitation of basic structure of ISO/IEC 11179, there has been a gap between development of CDEs and utilization of CDEs in clinical form with comprehensive representation in practical.

To break this obstacles, metadata experts revised ISO/IEC 11179 standard as adding Data Element Derivation and Derivation Rule tables to enhance inter-related data elements [25]. For example, a data element 'length of stay in a hospital' is derived by calculating the number of days from 'admission date' to 'discharge date'. In this case, ISO/IEC 11179-3 defines the Data Element Derivation table with the two input data elements of 'admission date' and 'discharge date' and the output data element of 'length of stay in a hospital'. The mathematical rule is managed in the Derivation Rule table. It is a simple but powerful idea to represent how multiple data elements are processed to produce (i.e., derive) a new data element based on what derivation rules. However, implementation of this idea is still in an early stage, only simple definitions and unspecific structures are provided in ISO/IEC 11179-3.

On the other hand, the CDISC ODM, the XML-based underlying standardized data model supporting the acquisition and exchange of metadata specifically related to clinical studies is also used to rectify the limitations of ISO/IEC 11179; however, it is not comprehensive enough to be implemented for CRF generation as importing elements directly [26].

Lin et al [27] also suggest to use the openEHR approach as modeling of CDE concept. Though openEHR has comprehensive structure with 2-level modeling approach, several limitations in the course of implementation of openEHR were identified in various studies such as immaturity of archetype modification operations, insufficient support for hierarchical archetypes as the granularity of archetype [28-29], and it has a burden of cost to develop and adopt since it is too complex to define.

Instead of utilization external data models, we proposed to extend the existed composite relationship as specifying three subtypes with constraints and additional new data element type for comprehensive understanding of the semantic relationships among the inter-related data elements.

Overcoming the challenges of understanding semantic relationships of data in form layer, and synchronizing different standardized data

This paper presents an in-depth description of ISO/IEC 11179 MDR standard based-CDE inter-related relationships for reflecting the structure of clinical forms including the neighboring CDEs without loss of the contextual semantic relationships between CDEs and clinical forms.

For reflecting the form level data into data element level data, two subtypes of relationships (e.g. *repeated* cCDE, *dictionary* cCDE) were developed and they provides benefits as following:

1) **Repeat property in a repeated cCDE can manage data of tabular format in clinical forms.** Since multiple value sets can be represented through repeat property, repeated cCDE is useful to manage sequential data in tabular format at one time, and to analysis how value is changed as time goes by or how related values are occurred in a certain time.
2) **Advantages for having the CDE in the variable relationship connected to a particular dictionary.** The data item, referencing a certain standard terminology is frequently appeared in a clinical form. A CDE in a variable relationship is the key to connect with a dictionary. In other word, observation data set in a variable relationship is joined to knowledge data in dictionary composite data element. It can help to gain rich contents of external terminology or dictionary. It can provide validation function as connecting external dictionary information.

Providing functional programming through description of developed constraints among aCDEs and sub-aCDEs in cCDEs

Through the development of the method to store and represent aCDE and their related semantic relationships, context information and constraints among

inter-related data items in clinical documents were able to be described and used for value set validation. Specially, described constraints were directly used for form-level data validation.

Advantages and expected effects as using metadata to build clinical documents

As we verified through the evaluation process, using metadata to build clinical documents can give us mainly three advantages. First, it prevents redundant data generation as reusing predefined and stored data elements from MDR, and secondly it ensures data integrity as data elements have information of comprehensive data description and data inter-relationship from defined CDE relationships and constraints. Thirdly, it provides the future possibility to integrate clinical data from various EMR systems.

Limitations and Future Work

So far, the proposed MDR based extended semantic relationships were mainly evaluated from a design perspective. Though we conducted a pilot study with 25 clinical documents from five different AMCs, it cannot be fully represented all possible clinical documents as the real status of clinical practices. As extended study, we require to examine our developed CDE relationships and constraints to verify whether those can cover all cases of clinical documents, and how generally those can be applied to other EMRs.

CHAPTER 2

Syntactic and semantic validation of clinical documents containing composite common data elements representing complex clinical metadata

INTRODUCTION

Clinical documentation was developed to track a patient's condition and to communicate the author's actions and thoughts to other members of the care team [30]. It is supporting structured data capture for multiple needs will enable the vision of ‘collecting once, using many’, and reduce the extra time and expense clinicians spend on data entry as enabling electronic capture and storage of clinical data [31].

For comprehensive semantic representation with clear definition of clinical data, ISO/IEC 11179 standard based Common Data Element (CDE) has been used to compose and build up a clinical document [1-3]. The CDEs are often called ‘content standards’ as they are disease-specific data elements that could be used widely across several domains and it is a popular and practical approach to identifying data standards, and they are stored in metadata registry (MDR).

Meanwhile, in the process of sharing clinical data through clinical documents, the data should be checked first for completeness and to ensure that no errors were introduced. The process of data validation significantly adds to the time, and complexity of the data sharing process, but is critical to maintain the integrity of clinical data and to ensure high-quality data.

The overarching goal of this work is to identify the scope of validation for clinical documents containing complex clinical metadata and to develop syntactic and semantic validation process. For the evaluation of feasibility of

the validation function, we examined how the practical clinical document with several semantic relationships are used to validate.

MATERIALS AND METHODS

1. Complex clinical metadata in clinical documents

For representing and managing complexity of clinical documents, we used ISO/IEC 11179 standard based CDEs and their relationships to build clinical documents. In this paper, we designated the complex metadata as having several semantic relationships of CDEs, developed in our previous study [19]. Figure 2-1 shows how complex metadata are created through following steps. We first extracted metadata from the clinical document as analyzing questions, possible responses, section information and other form-level information, and defined them as CDEs according to ISO/IEC 11179 standard.

The term CDE is an atomic data element that represent the lowest level of data detail as an atomic unit of data. When more than two atomic data elements are inter-related each other, we declared that they are in a semantic relationship.

Then, as observing form-level information including constraints among data items in the clinical document, we looked at whether data item allows to have multiple data types, or whether data items are describe in tabular format repeatedly. And if so, we defined them that they are CDEs in such a semantic relationships such as *variable* and *hybrid relationships*, respectively.

Beyond a single atomic CDE (aCDE), we also looked at whether aCDEs were needed to be grouped as referencing external standard terminologies or data models. And if so, we defined them that they are grouped as a composite common data element (cCDE) and sub-classified cCDEs as three sub-types such as *dictionary, repeated, and templated cCDEs*.

2. Restructure a validation process

In our previous study, we already defined the architecture of multi-layered validation [32], which was included syntactic and semantic validation processes including XML schematic validation, personal health record data model XSD-based validation, and CDE-based semantic validation. But, it was a simple process that only applied aCDE on a single standard data model, and it wasn't fully described that which validation attributes are used in such a process.

As adding attributes that related to several semantic relationships of aCDEs and cCDEs, we upgraded the simple architecture of multi-layered validation. We also specified the list of validation attributes to represent the detail of the validation. Table 2-1 shows the list of the summary of validation attributes from the semantic relationships of CDEs and ISO/IEC 11179 standard.

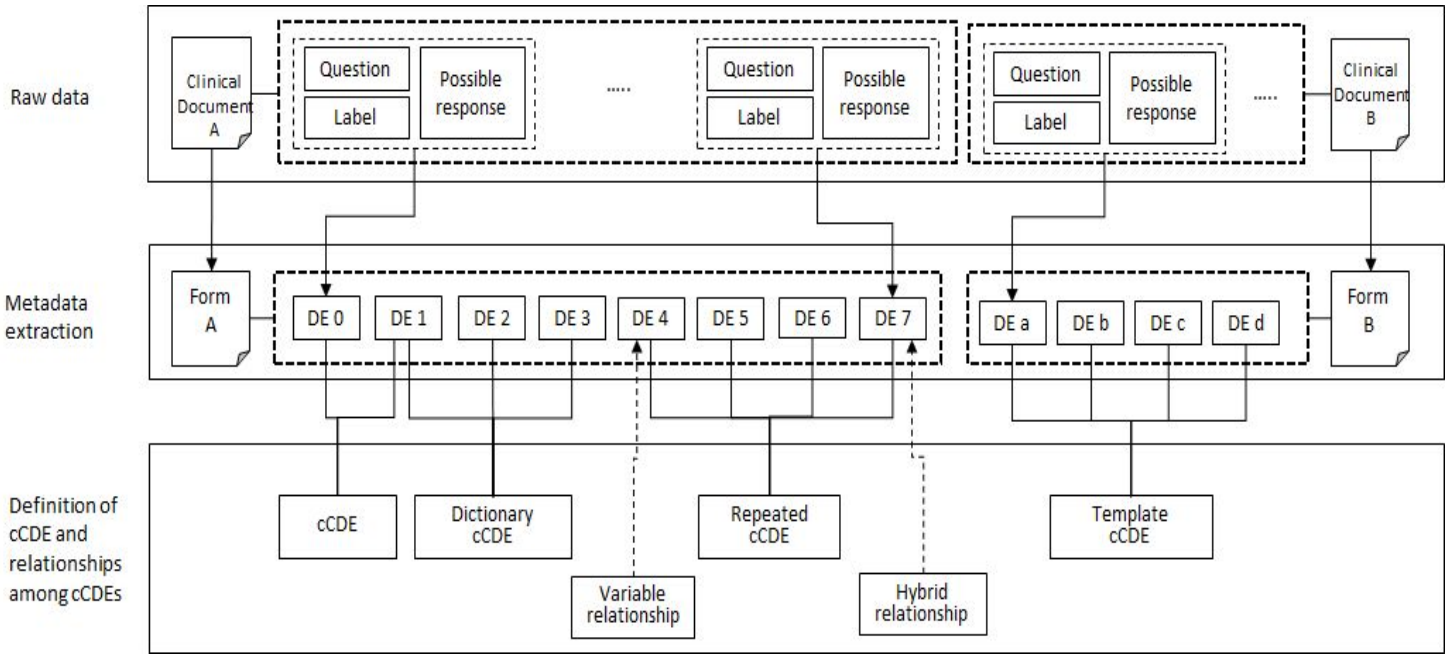

Figure 2-1. Creating clinical complex metadata for clinical document containing composite data elements. The process diagram of metadata extraction and composite data elements and related constraints.

Table 2-1 List of validation attributes.

Types of data element	Validation Category	#	Validation Attributes
aCDE	Value domain	1	Data type
		2	Min/Max value
		3	Unit of measure
		4	Permissible value
	CDE relationships	5	Hybrid
		6	Variable
	Constraints among aCDEs	7	Dependent
		8	Operated
		9	Required
cCDE	Dictionary cCDE	10	Dictionary information
	Template cCDE	11	Data model information
	Constraints among sub-aCDEs	12	Ordered
		13	Dependent
		14	Operated
		15	Required

RESULTS

Schematized procedure to build a clinical document with metadata

Before understanding of validation procedure, we simply schematized the procedure how the clinical document with metadata is made through a real example data to comprehend complex metadata (see Figure 2-2).

A clinician may develop the paper-based clinical form to record laboratory result data in tabular format as including five attributes such as lab panel name, lab test name, lab result value, indicator whether the lab result value is normal, and indicator whether lab test result is clinically significant (see Figure 2-2(A)).

In figure 2-2 (B), we can see that each attribute can be defined as aCDE according to the process of extracting metadata. All attributes are defined as five aCDEs; ‘DE70 Laboratory Panel’, ‘DE71 Laboratory Test Name’, ‘DE72 Laboratory Test Result’, ‘DE73 Lab Test Result Abnormal Indicator’, and ‘DE74 Lab Test Result Abnormal significance’. We also defined the one dictionary cCDE, ‘DE82 Composite Lab Test Result’ to include defined five aCDEs.

The sub-aCDEs of the cCDE are defined independently. However, there are certainly two rules, which can be defined as constraints and be used in validation process; 1) the prefix notation based constraint, (R DE71 DE72 DE73) that three sub-aCDEs such as ‘DE71’, ‘DE72’, and ‘DE73’ are required to have value set, 2) the another prefix notation based constraint, IIF([== DE73 ‘Abnormal’] DE74 DE74=null) that two sub-aCDEs are

dependent that when the value of ‘DE73’ is abnormal, ‘DE74’ can be activated. After all, these constraints can be defined and recorded in the cCDE. Meanwhile, we can also find another two cases to be defined as constraints for validation through a dictionary as one of sub-aCDEs in the dictionary cCDE, ‘DE71’ is linked to the dictionary as a key aCDE. The first case is that the value set of ‘DE70’ can be checked by the value of the ‘LabTest Name’ in dictionary cDE whether the value of ‘DE70’ is valid in the dictionary. And the other case is that the dictionary has the information of normal range for each lab test, and it can be used to check whether the values between ‘DE72’ and ‘DE73’ are correlated. For instance, for lab test ‘HCT’, it has 48% test result in DE72 and indicated it is normal value in DE73. In the lab test related dictionary, it shows that when the HCT is 48%, it is normal, so the values between ‘DE72’ and ‘DE73’ are valid each other.

After finishing the definition of metadata and their relationships and constraints, the clinical document can be built as adding complex metadata, which are defined aCDEs, cCDEs and their semantic relationships. In figure 2-2 (C), the left side of the figure is XML model based PHR data and the right side of the figure is the correlated metadata part in the data.

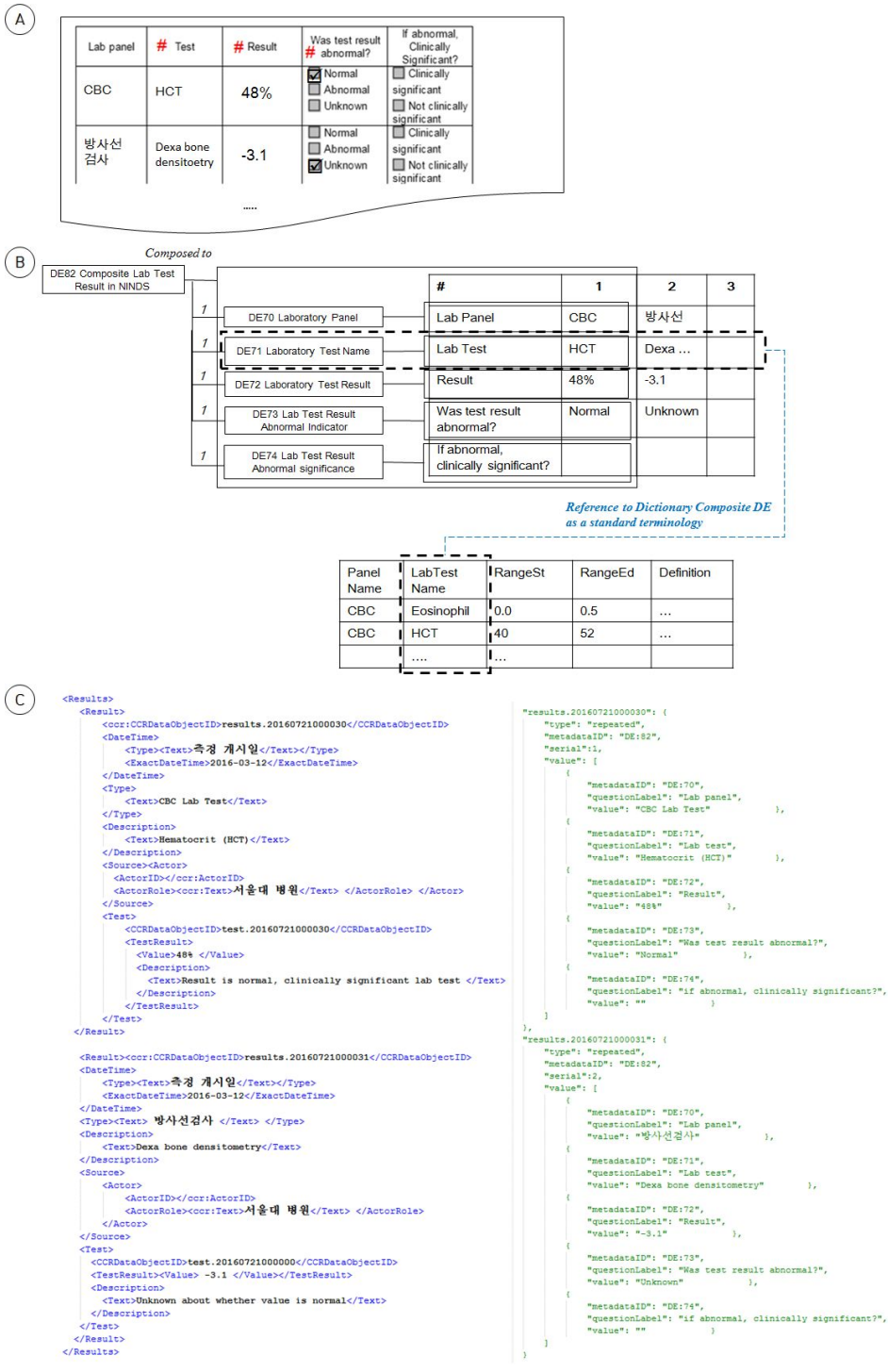

Figure 2-2. Three steps to build a clinical document with metadata.

Overview of validation procedure

We restructured a validation process with mainly two processes, syntactic and semantic validation processes. In figure 2-3, the validation procedure is started from syntactic validation process with two validation parts; 1) XML schematic validation, and 2) template cCDE based schematic validation. The reason to have XML schematic validation is that we assumed that clinical documents are XML based files, as those are the electronic documentation that works on computer-based systems.

Since the template cCDE is represented the each data model from standard model to proprietary models as metadata level, it contains information of hierarchical and subsumption relationships among data items of each data model and the data model itself as well. If the clinical document is followed a certain data model as being compliant with the template cCDE, the syntactic information from the template cCDE can be used to validate the clinical document. The number in the top left corner of the square box to represent each validation is correlated to the number of the list of validation attribute in Table 2-1. For instance, number eight is the template cCDE based validation.

Not like semantic validation, the invalidity of the structure of clinical documents can be significant problem to complete clinical documentation. So when there is syntactic invalidity, the validation system should give a message to modify the structure of the clinical document.

After the syntactic validation process, we eliminated the XML comment tags from the XML input file. The XML comment tags, which were started with

‘<!-’ and ended with ‘-->’, were used to include metadata information in the clinical document for ignoring the metadata part because it might not be compliant to the structure of the clinical document. The following four attributes were used to have brief metadata information: *Question, MetadataID, QuestionLabel,* and *Value*, developed in our previous study [6]. Especially, the *MetadataID* attribute is a semantic identifier that connects to the MDR to bring out rich contents of metadata information. With the found CDE IDs in the *MetadataID* attributes from the files, we conducted semantic validation process.

After eliminating comment tag, the semantic validation process is started. The semantic validation process has the order from cCDE to aCDE because cCDE is composed with aCDEs. Among sub-types of cCDE, we first checked whether cCDE is dictionary cCDE because it is referencing external resources. As using rich content of dictionaries, the dictionary cCDE based validation is conducted.

After dictionary cCDE based validation, we also checked whether cCDE has internal constraints among sub-aCDEs, which is corresponded to the number from 12 to 15 in Table 2-1. If there are constraints among sub-aCDEs, the validation process is conducted according to the defined constraints. After the validations of cCDE, each sub-aCDE in cCDE has checked whether it is valid according to validation rules of aCDE according to value domain information, which is corresponded to the number from 1 to 4 in Table 2-1. After cCDE

based validation, single aCDE based validation process, which are aCDE relationship based validation and value domain based validation, is conducted.

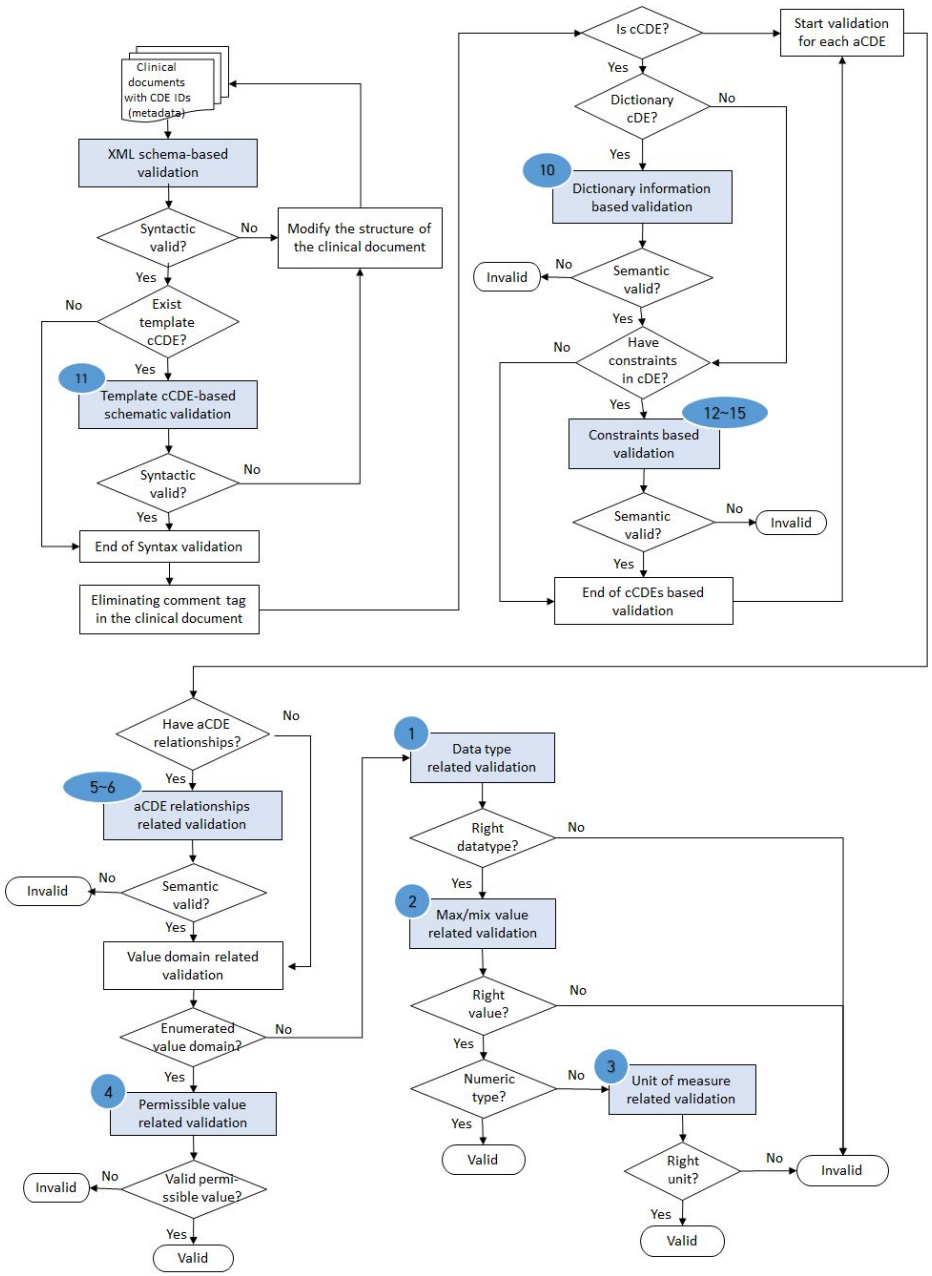

Figure 2-3. Process of syntactic and semantic validation of clinical documents containing composite data elements.

DISCUSSION

During the process of data sharing through clinical documents, the data should be checked first for completeness of clinical documents and to ensure that no errors were introduced. It is critical to maintain the integrity of clinical data and to ensure high-quality data.

When the clinicians or clinical researchers create clinical documents, they already have the rules and constraints that allow such information of the relationships between data in their mind. However, the current clinical systems does not have a methodology to cover and handle these data. There is basically only to determine whether the data is correlated the correct datatype.

Meanwhile, increasing use of CDEs in clinical documents led us to develop specific semantic relationships among CDEs for representing context, constraints and inter-related relationships among data in clinical documents, and we defined and stored these information in ISO/IEC 11179 MDR based metadata.

We first defined the complexity of clinical documents, represented by ISO/IEC 11179 standard based CDEs and their various semantic relationships. Basically, the major validation process is focused on value of aCDE according to defined value domain information, composed to datatype, min/max value, permissible value, and unit of measure.

Based on the defined semantic relationships and constraints among aCDEs, we also defined the list of validation attributes and designed the process of validation including four constraints such as operated, ordered, dependent,

and required, and several types of cCDEs such as template, dictionary. Through designed validation attributes and the structured process, we can expect high semantic effectiveness as establishing standard criteria to validate data for sharing data among the possible medical institutions to use ISO/IEC 11179 MDR based metadata in their data.

CHAPTER 3

Clinical Metadata Ontology: A Simple Classification Scheme for Common Data Elements of Clinical Data based on Semantics

INTRODUCTION

Data should be collected in a consistent manner by using standardized format for providing unified data collection, and facilitating data sharing and data integration. Effective combining and comparing data from multiple sites is required to facilitate patient care or research purpose. There has been several efforts trying to accomplish the standardized data, which is focused on specifying both the syntax and the semantics of clinical information.

One of major approach is to build a common data model; it can be classified as a top-down approach where a top-level knowledge model agreement is forced for the underlying data models of the interoperating parties for successful data exchange [1]. Representative data models are the HL7 Reference Information Model (RIM) and EN 13606 standards, that includes generic reference models of concepts and relationships (e.g. CEN/ISO 13606, openEHR Reference Model, or HL7 RIM) and more detailed models (e.g. openEHR Archetypes/Templates, or HL7 Detailed Clinical Models. For semantic interoperability, they are connected terms from terminology models (e.g. SNOMED, LOINC, etc.). However, a major problem of this top-down approach is that it takes a long time to adopt and reflect them at each different circumstances, and to revise the model with recent advances.

Another approach is the unified one between the top-down and bottom-up approaches by building a Metadata Registry (MDR) based on the international standard named ISO/IEC 11179. This standard specifies a metadata model for representing the common data elements (CDE) that is a logical data unit that

provides for the definitions of data, including an identifier, and response option values to indicate the value type, and detailed information to represent data concepts and its semantics [11]. The well-defined CDEs can be collected and reused as content standard. The CDEs are registered in MDR are designed whenever researchers require (bottom-up) as following pre-coordination strategy, and are aggregated and integrated by research community (top-down) as following post-coordination strategy. Since ISO/IEC 11179 approach has a concept of metadata (literally 'data of data'), it can encompass both approaches in the view of common data collection structure [14]. We finally adopted the ISO/IEC 11179 approach.

The anticipated benefits of facilitating ISO/IEC 11179 based standardized data are multiple, including 1) effective and rapid data collection as reducing burden on investigators to facilitate their participation in clinical research, 2) enriched data sharing and data aggregation by employing common forms, and standard definitions, and 3) improved data quality by providing unified data and its descriptions [11-12].

We also implemented ISO/IEC 11179 based metadata registry, called Biomedical Metadata Standard for Health (BMeSH) server [19, 32-33] for management of CDEs based clinical data, and developed more than 2 thousand CDEs.

The key aspect in facilitating CDE based standardized data is to search an appropriate CDE from MDR. For this purpose, the ISO/IEC 11179 is provided Classification Scheme (CS) structure for conceptual classifying and

identifying the data elements. Thus, when the communities or researchers decided to build a MDR or to register designed data elements into a MDR, they should also select or design the contents of CS using controlled vocabularies [34] or their own [16, 35]. However, most of MDR do not fully utilize or register the CS. Even some of MDRs support only two or three Classification Scheme Items, which are concept items in each CS, to classify their metadata [18].

Also, most CDE browsers that developed by each research project is not fully utilized CS and remain to adopt simple keyword search engine. The keyword-based search has the problem of ambiguity in the written natural language, for example documents containing synonyms of the query keywords will not be retrieved. Another shortcoming of the keyword search is cause by homonyms from natural language. For example, searching information on 'Radical' will bring up pages containing information about both 'Radical' (Extreme or drastic) and 'Radicle' (a vessel's smallest branch) even though the use is only interested only in one of them. An ontology based search approach can be one of solutions as it is considered semantically enhanced information retrieval method.

The objective of our study is to develop an ontology for managing CDEs, which is called Clinical Metadata Ontology (CMO) to enable retrieve and classify CDEs. We use clinical documents from HL7 templates [36] and 25 common documents from 5 hospitals in South Korea [2], and demonstrate

them for verifying the suitability of CMO to facilitate classify and integrate CDEs as a proper clinical data-organization scheme.

MATERIALS AND METHODS

CMO was developed using the General Formal Ontology method [37], which has a manual and iterative process with four steps: (1) defining the scope of CMO as conceptualizing first-level terms (or first-level classes) of CMO, (2) identifying CMO concepts, (3) assigning hierarchical relationships among CMO concepts, and (4) development of CMO properties (e.g., synonyms, preferred term, and definitions) for each CMO concept.

1. Defining the scope for CMO

A clinical document is a document of a patient's history and care. Every evidence or background of the care also can be documented. It is the most important source of information in clinical decision-making, the communication between health care providers, and legal issues.

Through investigating how clinical data is generated in clinical practices with clinical documents, we can assume which data elements are used and included. Since CDE has been used as composition of a clinical document, we analyzed clinical documents and identified CMO concepts.

At initial contact, a patient is registered and his/her health-related problem (history) is gathered as focusing on their current illness, symptoms, and chief complaint. Then health care providers provide diagnostic or therapeutic

procedures depending on the information the patient provided. The process of procedure and observation or testing is repeated until the end of treatment. During this interaction process the event such as admission, discharge, or adverse drug reaction can be occurred and usually is changed in the general environments of health care. We found that clinical information could be categorized with four main terms; (1) *Procedure*, (2) *Finding*, (3) *Event* and (4) *Description* and determined those terms as first-level terms of our ontology.

2. Identifying CMO concepts

A data element, which is the atomic unit of data, is associated with a data element concept (an abstract unit of knowledge for representing semantics) and value domain (representation of data including data type, permissible values) according to ISO/IEC 11179 standard.

To identify CMO terms, we used representative data element concepts of CDE from BMeSH. Especially, we selected Seoul National University Hospital (SNUH) clinical documents related CDEs among total CDEs in BMeSH for examining data element concepts. Most frequently used SNUH clinical documents, which were used more than 10 times between Jan.2010 and Aug. 2010 at each hospital department were only selected. 27,109 CDEs were extracted from 663 SNUH clinical documents.

We extracted common concepts from selected data element concepts as considering whether those are reasonable to be subordinated to first-level

terms of CMO, and chose them as CMO terms, the child terms of each first-level term. Those were reviewed and determined by three medical doctors and medical informatics researchers. For example, we classified *Description* into eleven child terms including *Advance Directives, Alerts, Assessment, Chief Complaint, Demographics, Encounter, General Weakness, History of Immunization, Past Medical History,* and *Present Illness* because these terms were readily accepted by most clinicians in SNUH to represent of Description. We conducted this repeated process that finding child terms until optimal semantic granularity was achieved.

3. Assigning relationships among CMO concepts

The structural foundation of CMO is formally a hierarchical tree structure, with a root value and subtrees of child nodes with a parent node. We assigned is_a relationship between CMO terms by following process. When there were terms that seem to be in a relationship between subordinates and superiors, it was determined that they were in is_a relationship by three medical doctors and medical informatics researchers.

During the process of developing hierarchical relationships, identified CMO terms have been modified as considering and reflecting a superordinate term. We called these terms as post-coordinated CMO terms. For instance, *Result of Physical Examination* in the first-level CMO term *Finding* has *Breast* as a child term. In this hierarchical structure, *Breast* means a result of physical

examination on breast, not anatomical structure of breast. We changed CMO term name from *Breast* to *Result of Physical Examination on Breast*.

4. Development of CMO properties

We created two CMO properties, synonyms, and definitions for each CMO term by referencing UMLS Meta-thesaurus and Wikipedia. UMLS has CUI to identify each unique concept and the terms having the same CUI can be grouped together as they are semantically equivalent [36].

As using UMLS CUI, we found synonyms, which were flagged in the relationship (RELA='same_as' or 'possibly_equivalent_to') column of the MRREL table and in the Term Type (TTY='SY') column of the MRCONSO table. We also found definitions, which were flagged in the definition (DEF) column of the MRDEF table. For definitions of CMO terms, which are not assigned any CUIs, two medical doctors were examined CMO terms and clarified them as considering a superordinate term manually.

For CMO unassigned UMLS CUI, we used Wikipedia or manually descripted to create synonyms, and definitions by medical doctors as experts. Especially, for the synonyms of the post-coordinated CMO term, we allowed combined type of synonyms with multiple UMLS concepts. For instance, *Result of Physical Examination on Breast* has *Physical Exam Result^Breast anatomy* or *Physical Exam^Breast structure*.

5. Evaluation scheme

We used two clinical document sets; (1) six documents from HL7 templates including Operation Note (2009), Consultation Note (2008), Discharge Summary (2009), History and Physical (2008), Procedure Note (2010), and Progress Note (2010), (2) 25 documents, which are five documents including Admission Note, Outpatient Note, Discharge Note, Emergency Note, and Operation Note from each five Korean hospitals including Seoul National University Hospital, Pusan National University Hospital, Ajou University Hospital, Chonnam National University Hospital, and Gachon University Gil Hospital, which have 126 and 55 CDEs, respectively.

For evaluating the suitability of CMO to facilitate classify and integrate CDEs, we first conducted CMO annotation for the extracted CDEs from two clinical document sets. The CMO annotation was performed by two independent nurses as considering annotation with most granular terms in CMO (if possible). Annotation with multiple CMO concepts for each CDE was allowed. Two administrators of medical records separately validated above two CMO annotation sets. To improve accuracy of CMO annotation, at last two medical informatics researchers confirmed the above four CMO annotation sets and rated the coverage of CMO following categories: adequate, too broad, and too specific.

To exam whether CMO can be one of classifications for similar clinical documents, we applied the Jaccard similarity index to estimate the similarity between each clinical document from two sets $J=|A \cap B|/|A \cup B|$, where A and

B are the numbers of CMO concepts of the corresponding clinical documents. We also examined whether one kind of clinical documents, personal health record (PHR) also can be classified by CMO.

RESULTS

CMO concepts

CMO has Clinical data element as a root term with four first-level classes. Total number of CMO concepts is 209. Finding has the largest number of child terms (n=78) among other first-level classes. Table 3-1 lists statistics of CMO concepts for each level under the first-level classes.

Table 3-1 Statistics of the CMO

First-level class	CMO level	No. of child terms for each level in each class	No. of child terms for each first-level class (%)
Description	1	1	62 (30.1)
	2	10	
	3	10	
	4	41	
Event	1	1	13 (6.3)
	2	10	
	3	2	
Finding	1	1	76 (35.4)
	2	16	
	3	34	
	4	22	
	5	3	
Procedure	1	1	58 (28.2)
	2	16	
	3	32	
	4	9	
Total			209

CMO provides 255 synonyms for 73 (35.4%) CMO terms, and 261 definitions for 206 (100%) CMO terms. Figure 3-1 shows specific statistics of CMO

properties. 102 (49.5%) CMO terms were matched to UMLS preferred terms. The main characteristic of UMLS unmatched CMO concepts was that they were the post-coordinated CMO term or too specific such as Medication History for Skin and Nero Exam on Cerebellum.

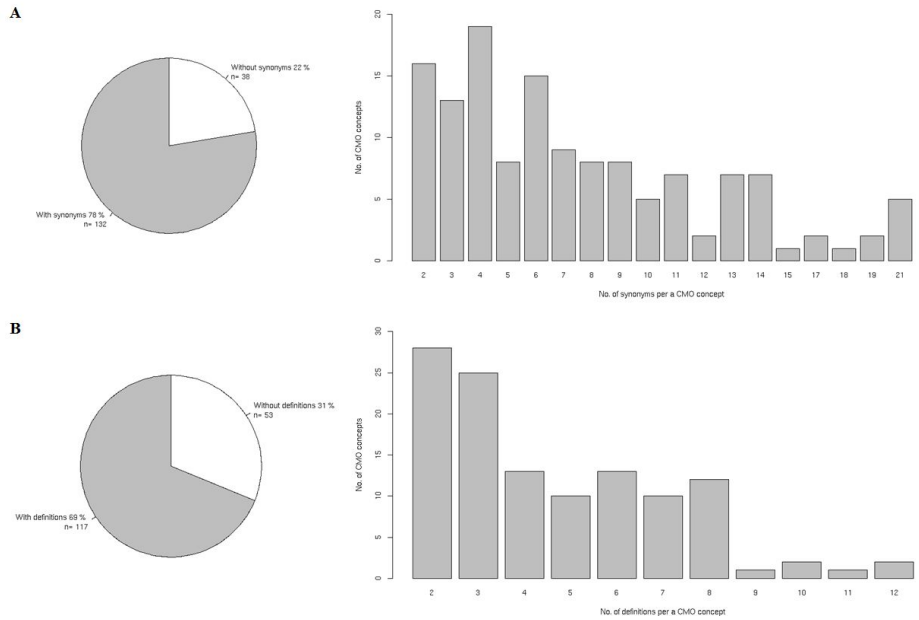

Figure 3-1. Distribution of CMO concepts with synonyms and definitions. (A) Distribution of CMO concepts with synonyms. Seventy-eight percent of CMO concepts have synonyms. (B) Distribution of CMO concepts with definitions. Sixty-nine percent of CMO concepts have definitions.

CMO web service

To help convenient access to CMO, we developed the CMO Browser, which provides CMO ID, preferred term, relative properties such as synonyms, definitions, parent term, and UMLS CUI (see Figure 3-2). We can explore

CMO terms and its properties as clicking CMO term in the hierarchical tree structure on the left side. CMO terms also can be searched through auto-completion function as entering particular terms in the text box next to 'Jump to'. The number in brackets next to the CMO term refers to the number of child terms.

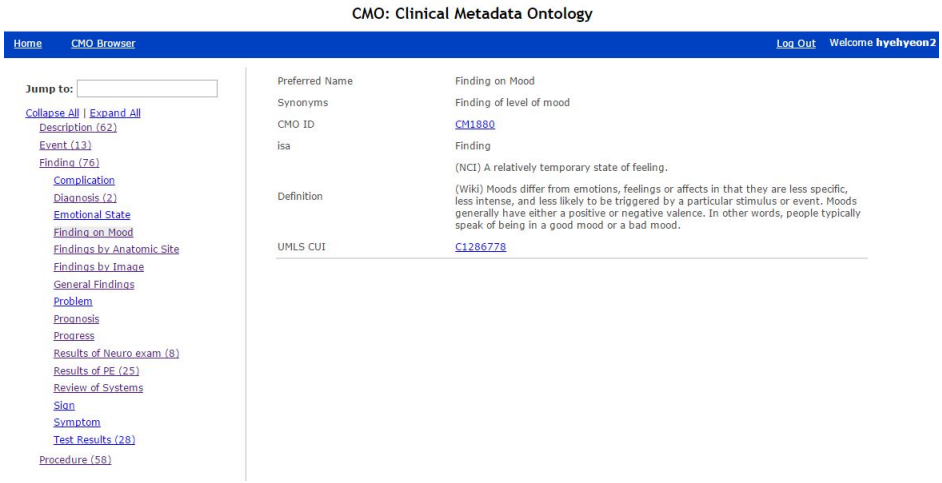

Figure 3-2. Web-based CMO Browser.

The abbreviation with three words in brackets in CMO definition indicates the source of definition, which are controlled vocabularies in UMLS. Table 3-2 lists the full names of sources of definition for each CMO term. A preferred term and a CMO ID are essential items, and the other properties are optional, which are provided when they have values. A UMLS CUI has URL link information to connect UMLS information.

Table 3-2 Full names of definition sources. UMLS is a compendium of many controlled vocabularies in the biomedical sciences. Among 177 controlled vocabularies, 11 controlled vocabularies were included for source of definitions for CMO concepts.

Abbreviation for source of CMO definition	Full names of controlled vocabularies in UMLS
AOT	Authorized Osteopathic Thesaurus (2003)
CHV	Consumer Health Vocabulary (2011)
CSP	CRISP Thesaurus (2006)
FMA	Foundational Model of Anatomy Ontology (v3.1)
GO	Gene Ontology (2010)
HL7V3.0	HL7 Vocabulary Version 3.0, (2011)
MEDLINEPLUS	MedlinePlus Health Topics (2011)
MSH	Medical Subject Headings, (2011 – 2013)
NCI	NCI Thesaurus (2013)
SNOMEDCT_US	US Edition of SNOMED CT (2011)
UWDA	University of Washington Digital Anatomist (v1.7.3)

Evaluation results

All extracted CDEs from two clinical document sets were annotated with CMO concepts. As observing how three categories for rating CMO coverage were classified, we can be clear for the definition of each category; 1) *too broad* means that first-level terms or general terms in second-level terms are used, 2) *too specific* means that few terminal node terms, having low frequency annotation are used, and 3) the rest CDEs were annotated adequately as belonging to *adequate* category.

Table 3-3 lists CMO coverage in two clinical document sets by rating three categories. 84.1%, 93.6% and 91.8% of CMO annotations in HL7 6 templates, 25 common documents and both were rated as adequate, respectively. CMO

could not cover specifically for too detailed CDEs such as '*estimated blood loss specify in procedure'*, and thus this kinds of CDEs were annotated with broad CMO concepts.

Table 3-3 CMO coverage in two clinical documents sets.

	HL7 6 templates (%)	25 common documents from 5 hospitals (%)	Total (%)
Adequate	106 (84.1)	523 (93.6)	629 (91.8)
Too broad	17 (13.5)	16 (2.9)	33 (4.8)
Too specific	3 (2.4)	20 (3.6)	23 (3.4)
Total	126	559	685

Figure 3-3 shows another evaluation result from the Jaccard similarity index of clinical documents by calculating the rate of the commonly annotated CMO concepts. Similar documents are collected such as operation note from various medical institutions, and documents from the same institutions are collected such as admission, discharge documents from Chonnam National University Hospital.

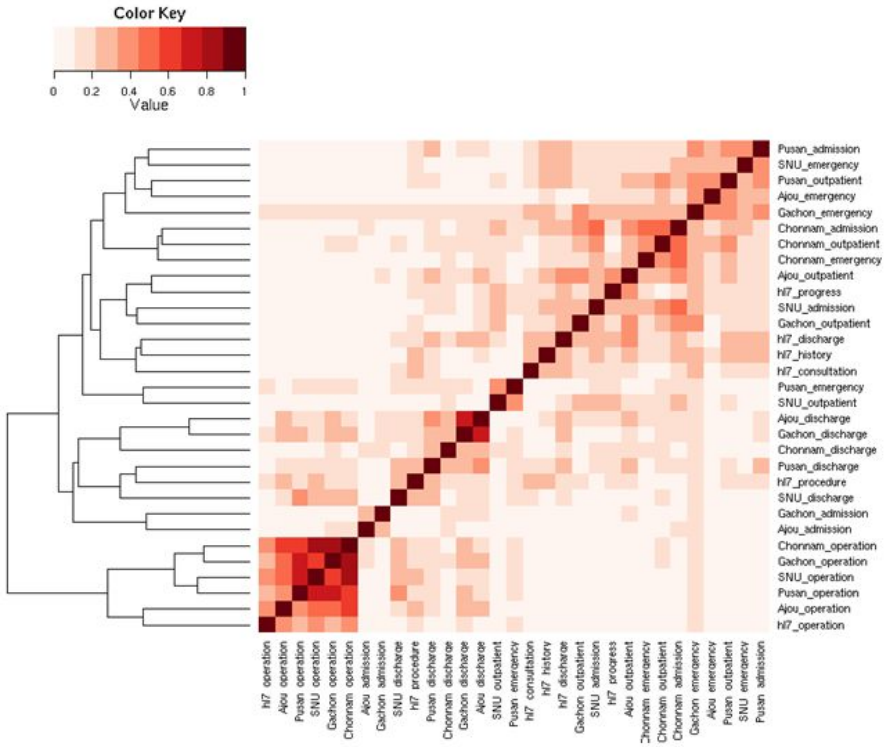

Figure 3-3. Jaccard similarity index among clinical documents.

Additionally we examined whether one kind of clinical documents, PHR also can be classified by CMO. Representative PHR model HL7 CCD and/or ASTM CCR, consisting of 13 sections were used to match with CMO concepts. Table 3-4 lists how 13 sections of PHR models are match to CMO concepts with its hierarchical structure having '|' as a delimiter.

Table 3-4 Three Sections of HL7 CCD and ASTM CCR mapped to CMO

CCD/CCR sections	CMO terms with hierarchical structure
Advance Directives	Description|Advance Directives
Alerts	Description|Alerts
Payers	Description|Demographics|Payor
Encounter	Description|Encounter
Immunization	Description|History of Immunization
Family History	Description|Past Medical History|Family History
Medications	Description|Past Medical History|Medication History
Social History	Description|Past Medical History|Social History
Problems	Finding|Problem
Results	Finding|Test Results
Functional Status	Finding|Test Results|Function Test Results
Vital Signs	Finding|Test Results|Vital Signs
Procedures	Procedure

DISCUSSION

Increasing use of CDE in clinical documents led us to develop CMO for the classification, annotation, and proper organization of CDEs in clinical use. For representing the semantics of CDEs for clinical documents, we adopted the concept of ISO/IEC 11179-based metadata to have the detailed descriptions of data held in publicly assessable data sets with its concepts. Through the proposed model, unambiguous semantics of all these components is formally defined. In this way, accurate common understanding and management of CDEs and reuse of these components are facilitated.

CMO is used to enable retrieval of CDE for multiple purposes including clinical research and sharing clinical data across organizations. In addition, CMO supports precise and comprehensive semantic annotation with UMLS as it has properties including synonyms and definitions for each CMO term by using UMLS. In other words, CMO concepts annotated with UMLS CUIs provide semantically rich content, since it is interoperable with the existing biomedical ontologies. In a broad sense, the meaning that CMO is adopted the UMLS preferred terms is that we consider CMO as essential subset of UMLS in the clinical medicine domain to classify CDE. Massive UMLS is rather inconvenient to use in practical way and has many unused terms, but CMO provides usable and practical terms.

Hierarchical structured CMO also enables further the navigation inside the categories of DEs in clinical documents. A manual evaluation on two different clinical document sets identified higher percentage of CMO mapping rate in

Finding class in 25 common documents. It seems that practical examples were represented more observational result related metadata in 25 common documents, unlike virtual items in HL7 templates. We also compared coverage of each level in each CMO class to verify how well CMO can be used for representing the level of detail with one assumption that if CDE is annotated with terminal concept, it means that the CDE is represented more specifically. We found that the 25 common documents contained more specific contents as three times more terminal terms were used.

Since CDEs have been used to develop CRFs and clinical documents, CMO can be one of informative classification schemes as classifying CDEs for clinical data in clinical documents or CRFs. In other words, researchers can retrieval and access these clinical documents or CRFs they needed by using CMO.

Further expansion of CMO concept and improvement of the CMO Browser are planned. First, we will improve the CMO term coverage with its contents and granularity as only less than half CMO concepts were used to annotate in our evaluation, for example 51 (25.5%) for HL7 6 templates, and 80 CMO terms (40%) for 25 common documents were used.

In summary, CMO as 1) a classification scheme for CDEs for clinical documents and CRFs, 2) an integration tool for CDEs from a diversity of clinical documents and CRFs, 3) a proper clinical data-organization scheme for CDEs for developing clinical information systems including PHRs, and 4) a component ontology expendably connected to other healthcare data domains

such as personal lifelog data, which is supported by MELLO [39], and personal genomic data, which is supported by Health Avatar Project [40].

REFERENCES

1. Richesson RL., Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007; 14(6):687-696.

2. Park YR., Kim HH, An EY, Kim HH, Kim JH, et al. Establishing semantic interoperability in the course of clinical document exchange using international standard for metadata registry. J Korean Med Assoc. 2012 Aug; 55(8):729-740.

3. Mohanty SK, Mistry AT, Amin W, Parwani AV, Pople AK, et al. The development and deployment of Common Data Elements for tissue banks for translational research in cancer–an emerging standard based approach for the Mesothelioma Virtual Tissue Bank. BMC cancer. 2008 Dec; 8(1):91.

4. Groft SC, Rubinstein YR. New and evolving rare diseases research programs at the National Institutes of Health. Public Health Genomics. 2013 Feb; 16(6):259–67.

5. Common Data Element (CDE) Resource Portal Website. https://www.nlm.nih.gov/cde/. Accessed Feb. 15, 2016.

6. Saver JL, Warach S, Janis S, Odenkirchen J, Becker K, et al. Standardizing the structure of stroke clinical and epidemiologic research data: the National Institute of Neurological Disorders and Stroke (NINDS) Stroke Common Data Element (CDE) project.

Stroke. 2012 Apr; 43(4):967-73.

7. Loring DW, Lowenstein DH, Barbaro NM, Fureman BE, Odenkirchen J, et al. Common data elements in epilepsy research: development and implementation of the NINDS epilepsy CDE project. Epilepsia. 2011 Jun; 52(6), 1186-1191.

8. Hachinski V, Iadecola C, Petersen RC, Breteler MM, Nyenhuis DL, et al. National Institute of Neurological Disorders and Stroke–Canadian stroke network vascular cognitive impairment harmonization standards. Stroke. 2006 Sep; 37(9):2220-41.

9. Grinnon ST, Miller K, Marler JR, Lu Y, Stout A, et al. National institute of neurological disorders and stroke common data element project–approach and methods. Clinical Trials. 2012 Jun;9(3):322-9.

10. Biering-Sørensen F, Alai S, Anderson K, Charlifue S, Chen Y, et al. Common data elements for spinal cord injury clinical research: a National Institute for Neurological Disorders and Stroke project. Spinal Cord. 2015 Apr; 53(4):265-77.

11. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc. 2011 May; 18(3):341-6.

12. Lin CH, Wu NY, Liou DM. A multi-technique approach to bridge electronic case report form design and data standard adoption. J Biomed Inform. 2015 Feb; 53:49-57.

13. NCI caDSR: Cancer Data Standards Registry and Repository CDE Browser Website. https://cdebrowser.nci.nih.gov/CDEBrowser/. Accessed Feb. 15, 2016.

14. Parkinson's Disease Biomarkers Program. National Institute of Neurological Disorders and Stroke Website. https://pdbp.ninds.nih.gov/. Accessed Feb. 12, 2016.

15. Roozenbeek B, Maas AI, Menon DK. Changing patterns in the epidemiology of traumatic brain injury. Nat Rev Neurol. 2013 Apr; 9(4):231-6.

16. Patel AA, Kajdacsy-Balla A, Berman JJ, Bosland M, Datta MW, et al. The development of common data elements for a multi-institute prostate cancer tissue bank: the Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. BMC Cancer. 2005 Aug; 21(5):108.

17. Ghitza UE, Gore-Langton RE, Lindblad R, Tai B. NIDA clinical trials network common data elements initiative: advancing big-data addictive-disorders research. Front Psychiatry. 2015 Mar; 3(6):33.

18. Pathak J, Wang J, Kashyap S, Basford M, Li R, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc. 2011 Jul-Aug; 18(4):376-86.

19. Park YR, Yoon YJ, Kim HH, Kim JH. Establishing semantic interoperability of biomedical metadata registries using extended

semantic relationships. Stud Health Technol Inform (Medinfo). 2013; 192:618-21.

20. Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. Methods Inf Med. 2006; 45(6):594-601.

21. NINDS Common Data Elements Website. https://commondataelements.ninds.nih.gov/. Accessed Feb. 15, 2016.

22. Ku HS, Kim S, Kim H, Chung HJ, Park YR, et al. DialysisNet: Application for Integrating and Management Data Sources of Hemodialysis Information by Continuity of Care Record. Healthc Inform Res. 2014 Apr; 20(2):145-51.

23. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford). 2016 Feb; 2016.

24. Dugas M, Dugas-Breit S. Integrated data management for clinical studies: automatic transformation of data models with semantic annotations for principal investigators, data managers and statisticians. PLoS One. 2014 Feb; 9(2):e90492.

25. ISO/IEC 11179. International Standard, International Electrotechnical Commission, Information technology — Metadata registries (MDR) — Part 3:Registry metamodel and basic attributes. https://webstore.iec.ch/preview/info_isoiec11179-

3%7Bed3.0%7Den.pdf, Publication date April 10, 2006.

26. Iberson-Hurst D. The CDISC Operational Data Model: Ready to Roll? Appl Clin Trials. 2004 Jul; 11(7):48–53

27. Lin CH, Fann YC, Liou DM. An exploratory study using an openEHR 2-level modeling approach to represent common data elements. J Am Med Inform Assoc. 2016 Jan;

28. Garde S, Hovenga E, Buck J, Knaup P. Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing. Int J Med Inform. 2007 Dec; 76 Suppl 3:S334-41.

29. Späth MB, Grimson J. Applying the archetype approach to the database of a biobank information management system. Int J Med Inform. 2011 Mar; 80(3):205-26.

30. Kuhn T, Basch P, Barr M, Yackel T. Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. Ann Intern Med. 2015 Feb; 162(4):301-3.

31. Richesson RL, DuLong D, Luigi Sison MSM, Goossen W, Huang W, et al. Common Data Elements for Clinical Documentation and Secondary Use: Diabe-DS Proof-of-Concept for “Collect Once, Use Many Times”. http://wiki.hl7.org/images/8/85/Whitepaper_Diabe_DS_Summary_of_Progress_and_Findings_Nov_2011.pdf Publication date Nov 30,

2011.

32. Park YR, Yoon YJ, Jang TH, Seo HJ, Kim JH. CCR+: Metadata Based Extended Personal Health Record Data Model Interoperable with the ASTM CCR Standard. Healthc Inform Res. 2014 Jan; 20(1):39-44.

33. Park YR, Kim JH. Metadata registry and management system based on ISO 11179 for Cancer Clinical Trials Information System. AMIA Annu Symp Proc 2006; 1056.

34. Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform. 2013 Oct; 46(5):784-94.

35. Min H, Ohira R, Collins MA, Bondy J, Avis NE, et al. Sharing behavioral data through a grid infrastructure using data standards. J Am Med Inform Assoc. 2014 Jul-Aug; 21(4):642-9.

36. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006 Jan-Feb; 13(1):30-9.

37. Obrst, L. Ontological architectures. Theory and Applications of Ontology: Computer Applications. Netherlands: Springer; 2010 ISBN: 978-90-481-8847-5.

38. Xu L, Furlotte N, Lin Y, Heinrich K, Berry MW, et al. Functional cohesion of gene sets determined by latent semantic indexing of

PubMed abstracts. PLoS One. 2011 Apr; 6(4):e18851.

39. Kim HH, Lee SY, Baik SY, Kim JH. MELLO: Medical lifelog ontology for data terms from self-tracking and lifelog devices. Int J Med Inform. 2015 Dec; 84(12):1099-110.

40. Kim JH. Health avatar: an informatics platform for personal and private big data. Health Inform Res. 2014 Jan; 20(1):1-2.

국문 초록

서론: 데이터 표준화는 다양한 연구들 간의 이해와 데이터 공유에 매우 중요한 역할을 한다. 이를 위해 여러 가지 방안이 마련되었는데 크게 표준 데이터 모델 기반의 하향식 방식과 표준 용어체계 기반의 의미론적 표준화를 지향한 상향식 방식으로 분류되고 있다. 하향식 방식은 정교히 개발된 상위모델을 제안하고 이를 준수함으로 명확한 의미전달과 표준화된 데이터 생성이 가능하다. 이는 데이터의 종합적이고 포괄적인 표현의 효과를 갖지만, 실제 운영과 관련해서는 상위 모델을 익히고 이를 준수하는 방향으로 변환하고 표현하는 것에 시간과 에너지 소비가 많으며, 모델에서 표현하지 못하는 데이터에 대해서는 모델의 버전 업그레이드를 기다려야 하는 등의 진행 속도와 운영 및 관리에 있어서 문제점이 제기되고 있다. 반면 상향식 방식은 복잡한 데이터 구조보다는 데이터의 정확한 의미를 표현하는 것에 의의를 두고, 명확한 정의와 데이터간의 상·하위 관계 및 분류 정보를 포함하여 매우 유용하다. 그러나 표준 용어체계 별로 post-coordination 방법론이 다르며, 새로 추가되는 개념 등에 대해서 표준별로 상이하게 표현될 수 있음으로 데이터 통합 및 활용에 어려움이 있다. 위의 2 가지 방식의 한계를 해결하는 middle-out 방식의 국제표준 ISO/IEC 11179 메타데이터 저장소 구성 및 관리 (Metadata

Registry, MDR)에 기반한 공통 데이터 엘리먼트 (Common Data Element, CDE)를 활용하는 방식을 채택하였다. MDR 기반 CDE 는 국제표준에서 정의한 데이터의 의미와 표현형에 대해 구체적인 구조화를 기반으로 잘 정의된 기본 데이터로써 임상문서를 구성하는 문항을 CDE 로 정의하고 이를 활용함으로 임상문서의 의미론적 상호 운용성을 지원한다. 임상문서는 개인건강과 관련된 임상정보를 수집하는 필수적인 도구이다. 임상문서의 구성은 기본적인 문항과 답항의 조합형이 열거된 것뿐만 아니라 문·답항들의 관계 및 문서의 성격을 포함하는 복잡한 관계로 되어있다. 이를 모두 표현하기 위해 CDE 뿐만 아니라 CDE 간의 관계들의 의미론적 정의가 필요하다.

그러나, CDE 간의 관계를 설명할 수 있는 구조가 없는 메타데이터 저장소 표준의 구조적인 한계가 CDE 로 임상문서를 구성하거나, CDE 기반으로 생성된 문서를 해석하는데 큰 장애물이 되고 있다. 비록 우리의 이전 연구에서 간단한 CDE 간의 관계를 정의하여 해결점을 제안하였지만 여전히 너무 단순하고 포괄적이라는 한계로 새롭게 추가적인 의미론적 관계 정의가 필요하게 되었다.

한편, 의료진-환자 혹은 의료진간의 임상 정보 공유를 위해 임상문서의 공유는 필수적이다. 하지만 임상문서 공유 전에 반드시 임상문서 데이터의 완전성과 오류 여부를 확인하는 임상문서 검증 과정이 수행되어야 한다. 임상문서 검증 항목 및 구조화된 검증

과정이 정립된 것이 없기 때문에 MDR 기반 CDE 를 포함하는 복잡한 임상문서를 검증하기 위한 방법론이 필요하게 되었다.

또한, 메타데이터 저장소에 저장된 방대한 CDE 중에서 적절한 CDE 를 찾아 임상문서에 활용하는 것은 효과적인 의미론적 데이터 교환을 위해 필수적이다. 이에 데이터 엘리먼트를 분류하고 찾기 위한 온톨로지 개발이 필요하게 되었다. .

방법: 우리는 데이터 엘리먼트간의 관계를 이해하기 위해 임상 실제 현장에서 활용되고 있는 공통 데이터 엘리먼트들을 검토하였다. 그리고 리뷰 과정으로 발견한 공통된 특징 즉 확장 관계로 새롭게 정의되기 필요한 경우들을 유스 케이스 시나리오를 작성함으로 공통 데이터 엘리먼트 관계들을 설계하였다. 한편 임상 문서 검증 과정 개발을 위해, 임상문서의 복잡성을 복잡한 메타데이터 기반의 임상문서로 정의하였다. 정의된 의미론적 메타데이터와 그 관계 기반의 임상문서를 검증하는 과정과 검증 속성들을 구체화하였다.

한편, 메타데이터 분류 및 통합을 위한 임상 메타데이터 온톨로지(CMO) 개발을 위해 다섯 단계의 구성단계를 가지는 공식적인 일반 온톨로지 개발 방법론을 차용하여 개발하였다; (1) 온톨로지 범위 정의, (2) 온톨로지를 구성하는 개념들의 식별, (3) 개념들간의 계층적 구조의 정의, (4) 속성 (예, 동의어, 대표어, 정의) 정의, (5) 온톨로지 평가.

결과: 기존 연구에서 개발된 Composite common data element(cCDE) 을 확장하여 3 개의 확장형으로 개발하였다. 관찰 임상 데이터에서 테이블 형식의 표현법으로 반복적으로 사용되는 데이터 형식을 Repeated cCDE 로 정의하였고, 외부 지식데이터 표현과 데이터 모델의 표현을 가능케 하는 Dictionary cCDE, Template cCDE 를 정의하였다. 또한 CDE 간 혹은 cCDE 을 구성하는 sub-CDE 간의 관계를 설명하기 위한 4 가지 제약조건 (Constraints)인 Dependent, Operated, Ordered 그리고 Required 를 정의하였다. 또한 여러 데이터 타입을 허용하는 공통 데이터 엘리먼트의 새로운 유형인 Hybrid 관계를 정의하였다.

한편, 복잡한 임상문서 구조를 메타데이터 저장소 기반의 공통 데이터 엘리먼트와 그 관계들로 표현함으로 이를 기반으로 구문론적 그리고 의미론적 임상문서 검증 과정을 개발하였다. 구문론적 검증은 XML 스키마 기반 검증, Template composite data element 기반 검증 이렇게 2 가지로 구성하였으며, 의미론적 검증은 general data element 검증 과정이전에 만약 데이터가 compoiste data element 과 연관되었다면 외부 데이터 기반의 검증을 지원하는 Dictionary composite data element 기반 검증과정과 Composite data element 내부에 저장된 제약조건 기반 검증과정을 우선적으로 수행하는 것으로 구성하고 일반 데이터 엘리먼트 검증에서는 값 부분의 규칙 정보를 포함하는 valuse domain 기반 검증과정을

수행하는 것으로 구성하였다. 실제 운영 가능한 검사결과 예제 데이터를 적용함으로 검증 과정의 실현 가능성을 평가하였다.

트리 구조 기반의 CMO 은 첫번째 상위 개념 (Description, Event, Finding, Procedure)으로 구성되어 전체 200 개의 용어로 구성된다. 151 CMO 용어 (76%)에 대해 1060 개의 동의어를 가지며, 137 CMO 용어 (69%)에 대해 400 개의 정의를 갖는다. 웹 기반의 CMO 브라우저와 CMO 매칭된 BMeSH 데이터 엘리먼트 브라우저를 제공함으로 CMO 의 편리한 접근과 실제 임상문서에 매칭되는 CMO 용어를 제공하여 CMO 의 이해를 돕는 것을 가능하게 한다 (http://www.snubi.org/software/cmo). CMO 는 (1) 임상문서의 데이터 엘리먼트들의 분류체계이며, (2) 다양한 임상문서의 데이터 엘리먼트들의 통합 도구이며, (3) PHR 과 같은 임상 정보 시스템을 개발에 활용 가능한 적합한 임상 데이터 구성 스키마이며, (4) MELLO 에 의해 지원되는 라이프로그, 개인 유전체 데이터와 같이 헬스아바타 프로젝트에서 지원되는 다양한 의료 데이터 도메인과 연결되는 확장형 온톨로지이다.

결론: 복잡한 임상문서를 메타데이터 저장소 기반의 공통 데이터 엘리먼트와 그 관계들로 표현함으로 의미론적 이해를 돕고 복잡한 임상 데이터 표현을 구조적으로 가능하게 함으로 상호 운용성 증진에 기여하였다. 크게 2 가지 관점에서 장점을 나타낼 수 있는데, 하나는 문서 레벨의 데이터 즉, 문서의 구조적 의미론적 관계 및

정의 관련 정보를 잃지 않고 데이터 엘리먼트 레벨로 표현하고 저장함으로써 의미론적 상호 운용성을 보장한다는 것이고, 둘째는 문서의 메타데이터화가 가능함으로 특히 composite 관계 및 template composite 관계 적용을 하게 됨으로 데이터 통합 및 데이터 모델 변환 등의 기능도 가능해졌다는 것이다. 같은 의료기관에서 사용되는 다양한 임상문서간 혹은 서로 다른 기관에서 사용되는 다양한 임상문서를 메타데이터와 그 관계들로 정의하고 임상문서를 통합하는 연구를 통해 메타데이터를 통한 데이터 재사용성과 통합의 가능성을 보임으로 메타데이터의 효과를 검증하였다. 또한 개발된 메타데이터 관계들을 기반으로 임상문서를 검증하는 과정을 구체화함으로 데이터 유효성과 무결성 완성에 기여하였다. 더불어 임상 데이터 엘리먼트 온톨로지를 개발함으로 메타데이터 저장소의 방대한 데이터 엘리먼트 중 적합한 데이터 엘리먼트를 찾는 것을 가능하게 되었습니다. 설계기반의 예비결과만을 가지는 본 연구의 한계에도 불구하고 향후 진행되는 실제 시스템 개발 및 운영에 기초 자료가 되어 메타데이터 저장소 기반의 데이터 표현과 공유 그리고 데이터 검증이 가능함으로 의미론적 상호운용성에 크게 기여할 것을 기대한다.

주요어 : 임상문서 교류, 메타데이터 레지스트리, 공통 데이터 엘리먼트, 의미론적 상호운용성

학 번 : 2010-30607