Technical advance articles Composite CDE: modeling composite relationships between common data elements for representing complex clinical data

CURRENT STATUS: UNDER REVISION

BMC Medical Informatics and Decision Making • BMC Series

Hye Hyeon Kim
Seoul National University College of Medicine

Yu Rang Park
Yonsei University College of Medicine

Ju Han Kim  juhan@snu.ac.kr
Seoul National University
Corresponding Author
ORCID: 0000-0003-1522-9038

DOI:
10.21203/rs.2.11646/v1

SUBJECT AREAS
Medical Informatics

KEYWORDS
Common Data Elements, Semantic Interoperability, Semantic Relationship, Metadata Registry
Abstract

Background

Semantic interoperability is essential for improving data quality and sharing. The ISO/IEC 11179 Metadata Registry (MDR) standard has been highlighted as a solution for standardizing and registering clinical data elements (DEs). However, the standard model has both structural and semantic limitations, and the number of DEs continues to increase due to poor term reusability. Semantic types and constraints are lacking for comprehensively describing and evaluating DEs on real-world clinical documents.

Methods

We addressed these limitations by defining three new types of semantic relationship (dependency, composite, and variable) in our previous studies. The present study further extended semantic types (hybrid atomic and repeated and dictionary composite common data elements CDEs) with four constraints: ordered, operated, required, and dependent.

For evaluation, we extracted all atomic and composite CDEs from five major clinical documents from five teaching hospitals in Korea. Metadata reusability and semantic interoperability in real clinical settings were comprehensively evaluated by applying the CDEs with our extended semantic types and constraints.

Results

All of the CDEs (n=1142) extracted from the 25 clinical documents were successfully integrated with a very high CDE reuse ratio (46.9%) into 606 CDEs (259 atomic and 20 unique composite CDEs), which improved the semantic integrity and interoperability without any semantic loss. Moreover, the most complex data structures from two CDE projects were successfully encoded with rich semantics and semantic integrity.

Conclusion

MDR-based extended semantic types and constraints can facilitate comprehensive
representation of clinical documents with rich semantics and improved semantic interoperability without semantic loss.

**Background**

Data harmonization and interoperability are essential for advancing biomedical research. These features can be achieved by representing clinical data in a standard format, and they are crucial for facilitating understanding and sharing data across diverse translational studies [1, 2]. A common data element (CDE) is defined as the fundamental unit of data containing information with a clear conceptualized meaning, and this representation is considered the correct approach for standardizing data and improving data quality and efficiency.

The ISO/IEC 11179 Metadata Registry (MDR) standard describes a method of standardizing and registering data elements (DEs) to make them understandable and shareable between studies and institutions. An MDR-based CDE collects data uniformly, allowing data interoperability between clinical studies since they are specified based on a metadata model that consists of specified sets of attributes that include delineating the definition, identification, representation, classification, and permissible values [3–5].

CDEs are increasingly being used by clinical researchers and in trials for harmonizing data collected across diverse studies. The use of standardized CDEs provides various benefits to investigators, including (1) rapid and efficient study start-up by enabling access to defined CDEs and case report forms (CRFs), and (2) enriched data sharing and data aggregation using standard definitions and forms [6].

The use of CDEs has recently been extended to clinical practice by using standardized CDEs for representing the clinical information in electronic health records (EHRs). For example, Newton *et al.* included phenotype data in EHRs using CDEs in order to facilitate EHR-driven genomic studies [7]. The National Institutes of Health have developed ISO/IEC
11179 MDR-based CDEs that provide a controlled terminology for data descriptors, and they encouraged clinical researchers to use CDEs in order to facilitate data harmonization [5]. CDEs have been adopted in numerous clinical domains including cancer, stroke, epilepsy, rare diseases, emergency medicine, and radiology for patient care, and research. Utilizing CDEs will facilitate secondary data use (i.e., ‘collect once and use many times’), which is an approach to standardization that spans silos in primary and secondary data use [8].

However, ISO/IEC 11179 MDR-based CDEs do not provide the ability to describe constraints on a CDE or relationships among different CDEs, instead merely focusing on single independent CDEs, which makes it difficult to either correctly compose or interpret CDEs on clinical documents [9–12]. Although the ISO 11179 standard describes a derived data element (DDE) [13] to detail the relationship between a DE, the rule controlling its derivation, and another DE from which it is derived, this approach is inherently limited by the DDE requiring one or more input DEs and itself becoming the output DE. For example, systolic blood pressure (SBP) and diastolic blood pressure (DBP) can be easily defined as two separate DEs annotated with standardized metadata conforming to the ISO/IEC 11179 MDR standard. However, these two DEs are only input DEs, and an output DE is needed as the DDE. Also, a constraint between the two DEs such as ‘SBP must be greater than DBP’ is usually described outside of the DEs because there is no straightforward for DEs to carry such information.

To address these challenges, in our previous study [9] we proposed three types of semantic relationships (variable, dependency, and composite relationships) representing semantic constraints or rules among multiple CDEs. These relationships can be described as follows: First, CDEs are in a variable relationship when they can be systematically derived from a base CDE by applying a standardized concept from a controlled vocabulary.
For example, the meanings of two CDEs for ‘normal value range of laboratory test, Albumin’ and ‘normal value range of laboratory test, Homocysteine’ are closely related, differing only in the laboratory test names of ‘Albumin’ and ‘Homocysteine.’ The variable relationship can systematically represent all of these variations as a single CDE, ‘DE: Normal value range of lab test $x,$’ by applying a controlled vocabulary such as LOINC. The variable relationship can therefore systematically reduce the number of CDEs required.

Second, a CDE is in a dependency relationship when the value of the CDE is determined by the value(s) of CDE(s). For example, the value of a certain CDE may be defined as the sum of the values of a set of CDEs in a questionnaire. Third, the composite relationship can be conveniently applied to integrate several interrelated CDEs into a composite CDE. For example, the medical history of a patient is likely to be more informative when body parts are correctly assigned, which can be achieved by grouping ‘DE: Body System for Medical History’ and ‘DE: Medical History Specify’ into the composite CDE of ‘DE: Medical History.’

However, we realized that our previous work supports relatively simple semantic relationships among CDEs and is not robust enough to cover many other specific challenges associated with CDEs used in clinical forms.

The present study further proposed extending semantic types (hybrid atomic and repeated and dictionary composite CDEs) and four semantic constraints (ordered, operated, required, and dependent) for correctly representing even more complex but essential semantic relationships between CDEs that are found in real-world clinical documents. We found useful patterns characterizing challenging cases that required further semantic definitions and descriptions as following 4 cases;

1.1 Data entries with multiple data types
A data type determines the type of data that can be entered and stored in a DE, and each DE contains only one data type [14]. However, we found that free-text-based data entry in many clinical documents stored in EHRs often allows multiple data types to be entered and stored in the same attribute. For example, a laboratory result for syphilis normally has a numeric data type that allows numeric values (e.g., ‘0.8’) as input, but this also often requires the entry of string or logical data such as ‘negative’ or ‘false’ as input. Sometimes creating two strictly separate CDEs for a laboratory result for syphilis (i.e., numeric and string) may cause greater confusion than if the data are harmonized and made interoperable, and sometimes it is better to allow either numeric or string data types in the same value domain. We created a value property (hybrid) to make it possible to ensure that conventionally multiple data types are available in the same CDE (i.e., numeric or string) in order to reduce confusion by explicitly defining the hybrid data type for CDEs.

1.2 Dictionary data entries

Data may refer to a particular controlled biomedical vocabulary for several reasons, such as adherence to standards, semantic enrichment for better understanding, and input validation for improving semantic integrity. A CDE referring to a controlled biomedical vocabulary was defined as being in a variable relationship in our previous study [9]. We extended the concept of the variable relationship to dictionary data entries in order to tightly link a set of CDEs with a real-world dictionary database via a controlled biomedical vocabulary using a ‘foreign key.’ This also ensured that the set of CDEs and tuples with rich attributes provided by the dictionary were linked with their proper data type definitions and value domains.
1.3 Tabular data entries with repeated data entry

Clinical data is frequently in a tabular format. A tabular data entry is an enclosed structure in which a composed set of DEs is repeated for repeated observations. For example, body weight and height may be measured for each patient when he/her visits for treatment. The set of data items such as body weight, height, and date of measurement should be collected both together and repeatedly. We created a value property (repeat) to ensure that the values that belong to the same set of CDEs are identified as such.

1.4 Required and derived data

Particular CDEs on a clinical document that are highly interrelated need to be defined by semantic constraints. For example, the value of a certain CDE that has a value other than null should be described by the required constraint. Derived values such as BMI (body mass index) can be automatically calculated from the values of body-weight and height CDEs.

Methods

2.1 DATA RESOURCE: 2 CDE PROJECTS AND 25 CLINICAL DOCUMENTS USED AT 5 TEACHING HOSPITALS

The National Institute of Neurological Disorders and Stroke (NINDS) CDE Project [15] is an ongoing effort to develop data standards for use in clinical research in neuroscience. It was initiated in 2006 to standardize data collection across neurological-disorder-related clinical studies funded by the NINDS. As at October 2016, the NINDS CDE project included 20 studies with 11,296 distinct CDEs. The NINDS CDEs are not fully compliant with ISO/IEC 11179, instead only providing simple DE descriptions and definitions. However, a part of
NIND CDEs that are registered in National Cancer Institute (NCI) cancer Data Standards Registry (caDSR) and reviewed by the NCI cancer Biomedical Informatics Grid project manager conforms fully with the ISO/IEC 11179 MDR standard. In the present study we used the part of the NINDS CDEs, which are 308 (3.1%) stroke and general CDEs of the NINDS that are registered in the caDSR. Selected CDEs within the context of their CRFs were explored for challenging cases requiring new semantic relationships.

The DialysisNet and Avatar Beans Project is a tablet- and phone-based mobile application developed by the Health Avatar Initiative [16]. The project started in 2013 and it has established clinical data standards for managing and harmonizing hemodialysis data across multiple medical institutions in Korea [17, 18]. This project aims to improve the management of chronic kidney disease and end-stage renal disease by using an integrated mobile application for data collection and documentation. The DialysisNet application was initially built upon 122 distinct hemodialysis-related CDEs based on CRFs from major four hemodialysis centers.

2.2 DESIGNATE KEY CONCEPTS

The CRFs and clinical documents from the two CDE projects incorporate all of the data collection items with CDEs. We first examined the CDEs to formalize the found 4 challenging cases. Figure 1 displays the formal relationship between atomic CDE (aCDE) and composite CDE (cCDE) with type-specific constraints. Since the core structure of a CDE is a name-value pair augmented by DE concept-domain and value-domain details, the aCDE is a single unambiguously described data item [18]. Our previous and simple-minded definition of cCDE as a set of interrelated aCDEs [9] was extended to include two new semantic relationships: dictionary and repeated cCDEs. We extracted aCDEs and cCDEs from five major clinical documents used at five teaching hospitals in Korea and applied by the extended the semantic types and constraints. We then mapped and integrated the
CDEs in order to comprehensively evaluate the metadata reusability and semantic interoperability in the clinical-practice setting.

2.3 EVALUATION SCHEME

For the purpose of evaluating the utility of the newly proposed semantic types and constraints, we collected 25 clinical documents used in clinical practice, comprising 5 documents covering admission notes, initial medical examination notes, discharge notes, emergency notes, and operation notes from each of 5 major teaching hospitals in Korea: Seoul National University Hospital, Ajou University Medical Center, Pusan National University Hospital, Gachon University Gil Hospital, and Chonnam National University Hospital [17]. The evaluation process consisted of the following three steps: CDE extraction, CDE integration, and construction of semantic relationships among the CDEs. We counted the numbers of CDEs generated in each step as a measure of the structural efficiency.

Results

3.1 OVERVIEW OF ALL TYPES OF SEMANTIC RELATIONSHIPS

To address the challenges described above, we defined aCDEs and cCDEs using three new semantic types (hybrid, dictionary, and repeated) and three new types of constraints (ordered, operated, and required) in addition to the existing two semantic relationships (dependent and variable relationships) defined in our previous study [9]. The newly defined composite semantic type replaced the old composite relationship that we defined previously [9].

Figure 1 displays aCDEs and cCDEs with their specific constraints. An aCDE can be constrained using variable and hybrid relationships by classifying them as variable aCDE and hybrid aCDE, respectively. The definition of cCDE as a set of interrelated aCDEs in our
previous study [9] was extended to include a clear definition, a separate identifier for 
reuse, and constraints among aCDEs inside a cCDE. CDEs can be classified into dictionary 
and repeated cCDEs. One of the existing semantic relationships, the dependent 
relationship in our previous study, was extended to four constraints: ordered, operated, 
required, and dependent. As shown in the lower-left box in Figure 1, the ordered constraint 
does not apply to an aCDE.

3.2 DATA ENTRIES WITH MULTIPLE DATA TYPES: Hybrid aCDE

A hybrid aCDE is a particular type of aCDE that allows a value domain with multiple (or 
hybrid) data types. Technically it includes several aCDEs having the same DE concept but 
different value domains. Figure 2A shows part of a hemodialysis report form from the 
DialysisNet and Avatar Beans Project. A time-tagged hybrid aCDE was applied to the Time 
attribute in a tabular data-entry format. The hybrid aCDE for Time (‘DE:47616 
Hemodialysis_Time_Hybrid_DE’) was derived from two aCDEs: ‘DE:43239 
Hemodialysis_Time_DE’ allowing a time data type, and ‘DE:47614 
Hemodialysis_Time_String_DE’ allowing an enumerated string data type supporting Finish 
and Start (Figure 2B). The hybrid aCDE can capture either a time or an enumerated string 
value, such as ‘DE:47616.

3.3 TABULAR DATA ENTRIES: Repeated cCDE

A repeated cCDE is a cCDE that captures data input multiple times in a tabular format. The 
definition of the repeated cCDE prevents the unnecessary creation of redundant CDEs. A 
repeated cCDE efficiently captures and displays changes in input values over a certain 
time span, as shown in Figure 2A. We first grouped eight aCDEs (i.e., DE:47616, 
DE:43340, DE:43197, DE:43195, DE:43155, DE:43092, DE:43372, and DE:43166) to create 
a cCDE, and then assigned them as a repeated relationship to create a repeated cCDE 
(‘DE:47575 Hemodialysis_Repeated_Composite_DE’) (Figure 3). As shown in Figure 2, 
DE :47616 is a hybrid aCDE contained in the repeated cCDE (DE:47575).
3.4 DICTIONARY DATA ENTRIES: Dictionary cCDE

Our previous study [9] defined a variable CDE as a CDE containing a variable that refers to a controlled biomedical vocabulary. Similarly, we defined a dictionary cCDE as a cCDE containing a variable aCDE with a variable that refers to the corresponding attribute as the primary key of a dictionary table. This approach provides a way to encode an entire dictionary table as well as a controlled vocabulary into a single dictionary cCDE, and thereby capture comprehensive biomedical knowledge from a database. A dictionary cCDE provides a useful means to apply relevant attributes of a dictionary database to constrain and validate input values to the dictionary cCDE.

Figure 4A displays a typical data-entry document for laboratory test results in a tabular format. The ‘Electrolyte Laboratory Tests’ form from ‘Recommended Labs for Stroke’ of the NINDS CDE project [19] consists of six attributes, including the laboratory test name, laboratory test result, unit of the laboratory test result, an indicator for whether the laboratory test result is abnormal, and another indicator for whether the laboratory test result is clinically significant when the laboratory test result is abnormal. Figure 4B shows a part of the structured NINDS ‘Electrolyte Laboratory Tests Dictionary’ reference table. The Unit of Result attribute supports multiple units that are delimited by ‘^’. The Normal Range attribute is also separated according to the Unit of Result and is represented in JSON (Javascript object notation)-type encoding.

A dictionary cCDE can systematically capture the entire ‘Electrolyte Laboratory Tests’ data-entry document ‘DE:47571Laboratory_Test_NINDS_Composite_DE,’ which is composed of six aCDEs (Figure 4C ) that include a variable aCDE for Test, ‘DE:43938 Laboratory_Finding_Test_Name_DE,’ which functions as the foreign key to refer to the primary key, and ‘Lab Test Name’ of the ‘Electrolyte Laboratory Tests Dictionary’ table (Figure 4B ).
Now that the dictionary cCDE (DE:47571) is related to the NINDS ‘Electrolyte Laboratory Tests’ dictionary table via the variable aCDE (DE:43938), it provides a means to evaluate the validity of an input value to Result and Units for Result for a Test ['Sodium (Na+)'] value of 138 mEq/L, with respect to the Normal Range (i.e., 135−145 mEq/L) provided by the dictionary table. The input value of Was test result abnormal? can also be input automatically using the biomedical knowledge provided by the dictionary table. Moreover, when the value of Was test result abnormal? (DE:47566) is ‘Abnormal,’ the value of If abnormal, Clinically Significant? (DE:44135) can automatically be constrained to contain a value other than null. This constraint can be encoded by a Dependent Rule, as shown in Figure 4C.

Figure 4C shows how a dictionary cCDE accompanied by its constraint rules are defined. For the two evaluation cases listed in Figure 4B, both a Dictionary Rule and a Dependent Rule are defined by symbolic logic (or pseudocode) with the accompanying Descriptions. A Dictionary Rule defines how to use biomedical knowledge contained in a dictionary table, and a Dependent Rule defines the interrelatedness of aCDEs in a cCDE by using dependent constraint relationships.

3.5 DERIVED DATA : Constraints

We defined four constraints that support the creation of a robust clinical document by specifying the interrelationship among many aCDEs. We defined four classes of operators: assignment, arithmetic, logical, and relational. Order can only be applied to aCDEs contained in a cCDE. However, the other tree constraints ( operated, required, and dependent) can be applied to independent aCDEs on a document and those contained in a cCDE (Figure 1). We created a symbolic logic with prefix notation [20] (Table 1 in the Supplementary Files) to describe the order of operations and to formulate constraints. More practical examples are shown in Figure 5 to demonstrate how constraints are applied.
to a *repeated* cCDE as well. The four constraints are described as follows:

Operated. Table 1A presents the standard BMI formula \[\text{BMI (in } \text{kg/m}^2) = \text{weight} / (\text{height} \times \text{height})\] in a prefix notation as \((/ \text{CDE30 CDE31 CDE31 100 100})\), where CDE30 and CDE31 represent *Body Weight Value* in kg and *Body Height Value* in cm, respectively. Both the ‘cm’ and ‘m’ units of measurements can be supported by applying an IF conditional statement to manage different units: \((\text{IF } (= \text{CDE31.unit_of_measure 'm'}) / (/ \text{CDE30 CDE31 CDE31 100 100}))\).

Required. A *Required* constraint applied to an aCDE means that the aCDE must have a value other than null. Table 1B lists the demographic information of a clinical document constraining ‘*Patient Age (CDE40)*’ and ‘*Gender (CDE41)*’ as *required* by the statement \((\text{Required CDE40 CDE41})\).

Dependent. It might be necessary to dynamically enable or disable a certain aCDE according to the value(s) of other aCDE(s). For example, a gender-specific CDE might only be applied to subjects of the applicable gender. Table 1C presents an example for checking whether a patient is a current \((\text{CDE20})\) or past \((\text{CDE21})\) smoker in order to obtain the age when tobacco use was started \((\text{CDE22})\). A nonsmoker can conveniently skip \(\text{CDE22}\) if \((= \text{CDE20 CDE21 'No'})\) by setting the value of CDE22 as null. In other words, a rule such as \((\text{IF } (\text{or } (!= \text{CDE20 'Yes'}) (!= \text{CDE21 'Yes'})) \text{CDE22 NULL})\) can be imposed. Another constraint can be imposed to check illogical input values such as \((= \text{CDE20 CDE21 'Yes'})\) if necessary.

Order. The ordering of aCDEs (especially in a cCDE) is important for certain conditions and contexts. CDEs in Table 1C can be ordered by a constraint statement such as \((\text{Ordered CDE20 CDE21 CDE22})\).

### 3.6 EVALUATION STUDY

To evaluate the usefulness of our newly extended composite semantic relationships, we
applied them to CDEs that were systematically extracted from five major clinical documents used at five teaching hospitals in Korea. The evaluation process consisted of the following steps: CDE extraction, CDE integration by using the new aCDEs and cCDEs, and semantic enrichment. We compared the number of CDEs extracted and integrated in each evaluation steps as a measure of the structural and semantic efficiency of DEs on clinical documents.

We first extracted 84, 48, 70, 83, and 37 CDEs from the following 5 clinical documents used at Hospital A: admission note, initial medical examination note, discharge note, emergency note, and operation note, respectively. We found that 95 (29.5%) of the 322 CDEs were reused in at least 2 of the 5 clinical documents, resulting in 227 unique aCDEs.

We then created clinically relevant cCDEs and applied semantic relationships to them. Of the 84 aCDEs extracted from admission notes at Hospital A, 55 were successfully captured by 10 created cCDEs. Finally, 16 cCDEs successfully captured 110 (48.5%) of the 227 unique CDEs, such that 133 (=16 + 117) CDEs (41.3%) were sufficient to represent the initial 322 CDEs extracted from the 5 clinical documents used at Hospital A (Table 2 in the Supplementary Files).

In the CDE extraction step, we found that applying CDE is an effective way to reduce redundant CDEs (22.2~37.9%) at each hospital. This means that there were many CDEs shared across the five different documents used at each hospital. We found that an even higher CDE reduction rate of 48.7% could be achieved by integrating the information for all five hospitals, which indicates that various CDEs were commonly used across the different hospitals. The CDE integration step involved integrating aCDEs into clinically relevant cCDEs to further structure the clinical documents, and then integrating the cCDEs across different clinical documents. For example, when a vital sign related cCDE contained three aCDEs (‘body weight,’ ‘body temperature,’ and ‘blood pressure’) and another vital
sign related cCDE contained an additional aCDE (‘description the reason of unstable vital sign’), we integrated this into a vital-sign cCDE comprising four aCDEs. The application of these three steps constantly decreased the number of CDEs. Supplementary Tables S1–S3 list the cCDEs and how they were distributed in each document at each hospital. These tables also provide a detailed view of how the 20 unique cCDEs comprised 327 sub-aCDEs. The integrated CDEs not only reduced the number of CDEs, with a reuse ratio of up to 46.9% [= (1142 – 20 – 586)/1142] (Table 2), but also greatly improved the semantic accuracy and interoperability.

a Number of CDEs extracted from each clinical document from each hospital
b Number of cCDEs created for each clinical document
c Number of aCDEs contained in bCDEs
d Number of remaining aCDEs that are not contained in any of the cCDEs in each clinical document
e Total number of CDEs consisting of bCDEs and dCDEs that are not contained in any of the cCDEs in each clinical document
f Number of unique CDEs across the five clinical documents
g Reuse ratio of CDEs across the five documents

We found that the compositions of the clinical documents differed quite markedly across the included hospitals. The clinical documents at Hospitals P and S contained the largest (n=266) and smallest (n=31) numbers of independent DEs, respectively. We also found that even the same clinical documents showed huge variations in DE numbers, such as with the number of admission notes varying from 12 at Hospital S to 204 at Hospital P. Hospital P also had the largest number of aCDEs for initial medical examination notes (n=123), while Hospital A had the largest number of aCDEs for emergency notes (n=83)
and operation notes (n=37).

We also applied constraint rules for the five clinical documents used at the five hospitals (Table 3 in the Supplementary Files). We could not determine if a DE was a hybrid cCDE, partly due to the lack of actual input values and partly due to poor descriptions of the response values for the clinical documents. We designated the cCDEs as general cCDEs to distinguish them from repeated and dictionary cCDEs. A cCDE was on average reused twice among the five documents by the hospitals. We also found that the clinical documents at Hospital A were the best structured and contained the greatest detail, with more cCDEs and constraint rules compared to the documents at the other hospitals.

Discussion

4.1 COMPARISON WITH RELATED STUDIES

Standardizing data using CDEs based on ISO/IEC 11179 is clearly one of most effective ways to harmonize data collected from various clinical studies. This approach provides the following advantages: (1) providing a consistent data collection tool, and (2) improving the study quality and reducing the cost of data entry, cleansing by having uniform data. However, the inherent limitation of ISO/IEC 11179 not providing a data structure for representing interrelationships among CDEs has resulted in a gap between the development of CDEs and their utilization on clinical forms for comprehensive representations.

To overcome this obstacle, ISO/IEC 11179 provides DDEs to enhance interrelated DEs. A DDE is a DE whose values are derived through a transformation of the values of one or more source DEs. For example, the DDE of the ‘length of stay in a hospital’ is derived from two independent DEs that calculate the number of days from two input DEs: ‘admission date’ and ‘discharge date.’ However, this strategy is far from sufficient to cover all of the use cases of interrelated DEs that we describe in Background.
Table 4 (in the Supplementary Files) compares the DDE and our CDE semantic relationships. The value of a DDE is derived from input DE(s). Our CDE semantic relationship provides rich semantics for creating aCDEs and cCDEs that feature repeat and dictionary properties, supporting references to outside biomedical resources as described in Table 4. The relatively simple-minded concept of the DDE may be inadequate to cover various CDE semantic relationships, since a DDE covers only two constraints: Operated and Ordered.

There have also been efforts to address the issues of interrelated DE(s) by applying external data models. The CDISC (clinical data interchange standards consortium) ODM (operational data model), which is an XML-based standardized data model that supports the acquisition and exchange of metadata specifically related to clinical studies, can also be used to overcome the limitations of ISO/IEC 11179; however, it is not sufficiently comprehensive to generate CRFs by importing elements directly [21, 22]. Lin et al. also suggest using the openEHR approach for modeling CDEs [23]. Though this approach provides a comprehensive structure with two-level modeling, several limitations when implementing openEHRs have been identified in various studies, such as immaturity of archetype modification operations, insufficient support for hierarchical archetypes due to their granularity [24, 25], and the cost burden of development and adoption due to the complexity of defining openEHRs. Therefore, instead of utilizing external data models, we propose improving and extending the existing composite relationship by specifying two subtypes of aCDE, three subtypes of cCDEs, and four constraints to take advantage of utilizing CDEs and related technologies.

4.2 OVERCOMING THE CHALLENGES OF UNDERSTANDING SEMANTIC RELATIONSHIPS OF FORM-LEVEL DATA

This paper has presented an in-depth evaluation of the ISO/IEC 11179 MDR standard-based
CDE semantic interrelationships in the context of formalizing clinical document structures. For converting form-level data into DE-level data, two cCDEs (repeated and dictionary cCDEs) and their related constraints were developed, which provide the following benefits: Repeated cCDEs support clinical data management in a tabular format in a clinical document. Since multiple value sets are supported to be represented in a unified tabular format, a repeated cCDE is useful for managing sequential data entry in a tabular format and for analyzing how the values change over time. A repeated cCDE enables standard MDR-based CDE-level descriptions and evaluations of clinical data entry in a tabular format. Dictionary cCDEs enable biomedical knowledge to be brought from a dictionary database via a variable aCDE. Data items referencing a certain standard terminology appear frequently on clinical forms. A dictionary cCDE can help to include rich semantics from externally managed biomedical terminologies and/or dictionaries, with rich attributes being applied for input data validation. Four different types of constraints enable rich evaluations of input values. A prefix notation with functional logic programming can be applied for evaluating user-defined constraints in order to ensure contextual correctness and interrelationships among data items on clinical document.

4.3 ADVANTAGES OF USING CDES AND CDE RELATIONSHIPS FOR BUILDING CLINICAL DOCUMENTS

As verified in the evaluation part of this study, building clinical documents with CDEs can provide three major advantages. First, it prevents the generation of redundant data by facilitating predefined and registered CDEs to the MDR. Second, it ensures semantic data integrity since an MDR-based CDE has comprehensive and standardized metadata attributes for data description and the proposed cCDE provides a means to encode rich
constraints for inter-CDE relationships. The health data of a patient that are fragmented, dispersed, and duplicated in a variety of clinical documents across different medical centers should be integrated, and mapping data items to CDEs facilitates data integration and semantic interoperability across different clinical documents. Third, clinical data exchange and sharing can be greatly facilitated by this approach.

4.4 LIMITATION AND FUTURE WORK

This exploratory study was conducted with 25 clinical documents obtained from 5 different teaching hospitals in Korea, and so the findings cannot represent the full breadth of clinical documents. Our development of new aCDEs and cCDEs has relied on our limited experience with three studies using the NINDS CDE project (focused on stroke), the DialysisNet CDE project (focused on hemodialysis care). And all evaluations are as well done only on form data, such as 25 major clinical documents obtained from 5 teaching hospitals. The generalizability of the proposed CDEs and their relationships to other biomedical domains remain to be investigated. We also couldn’t measure the data quality (DQ), though the DQ is one of the aspects of interoperability that reveals in the process standardizing EMRs. The data element concept part, which are composed by the object, and the property in ISO/IEC 11179 can be used to meet the DQ measure. We will measure it in future work as well.

Conclusion

The sharing and understanding of data from multiple different domains can be facilitated by standardization. An MDR-based CDE is considered a type of standardized data with specified concept and value domains. However, ISO/IEC 11179 MDR-based CDEs do not provide the ability to describe constraints on a CDE or relationships among different CDEs, instead merely focusing on single independent CDEs, which makes it difficult to either correctly compose or interpret CDEs on clinical documents. We developed MDR-based
extended semantic types and constraints, and it can facilitate comprehensive representation of clinical documents with rich semantics and improved semantic interoperability without semantic loss.

**Abbreviations**

aCDE Atomic CDE

BMI Body mass index

cCDE Composite CDE

CDE Common data element

CDISC Clinical data interchange standards consortium

CRF Case report form

DDE Derived data element

DE Data element

DQ Data Quality

EHR Electronic health record

JSON Javascript object notation

MDR Metadata registry

NINDS National institute of neurological disorders and stroke

ODM Operational data model

**Declarations**

**6.1 ETHICS APPROVAL AND CONSENT TO PARTICIPATE**

Not Applicable. To give you more description, we have not used any of patients' data. The data described in the Methods section are metadata, which is data about data including data 'specifications' and 'definitions'. We have had no chance of using patients' private and/or personal information at all in writing the manuscript.

**6.2 CONSENT TO PUBLISH**
6.3 AVAILABILITY OF DATA AND MATERIAL

Not Applicable

6.4 COMPETING INTERESTS

None of the authors has conflicts of interest with other persons or organizations that could inappropriately influence their work.

6.5 FUNDING

This work was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health and Welfare (HI13C2164).

References

1. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007; doi:10.1197/jamia.M2470.

2. Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The clinical document architecture and the continuity of care record: a critical analysis. J Am Med Inform Assoc. 2006; doi:10.1197/jamia.M1963.

3. Mohanty SK, Mistry AT, Amin W, 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; doi:10.1186/1471-2407-8-91.

4. Groft SC, Rubinstein YR. New and evolving rare diseases research programs at the National Institutes of Health. Public Health Genomics 2013; doi:10.1159/000355929.

5. NIH Common Data Element (CDE) Repository Website. https://www.nlm.nih.gov/cde/. Accessed Jun. 12, 2019.

6. Saver JL, Warach S, Janis S, 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; doi:10.1161/STROKEAHA.111.634352.

7. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013; doi:10.1136/amiajnl-2012-000896.

8. Nahm M, Walden A, McCourt B, et al. Standardising clinical data elements. Int J Funct Inform Personal Med. 2010; doi:10.1504/IJFIPM.2010.040213.

9. Park YR, Yoon YJ, Kim HH, Kim JH. Establishing semantic interoperability of biomedical metadata registries using extended semantic relationships. Stud Health Technol Inform. 2013;192:618-21.

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

11. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc. 2011; doi:10.1136/amiajnl-2011-000107.

12. 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.

13. NCI caDSR Wiki, CDE Curation Tool User Guide- Creating Derived Data Element Website. https://wiki.nci.nih.gov/display/caDSR/10+-+Creating+Derived+Data+Elements/. Accessed Jun. 12, 2019.
14. Data type in Wikipedia. https://en.wikipedia.org/wiki/Data_type/. Accessed Jun. 12, 2019.

15. NINDS Common Data Elements Website. https://comondataelements.ninds.nih.gov/. Accessed Apr. 15, 2017.

16. Ku HS, Kim S, Kim H, Kim JH. DialysisNet: Application for Integrating and Management Data Sources of Hemodialysis Information by Continuity of Care Record. Healthc Inform Res. 2014; doi:10.4258/hir.2014.20.2.145.

17. Park YR, Kim H, An EY, et al. Establishing semantic interoperability in the course of clinical document exchange using international standard for metadata registry. J Korean Med Assoc. 2012; doi:10.5124/jkma.2012.55.8.729.

18. Kim JH. Health Avatar: an informatics platform for personal and private big data. Healthc Inform Res 2014; doi:10.4258/hir.2014.20.1.1.

19. NINDS Common Data Elements Website.
   https://www.comondataelements.ninds.nih.gov/Doc/Stroke/F1168_Laboratory_Tests _Permissible_Values_for_Stroke.xlsx. Accessed Jun. 12, 2019.

20. Wikipedia. https://en.wikipedia.org/wiki/Polish_notation. Accessed Jun. 12, 2019.

21. Ngouongo SM, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research?. J Biomed Inform. 2013; doi:10.1016/j.jbi.2012.11.008.

22. Iberson-Hurst D. THE CDISC OPERATIONAL DATA MODEL: READY TO ROLL? Appl Clin Trials. 2004;13:48–53.

23. 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; doi:10.1093/jamia/ocv137.

24. 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; doi:10.1016/j.ijmedinf.2007.02.004.

25. Späth MB, Grimson J. Applying the archetype approach to the database of a biobank information management system. Int J Med Inform. 2011; doi:10.1016/j.ijmedinf.2010.11.002.

Additional File Details

Supplementary Table S1. List of general, dictionary, and repeated cCDEs with the numbers of operated, required, dependent, and ordered constraints extracted from five clinical documents used at five teaching hospitals in Korea.

Supplementary Table S2. Distribution of aCDEs and cCDEs extracted from five clinical forms used at five teaching hospitals in Korea.

Supplementary Table S3. List of 327 aCDEs comprising 20 cCDEs. The order of the cCDEs is identical to that in Supplementary Table S1.

Tables

Due to technical limitations, tables 1 through 4 are only available as a download in the supplemental files section.

Figures
Figure 1. Overview of the formal relationship between aCDE and cCDEs with extended semantic types and CDE-type specific constraints.
Figure 2. An example hybrid aCDE from a hemodialysis report. (A) The hemodialysis table of the DialysisNet Project has a tabular data-entry format, where Time (DE:47616) allows two different data types: time and an enumerated string. (B) The hybrid aCDE (DE:47616) contains two aCDEs (DE:43239 and DE:47614) in a hybrid relationship (http://chmr2.snubi.org:8083/chmr/data_element_view.jsp?id=28476).

Figure 2

An example hybrid aCDE from a hemodialysis report. (A) The hemodialysis table of the DialysisNet Project has a tabular data-entry format, where Time (DE:47616) allows two different data types: time and an enumerated string. (B) The hybrid aCDE (DE:47616) contains two aCDEs (DE:43239 and DE:47614) in a hybrid relationship (http://chmr2.snubi.org:8083/chmr/data_element_view.jsp?id=28476).
Figure 3. Example of the composition of a repeated cCDE from a hemodialysis report form. A repeated cCDE, ‘DE:47575 Hemodialysis_Repeated_Composite_DE,’ composed of eight aCDEs from a tabular data-entry format (Figure 2A) for the DialysisNet hemodialysis project (http://chmr2.snubi.org:8083/chmr/data_element_view.jsp?id=28449).

Figure 3

Example of the composition of a repeated cCDE from a hemodialysis report form. A repeated cCDE, ‘DE:47575 Hemodialysis_Repeated_Composite_DE,’ composed of eight aCDEs from a tabular data-entry format (Figure 2A) for the DialysisNet hemodialysis project (http://chmr2.snubi.org:8083/chmr/data_element_view.jsp?id=28449).

| Date Collected (MM/DD/YYYY) | Test            | Result | Units for Result | Was test result abnormal? | If abnormal, Clinically Significant? |
|-----------------------------|-----------------|--------|------------------|---------------------------|-------------------------------------|
| 01/09/2016                  | Sodium (Na⁺)    | 138    | mEq/L            | Normal                    | Not clinically significant           |
| 01/09/2016                  | Potassium (K⁺)  | 5.3    | mEq/L            | Normal                    | Clinically significant               |
Figure 4. Creation of a dictionary cCDE for a CRF. (A) The ‘Electrolyte Laboratory Tests’ table on a clinical document is provided as an example tabular data-entry document to capture laboratory test results for sodium (Na⁺) and potassium (K⁺) along with two clinical evaluation attributes. (B) We constructed the ‘Electrolyte Laboratory Tests Dictionary’ table by extracting the relevant attributes from the CDEs defined in the ‘Recommended Labs for Stroke’ from the NINDS CDE project. (C) The dictionary cCDE (DE-47571) consists of six aCDEs that include a variable aCDE (DE-45935) that relates the dictionary cCDE to the dictionary table in Figure 4B. Two rules for clinical evaluation are presented (http://chner2.snomed.org:8083/chmr/data_element_view.jsp?id=28445).

Figure 4
Creation of a dictionary cCDE for a CRF. (A) The ‘Electrolyte Laboratory Tests’ table on a clinical document is provided as an example tabular data-entry
document to capture laboratory test results for sodium (Na+) and potassium (K+) along with two clinical evaluation attributes. (B) We constructed the ‘Electrolyte Laboratory Tests Dictionary’ table by extracting the relevant attributes from the CDEs defined in the ‘Recommended Labs for Stroke’ from the NINDS CDE project. (C) The dictionary cCDE (DE:47571) consists of six aCDEs that include a variable aCDE (DE:43938) that relates the dictionary cCDE to the dictionary table in Figure 4B. Two rules for clinical evaluation are presented (http://chmr2.snubi.org:8083/chmr/data_element_view.jsp?id=28445).
Figure 5. Encoding *Operated, Ordered, Required,* and *Dependent* constraints for a repeated cCDE. (A) A ‘Medical History’ clinical document presented in a tabular format containing six attributes. (B) A repeated cCDE is created with the corresponding six aCDEs along with four constraint rules: (1) the start date (DE:47618) should be earlier than the end date (DE:47619): (< DE:47618 DE:47619); (2) all attributes are required to have values other than null, except for the end date (DE:47619): (Required DE:37059 DE:47621 DE:31106 DE:47618 DE:44078); (3) when a certain medical history is not ongoing (DE:44078), the end date (DE:47619) cannot be obtained, and vice versa: (IF (!= DE:44078 'Yes') DE:47619 NULL); and (4) aCDEs can be ordered according to a constraint statement such as (Ordered DE:37059 DE:47621 DE:31106 DE:47618 DE:44078 DE:47619) (http://chmr2.srnihi.org:8083/chmr/data_element_view.jsp?id=28477).

Figure 5

Encoding Operated, Ordered, Required, and Dependent constraints for a repeated cCDE. (A) A ‘Medical History’ clinical document presented in a tabular format
containing six attributes. (B) A repeated cCDE is created with the corresponding six aCDEs along with four constraint rules: (1) the start date (DE:47618) should be earlier than the end date (DE:47619): (< DE:47618 DE:47619); (2) all attributes are required to have values other than null, except for the end date (DE:47619): (Required DE:37059 DE:47621 DE:31106 DE:47618 DE:44078); (3) when a certain medical history is not ongoing (DE:44078), the end date (DE:47619) cannot be obtained, and vice versa: (IF (!= DE:44078 'Yes') DE:47619 NULL); and (4) aCDEs can be ordered according to a constraint statement such as (Ordered DE:37059 DE:47621 DE:31106 DE:47618 DE:44078 DE:47619)

(http://chmr2.snubi.org:8083/chmr/data_element_view.jsp?id=28477).

Supplementary Files

This is a list of supplementary files associated with the primary manuscript. Click to download.

Supplementary Tables.pdf
Table 1.jpg
Table 3.jpg
Table 4.jpg
Table 2.jpg