RECEIVED 13 November 2014
REVISED 27 July 2015
ACCEPTED 28 July 2015
PUBLISHED ONLINE FIRST 23 January 2016

An exploratory study using an openEHR 2-level modeling approach to represent common data elements

Ching-Heng Lin[1], Yang-Cheng Fann[2], Der-Ming Liou[1]

ABSTRACT

Background and Objective In order to facilitate clinical research across multiple institutions, data harmonization is a critical requirement. Common data elements (CDEs) collect data uniformly, allowing data interoperability between research studies. However, structural limitations have hindered the application of CDEs. An advanced modeling structure is needed to rectify such limitations. The openEHR 2-level modeling approach has been widely implemented in the medical informatics domain. The aim of our study is to explore the feasibility of applying an openEHR approach to model the CDE concept.

Materials and Methods Using the National Institute of Neurological Disorders and Stroke General CDEs as material, we developed a semiautomatic mapping tool to assist domain experts mapping CDEs to existing openEHR archetypes in order to evaluate their coverage and to allow further analysis. In addition, we modeled a set of CDEs using the openEHR approach to evaluate the ability of archetypes to structurally represent any type of CDE content.

Results Among 184 CDEs, 28% (51) of the archetypes could be directly used to represent CDEs, while 53% (98) of the archetypes required further development (extension or specialization). A comprehensive comparison between CDEs and openEHR archetypes was conducted based on the lessons learnt from the practical modeling.

Discussion CDEs and archetypes have dissimilar modeling approaches, but the data structure of both models are essentially similar. This study proposes to develop a comprehensive structure to model CDE concepts instead of improving the structure of CED.

Conclusion The findings from this research show that the openEHR archetype has structural coverage for the CDEs, namely the openEHR archetype is able to represent the CDEs and meet the functional expectations of the CDEs. This work can be used as a reference when improving CDE structure using an advanced modeling approach.

Keywords: common data element, openEHR archetype, modeling approach

RESEARCH AND APPLICATIONS

INTRODUCTION

In recent years, electronic data capture (EDC) systems have been widely adopted in the clinical research domain. We expect the next generation of clinical trials to be "eClinical Trials."[1] EDC systems do not only enhance the quality of data collection but also substantially reduce the workload and time needed to carry out such work.[2] EDC systems also facilitate data collection; however, there is still an urgent need among researchers for data harmonization and data interoperability, which would greatly facilitate any subsequent data aggregation, analysis, and sharing.

Data in standard format are often more valuable and have greater availability; this is especially true for data produced by multicenter research studies. The common data element (CDE) is a data element that collects and stores data uniformly across institutions and studies. The National Institutes of Health (NIH) has developed CDEs that serve as a controlled vocabulary of data descriptors.[3] The main purpose of CDEs is to standardize data and facilitate any follow-up comparisons of data collections across multiple studies. However, there are several concerns that limit the use of CDEs. One concern is that CDE content has a poor level of standardization in terms of definition. Albright et al[4] conducted a survey of academic vascular neurologists in order to compare their definition of CDEs with standardized definitions, and the results indicated that standardized definitions of CDEs are needed. In addition, related CDEs, namely those that are ISO/IEC (International Organization for Standardization and International Electrotechnical Commission) 11179 standard compliant, were found to be groupable into object classes; nevertheless, no higher-order groupings were found for the isolated CDEs. Grouping of CDEs allows them to be presented as a rational sequence of questions in a case report form (CRF). Accordingly, users are forced to select an appropriate CDE individually, even when using an electronic CRF design. An improved architecture is needed to address these aforementioned limitations.

The openEHR foundation, which is an international nonproprietary foundation, aims to enable the development of open, future-proof specifications and software for use in electronic health record (EHR) systems. These include the archetypes and the ISO-approved standard European Committee for Standardization (CEN)/European Standard (EN) 13606,[5] which are subsets of the openEHR specification.[6,7] EHR systems are essential to improving the quality of patient care as well as being a great help to clinical researchers.[8] The openEHR is a major standard in this area; therefore, Grade and Knaup et al explored the feasibility of using the openEHR approach to support multicenter clinical studies. They concluded that the openEHR is suitable because it enables the use and sharing of routine data during multicenter clinical studies.[9,10] To improve data interoperability between EHRs and the EDC system, Kohl et al[11] implemented the open Study Data Management System (openSDMS) that is based on the openEHR archetypes. The open Study Data Management System demonstrated that the openEHR archetypes are suitable for documenting medical data during clinical trials. The above studies illustrate that the

Correspondence to Der-Ming Liou, No. 155, Sec. 2, Li-Nong St, Beitou District, Taipei City 112, Taiwan (ROC); dmliou@ym.edu.tw. For numbered affiliations see end of article.

©The Author 2016. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved. For Permissions, please email: journals.permissions@oup.com.

openEHR approach is not only able to support data collection but also facilitates medical data interoperability during clinical research.

In order to find a more advanced architecture that will address the limitations of CDEs, this study's aim is to explore the feasibility of applying the openEHR approach to modeling of the CDE concept. Taking the National Institute of Neurological Disorders and Stroke (NINDS) CDEs[12] as our material, we developed a semiautomatic mapping tool that was able to map CDEs to existing archetypes and allow evaluation of their reuse rate. In addition, we modeled a series of archetypes based on Protocol Experience CDEs, which are called common data archetypes, in order to evaluate the archetype's ability to support CDE representation.

BACKGROUND

The common data element and its architecture

The NIH supports the CDE initiative, which defines a CDE as a data element that is common to multiple data sets across different sites, registries, and diseases.[13] Toward this end, it is necessary to ensure that the data are defined in the same way. Many NIH-supported studies are encouraged to use CDEs for harmonization, the sharing of results and the exchange of information.[12,14–16] The NIH CDE contains 28 diverse subject areas, including study design (study details, patient contact information), diseases (cancer, diabetes, cardiovascular), body parts (ocular, oral, skin, and bone), and substances used (alcohol, tobacco, nutrition).[17]

Essentially, the core structure of a CDE is a name-value pair that is augmented by data element details, data element concept details, value domains, permissible values, and property/object class information. A unified modeling language class diagram describing CDE content is presented in figure 1A. To manage and standardize CDEs, the National Cancer Institute (NCI) developed the Cancer Data Standards Repository (caDSR), which contains over fifty thousand data elements from different institutes or projects.[18] The caDSR implemented the ISO/IEC 11179 standard for metadata registries when including CDEs in the repository.[19] ISO/IEC 11179 is an international standard for representing metadata in a metadata registry[20]; and in order to do this, it maintains a precise semantic structure for the data element. However, ISO/IEC 11179 is not intended to cover all peculiarities of metadata and is also not required to address every potential use; as a result, several limitations of this standard still need to be considered when applying it to clinical research. First, ISO/IEC 11179 does not provide a suitable structure for the semantic and syntactic relationships between related concepts; and, as a consequence, it can be difficult for users to group CDEs into identical concepts.[21–23] Second, this standard is also unable to represent interelement constraints; for example, a calculation giving a body mass index must result in a positive number.[22] As a result, the caDSR implementation of this standard is not that rigorous.[23] Third, a structure that can be used for representing concept synonyms and interconcept relationships in order to support controlled terminologies is not present in this standard.[23] Ngouongo et al[24] assessed the ability of ISO/IEC 11179 edition 3 to represent the items of the Operational Data Model (ODM) from the Clinical Data Interchange Standard Consortium (CDISC). The assessment results showed that ISO/IEC 11179 only allows single measure units and is only able to support selfreferencing associations between elements in the same class. In general, this standard is unable to cover ODM or ISO 13606.

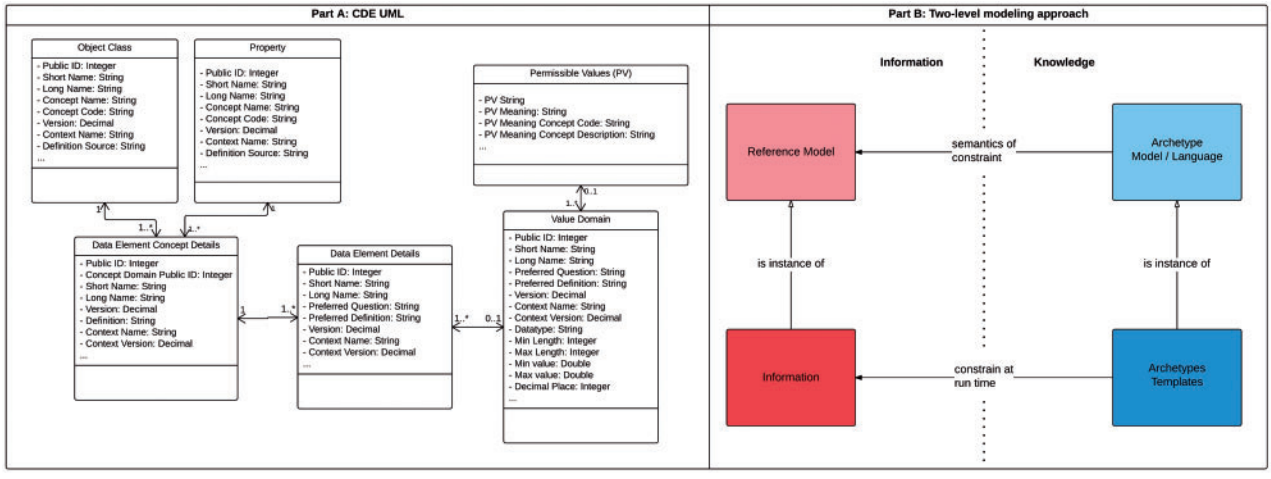
The openEHR

The 2-level modeling approach

In essence, openEHR uses a modeling approach referred to as 2-level modeling.[25] This modeling approach separates clinical knowledge from the information model rather than trying to capture all required information in a single-level data model.[26] Figure 1B, which has been modified from previous studies,[27,28] illustrates the 2-level modeling approach. On the information side, the reference model, which is a generic and underlying model, defines the semantics and structure of the information and deals with interoperability at the syntactic level. Information items, such as laboratory results, are instances of the reference model. Meanwhile, information items are constrained by archetypes, an example of which is the blood pressure archetype; these are instances of archetype models on the knowledge side. The archetype model enables the creation and description of archetypes and templates that are able to represent the clinical domains.[27]

Specifically, openEHR allows the separation of knowledge-modeling from system-development tasks—namely, the domain experts can model and update the medical concept without there being a need for software modification. At the same time, the system developer can

Figure 1: (A) The UML class diagram of CDE content. Some attributes have been edited by shortening the content details. (B) An overview of the openEHR 2-level modeling approach.
UML, Unified Modeling Language; CDE, common data element; ID, identification; EHR, electronic health record.



start on the software development before the domain-modeling task is completed. Compared with the single modeling approach, the 2-level modeling approach enables the clinical system to be independent of domain concept creations or changes by being only bound to the stable reference model.[29] Thus, it is able to meet the design needs of the system, which then allows both system integration and semantic interoperability.[28]

To date, this approach has been widely applied to a range of different research fields, including EHR system development,[30] EHR standard modeling,[31,32] biobank information system development,[33] clinical statement modeling,[34–36] clinical guideline representation,[37,38] and semantic interoperability research.[39]

The archetype model

An archetype represents a distinct domain-level concept or a data collection item. It specifies constraints on the data structures in the reference model and is used to address semantic interoperability.[27] According to their reference model class, archetypes can be classified as follows[34,40]:

- The *composition archetype* represents the theme of a clinical document; as the unit of contribution, it can contain 1 or more *section archetypes* and *entry archetypes*.
- The *section archetype* corresponds to the clinical document heading and contains 1 or more *entry archetypes* that assists with human navigation.
- The *entry archetype* is the semantic unit of information and can be divided into the following types:
    - the *observation archetype* represents direct clinical observations, measurements, or the experiences of the subject, eg, their pulse rate;
    - the *evaluation archetype* represents an assessment or a clinically interpreted finding, eg, a risk assessment;
    - the *instruction archetype* represents a clinical or medical order; these are statements about what should happen, eg, a medical instruction; and
    - the *action archetype* represents clinical activity records and indicates what was done, eg, a surgical procedure.
- The *cluster archetype* represents compound entries. This is a reusable archetype and can be used within any *entry archetype* or within another *cluster archetype*.
- The *element archetype* is used to model a single item. This is also reusable in a similar manner to that of the *cluster archetype*.

The different archetypes can be mapped to specific components of a CRF; for example, the *composition archetype* can be used when presenting the whole CRF, while the *section archetype* can be used as section building blocks. Finally, entry items of the CRF can be represented with *entry archetypes*.[22,41]

Archetypes can be expressed using the archetype definition language (ADL). The ADL is a formal syntax used for expressing archetypes that was developed by openEHR and has also been adopted by EN 13606.[42] The ADL is designed as a human readable syntax that is easy to understand as well as a computer-processible syntax that can be hand-edited using a normal text editor or a specific ADL editor. Each archetype is described in 4 required sections. The first 3 are the archetype section, which indicates the unique identifier of an archetype; the language section, which includes language details; and the definition section, which describes the definition of this archetype such as object structure, constraints, or value ranges. The last required section is ontology, which includes semantic information like terminology definitions and term binding.[43] In addition to the required sections above, the description section is an important block of ADL as well; this section contains metadata such as the purpose of the item, the authorship lifecycle, use/misuse information, and keywords.

Numerous openEHR models have been developed by domain experts worldwide. To ensure that released archetypes are of the highest clinical and technical quality that will enable international involvement, the openEHR Foundation has established the openEHR Clinical Knowledge Manager (CKM).[44] This is an international online repository of clinical knowledge artifacts that included 465 archetypes of various types when we accessed it in October 2014. Since then, these archetypes have been represented in the ADL. In this context, the repository allows manual searches for existing archetypes via clinical specialties, clinical purpose, or an available keyword string matching function. The CKM also supports a formal community review process that is initiated for each archetype. Once consensus has been reached on the clinical content and design, the archetype is then published in the CKM.[45]

Applying the archetype approach to the CDE

In this section, we have used blood pressure measurement as an example to further illustrate the notion of applying the archetype approach to the CDE. The concept of a blood pressure measurement is composed of at least a systolic measurement and a diastolic measurement. Figure 2 shows the blood pressure measurement concept as represented by both the CDE approach and the archetype approach. There are 2 CDEs—one represents the diastolic measure, while the other represents the systolic measure. Using the archetype approach, the root node of the blood pressure concept is an entry archetype that comprises 2 element archetypes—the diastolic measure and the systolic measure—via a cluster archetype called blood pressure measurement. In terms of concept grouping, the CDE merely uses the object name and object public identification (highlighted in bold text in figure 2) to indicate their higher-level concept. In contrast, the archetype approach has a strong hierarchy that brings about a grouping of these two individual but interrelating concepts.

A higher concept element, such as the entry archetype, helps us to manage and reuse related concepts; however, the CDE lacks this feature. Thus, there is an inherent advantage to the 2-level modeling approach. As we mentioned before, the most outstanding feature of this modeling approach is the separation of the information model from the knowledge model; in this context, the reference model provides a stable and clear hierarchy that is able to achieve the higher grouping purpose.

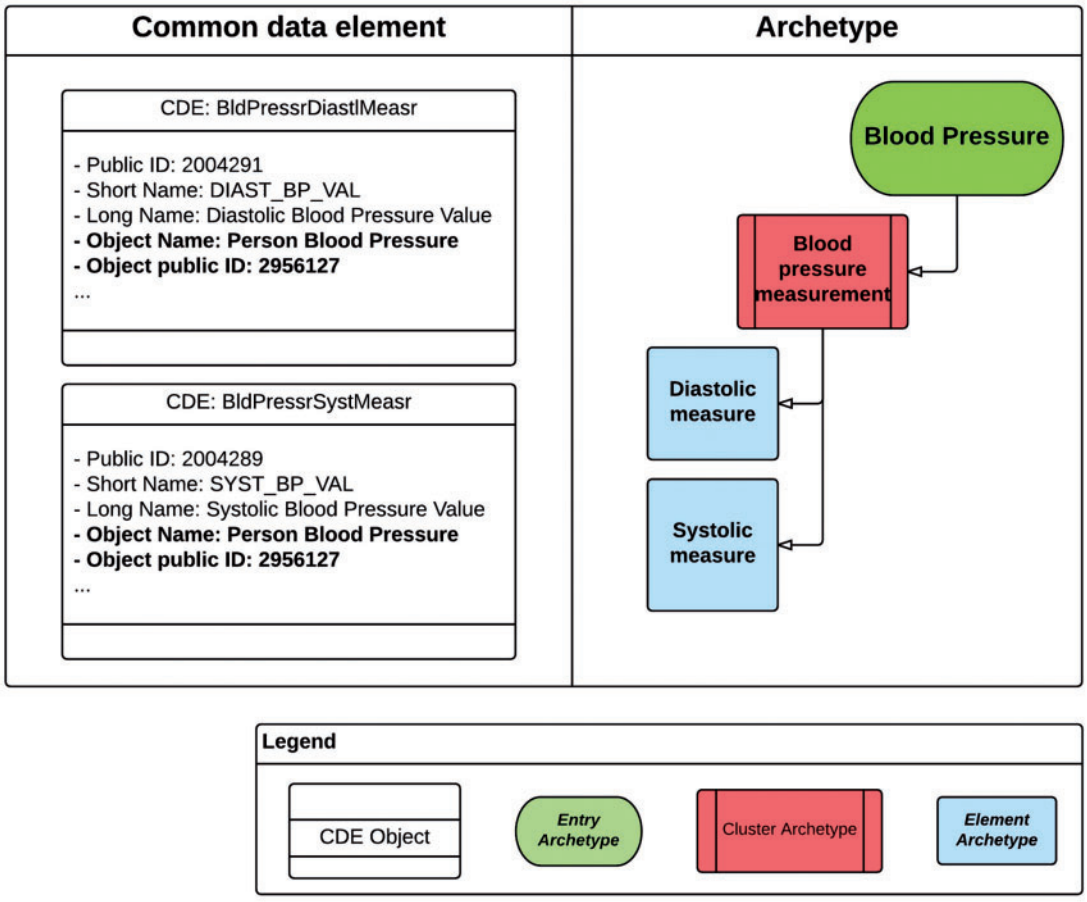
MATERIALS AND METHODS

In this study, we initially investigated the coverage that CDEs provide with respect to archetypes. The aim of this step is to avoid any unnecessary work involved in recreating parallel archetypes and to avoid the risks associated with overlapping archetype development that might obstruct interoperability.[46] Furthermore, we also built a new archetype in order to explore the feasibility of applying openEHR archetypes to represent CDEs.

Study material

A number of clinical CDEs for a variety of different purposes has been developed, such as the NINDS CDE project,[12] the Cancer Bioinformatics Grid,[47] the Parkinson Disease Biomarker Program,[48,49] and others.[50–52] This study used the General CDEs of the NINDS as our primary study material for two reasons. First, the General CDEs have not been developed by a disease-specific working group; rather, the General CDEs involve CDEs across diseases and have been harmonized with other relevant

Figure 2: The Blood pressure concept represented by both CDE and archetype approach
CDE, common data element.



clinical data standards. Second, there is often no or not enough archetypes resources when covering disease-specific CDEs. Therefore, our preliminary study chose wider general-domain CDEs as material. The General CDEs can be categorized into 7 domains: Assessments and Examinations, Outcomes and End Points, Participant/Subject Characteristics, Participant/Subject History and Family History, Protocol Experience and Safety Data, together with Treatment/Intervention Data. We downloaded a total of 189 CDE concepts via the NCI's CDE browser.[18]

In addition, we adopted the CKM[44] as the archetype source. Since the purpose of CDEs is to standardize data entries, it is reasonable not to include the *composition archetype* and *section archetype* classes. Finally, a total of 349 archetypes were included in our study.

Searching for existing archetypes

The openEHR CKM provides an online string search function; however, the number of CDEs in our study was quite large, so it was not efficient to manually search for reusable archetypes via the CKM. To improve our mapping efficiency, we developed a semiautomatic mapping tool that used an embedded information retrieval (IR) technique to retrieve and rank the CDE relevant archetypes; this used Apache Lucene (Apache Software Foundation, Forest Hill, Maryland), which is an open-source and full-feature text search engine.[53]

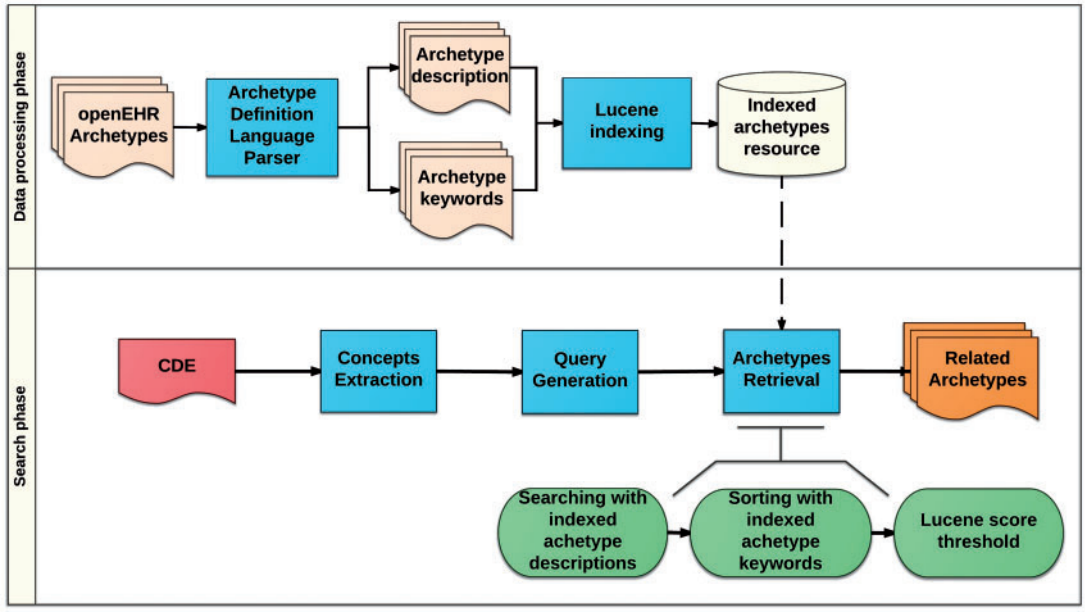
An overview of our IR strategy is presented in figure 3. During the data processing phase, we created an indexed archetypes source that acted as a search source. First, each archetype was parsed by the ADL parser[54] that extracted the description and keywords. In the next step, we indexed those description and keywords files using Lucene and stored them in an indexed archetypes resource that was to be used at a later stage.

In order to retrieve relevant candidate archetypes for a given CDE, a series of steps were developed that included concept extraction, query generation, and IR (see the search phase in figure 3). For each NINDS CDE, there is an attribute called the data element concept that in short form represents a CDE definition. Our tool extracted the concept terms from the input CDE and then took those terms as queries for the next query generation step. After query generation was completed, 3 subprocesses were performed. First, our tool launched the Lucene IR function targeting the archetype descriptions with the aim of retrieving the candidate archetypes. Next, our tool narrowed down the search scope to candidate archetypes within the archetypes' keywords; the candidate archetypes were then sorted and filtered based on their similarity score.[55] Finally, the results were forwarded to our domain experts for further mapping.

Developing archetypes for the NINDS CDE

To model the NINDS CDEs with archetypes, we utilized the openEHR archetype editor (2.2.905 beta),[56] which offers a graphical user interface and supports archetype creation and editing, as the modeling tool. The openEHR also provides a simple decision-making algorithm[57] that helps to decide which reference model class should be made into an archetype.

Figure 3: An overview of the semiautomatic mapping workflow
EHR, electronic health record; CDE, common data element.



We initially analyzed the CDE structure and characteristics. As mentioned in the Background, because the CDE structure is compliant to ISO/IEC 11179, some basic aspects of the available information needed for archetype modeling are sufficient. These include concept name, definition, data type, value domain, and other metadata. For example, the definition of the CDE can be used for purpose, description, and usage of the archetype, while the constraints of an archetype can be referred to as the value domain of the CDE. Using this information directly reduces the information collection effort needed to identify concepts before modeling takes place.

The CDE is used to standardize the individual data entry (or data field in a CRF); therefore, its granularity is very finely defined. In contrast to an *entry archetype*, each CDE is much more likely to be considered to be an attribute within an archetype. However, we may lose the object properties of a given CDE if we merely convert it into an attribute within an *entry archetype*. Furthermore, if this is done, the CDE does not benefit from the 2-level modeling approach. Therefore, in order to resolve this issue, we felt it necessary to model each CDE as an *element archetype* rather than an *entry archetype*.

As we mentioned in the Introduction, the lack of higher-order elements when grouping isolated CDEs is a major issue. In order to address this limitation, we identified the higher-level archetypes of the CDE *element archetypes* by referring to the hierarchy in the NCI's CDE browser; this helps the user to easily navigate the CDEs. For instance, 2 CDE *element archetypes*, "off treatment date" and "off treatment category," under the Protocol Experience domain, can be embedded into an *entry archetype* called "off treatment." We further converted the Protocol Experience domain into a *section archetype* that contained the "off treatment" *entry archetype*. By this approach, isolated CDEs could be formed that have a clear structure.

Evaluation methods

In order to evaluate the coverage of the NINDS Generation CDEs by related existing archetypes (Equation 1), 2 domain experts (1 clinical researcher and 1 medical informatics specialist) cooperated to perform the mapping tasks on all CDEs, using our semiautomatic mapping tool. For each input CDE, our tool provided up to 5 archetypes as a suggestion list, even if there were more archetypes returned. If there were no suitable archetypes in the suggestion list, domain experts would carry out a manual search on the openEHR CKM. When an archetype was found to cover a CDE, the domain experts would further define its reusable type. If an archetype was a more general concept than the input CDE, its reusable type would be defined as a specialization. If an archetype was able to cover the input CDE by adding missing items, its reusable type would be defined as extended. Finally, when an archetype could be reused without any modification, its reusable type would be defined as directly reusable.[33,34]

$$\text{coverage} = \frac{\text{the number of CDEs which can be represented by related existing archetypes}}{\text{total of CDEs}} \tag{1}$$

Three measures were used to describe the performance of our semiautomatic mapping tool; these were precision (Equation 2), recall (Equation 3), and the *F* measure (Equation 4). To calculate these values, the archetypes retrieved were counted and categorized as true positive (the retrieved archetype is related to the given CDE), true negative (no archetype is retrieved and the given CDE has no related archetype), false positive (the retrieved archetype is not related to the given CDE), and false negative (no archetype is retrieved but the given CDE has related archetypes).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

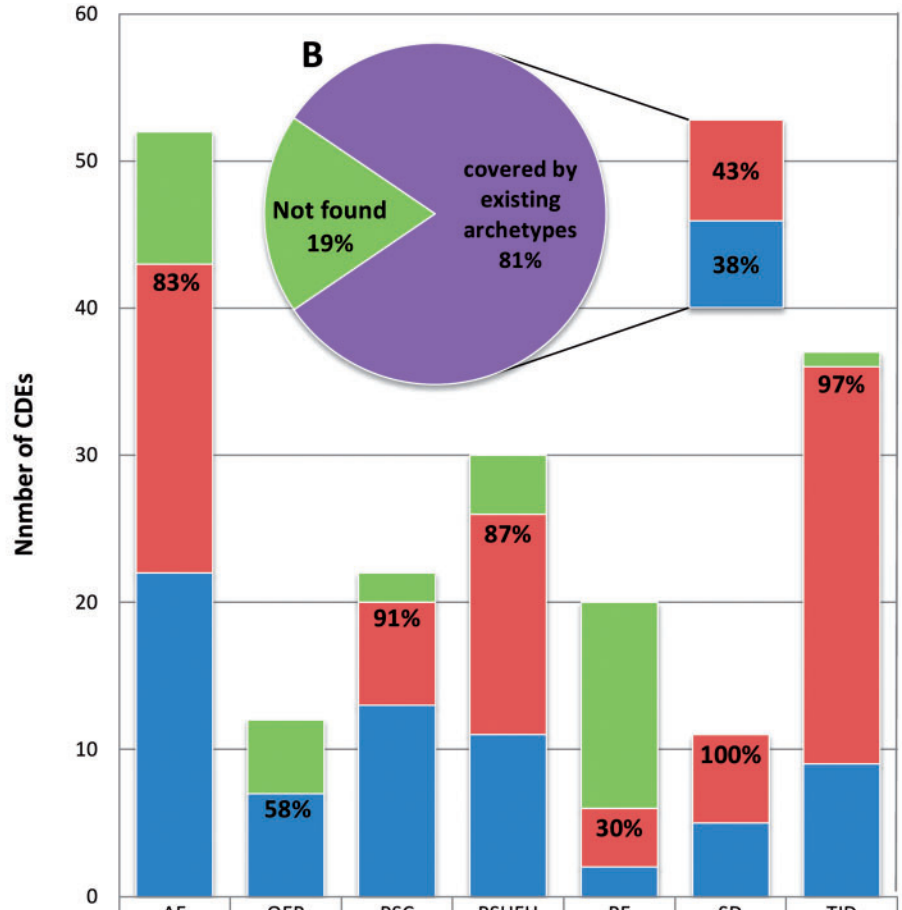
$$F\ \text{measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

RESULTS

Evaluation of coverage and reusable type

In this study, 349 entry, cluster, and element archetypes were evaluated for their coverage of a set of 184 unique NINDS General CDEs. Figure 4 illustrates the statistics obtained in terms of existing archetypes searched. Part A shows the stacked column chart of search results for the 7 domains of the General CDEs. The percentage of each bar indicates the coverage ratio as defined in Equation 1. The highest coverage (100%) was found for the Safety Data domain because those CDEs were relevant to the adverse events that are covered by openEHR-EHR-EVALUATION.adverse_reaction.v1. By way of contrast, the Protocol Experience domain had the lowest coverage (30%, 6 of 20), since the existing archetypes from the CKM are rarely able to cover certain clinical protocol concepts such as participant enrollment, participant ineligibility, and inclusion/exclusion criteria. These results also indicate that there is a lack of support for protocol design within the archetypes. Across all the General CDEs, using our semiautomatic mapping tool to help the user, it was possible to find mapped archetypes for 38% (69 of 184) of the CDEs, while about 43% (80 of 184) of the CDEs needed further searching using openEHR CKM. Overall, 81% (149 of 184) of the General CDEs could be covered by existing openEHR archetypes (see figure 4B). In this study, we did not include *composition archetypes* and *section archetypes*. Therefore, the coverage may have been less if the evaluation was performed completely on the CKM browser. For example, the CDE "Measurement Vital Signs Occurrence Date" could be covered by the archetype openEHR-EHR-SECTION.vital_signs.v1, but it is not counted as part of the coverage because it is a *section archetype*. A further evaluation of reusable types within the mapped archetypes is presented in table 1. In general, about 28% (51 of 184) of the mapped archetypes can be directly used, and 53% (98 of 184) need either further specialization or an extension in order to cover the General CDEs.

Figure 4: (A) A stacked column chart of the search results. The percentage of each bar indicates the coverage ratio as defined in Equation 1. In the table: AE, Assessments and Examinations; OEP, Outcomes and End Points; PSC, Participant/Subject Characteristics; PSHFH, Participant/Subject History and Family History; PE, Protocol Experience; SD, Safety Data; TID, Treatment/Intervention Data. (B) A pie chart of the existing archetype coverage in relation to all National Institute of Neurological Disorders and Stroke General common data elements.



	AE	OEP	PSC	PSHFH	PE	SD	TID
Not found	9	5	2	4	14	0	1
Manually search	21	0	7	15	4	6	27
Semi-automatic mapping tool	22	7	13	11	2	5	9

Table 1: Statistics describing the reusable types of existing archetypes

CDE classes	Reusable types of existing archetypes, No. (%)			
	Directly used	Specialization	Extended	Not found
Assessments and examinations (*n*=52)	30 (58)	3 (6)	10 (19)	9 (17)
Outcomes and end points (*n*=12)	2 (16)	0 (0)	5 (42)	5 (42)
Participant/subject characteristics (*n*=22)	6 (27)	11 (50)	3 (14)	2 (9)
Protocol experience (*n*=20)	2 (10)	0 (0)	4 (20)	14 (70)
Safety data (*n*=11)	4 (36)	0 (0)	7 (64)	0 (0)
Treatment/intervention Data (*n*=37)	0 (0)	36 (97)	0 (0)	1 (3)
Participant/subject history and family history (*n*=30)	7 (23)	0 (0)	19 (63)	4 (13)
	Total			
	51 (28)	50 (27)	48 (26)	35 (19)

	CDE Classes (N=184)							
Reusable Types of Existing Archetypes	Assessments and Examinations (*n*=52)	Outcomes and End Points *n*=12)	Participant/ Subject Characteristics (*n*=22)	Protocol Experience (*n*=20)	Safety Data (*n*=11)	Treatment/ Intervention Data (*n*=37)	Participant/ Subject History and Family History (*n*=30)	Total Reusable Types of Existing Archetypes
Directly used, No. (%)	30 (58)	2 (16)	6 (27)	2 (10)	4 (36)	0 (0)	7 (23)	51 (28)
Specialization, No. (%)	3 (6)	0 (0)	11 (50)	0 (0)	0 (0)	36 (97)	0 (0)	50 (27)
Extended, No. (%)	10 (19)	5 (42)	3 (14)	4 (20)	7 (64)	0 (0)	19 (63)	48 (26)
Not found, No. (%)	9 (17)	5 (42)	2 (9)	14 (70)	0 (0)	1 (3)	4 (13)	35 (19)

More detailed mapping results can be seen online in supplementary appendix A. It should be noticed that some of the mappings between the CDEs and archetypes are many-to-one mappings. This can be attributed to the difference in granularity between the 2 systems; in other words, an archetype can be compared to a maximum data set and a CDE to a single, specific entry.

The performance of the semiautomatic mapping tool

The performance matrix obtained from our semiautomatic mapping tool is provided in table 2. Among the 184 General CDEs, 3 values were measured—precision, which was 0.30 overall; recall, which was 0.69 overall; and the *F* measure, which was 0.42 overall.

For each domain, the highest *F* measure occurred in the Outcomes and End Points domain. In contrast, the Protocol Experience had the lowest *F* measure, which was caused by low precision. As we used an IR approach, our tool was sensitive to the existing archetype source and the concept terms of the input CDE. The reason for the low precision in the Protocol Experience domain was a lack of suitable archetypes among the existing archetype source (this domain had the lowest coverage). The lack of a semantic extension process resulted in our tool finding it hard to identify suitable archetypes in some cases. For example, the Treatment/Intervention Data domain had low precision because of the limitations imposed by the syntax search. Most Treatment/Intervention Data CDEs are related to the study agent, which can be semantically extended to either study medication or study drug.

An example of a common data archetype model

To better understand the feasibility of using archetypes to model CDEs, we modeled a set of Protocol Experience CDEs that had the lowest coverage. We used a mind map to illustrate the Protocol Experience archetype model, which is presented in figure 5. A total of 18 new *element archetypes* were created to represent the existing Protocol Experience CDEs (see the rectangles). These *element archetypes* were further included in 7 new *entry archetypes* that represent higher-level concepts (see the ovals). Two CDEs could be covered by the openEHR-EHR-ACTION.informed_consent.v1. Therefore, we extended this *entry archetype* by adding a new attribute and directly using the other attribute (see the rectangle). Finally, we created a *section archetype* that includes the 8 *entry archetypes* used for concept grouping purposes. This modeling example can be downloaded from online supplementary appendix B.

DISCUSSION

Comparison with related studies

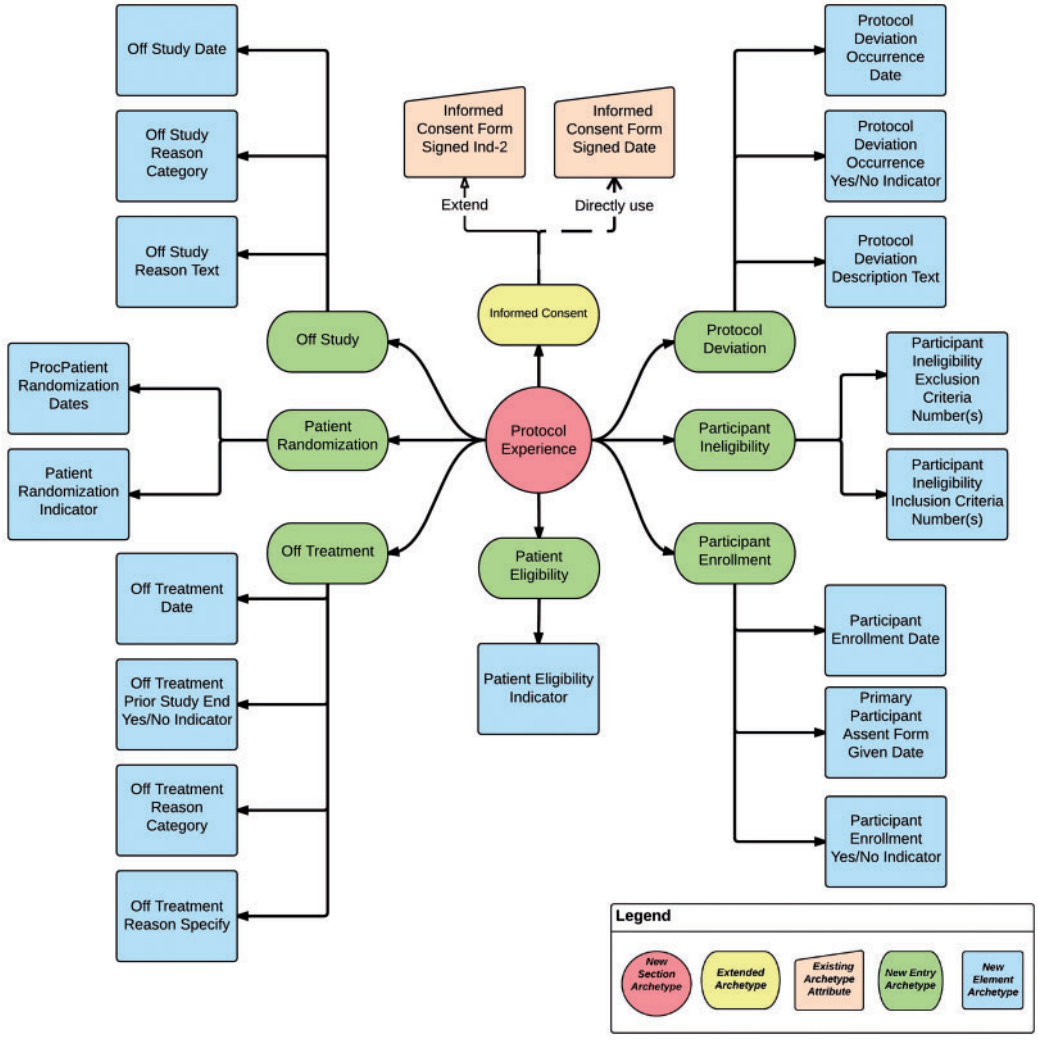
CDEs have been developed over a long time period. However, due to the limitations of their basic structure, the development of CDEs has encountered a bottleneck. To break through the barrier, Park et al[21] tried to address the structure limitations of ISO/IEC 11179 by extending the standard model via the addition of structural and semantic extensions. On the other hand, the CDISC ODM, which is a XML-based underlying data model, supports the acquisition and exchange of metadata specifically related to clinical studies and clinical facts about subjects.[58] The CDISC ODM also tries to rectify the limitations of ISO/IEC 11179;

Table 2: The performance of the semiautomatic mapping tool

CDE Classes	Performance in terms of the semiautomatic mapping tool							
	TP	FP	TN	FN		Precision	Recall	*F* measure
Assessments and examinations	35	91	2	9		0.28	0.80	0.41
Outcomes and end points	14	12	0	9		0.53	1.00	0.69
Participant/subject characteristics	18	34	2	5		0.35	0.78	0.48
Protocol experience	4	16	9	3		0.20	0.57	0.30
Safety data	5	12	0	0		0.29	1.00	0.45
Treatment/intervention data	18	41	0	20		0.30	0.47	0.37
Participant/subject history and family history	14	47	3	3		0.23	0.82	0.36
	Total					**Average**		
	108	253	16	49		0.30	0.69	0.42

Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Figure 5: A mind map of the Protocol Experience common data archetype model



RESEARCH AND APPLICATIONS

however, it is not comprehensive enough for a EDC system that directly uses imported elements for CRF generation.[22] Instead of improving the structure of ISO/IEC 11179, we propose to develop a more comprehensive structure to model CDE concepts as an alternative solution.

The issues related to archetype mapping

Thanks to the effort of various domain experts, the number of CDEs is large and continues to grow. Even though $< 20\%$ of the CDEs need new archetypes to be developed, it still would be a labor-intensive task to transfer the structure of the CDEs, since more than half of the mapped archetypes still require further modification. A certain proportion of development time was spent searching for existing archetypes. A tool to automatically search and map existing archetypes would significantly save in terms of modeling effort, especially because there is a trend towards growth in terms of archetypes.[34] In our study, we developed a semiautomatic mapping tool that was able to help domain experts with the mapping task. Our findings show that, for the total of 184 various class CDEs, the domain experts were able to use our tool to identify 38% of existing archetypes with an F measure of 0.42. Notwithstanding this, the performance of this tool did not satisfy the users completely. More advanced natural language processing techniques, such as the similarity detection approach based on semantic nets and corpus statistics at the semantic level,[59] might be expected to improve performance. However, we are not at all optimistic that such an improvement is possible because of the limitations of these natural language processing techniques. In this area, integration of an information model and terminology has been shown to facilitate semantic interoperability.[60] Unfortunately, only 9% (33 of 349) of archetypes were found to link to standard terminologies in our data source, which was loaded from the CKM. The lack of standard terminology binding is likely to limit effective mapping; therefore, more research effort is needed in the area of archetype terminology binding.[61–63]

A comparison between CDEs and archetypes

A comparison was conducted based on earlier literature and the lessons learnt from the practical modeling of the Protocol Experience CDEs (see table 3). CDEs and archetypes have dissimilar modeling approaches. The CDE concept is represented by a single general-purpose structure, while the archetype uses a 2-level model. On the information side, the CDE structure merely supports data registry, and it does not provide a specific structure to manage the grouping of concepts. This limitation makes CDEs weak when presenting data entries at the group level and forces the user to pick CDEs one by one for each study question when creating a design.[64] In contrast, the reference model underlying an archetype provides a clear hierarchical class structure that enables the presentation of questions at the group level.[22] For example, a *section archetype* can include many *entry archetypes*, and many *element archetypes* can be included in an *entry archetype*. In general, the data structure of both models is essentially similar, namely both have name-value pairs at their core, but archetypes additionally support structural containment.

Compared with CDEs, which need an external EDC system to support CRF design, the openEHR already provides the Template Designer[65] for assisting template designs using archetypes. The templates allow archetypes to be further constrained to suit a particular study topic, which also partially supports data validation. Figure 6 shows the Protocol Experience form obtained from the Template Designer. We are a little concerned regarding the question text display in the archetype template. Unlike CDEs, which have an attribute to record the preferred question text, the archetype display uses the names of concepts and items. Users are able to change the question text during template design, which can be quite inconvenient. The result is that it is quite hard to establish a study question library for reuse and sharing.

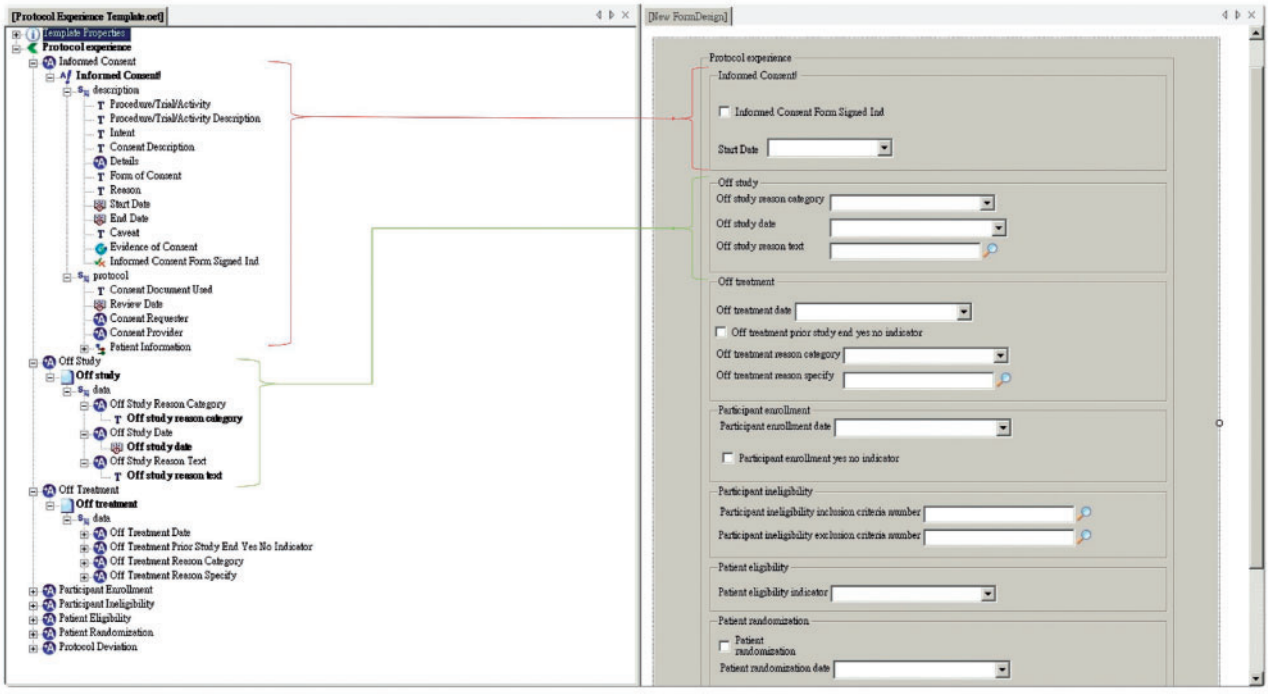
Table 3: Feature comparison of common data elements and openEHR archetypes

	Common Data Elements (CDEs)	openEHR Archetypes
Modeling approach	Single-level model	Two-level model
Concept grouping	Does not provide a certain structure for concept grouping.	The reference model classes provide a clear hierarchical structure.
Case report form design	Needs an external EDC system to support form design. User needs to pick the CDEs one by one for each question.	Possible via openEHR's Template Designer. By picking a higher-level archetype, eg, *section archetype*, the user can create a set of questions in a rational sequence.
Degree of constraint and data validation	Loosely constrained. Needs an external EDC system to support data validation.	Strictly constrained by the reference model. The openEHR template can partially support data validation.
Database design	Compatible with any database schema as long as an association between data entry and CDE is created.	Implemented as an XML database
Data query language	SQL	AQL
Support for terminology binding	Concept and any permissible values are bound by the NCI thesaurus.	All terms in an archetype can be well defined and bound using standard terminology.
Support for concept specialization and extension	Does not allow concept specialization or extension.	Fully supports concept specialization and extension.
Concept governance	Central governance by NCI caDSR.	Central governance by openEHR CKM.
Internationalization	Only English. Does not provide any mechanism for other languages.	The archetype allows translation and thus supports multiple languages.
Concept import/export	UML format and CSV format	ADL format and XML format

Abbreviations: EHR, electronic health record; EDC, electronic data capture; SQL, Structured Query Language; AQL, ArangoDB Query Language; NCI, National Cancer Institute; caDSR, Cancer Data Standards Registry and Repository; CKM, Clinical Knowledge Manager; UML, Unified Modeling Language; CSV, Comma Separated Values; ADL, archetype definition language; XML, Extensible Markup Language.

How to persistently store data is likely to determine whether an information model is practical for a clinical research system. CDEs are used to identify each single data entry during data collection and data harmonization and therefore can be made compatible with any data base schema as long as an association between the data entry and

Figure 6: A protocol experience form obtained via the openEHR Template Designer
EHR, electronic health record.



the CDE has been created. In terms of archetype, the persistence layer is constructed via its reference model, and the database is often implemented as an XML-database.[66] In order to search and retrieve clinical data found in archetype-based records, the openEHR provides the Archetype Query Language, which utilizes archetype path syntax and expresses the query at the semantic level.[67]

As we have mentioned before, semantic interoperability can be built by integrating the information model and standard terminology. The ADL, which is part of the archetype specification, provides a mechanism that allows its archetype term codes (local terms) to be bound to standard terminology.[62] However, CDEs merely allow their concepts and permissible values to be annotated. Well-defined relationships between related concepts are also able to help improve semantic interoperability. Archetypes fully support concept specialization as well as extension to more specific concepts mechanistically; in contrast, CDEs do not have such a mechanism, which results in concept isolation. In addition, the quality of any content is the key to interoperability. A comprehensive information model management repository should be able to help ensure a high-quality model content, which will enable model sharing.[68] Both CDEs and openEHR use a central repository to manage their information models and provide a review mechanism for the content—one is the NCI caDSR and the other is the openEHR CKM. In relation to content, internationalization is also key to the growth and spread of standards. In this context, the archetype is able to support multiple languages with respect to both concept definition and local terms using the ADL. In contrast, CDEs are only available in English, and there is no easy mechanism for translation into other languages.

Limitations and future work

As an exploratory study, the present study has a number of limitations. We only took the NINDS General CDEs as our starting material; therefore, the coverage and the archetype reusability ratio may not completely represent the whole CDE data set. For this reason, it is not recommended to extrapolate our experience to the whole corpus of CDEs. A better mapping strategy is needed to allow for a wide-ranging evaluation. Nevertheless, while the openEHR has been widely implemented in various medical systems, in this study, it was not possible to integrate our common data archetypes into an EDC system. To do further systemic analysis and demonstrate the superiority of the 2-level modeling approach for system development, a new generation EDC system that has been integrated with common data archetypes and supports clinical research will be part of our future work.

CONCLUSIONS

This study investigated the feasibility of applying an advanced modeling structure to represent CDE concepts. The results showed that there is considerable work needed with respect to CDEs because more than half of the mapped archetypes need to be modified. A comprehensive comparison was conducted after practical modeling of the Protocol Experience CDEs. Our results indicate that the openEHR archetype approach is able to comprehensively represent the CDEs and cover their functionality. The preliminary results of the present research can be used as a reference for future development of next generation CDEs.

CONTRIBUTORS

LCH conceived the idea of the study, implemented the prototype system, developed the approach, and drafted the manuscript. FYC and LDM provided key support, coordinated cooperative organizations, and input practical concerns to the study. All authors contributed to the review of the manuscript and approved the final version.

FUNDING

This work was supported by the Taiwan Ministry of Science and Technology (MOST) grant no. MOST 103-2221-E-010-004.

COMPETING INTERESTS

None.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Yamamoto K, Yamanaka K, Hatano E, *et al*. An eClinical trial system for cancer that integrates with clinical pathways and electronic medical records. *Clin Trials* 2012;9(4):408–417.
2. Kush R. Electronic data capture—pros and cons. *BioExecutive Int* 2006;2(6):S48–S52.
3. Common Data Element (CDE) Resource Portal. http://www.nlm.nih.gov/cde/. Accessed August 22, 2014.
4. Albright KC, Martin-Schild S, Bockholt HJ, *et al*. No consensus on definition criteria for stroke registry common data elements. *Cerebrovasc Dis Extra* 2011;1(1):84–92.
5. The CEN/ISO EN13606 standard. EN 13606 Association Website. http://www.en13606.org/the-ceniso-en13606-standard. Accessed September 27, 2014.
6. Kalra D, Beale T, Heard S. The openEHR foundation. *Stud Health Technol Inform* 2005;115:153–173.
7. Schloeffel P, Beale T, Hayworth G, *et al*. The relationship between CEN 13606, HL7, and openEHR. *HIC 2006 and HINZ 2006: Proceedings* 2006:24.
8. Powell J, Buchan I. Electronic health records should support clinical research. *J Med Internet Res* 2005;7(1):e4.
9. Garde S, Knaup P, Schuler T, *et al*. Can openEHR archetypes empower multicentre clinical research? *Stud Health Technol Inform* 2005;116:971–6.
10. Knaup P, Garde S, Merzweiler A, *et al*. Towards shared patient records: an architecture for using routine data for nationwide research. *Int J Med Inform* 2006;75(3):191–200.
11. Kohl CD, Garde S, Knaup P. Facilitating secondary use of medical data by using openEHR archetypes. *Stud Health Technol Inform* 2009;160(Pt 2):1117–1121.
12. Stone K. NINDS common data element project: a long-awaited breakthrough in streamlining trials. *Ann Neurol* 2010;68(1):A11–A13.
13. Common Data Element Definition. US National Library of Medicine Website. http://www.nlm.nih.gov/cde/glossary.html#cdedefinition. Accessed September 25, 2014.
14. Parkinson's Disease Biomarkers Program: PDBP. National Institute of Neurological Disorders and Stroke. https://pdbp.ninds.nih.gov/. Accessed September 25, 2014.
15. Pathak J, Pan H, Wang J, *et al*. Evaluating phenotypic data elements for genetics and epidemiological research: experiences from the eMERGE and PhenX Network Projects. *AMIA Summits Trans Sci Proc* 2011;2011:41.
16. Min H, Ohira R, Collins MA, *et al*. Sharing behavioral data through a grid infrastructure using data standards. *J Am Med Inform Assoc* 2014;21(4):642–649.
17. Subject Areas of NIH CDE Initiatives. US National Library of Medicine Website. http://www.nlm.nih.gov/cde/subject_areas_1.html. Accessed May 12, 2015.
18. National Cancer Institute common data element browser. National Cancer Institute Website. https://cdebrowser.nci.nih.gov/CDEBrowser/. Accessed October 9, 2014.
19. Warzel DB, Andonyadis C, McCurry B, *et al*. Common data element (CDE) management and deployment in clinical trials. *AMIA Ann Symp Proc* American Medical Informatics Association; Washington, DC, 2003.
20. ISO/IEC 11179, Information Technology – Metadata registries (MDR). ISO/IEC JTC1 SC32 WG2 Development/Maintenance Website. http://metadata-standards.org/11179/. Accessed September 29, 2014.
21. Park YR, Yoon YJ, Kim HH, *et al*. Establishing semantic interoperability of biomedical metadata registries using extended semantic relationships. *Stud Health Technol Inform* 2013;192:618–621.
22. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. *J Am Med Inform Assoc* 2011;18(3):341–346.
23. Nadkarni PM, Brandt CA. The common data elements for cancer research: remarks on functions and structure. *Methods Inform Med* 2006;45(6):594.
24. Ngouongo S, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? *J Biomed Inform* 2013;46(2):318–327.
25. What is openEHR? openEHR Foundation Website. http://www.openehr.org/what_is_openehr. Accessed September 27, 2014.
26. Bird L, Goodchild A, Tun ZZ. Experiences with a two-level modelling approach to electronic health records. *J Res Pract Inform Technol* 2003;35(2):121–138.
27. Beale T. Archetypes: constraint-based domain models for future-proof information systems. In: Proceedings of the OOPSLA 2002 conference, Northeastern University, Boston, Seattle, Washington, USA; 2002:16–32.
28. Hovenga E, Garde S, Heard S. Nursing constraint models for electronic health records: a vision for domain knowledge governance. *Int J Med Inform* 2005;74(11):886–898.
29. Maldonado JA, Costa CM, Moner D, *et al*. Using the ResearchEHR platform to facilitate the practical application of the EHR standards. *J Biomed Inform* 2012;45(4):746–762.
30. Chen R, Klein GO, Sundvall E, *et al*. Archetype-based conversion of EHR content models: pilot experience with a regional EHR system. *BMC Med Inform Decis Mak* 2009;9(1):33.
31. Moner D, Moreno A, Maldonado JA, *et al*. Using archetypes for defining CDA templates. *Stud Health Technol Inform*. 2012;180:53–57.
32. Browne E. openEHR Archetypes for HL7 CDA Documents. https://openehr.atlassian.net/wiki/display/stds/openEHR+Archetypes+for+HL7+CDA+Documents. Accessed December 16, 2014.
33. Späth MB, Grimson J. Applying the archetype approach to the database of a biobank information management system. *Int J Med Inform* 2011;80(3):205–226.
34. Buck J, Garde S, Kohl CD, *et al*. Towards a comprehensive electronic patient record to support an innovative individual care concept for premature infants using the openEHR approach. *Int J Med Inform* 2009;78(8):521–531.
35. Braun M, Brandt AU, Schulz S, *et al*. Validating archetypes for the Multiple Sclerosis Functional Composite. *BMC Med Inform Decis Mak* 2014;14(1):64.
36. Hägglund M, Chen R, Koch S. Modeling shared care plans using CONTsys and openEHR to support shared homecare of the elderly. *J Am Med Inform Assoc* 2011;18(1):66–69.
37. Marcos M, Martínez-Salvador B. Towards the interoperability of computerised guidelines and electronic health records: an experiment with openEHR archetypes and a chronic heart failure guideline. Springer: Knowledge Representation for Health-Care; 2011:101–113.
38. Anani N, Chen R, Moreira TP, *et al*. Retrospective checking of compliance with practice guidelines for acute stroke care: a novel experiment using openEHR's Guideline Definition Language. *BMC Med Inform Decis Mak* 2014;14(1):39.
39. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. *J Biomed Inform* 2010;43(5):736–746.
40. Introduction to Archetypes and Archetype classes. openEHR Foundation Website. http://www.openehr.org/wiki/display/healthmod/Introduction+to+Archetypes+and+Archetype+classes. Accessed October 6, 2014.
41. Beale T, Herd S. openEHR Architecture Overview. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CBwQFjAA&url=https%3A%2F%2Fgithub.com%2FopenEHR%2Fspecifications%2Fraw%2FRelease-1.0.1%2Fpublishing%2Farchitecture%2Foverview.pdf&ei=LnlqVMijMpS68gXZ_YHoCw&usg=AFQjCNGOJBHsdPTeLWOI_RJYg6gr5a-aNA&sig2=qpYUDEGzUz8pjTf9YkO4Eg&cad=rja. Published 12 Apr 2007. Accessed October 3, 2014.
42. openEHR ADL 1.5. openEHR Foundation Website. http://www.openehr.org/downloads/ADLworkbench/learning_about. Accessed October 10, 2014.
43. Archetype Definition Language 1.4 (ADL) - openEHR. openEHR Foundation Website. http://www.openehr.org/releases/1.0.1/architecture/am/adl.pdf. Accessed February 3, 2015.

44. openEHR Clinical Knowledge Manager. openEHR Foundation Website. http://www.openehr.org/ckm/. Accessed October 9, 2014.
45. Clinical Knowledge Manager - Ocean Informatics. Ocean Informatics Website. https://oceaninformatics.com/files/solutions/CKM_brochure_2012.pdf. Accessed February 10, 2015.
46. Garde S, Knaup P, Hovenga EJ, *et al*. Towards semantic interoperability for electronic health records–domain knowledge governance for open EHR archetypes. *Methods Inform Med* 2007;46(3):332–343.
47. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. *Methods Inf Med* 2009;48(1):45–54.
48. Parkinson's Disease Biomarkers Program. National Institute of Neurological Disorders and Stroke Website. https://pdbp.ninds.nih.gov/index.jsp. Accessed November 11, 2013.
49. Kuehn BM. Parkinson Biomarker Program. *JAMA* 2013;309(8):759.
50. Buxton AE, Calkins H, Callans DJ, *et al*. ACC/AHA/HRS 2006 key data elements and definitions for electrophysiological studies and procedures: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (ACC/AHA/HRS writing committee to develop data standards on electrophysiology). *J Am College Cardiol* 2006;48(11):2360–2396.
51. Richesson R, Mon D, Kallem C. *A Strategy for Defining Common Data Elements to Support Clinical Care and Secondary Use in Clinical Research, in 2010 AMIA Clinical Research Informatics Summit*. San Francisco, 2010.
52. Kuperman GJ, Blair JS, Franck RA, *et al*. Developing data content specifications for the nationwide health information network trial implementations. *J Am Med Inform Assoc* 2010;17(1):6–12.
53. Gospodnetić O, Hatcher E. *Lucene in action*. Manning Publications, Greenwich, CT; 2005.
54. Chen R, Klein G. The openEHR Java reference implementation project. *Stud Health Technol Inform* 2007;129(1):58.
55. Lucene TFIDFSimilarity score. Apache Lucene Website. https://lucene.apache.org/core/4_3_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html#TFIDFSimilarity. Accessed October 10, 2014.
56. Archetype Editor Home. openEHR Foundation Website. http://www.openehr.org/downloads/archetypeeditor/home. Accessed October 10, 2014.
57. An algorithm for deciding which class to use. openEHR Foundation Website. http://www.openehr.org/wiki/download/attachments/786529/Decision%20algorithm.gif?version=3&modificationDate=1195616905000&api=v2. Accessed October 11, 2014.
58. Iberson-Hurst D. THE CDISC OPERATIONAL DATA MODEL: READY TO ROLL? *Appl Clin Trials*. 2004;13(7):48–53.
59. Li Y, McLean D, Bandar ZA, *et al*. Sentence similarity based on semantic nets and corpus statistics. *IEEE T Knowl Data Eng* 2006;18(8):1138–1150.
60. Markwell D, Sato L, Cheetham E. Representing clinical information using SNOMED clinical terms with different structural information models. Proceedings of KR-MED May 31-June 2 2008; Phoenix, Arizona, USA 2008.
61. Meizoso García M, Iglesias Allones JL, Martínez Hernández D, *et al*. Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations. *Int J Med Inform* 2012;81(8):566–578.
62. Yu S, Berry D, Bisbal J. Clinical coverage of an archetype repository over SNOMED-CT. *J Biomed Inform* 2012;45(3):408–418.
63. Sundvall E, Qamar R, Nyström M, *et al*. Integration of tools for binding archetypes to SNOMED CT. *BMC Med Inform Decis Mak* 2008;8(Suppl 1):S7.
64. caDSR Form Builder. National Cancer Institute Wiki Website. https://wiki.nci.nih.gov/display/caDSR/caDSR+Form+Builder#caDSRFormBuilder-ToolOverview. Accessed November 11, 2013.
65. openEHR Template Designer. Ocean Informatics Website. https://oceaninformatics.com/solutions/knowledge_management. Accessed October 27, 2014.
66. Freire SM, Sundvall E, Karlsson D, Lambrix P. Performance of XML Databases for Epidemiological Queries in Archetype-Based EHRs. *Scand Conf Health Inform* 2012:51–57.
67. Archetype Query Language Description. openEHR Foundation Website. http://www.openehr.org/wiki/display/spec/Archetype+Query+Language+Description#ArchetypeQueryLanguageDescription-WhatisAQL?. Accessed October 22, 2014.
68. Garde S, Hovenga EJ, Gränz J, *et al*. Managing archetypes for sustainable and semantically interoperable electronic health records. *Electronic J Health Inform* 2007;2(2):e9.

AUTHOR AFFILIATIONS

[1]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan

[2]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA