Simplifying Complex Clinical Element Models to Encourage Adoption

Robert R. Freimuth, PhD*, Qian Zhu, PhD*, Jyotishman Pathak, PhD, and Christopher G. Chute, MD, DrPH
Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

Clinical Element Models (CEMs) were developed to provide a normalized form for the exchange of clinical data. The CEM specification is quite complex and specialized knowledge is required to understand and implement the models, which presents a significant barrier to investigators and study designers. To encourage the adoption of CEMs at the time of data collection and reduce the need for retrospective normalization efforts, we developed an approach that provides a simplified view of CEMs for non-experts while retaining the full semantic detail of the underlying logical models. This allows investigators to approach CEMs through generalized representations that are intended to be more intuitive than the native models, and it permits them to think conceptually about their data elements without worrying about details related to the CEM logical models and syntax. We demonstrate our approach using data elements from the Pharmacogenomics Research Network (PGRN).

Introduction

Data normalization requires transforming information into a common semantic and syntactic representation. Normalization projects are often conducted within a defined community, consortia, or network in an effort to improve data interoperability among members. These efforts are often conducted retrospectively and include a review of data dictionaries to identify groups of data elements that share common semantic meaning[1,2]. Once sufficiently similar data elements have been identified a new data element is proposed as a local "standard" for use within the defined research context. While local standards may facilitate the collection and analysis of data for a given purpose or project, the narrow scope in which they were defined often prevents their reuse in other contexts, thereby leading to the development of additional context-specific standards. The proliferation of local data standards does not address barriers to interoperability on a larger scale[3,4]. Even in cases where local standards are utilized, data sets generated by a research study often remain in a standalone data repository that is not integrated with other clinical systems because of difficulties aligning the data elements to a given EMR data model. If the data is not integrated and accessible, researchers will have limited ability to discover, mine, and reuse the data.

These issues may be addressed by Clinical Element Models (CEMs), which are hierarchical, logical models of clinical data that can be readily aligned to EMR data models[5,6]. Our previous work[2] demonstrated the use of CEMs to standardize pharmacogenomics data elements collected from the Pharmacogenomics Research Network (PGRN)[7]. In parallel, the SHARPn project demonstrated how CEMs can be used as an implementation-independent means for exchanging clinical data[8,9]. Together, these projects illustrated how CEMs can be used to avoid the creation of local data standards and to decompose precoordinated data elements into forms that better fit EMR data models.

The CEM standard is quite complex, however, and specialized knowledge and training is required to implement the models appropriately. Therefore, the same complexity that makes the standard a robust and valuable resource becomes a barrier to investigators that lack the knowledge to adopt it. The goal of this project was to address that limitation by developing a system that would enable investigators to utilize the CEM standard for standardized data representation without first requiring them to acquire specialized knowledge about the specification.

To accomplish this goal we introduced a layer of abstraction over the CEMs that allows investigators to think conceptually about their data elements without getting bogged down in implementation details. This layer of abstraction, termed "Patterns", also illustrates how "primitive" data elements within CEMs can be used to construct more complex, semantically precoordinated data elements. We demonstrate our approach using use cases from the Pharmacogenomics Research Network (PGRN), but this method is also applicable to other data standards and scientific domains[3].

Materials and Methods

The "Pattern" is at the center of our proposed approach. Patterns group together the CEM attributes that are necessary to represent a given abstract data element, which eliminates the need for an investigator to examine the

underlying models and determine themselves how to utilize the standard. The Pattern meta-model, the creation of patterns for the PGRN, and the evaluation of this approach are described below.

Clinical Element Models (CEMs)

GE and Intermountain Healthcare developed a large library of CEMs, which are hierarchical, logical models of clinical data. CEMs are defined using the Constraint Definition Language (CDL)[5] and are available through the CEM Browser[4] in both CDL and XSD format. The CEMs that were used for this project were loaded into a local database to support aggregation and mapping at the attribute level. Figure 1 shows a portion of the SHARPn "Patient" CEM. The hierarchical model includes composite attributes (e.g., PersonName) that contain atomic attributes (e.g., FamilyName) and HL7 V3 datatypes (e.g., ST). Each attribute has a "key" (e.g., PersonName_KEY_ECID), a coded term that provides semantic meaning for the attribute, and cardinality (e.g., 1..M).

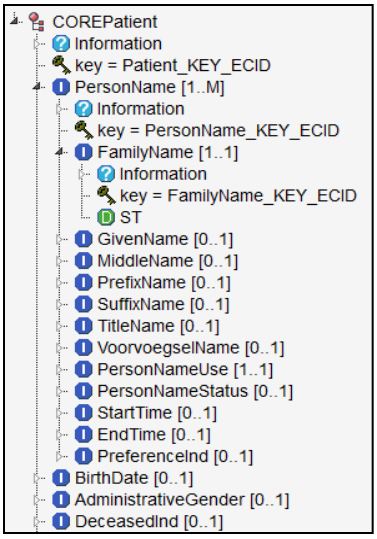

Figure 1: A portion of the SHARPn "Patient" CEM

Pattern Meta-Model

Patterns group all of the attributes that are necessary to represent a given abstract data element into a single container, thereby eliminating the need for an investigator to examine the underlying CEMs and try to determine themselves how to represent their data using the standard models. A conceptual diagram of the Pattern meta-model is shown in Figure 2.

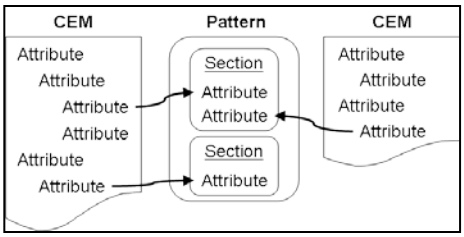

Figure 2: A conceptual diagram of the CEM Pattern meta-model. Not all components are shown (see text).

A Pattern has a name and description, and is composed of one or more Sections. Sections are logical groupings of one or more related CEM attributes. Sections have a name, definition, and an ordered list of attributes. Each attribute in a Section has a display name, datatype, and description. Sections can contain attributes from different CEMs or from different branches within a CEM hierarchy.

The ability to group attributes from structured, logical models into arbitrary sections enables the creation of conceptual abstract data elements, which can then be instantiated as study-specific data elements. It is important to note that none of the semantics of the underlying models are altered. Detailed mappings are maintained between the elements used in a Pattern and their respective source models. These mappings, as well as the source models themselves, can be hidden to simplify the information that is presented to the consumer.

Pattern Creation

The creation of a Pattern by a knowledge engineer includes three steps: the identification of an abstract data element, the identification of CEM attributes that capture the semantics of the data element, and the definition of the Pattern itself. This process helps to transform study-specific data elements into standardized representations that are more likely to be EMR-compliant and usable for secondary purposes.

The first step in creating a reusable abstract data element is to identify study-specific elements that can be generalized. It is common for investigators to define data elements based on a hypothesis or data analysis plan, as it is convenient to think about data in terms of how it will be used. Unfortunately, this often leads to the generation of study-specific, non-reusable data that is difficult to enter into an EMR due to excessive semantic pre-coordination. For example, it is common to collect age-based data in clinical trials. In previous studies we observed data elements derived from questions such as "how old was the patient when diagnosed with diabetes" and "at what age did you first experience palpitations"[1,2]. The answer to each of these questions, representing the age of the patient in years, is captured as a single integer. Both of these data elements can be generalized as the abstract data element "age at diagnosis of disease", where the specific disease is not specified until the data element is instantiated.

Once an abstract data element has been identified it must be decomposed into atomic concepts that are represented as attributes in CEMs. In our experience, this almost always results in the expansion of the number of data elements that need to be captured because data elements used in research tend to have some degree of semantic pre-coordination. For example, the abstract data element "age at diagnosis of disease" requires several attributes from

two different CEMs. The concept of "disease" is represented by a term from a controlled terminology (e.g., ICD, SNOMED-CT), and is captured in an attribute from the Disease/Disorder CEM. The concept of "age at diagnosis" is the result of a calculation that computes the difference between the date of diagnosis (from the Disease/Disorder model) and the patient's date of birth (from the Patient model) (Figure 3). Therefore, a minimum of three attributes are required for this abstract data element: disease code, date of diagnosis, and patient date of birth.

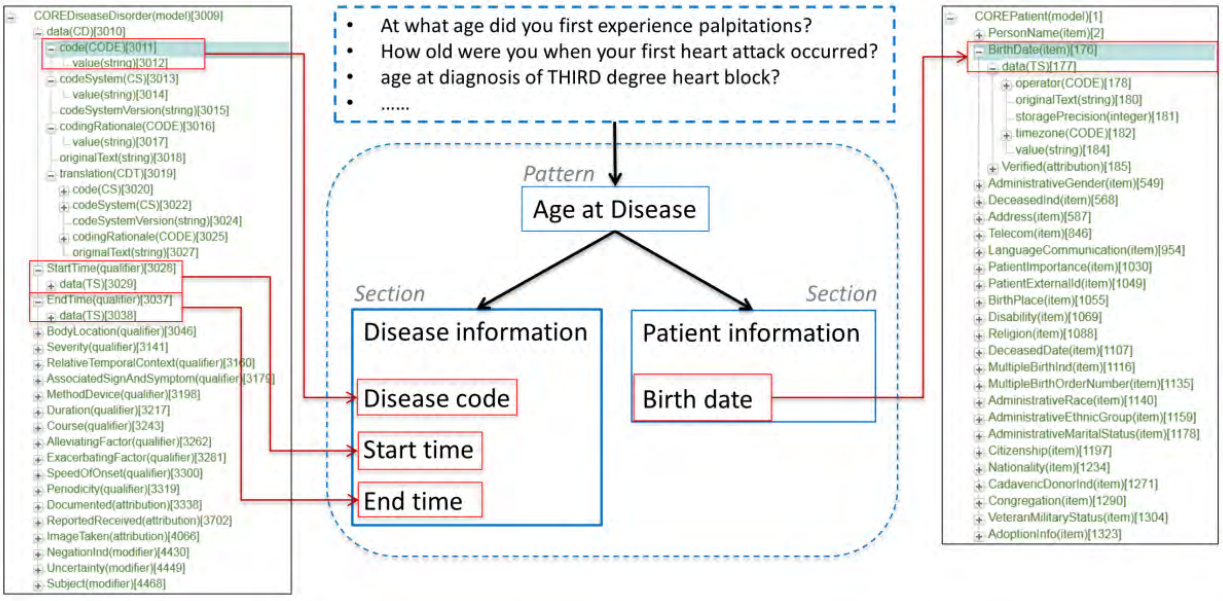

Figure 3: Identification of attributes for a Pattern. Attributes are selected from the Disease/Disorder model (left) and the Patient model (right) for the "Age at Disease" Pattern. See text for details.

A Pattern can be created once attributes have been identified. Since Patterns are intended to be intuitive, conceptual representations it may be necessary to provide user-friendly names or definitions instead of using the formal, technical ones that are part of the underlying logical model. It is important to note that the semantic meanings of the attributes are not changed and that a precise mapping is maintained between a CEM attribute and its representation within a Pattern.

Additional attributes can be added to increase the applicability of the Pattern to other, semantically similar data elements. This is consistent with the notion of creating a representation of a generalized, abstract data element. For example, the abstract data element "age at remission of disease" is closely related to "age at diagnosis of disease". The only difference between the two is whether the start date (for diagnosis) or end date (for remission) of the disease is used in the calculation of age. Therefore, it is reasonable to add this attribute to the pattern, further abstracting the pattern to simply "age at disease" (Figure 3). Note that not all attributes within a Pattern need to be used for a specific implementation, so instantiating a data element that captures age at diagnosis would not require the disease end date to be populated.

After a Pattern has been created, documentation is added to explain the purpose of the Pattern, the logical model(s) it was derived from, and how it should be instantiated for a particular implementation or research project.

User Interface

A web-based graphical user interface was developed to facilitate pattern creation (Figure 4). The figure illustrates the workflow for creating a Pattern with the web-based user interface. A new pattern is given a name and introductory text (which may include HTML tags). One or more sections are then created, which are populated with attributes that represent the semantics of the abstract data element. Attributes are selected from one or more CEMs that are stored in a local database. Attributes can be given a user-friendly display name, description, and order, which is displayed when the user views the completed pattern detail screen.

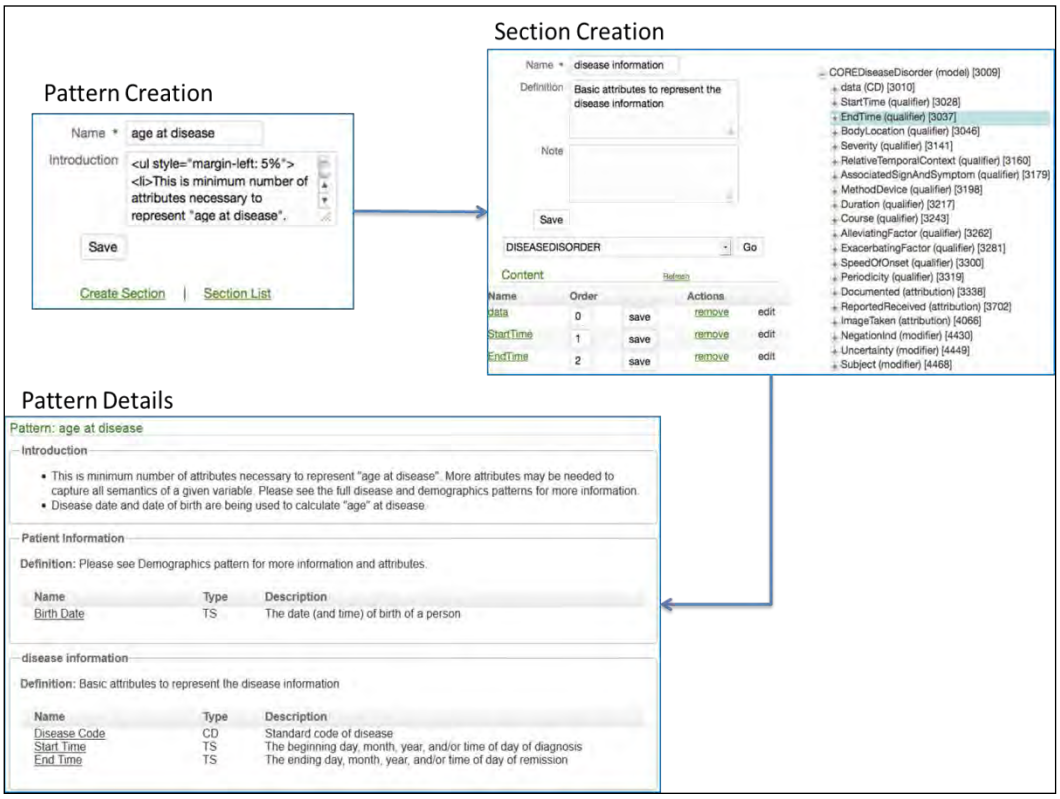

Figure 4: User interface for Pattern creation. The figure illustrates the workflow for creating a Pattern with the web-based user interface. The "age at disease" pattern is shown as an example.

Application and Evaluation

We evaluated this approach by extending our previous work on the semantic harmonization of data dictionaries from the Pharmacogenomics Research Network (PGRN)[7]. Due to the large size of the data set, RRF and QZ grouped PGRN data elements into arbitrarily-defined categories[2] to facilitate review and pattern identification. The data elements in each category were carefully reviewed and candidate patterns were identified, created, and documented using the process described above for each group of semantically similar data elements. This process was repeated until the majority of the PGRN data elements were mapped to a Pattern and a variety of Patterns were created.

The ability to use Patterns to capture the semantics represented by the PGRN data elements was evaluated by reviewing the mappings between each data element and its respective Pattern. In particular, QZ and RRF determined how each attribute would be used, and how they might be combined, to represent a given PGRN data element. The results were recorded as instructions for how the Pattern could be used to represent a given data element. If the initial definition of a Pattern could not fully represent a PGRN data element, the Pattern was extended as needed.

Results

The Pattern approach developed by this project was intended to provide an intuitive way for investigators to use standardized but highly complex CEMs for data collection without first acquiring specialized knowledge. To determine whether this approach could successfully represent the semantics of data elements used by research studies while still maintaining a link to the underlying formal models, which is necessary to enable transformation of the data into CEM syntax, we performed two evaluations. First, the mappings of 2,089 PGRN data elements that were previously mapped directly to CEMs were examined to verify that the semantics were retained when the data elements were mapped to one of the 16 Patterns created for this study (Table 1). All mappings were found to be semantically complete. Second, to demonstrate that the Pattern approach significantly reduced the complexity of the models by retaining only those attributes that were required to express the semantics of the data elements, the reduction in the number of attributes was quantified for each Pattern (Table 1). All Patterns reduced the number of attributes that an adopter would have to consider by >90%.

Pattern	Source CEMs	Number of Attributes		
		Source CEMs	Pattern	Difference (%)
Address	Patient	96	8	-88 (-92%)
Person Identifier	Patient	96	10	-86 (-90%)
Telecom	Patient	96	6	-90 (-94%)
Demographics	Patient, Primary Cause of Death	100	8	-92 (-92%)
Age at Disease	Disease/Disorder, Patient	221	4	-217 (-98%)
Disease	Disease/Disorder	125	12	-113 (-90%)
Disease History	Disease/Disorder	125	4	-121 (-97%)
Family History of Disease	Disease/Disorder, Personal Relationship Type	125*	6	-119 (-95%)
Drug Administration	Noted Drug	113	17	-96 (-85%)
Drug Admin. History	Noted Drug	113	5	-108 (-96%)
Laboratory Observation (Coded Result)	Lab Observation Coded	187	6	-181 (-97%)
Laboratory Observation (Quantitative Result)	Lab Observation Quantitative	190	6	-184 (-97%)
Blood Pressure	Systolic BP Meas., Diastolic BP Meas.	115	3	-112 (-97%)
Mean Arterial Pressure	Mean Arterial Pressure Meas.	115	2	-113 (-98%)
Heart Rate	Heart Rate Meas.	33	3	-30 (-91%)
Height Weight Measurement	Height Meas., Body Weight Meas., Body Mass Index Meas.	76	4	-72 (-95%)

Table 1: Patterns created for this study. The count of attributes from source models included the CEM elements of type item, modifier, qualifier, and attribution. Meas = Measurement. *The Personal Relationship Type model was not finished at the time of writing and therefore was excluded from the count.

Discussion and Conclusion

While CEMs provide a robust framework for the semantic representation and exchange of clinical data, their complexity can be a barrier to adoption. To facilitate the use of CEMs by investigators we sought to develop an approach that would present CEM content in a way that is more intuitive to non-informaticians while still retaining the full semantics of the underlying model. The Pattern approach permitted the creation of simplified, conceptual representations by hiding the complexity of the underlying CEMs that was not necessary for capturing the semantics of the data elements. Furthermore, the process of identifying generalized abstract data elements from groups of semantically similar, study-specific data elements resulted in the creation of relatively few Patterns, indicating that this approach is likely to scale well for all but the most highly-specialized data elements.

Current limitations of this approach include difficulty capturing highly qualified data elements (such as recording episodes of a disease that are associated with a specified condition) and modeling transformations that lead to derived values (e.g., logarithm), both of which are also limitations of the underlying CEM specification. These were addressed by using existing CEM attributes where possible (e.g., exacerbating factor) and by representing the pre-calculated value, respectively. Discussions with the model owners were required to establish consistent practices for representing complex concepts (e.g., pedigrees) and to extend the CEMs when needed, which may limit scalability. These issues could be mitigated by improved CEM documentation and tooling that supports community authoring.

To use CEMs directly, specialized knowledge is required to understand both the technical specification and how to interpret and implement the models themselves. The Pattern approach developed in this study eliminates those requirements for investigators by prespecifying, through hidden mappings to the underlying models, which attributes should be used to capture each aspect of a precoordinated, study-specific data element. The highly simplified models and intuitive instructions provided by Patterns lower the barrier for investigators to use standardized CEMs for data collection. The mappings between Pattern and CEM attributes permit the transformation of data into CEM syntax, which can then be used for data exchange or integration into an EMR. This approach may encourage the adoption of CEMs for data collection, thereby reducing the need for retrospective data normalization efforts.

Acknowledgment

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network). The authors thank Dr. Thomas Oniki for critical clarifications regarding the implementation of CEMs and Mr. Zonghui Lian for providing technical support.

*RRF and QZ contributed equally to this work.

References

1. Zhu Q, Freimuth RR, Lian Z, et al. Harmonization and semantic annotation of data dictionaries from the Pharmacogenomics Research Network: A case study. J Biomed Inform. 2013; 46(2):286-293.
2. Zhu Q, Freimuth RR, Pathak J, Chute CG. Using Clinical Element Models for pharmacogenomic study data standardization. Proc 2013 AMIA Summit on Clinical Research Informatics. 2013. 292-296.
3. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007; 14(6):687-696.
4. Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nat Genet. 2012; 44(2):121-126.
5. James A. Qualibria Constraint Definition Language (CDL) language guide. Oct. 22, 2010.
6. CEM Browser [Internet]. [cited 2013 October 10]. Available from: http://www.clinicalelement.com
7. Pharmacogenomics Research Network (PGRN) [Internet]. [cited 2013 October 10]. Available from: http://www.pgrn.org
8. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform. 2012 Aug; 45(4):763-71.
9. Tao C, Jiang G, Oniki TA, et al. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc. 2013; 20:554-562.