Using Clinical Element Models for Pharmacogenomic Study Data Standardization

Qian Zhu, PhD\textsuperscript{1}, Robert R Freimuth, PhD\textsuperscript{1}, Jyotishman Pathak, PhD\textsuperscript{1}, Christopher G Chute, MD, DrPH\textsuperscript{1}

\textsuperscript{1}Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract

Standardized representations for pharmacogenomics data are seldom used, which leads to data heterogeneity and hinders data reuse and integration. In this study, we attempted to represent data elements from the Pharmacogenomics Research Network (PGRN) that are related to four categories, patient, drug, disease and laboratory, in a standard way using Clinical Element Models (CEMs), which have been adopted in the Strategic Health IT Advanced Research Project, secondary use of EHR (SHARPn) as a library of common logical models that facilitate consistent data representation, interpretation, and exchange within and across heterogeneous sources and applications. This was accomplished by grouping PGRN data elements into categories based on UMLS semantic type, then mapping each to one or more CEM attributes using a web-based tool that was developed to support curation activities. This study demonstrates the successful application of SHARPn CEMs to the pharmacogenic domain. It also identified several categories of data elements that are not currently supported by SHARPn CEMs, which represent opportunities for further development and collaboration.

1. Introduction

Pharmacogenomics is a multidisciplinary and data-intensive science that requires increasingly clear annotation and representation of phenotypes to support data integration. The Pharmacogenomics Research Network (PGRN)\textsuperscript{[1]} is a collaborative partnership of research groups funded by the U.S. National Institutes of Health to perform pharmacogenomics studies that aim to discover and understand how genomic variation contributes to an individual’s response to medication. The variety of disease phenotypes are studied in the PGRN, as well as differences in clinical systems in use at each PGRN site, lead to data that is heterogeneous, non-standardized, and institution-specific. This not only hinders the data aggregation among collaborating sites on a given study, but also complicates or prevents secondary use of the data (e.g., in meta-analyses). To help overcome these issues we performed a survey of data dictionaries from PGRN research sites with the goal of identifying overlapping as well as non-overlapping data elements (DEs) among sites, as well as proposing standards that establish both a common semantic meaning and representation for the data.

The Clinical Element Model (CEM)\textsuperscript{[2-3]} was designed to provide a consistent architecture for representing clinical information in electronic health record (EHR) systems. GE and Intermountain Healthcare have developed a version of the CEM representation, defined in the Constraint Definition Language (CDL)\textsuperscript{[2]}. CEMs in CDL have been adopted in the Strategic Health IT Advanced Research Project, secondary use of EHR (SHARPn)\textsuperscript{[4]} as the common unified information model to ensure unambiguous data representation, interpretation, and exchange within and across heterogeneous sources and applications.

Pharmacogenomic study data includes genomics data elements as well as phenotyping clinical data elements, which are the target of CEMs. This study focused on the clinical portions of the PGRN data dictionaries, and we attempted to standardize pharmacogenomics study data by adopting four SHARPn CEMs: Patient (which includes demographics), Noted Drug (for medication administrations), Laboratory Observation (test results), and Disease/Disorder (diagnoses and problem lists). In order to utilize these CEMs, we first annotated and grouped PGRN metadata into different categories by semantic type. For each category, we semi-automatically mapped the data elements to CEM attributes using a graphical user interface.
interface (GUI), which allowed us to represent them from a standard CEM point of view. The details will be introduced in the following sections.

2. Materials

2.1 PGRN data

Data dictionaries were collected from PGRN research sites in a wide variety of formats, including text documents, spreadsheets, and relational database schemas. To accommodate the granularity of information provided for each data element, the data dictionaries were reformatted and loaded into local database to facilitate further analysis, as described elsewhere [5].

2.2 Clinical Element Models (CEM)

The complete details of the four CEMs utilized in this study are available at [6]. Each CEM model represents as a hierarchical tree with parent and child relationship, such as person name and family name, shown in Figure 1. To facilitate mapping, we assigned a unique identifier to each CEM attribute and loaded the CEM models into a MySQL database. The hierarchical structure of the CEMs was maintained by tracking the parent of each attribute. The CEM attributes were the targets of the mapping process. Each PGRN DE was mapped to one or more CEM attributes semi-automatically.

![Figure 1. Four CEM models used as references in this study (not all attributes are shown)](image)

3. Methods

3.1 Semantic annotation for PGRN data

To improve the efficiency of the mapping and standardization process, each data element description was decomposed and mapped to standardized terminologies using the Bioportal API [7]. The semantic types assigned to each decomposed component were used to group DEs into categories, as described previously[5]. The four categories: demographics, medication, disease and laboratory were the focus of this study.

3.2 Data standardization using CEMs

We attempted to represent each DE from the four categories listed above using CEMs. This was accomplished by breaking down the semantically precoordinated DEs into components that are more context-neutral and therefore more reusable. The mapping process was guided by the mapping rules, described below.

3.2.1 Predefined mapping guidelines

One of the challenges that we faced is harmonizing heavily pre-coordinated DEs from disparate studies across the PGRN network. To overcome this, we established a process by which we decomposed each DE into its constituent components, and then attempted to map each component to an attribute from a CEM. As a result, some DEs were mapped to multiple CEM attributes. To maintain consistent mappings during variable decomposition, we established a set of rules (Table 1) that linked components of the pre-coordinated DE to discrete attributes in the CEMs.

PGRN drug related DEs containing temporal qualifiers, associated particular events and / or disease information were mapped to CEM according to the corresponding mapping rule. For example, DEs related
to drug administration, including dose, route, and strength, were mapped to the corresponding attributes from the drug administration CEM. More complex drug-related DEs contained temporal or event-based qualifiers, or terms related to disease treatment. Similarly, DEs related to diseases often contained temporal qualifiers and/or patient information. Elements related to lab observations were primarily focused on test results and were mapped to attributes based on their datatype (quantitative or coded). For example, “Bicarbonate (mMol / L)” records a quantitative value for bicarbonate test, while “Urine Pregnancy Test” records a coded value (positive or negative).

Table 1. Mapping rules defined for each CEM

| CEM Models      | Data Elements            | Mapping Rules                                |
|-----------------|--------------------------|----------------------------------------------|
| Patient         | Patient Demographics     | Corresponding patient CEM attributes         |
| Noted Drug      | Drug + Temporal          | Drug Administration + Duration + StartTime + EndTime |
|                 | Drug + Age               | Drug Administration + StartTime + BirthDate  |
|                 | Drug + Event             | Drug Administration + Events (Procedure / Lab test / Disease) |
| Disease Treatment| Drug Administration     | Drug Administration + Disease code         |
|                 | Drug Code + Drug + Drug Dose / Route Method Device / Status Change |
| Disease/Disorder| Medical History          | Disease Code + + StartTime + EndTime       |
|                 | Sign and Symptom         | Corresponding disease CEM attributes        |
|                 | Disease + Age            | Disease Code + StartTime + BirthDate        |
| StandardLabObs  | Specimen Collection      | Lab Code + Specimen Collection Time         |
|                 | Lab Test Result          | Lab Code + Lab Coded Value                  |
|                 | Lab Test Status          | Lab Code + Lab Result Status                |

Figure 2 illustrates the breakdown of a complex DE into CEM attributes (refer to the “age at disease” mapping rules from Table 1). In this example, the DE representing the age of a patient when diagnosed with peripheral vascular disease is separated into three components. “Peripheral Vascular Disease” is mapped to the Disease/Disorder CEM and is represented by a SNOMED-CT code (which is stored in the “Data” attribute). The concept of “age” is calculated using the “BirthDate” from the Patient model and the disease date, which is recorded in the “StartTime” attribute in the Disease/Disorder CEM. As a result, the DE “age at Peripheral Vascular Disease” was ultimately represented by three attributes (BirthDate, StartTime and Data) from two CEM models (patient and disease/disorder models), and one calculation (age).

3.2.2 CEM mapping management Graphic User Interface (GUI)

To facilitate the CEM mapping process, a web application was developed that allows the curator to conduct the mapping task (Figure 2). With this GUI, the curator can browse PGRN DEs, search for appropriate CEM attributes from the CEM database, determine standardization workflow status, and submit the mapping results to the database.
4. Results

4.1 CEM Tree

Each CEM model is a hierarchical tree, as shown in Figure 1, that consists of attributes and associated datatypes. We stored all CEMs in a local repository; an example of the database schema is shown in Figure 4. Each CEM attribute and datatype was assigned a unique ID. The parent ID of each element was recorded to maintain a record of the attribute hierarchy.

The Patient model contained 1,341 entries, which included CEM attributes and complex data types. Similarly, there were 1,667 entries from the Noted Drug, 1,659 entries from the Disease/Disorder, and 2,325 entries from the Laboratory Observation models. These attributes (nearly 7000 total) formed the standard elements that were used as "targets" for the CEM mapping process.

4.2 CEM mapping results

We collected 4,484 DEs from 9 PGRN groups. We successfully mapped 1861 (40%) DEs to one or more of the four CEMs used as standards for this study: 484 (10.8%) mapped to the Laboratory Observation CEM, 551 (12.3%) to Noted Drug, 601 (13.4%) to Disease Disorder, and 301 (6.7%) to Patient.

Figure 5 presents a summary of the CEM mapping results. Of the 1,861 mapped DEs, 1,318 DEs were mapped to a single CEM attribute, such as “race” or “gender” (which mapped to “Administrative Race” and “Administrative Gender”, respectively from the “Patient” model). Conversely, 543 DEs were more complex and required mappings to multiple CEM attributes, such as “age at Peripheral Vascular Disease” (which mapped to three CEM attributes from two CEM models, as shown in Figure 2). Of the 1,861 DEs,
the semantics of 1,295 were judged to be represented completely by the mappings. The mappings of 566 DEs captured only a portion of the semantics of the original DEs, however. In these cases, additional CEM attributes are needed to fully capture the content of original DE. For example, the “marital status” attribute in the “Patient” CEM can be used to represent the current marital status of a patient, but an extension of the model is needed to represent the DE “Change in marital status of your parents”.

Of the 4,484 DEs, 2,623 (58%) were not mapped because they were unrelated to the four CEMs used in this study or represented site-specific DEs (e.g., workflow tracking), or because they represented generic or under-defined entities (Figure 5).

![Figure 5. CEM mapping summary (research focus in this study shown in the red dashed rectangle)](image)

5. Discussions and Conclusion

Data collected from PGRN network is site and study specific design by using different semantics and syntactics, hence, it makes harder to reuse and share the data within or out of the network, and the pace of advancement and discovery will be slowed down obviously. Fortunately, data and metadata standards can help to mitigate such major barriers. In this study, we evaluated the ability of four SHARPn CEMs (Patient, Noted Drug, Laboratory Observation, and Disease/Disorder) to serve as a standardized representation for data collected by the PGRN. Approximately 40% of the PGRN DEs have been successfully represented by these four CEMs. Of the remaining 60% of DEs, 41% represented clinical findings, 33% represented quality of life or cognitive assessment measures, and 26% were concepts such as adverse events, clinical procedures, pharmacokinetics/pharmacodynamics, and genomics, none of which are currently represented by SHARPn CEMs. These are clear areas for further work, and we are currently collaborating with SHARP to extend existing and develop new CEMs to fill these gaps. In addition, we continue to establish and document best practices related to the adoption and use of CEMs. These efforts will enable data managers to more easily implement CEMs when designing project-specific data dictionaries, which will ultimately result in data that is more consistent, comparable, and reusable.

Acknowledgment

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network). We thank Mr. Zonghui Lian for his help on technical support.

References

1. PGRN network, http://pgrn.org/display/pgrnwebsite/PGRN+Home
2. James A. Qualibria Constraint Definition Language (CDL), Language Guide; 2010 Oct. 22, 2010
3. GE/Intermountain Clinical Element Model, http://intermountainhealthcare.org/CEM/
4. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform. 2012.
5. Zhu Q, Freimuth R, Lian Z, Bauer S, Durski M, Tao C, Pathak J, Chute C. Harmonization and semantic annotation of data dictionaries from the Pharmacogenomics Research Network: a case study, J Biomed Inform. 2012 Nov 29.
6. SHARP, http://informatics.mayo.edu/sharp/index.php/Main_Page
7. BioPortal REST API, http://www.bioontology.org/wiki/index.php/BioPortal_REST_services