



Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience

Jyotishman Pathak,¹ Janey Wang,² Sudha Kashyap,² Melissa Basford,² Rongling Li,³ Daniel R Masys,² Christopher G Chute¹

► Additional materials are published online only. To view these files please visit the journal online (www.jamia.org).

¹Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

³National Human Genome Research Institute, Bethesda, Maryland, USA

Correspondence to

Jyotishman Pathak, Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, USA; pathak.jyotishman@mayo.edu

Received 23 December 2010
Accepted 7 April 2011
Published Online First
19 May 2011

ABSTRACT

Background Systematic study of clinical phenotypes is important for a better understanding of the genetic basis of human diseases and more effective gene-based disease management. A key aspect in facilitating such studies requires standardized representation of the phenotype data using common data elements (CDEs) and controlled biomedical vocabularies. In this study, the authors analyzed how a limited subset of phenotypic data is amenable to common definition and standardized collection, as well as how their adoption in large-scale epidemiological and genome-wide studies can significantly facilitate cross-study analysis.

Methods The authors mapped phenotype data dictionaries from five different eMERGE (Electronic Medical Records and Genomics) Network sites studying multiple diseases such as peripheral arterial disease and type 2 diabetes. For mapping, standardized terminological and metadata repository resources, such as the caDSR (Cancer Data Standards Registry and Repository) and SNOMED CT (Systematized Nomenclature of Medicine), were used. The mapping process comprised both lexical (via searching for relevant pre-coordinated concepts and data elements) and semantic (via post-coordination) techniques. Where feasible, new data elements were curated to enhance the coverage during mapping. A web-based application was also developed to uniformly represent and query the mapped data elements from different eMERGE studies.

Results Approximately 60% of the target data elements (95 out of 157) could be mapped using simple lexical analysis techniques on pre-coordinated terms and concepts before any additional curation of terminology and metadata resources was initiated by eMERGE investigators. After curation of 54 new caDSR CDEs and nine new NCI thesaurus concepts and using post-coordination, the authors were able to map the remaining 40% of data elements to caDSR and SNOMED CT. A web-based tool was also implemented to assist in semi-automatic mapping of data elements.

Conclusion This study emphasizes the requirement for standardized representation of clinical research data using existing metadata and terminology resources and provides simple techniques and software for data element mapping using experiences from the eMERGE Network.

INTRODUCTION

A principal goal of genetic research is to identify specific genotypes that are associated with human phenotypes. With recent advances in genotyping technologies, even though it has become possible to

systematically ascertain large numbers of sequence variants (eg, single-nucleotide polymorphisms) for the complete genome of an individual, our ability to fully understand the genetic basis of common diseases is significantly hindered by our inability to precisely specify the phenotypes (ie, the outward physical manifestation of the genotypes).¹ In particular, phenotyping at large varies greatly between medical specialties and different organizations, and lacks the systematization and throughput compared with large-scale genotype studies.

Hence, to address this important requirement, the US National Institutes of Health has recently funded projects, such as the eMERGE Network (Electronic Medical Records and Genomics²), that correlate whole genome scans with phenotype data extracted from electronic medical records (EMRs). An integral part of such efforts is to standardize the representation of phenotypic data in a dataset. However, in practice, biomedical research applications and clinical systems are developed independently of each other, and therefore do not have a common data representation structure or a data dictionary. This results in the creation of data elements (DEs) anew for each particular study, even for commonly collected data such as demographics, laboratory results, and vital statistics. Clearly, to facilitate the interpretation of such data, it is vital to provide not only a data dictionary to accompany the data, but also appropriate mapping of the DEs to controlled vocabularies and terminological resources to promote secondary reuse and standardization of DEs.

The overarching goal of this work is to investigate how a limited set of DEs, as part of the eMERGE Network studies, can be mapped to standardized metadata repositories and biomedical vocabularies. In particular, we studied mapping of multiple phenotype data dictionaries (dementia, peripheral arterial disease, low high-density lipoprotein (HDL), cataract, type 2 diabetes, and long QT syndrome) from five eMERGE consortium members to existing metadata and terminological resources including National Cancer Institute's (NCI's) caDSR (Cancer Data Standards Registry and Repository),³ Clinical Data Interchange Standards Consortium's (CDISC's) SDTM (Study Data Tabulation Model),⁴ NCI thesaurus (NCI-T),⁵ and SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms).⁶ We also implemented an open-source web services-based toolkit and repository called eleMAP⁷ for semi-automatic mapping of DEs to standardized biomedical vocabularies and metadata registries.

Table 1 List and definitions of abbreviations used

Abbreviation	Full form
eMERGE	Electronic Medical Records and Genomics
EMR	Electronic Medical Record
DE	Data Element
UMLS	Unified Medical Language System
SNOMED CT	Systemized Nomenclature of Medicine-Clinical Terms
NCI	National Cancer Institute
caDSR	Cancer Data Standards Registry and Repository
SDTM	Study Data Tabulation Model
HL7	Health Level 7 (http://www.hl7.org)
CDISC	Clinical Data Interchange Standards Consortium (http://cdisc.org)
IHE	Integrating the Healthcare Enterprise (http://www.ihe.net)

The rest of the article is organized as follows. We begin with a list of abbreviations (table 1) and the terminology (table 2) used throughout. In the next section, we provide background information about data element mapping. We then present the methods for mapping eMERGE Network data dictionaries to standardized metadata and terminological resources and summarize our findings. Finally, we discuss the implications of our investigation along with strengths and limitations and provide a conclusion.

BACKGROUND

Data standards, in essence, are consensual specifications for the representation of data from different sources or settings. Standards facilitate sharing, portability, and reusability of data.⁸ The notion of standardized data includes specifications for both data fields (ie, the ‘variables’) and value sets (ie, the ‘permissible values’) that encode the data within these fields. Although the current focus of data standards relevant to clinical research is primarily on regulated research (eg, clinical trials, safety reporting), it is important to note that clinical research encompasses other types of research, including observational, epidemiological, molecular, and biology (eg, biomarkers for diseases). Consequently, to facilitate the sharing of patient data and enable interoperability between healthcare and clinical research, it becomes important to permeate data standards in clinical practice as well as to make standardization of data in clinical research a high priority.

Toward this end, the process of mapping and standardizing DEs from clinical research studies to terminological and metadata resources can be broadly classified under the well-studied

problem of ‘schema alignment’, where the basic premise is to determine relations (eg, equivalence, subsumption) between existing pairs of elements (ie, the variables and instances) across the schemas. Numerous approaches have been proposed for this problem over the past several years (see surveys^{9–11}), ranging from simple lexical and structural alignment techniques^{12–14} to more complex machine learning-based methods.^{15–16} For instance, Ghazvinian *et al*^{12–13} developed the LOOM (Lexical OWL Ontology Matcher) algorithm which performs simple string comparisons between preferred names or synonyms for the concepts to identify similarity. Interestingly, LOOM outperformed most of the complex schema (and ontology) matching algorithms and showed significant performance enhancements. Similarly, Mougín *et al*¹⁴ demonstrated the benefits of lexical matching techniques to harmonize DEs to entries in a terminological resource. At the other end of the spectrum, Eckert and colleagues¹⁶ developed automatic machine learning techniques on an ensemble of schema matchers that learn rules for the correctness of a correspondence based on the output of different matchers, and additional information about the nature of elements to be matched. The authors demonstrate that their tool systematically outperforms existing matching tools. A similar approach was implemented in the GLUE system,¹⁵ which applies a meta-learning algorithm for generating matching hypotheses on the basis of multiple local classifiers that are trained on different aspects of the schemas to be matched.

In addition to the above work, there are several national efforts, under the rubric of Meaningful Use, spearheaded by the Office of the National Coordinator as part of the Health IT Standards Committee Vocabulary Task Force¹⁷ for data standards and harmonization. Of particular relevance to our study is the clinical research interoperability specification,¹⁸ which covers clinical research as it interoperates with healthcare systems, particularly EMR systems. Based on standards from healthcare (HL7 and IHE) and research (CDISC), this specification focuses on exchange of a core set of patient-level information between EMRs and clinical research systems. Specifically, this specification identifies and defines a library of DEs (referred to as module categories) that may be used by clinical and healthcare systems for standards-based exchanges of information. Tables 3 and 4 show examples of value set definition for current and relevant historical vital signs.

Informed by existing research and standards efforts, for the purposes of this work, we investigate traditional lexical approaches for mapping DEs and value sets to standardized

Table 2 Glossary of key terms and definitions used

Term	Definition
Metadata	Metadata can be defined as data about data. It can include aspects such as the creator of the data, time and date of creation, its purpose, etc. Several metadata standards and models have been proposed by standards organizations such as Health Level 7 (HL7) and World Wide Web Consortium (W3C).
Data element (DE)	A DE can be defined as an atomic unit of data with precise meaning and semantics that is built on standard structures having a unique meaning and distinct units or permissible values. ³ An example of a DE is ‘person gender type’, where ‘male’ and ‘female’ can be permissible values.
Terminology	A terminology is the collection of terms and concepts and their use in a particular domain. It could be a simple list of terms describing a category, such as ‘body parts’.
Controlled vocabulary	A controlled vocabulary organizes a collection of terms to reduce ambiguity and facilitate information retrieval. They include subject indexing schemes, subject headings, thesauri and taxonomies.
Ontology	An ontology is a special type of terminology that provides a formal representation of knowledge based on a set of concepts in a particular domain, along with the relationships that exist between those concepts.
DE harmonization	DE harmonization is the process of comparing conceptual and logical data representation models to identify similarities and dissimilarities.
Value set	A value set can be defined as a list of possible values for a specific purpose. In the context of terminologies, a value set is a uniquely identifiable set of valid values that can be resolved at a given point in time to an exact set (collection) of codes, and are often used as permissible values for a DE.
Data dictionary	A data dictionary is a collection of descriptions of the data objects or items stored in a dataset or a database.

Table 3 Vital signs data mapping table (adopted from HITSP/C83, Version 1.1)

Identifier	Name	Definition	Constraints
14.01	Vital sign results identifier	An identifier for this specific vital sign observation	
14.02	Vital sign results date/time	The biologically relevant date/time for the vital sign observation	
14.03	Vital sign result type	A coded representation of the vital sign observation performed	C83-[DE-14.03-1] Vital signs SHOULD be coded as specified in HITSP/C80 Section 2.2.3.6.4 Vital Sign Results Type.
14.04	Vital sign result status	Status for this vital sign observation, eg, complete, preliminary	
14.05	Vital sign result value	The value of the result, including units of measure if applicable	
14.06	Vital sign result interpretation	An abbreviated interpretation of the vital sign observation, eg, normal, abnormal, high, etc	
14.07	Vital sign result reference range	Reference range(s) for the vital sign observation	

metadata and terminology resources, since such approaches have been demonstrated to outperform most of the advanced algorithms in both precision and recall by Ghazvinian *et al.*¹² In particular, the latter authors selected the best of the breed tools and algorithms from the Ontology Alignment Evaluation Initiative (OAEI; <http://oaei.ontologymatching.org>), which is an annual ontology mapping and alignment competition, and concluded that either the advanced algorithms are not publicly available and do not scale to the size of biomedical ontologies, or perform poorly in terms of precision and recall compared with simple lexical matching approaches. Furthermore, ontology alignment techniques that are purely based on description logics (DL)¹⁹ are not relevant for our study because the notion of a DL ‘class’ or a DL ‘role’ is not applicable in the context of DEs and value sets. Consequently, rather than developing new methods, the main contribution of our work is to evaluate the applicability of existing lexical-based approaches for mapping DEs and value sets modeled by eMERGE Network members to standardized metadata and terminology resources.

Table 4 Vital signs result value set (adopted from HITSP/C80, Version 2.0)

Concept Code	Concept Name	Definition	Usage Note	Code System Name
9279-1	Respiratory Rate	Breaths:NRat:Pt:Respiratory system:Qn:		LOINC®
8867-4	Heart Rate	Heart beat:NRat:Pt:XXX:Qn:		LOINC®
2710-2	O2 % BldC Oximetry	Oxygen saturation:SFr:Pt:BldC:Qn:Oximetry		LOINC®
8480-6	BP Systolic	Intravascular systolic:Pres:Pt:Arterial system:Qn:		LOINC®
8462-4	BP Diastolic	Intravascular diastolic:Pres:Pt:Arterial system:Qn:		LOINC®
8310-5	Body Temperature	Body temperature:Temp:Pt:Patient:Qn:		LOINC®
8302-2	Height	Body height:Len:Pt:Patient:Qn:		LOINC®
8306-3	Height (Lying)	Body height*lying:Len:Pt:Patient:Qn:		LOINC®
8287-5	Head Circumference	Circumference.occipital-frontal:Len:Pt:Head:Qn:Tape measure		LOINC®
3141-9	Weight Measured	Body weight:Mass:Pt:Patient:Qn:Measured	Body Weight (Measured)	LOINC®

MATERIALS AND METHODS

eMERGE Network phenotype data dictionaries

The eMERGE Network² is a national consortium formed to develop, disseminate, and apply approaches to research that combine DNA biorepositories with EMR systems for large-scale, high-throughput genetic research. At present, there are five participating centers in the consortium, and each center has proposed studying the relationship between genome-wide genetic variation and one or more common human trait: Group Health Cooperative (dementia), Marshfield Clinic (cataract and low HDL), Mayo Clinic (peripheral arterial disease), North-western University (type 2 diabetes), and Vanderbilt University (long QT syndrome).

At the crux of eMERGE is the development of tools and algorithms for extracting phenotypic data and representing actual healthcare events from the EMR systems at each institution in a consistent and comparable fashion. However, owing to lack of common EMR systems as well as standardization of EMR data across the institutions, one of the goals of eMERGE is to develop phenotype data dictionaries (per institution) that are mapped to standardized metadata and terminological resources. It is expected that this will not only facilitate consistent and interoperable representation of healthcare data, but also enable exchange of data across institutional boundaries.

In particular, this process involved each individual site first preparing a data dictionary for the phenotype data to submit to dbGaP²⁰ using ‘local’ (ie, institution-specific) DEs (see table 5 for an example). Normalization (eg, removing underscores, spaces) of the DEs was carried out to bring more uniformity. As expected, some of the DEs, such as subject gender, were common for all the eMERGE studies, whereas others, such as age of first cataract surgery, were specific to a particular study. Furthermore, the value sets for the DEs were either enumerated (subject gender can be male, female, or unknown) or non-enumerated (age of first cataract surgery is a continuous variable). Our overall goal for this study was to analyze the site- and phenotype-specific data dictionaries and develop a tool that can assist in mapping and harmonizing the DEs, including permissible values, to standardized metadata and terminology resources.

Terminology and metadata repository resources

In practice, a metadata repository stores information about the DEs, such that the terminologies are used to represent the

Table 5 Snapshot of Northwestern University's type 2 diabetes data dictionary

Variable	Description	Type	Units	Permissible values
Subject_ID	Deidentified subject's ID	Integer		
Enrollment_age	Age at enrollment in DNA biorepository	Integer		
Case_control	Is the subject a case or a control?	Encoded		0=Case; 1=Control
Sex	Subject's gender	Encoded		M=Male; F=Female; U=Unknown
Race	Subject's race	Encoded		0=African American; 1=American Indian; 2=Asian; 3=White; 4=Native Hawaiian; 5=Other; 6=Unknown; 7=Missing
Weight	Subject's weight in kilograms	Decimal	kg	
Height	Subject's height in centimeters	Decimal	cm	
Glucose_measurement	Subject's random glucose level	Decimal	mg/dl	

concept domain of the DE as well as its permissible values. In this study, we investigate four terminology and metadata resources for mapping of eMERGE phenotype DEs. Our selection of these resources was guided by several factors (four A's): availability (freely and publicly); accessibility (programmable); adoption (by the clinical research community, government and healthcare industry); and appropriateness (relevancy to eMERGE). We discuss them briefly in the following sections.

Cancer Data Standards Registry and Repository

The caDSR,³ developed by the NCI, defines a comprehensive set of standardized metadata descriptors for cancer research data for use in information collection and analysis for clinical research questions. It provides a database and a set of application programming interfaces (APIs) to create, edit, deploy, and find common data elements (CDEs). The caDSR is based on the ISO/IEC 11179 model for metadata registration, and uses this standard for representing information about names, definitions, permissible values, and semantic concepts for the CDEs. Over the past several years, various NCI offices and partner organizations have developed the content for caDSR by registration of DEs based on data standards, data collection forms, databases, clinical applications, data exchange formats, UML models, and vocabularies.

NCI thesaurus (NCI-T)

The NCI-T⁵ is a reference terminology and biomedical ontology, developed by the NCI, that has a broad coverage of the cancer domain, including: cancer-related diseases, findings, and abnormalities; anatomy; agents, drugs, and chemicals; genes and gene products; and so on. It covers vocabulary for clinical care, translational and basic research, and public information and administrative activities, and provides definitions, synonyms, and other information on nearly 10 000 cancers and related diseases, 8000 single agents and combination therapies, and a wide range of other topics related to cancer and biomedical research. Comprising ~80 000 concepts, NCI-T is published monthly and is edited by a group of about 15 domain expert editors. The terminology is accessible through web browsers, directly through the LexEVS API,²¹ or by download in several formats, including OWL-DL.²²

Study Data Tabulation Model terminology (SDTM-T)

The SDTM,⁴ developed by the CDISC, is a standard for submitting data tabulations of human clinical trial studies to the Food and Drug Administration (FDA) in support of marketing applications. It is built around the concept of observations collected about subjects who participated in a clinical study, where each observation can be described by a series of variables, corresponding to a row in a dataset or table. A key element of SDTM is the SDTM-T, which defines a standard list of values for use in the clinical data lifecycle. Maintained and distributed as part of NCI-T, the SDTM-T comprises 50 code lists with ~2200 terms covering demographics, interventions, findings, events, trial design, units, frequency, and ECG terminology.

Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT)

The SNOMED CT⁶ is a comprehensive clinical terminology covering most areas of clinical information such as diseases, findings, procedures, microorganisms, and pharmaceuticals. It is concept-oriented and has an advanced structure based on DL¹⁹ that meets most accepted criteria for a well-formed, machine-readable terminology. Comprising more than 300 000 concepts, SNOMED CT is one of the designated data standards for use in US Federal government systems for electronic exchange of clinical health information.

Methods: DE mapping

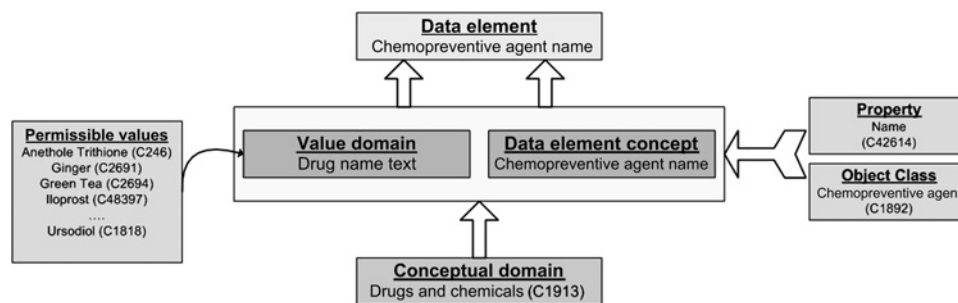
Our methods can be summarized as follows. Once the DEs have been collected and extracted from the eMERGE data dictionaries, we first apply simple lexical methods to find a direct correspondence between the variable and value set (ie, the permissible values) of the DE to CDEs in the caDSR as well as to biomedical concepts in terminologies, such as the SNOMED CT. Note that when mapping to concepts in biomedical terminologies, our lexical matching technique primarily takes into account pre-coordinated concepts.^{23 24} Hence, in situations where such lexical queries do not result in a match, we attempt post-coordination—that is, we compose new concepts by qualifying existing pre-coordinated concepts—to reflect the intended semantics of the DEs. We describe the details of our methods in the following sections.

DE mapping via pre-coordination

Our approach to mapping DEs to pre-coordinated terms and concepts from standardized biomedical terminologies and metadata resources is as conservative as possible. We first try to find an exact string match for the DE variable. If no match is found, we perform an approximate search by normalizing the original search string (eg, eliminating underscores, hyphen variations) as well as adding a wildcard (*) to the beginning and end of the string. The entire process is automated, and the search stops as soon as a match is found. Furthermore, if the DE has an enumerated list of permissible values (in the data dictionary) for its value set, we repeat the above procedure to find corresponding terms and concepts.

For querying the caDSR, we use the caDSR HTTP API, which allows an application to connect to caDSR remotely and search the database. The API provides various forms of functions for querying the caDSR, and returns the results in a well-formed XML document. As mentioned above, the caDSR is based on the ISO/IEC 11179 model for metadata registration and, as a result, decomposes the essence of a DE in well-formed parts, separating the conceptual entity (DE concept) from its physical representation in a database (value domain). The DE concept may be

Figure 1 caDSR and ISO/IEC 11179 model for metadata registries.



associated with an object class and a property, and the value domains have a list of permissible values (figure 1). Consequently, our searches for appropriate string matches were restricted to the DE concept and permissible values of the CDEs in the caDSR.

For querying the biomedical terminologies, we use web services provided by the National Center for Biomedical Ontology (NCBO) BioPortal²⁵ to find appropriate terminology concepts that can be mapped to the DE variable and DE permissible values. These services allow clients to access all BioPortal ontologies (their different versions and metadata for those versions), search for terms in all ontologies in BioPortal, and access information about any ontology concept in BioPortal (its definition, synonyms, and other properties). Specifically, we implemented a client application that invoked the BioPortal RESTful web services, which provides a light-weight, efficient resource for searching for and retrieving appropriate biomedical terms. Although BioPortal contains ~200 biomedical terminologies and ontologies, our searches were restricted to NCI-T, SNOMED CT, and SDTM-T.

DE mapping via post-coordination

Post-coordination is the process of explicitly and meaningfully combining concepts in a terminology to express a distinct semantic phenomenon. It describes the representation of a concept using a combination of two or more concepts, thereby potentially improving the domain coverage. Consequently, most recent generations of terminologies have increasingly begun to support post-coordination. However, use of post-coordination also has some essential limitations such as²⁶ (1) creation of semantically meaningless concepts by combining two or more meaningful concepts, (2) composition of unrecognized duplicate concepts, and (3) inefficiency in creating complex concepts from simpler concepts and qualifiers. Such limitations are not associated with a given usage or type of terminology, but rather with the process of composing complex concepts from multiple simpler concepts and modifiers.

Within the scope of our work, it was evident that lexically searching for relevant pre-coordinated terminology terms and concepts for DEs such as age of first cataract surgery would not yield any matches (see the Results section for more details). As a result, we investigated post-coordination to improve mapping for those DEs that could not be mapped by the lexical matching techniques.

In particular, from our list of four terminology and metadata resources outlined above, we investigated post-coordination in caDSR and SNOMED CT. Specifically, caDSR currently supports the use of multiple, ordered qualifiers with Boolean operators to fully capture complex semantics through post-coordination of atomic concepts. For example, consider a DE subject birth date. The object class (based on the ISO/IEC 11179 model) is subject,

and the property has to represent the date of birth. There are two possible ways to represent this property: one option is to create a specific concept for date of birth, and the alternative is to use two concepts—a primary concept (date) and a qualifier concept (birth). The latter case of post-coordination is commonly used in the caDSR for defining object classes and properties for a CDE. Within caDSR, the strategy is currently based on a determination of the utility to the controlled terminology (generally NCI-T) of the pre-coordinated concept. If the new concept is considered valuable to the terminology where it would reside, the pre-coordinated concept is used; if not, post-coordination is used. This has the additional benefit of preventing unnecessary growth in the size (and hence complexity) of the terminology.

For specifying post-coordinated concepts in SNOMED CT, a lightweight syntax called the SNOMED composition grammar is used. Box 1 shows a snippet of the grammar in Augmented Backus-Naur Form (ABNF) which provides a formal system of a language to be used as a bidirectional communications protocol (as defined in the internet Standard 68 (STD 68), RFC 5234). For instance, the following expression describes severe asthma:

```
195967001|asthma|:
246112005|severity|=24484000|severe|
such that, 195967001, 246112005, and 24484000 are SNOMED
CT concept identifiers. Similarly, a post-coordinated expression
for severe pain in the left thumb can be represented as:
53057004|hand pain|:
363698007|finding site|=
(76505004|thumb structure|: 272741003|laterality|=7771000|left|),
246112005|severity| = 24484000|severe|
```

eleMAP: a web-based toolkit for DE mapping

To assist in our process for DE mapping, we developed a web services-based tool called eleMAP (for ‘data element mapping’) which implements the technique outlined in the above section

Box 1 Augmented Backus-Naur Form (ABNF) definition of SNOMED CT compositional grammar (<http://www.ihtsdo.org/publications/draft-for-review-and-trial-use>)

- ▶ expression=concept *('+' concept) [':' refinements]
- ▶ concept=conceptId ['|' 'term' '|']
- ▶ refinements=(attributeSet * attributeGroup)/1*attributeGroup
- ▶ attributeGroup='{attributeSet}'
- ▶ attributeSet=attribute *(',' attribute)
- ▶ attribute=attributeName='attributeValue
- ▶ attributeName=attributeNameId ['|' 'term' '|']
- ▶ attributeValue=concept/'('expression)'

(at present the tool cannot handle post-coordination). In particular, we developed a RESTful interface that queries the caDSR and NCBO BioPortal REST services to determine a potential list of DEs and permissible values that can be mapped. The tool is built to provide access and support to two different (although not necessarily mutually exclusive) groups of users: the ‘consumers’ who are primarily interested in browsing mapped DEs from multiple phenotyping studies, and reusing them in their own study; and the ‘curators’ who are primarily interested in creating new DEs relevant for a particular study and adding/submitting them to the caDSR metadata repository and identifying concepts for describing their content using existing controlled terminologies. In its current implementation, eleMAP allows users to export the mapped DEs as either an Excel file or an XML file that is conformant to the ISO/IEC 11179 metadata registry standard. More details about the tool are available at <http://www.gwas.net/eleMAP>.

For the mapping process using eleMAP, we followed a simple operational workflow (authors JP and JW were primarily involved in these steps).

- i. For a particular study and a given DE that is being mapped, first perform a simple string search for the DE variable to find similar matches within the eleMAP repository. If a match is found, the user has the option to reuse the mappings that were defined previously (perhaps for another study).
- ii. However, if no matches were found, or the mappings are incorrect or do not capture the semantics and context for the DE under consideration, the user has the option to simultaneously query the caDSR, NCI-T, and SNOMED CT.
- iii. Depending on the search results, either the user is presented with relevant DEs or concepts that can be mapped to the DE variable and permissible values (where applicable), or no mapping can be performed.

The final mapped data dictionaries for all eMERGE sites were reviewed for accuracy and consistency by JP and JW with a κ agreement of ≥ 0.75 . Furthermore, the mapped data dictionaries for each individual eMERGE site were independently reviewed by clinical and genomics investigators (who are not coauthors of this article). Several communication channels (email, instant messaging, weekly teleconferences) were used to gather feedback on the mapped DEs and to ensure that the intended semantics of the eMERGE DEs were adequately reflected during the mapping process. Special emphasis was given to discussing the partial matches (see under Mapping via pre-condition) because of the nature of its approximate semantic representation. Overall consensus across all eMERGE sites was achieved before the mapped data dictionaries were made publicly available via eleMAP. Note that, in this study, the actual mapping process was performed entirely ‘semi-automatically’ for pre-coordinated expressions: user searches for the DE variable using eleMAP manually; eleMAP automatically queries caDSR and NCBO BioPortal, and the results presented by eleMAP are again manually evaluated. Only in situations where, for example, caDSR CDEs had to be curated by adding new content to the caDSR, was manual intervention required. Post-coordination, on the other hand, required manual mapping, since automated creation and analysis of post-coordinated expressions were not implemented in eleMAP.

RESULTS

DE collection

Analysis of the data dictionaries from the five eMERGE sites resulted in the extraction of 157 DEs, of which nine were repeated across all studies. Examples of overlapping DEs include body mass

index, race and gender, whereas study-specific DEs include history of coronary heart disease and QRS complex duration.

Mapping via pre-condition

We define two categories of mapping: proper and partial matches. A proper match indicates that the semantics of the DE are adequately represented in the mapped element—that is, for a proper match, the DE variable and permissible values should be semantically equivalent. For example, the caDSR CDE research case identifier (caDSR ID=2181644; caDSR definition=‘Identifier assigned to a clinical trial subject’) adequately represents the semantics of the eMERGE DE subject ID. Partial matches, on the other hand, indicate that the semantics of the DE are not incompatible, but also not entirely equivalent. For example, caDSR CDE Statin Use Ind-2 (caDSR ID=2480494; caDSR definition=‘Does the participant have a previous history of statin use?’) only approximates the semantics of the DE Statin Use ABI Date (ie, ‘Was statin used on the date when ankle–brachial index was measured?’). Note that, while the major emphasis of this study was to map the DEs and value sets to existing biomedical metadata and terminology standards, in cases where the mapping was partial, we modified the eMERGE DE with its value set (if applicable) for harmonization such that (1) the modified DE still retains the original intended semantics, and (2) the modified DE, and its corresponding value set if applicable, can be mapped to an existing standard. As an example, the label and textual definition of eMERGE DE case–control status was modified, and eventually mapped to the caDSR CDE clinical study participant phenotype status Case-ControlStatus (caDSR ID=2529082). Overall, we modified six (out of 157) eMERGE DEs such that they can be effectively mapped to caDSR and other resources as described below.

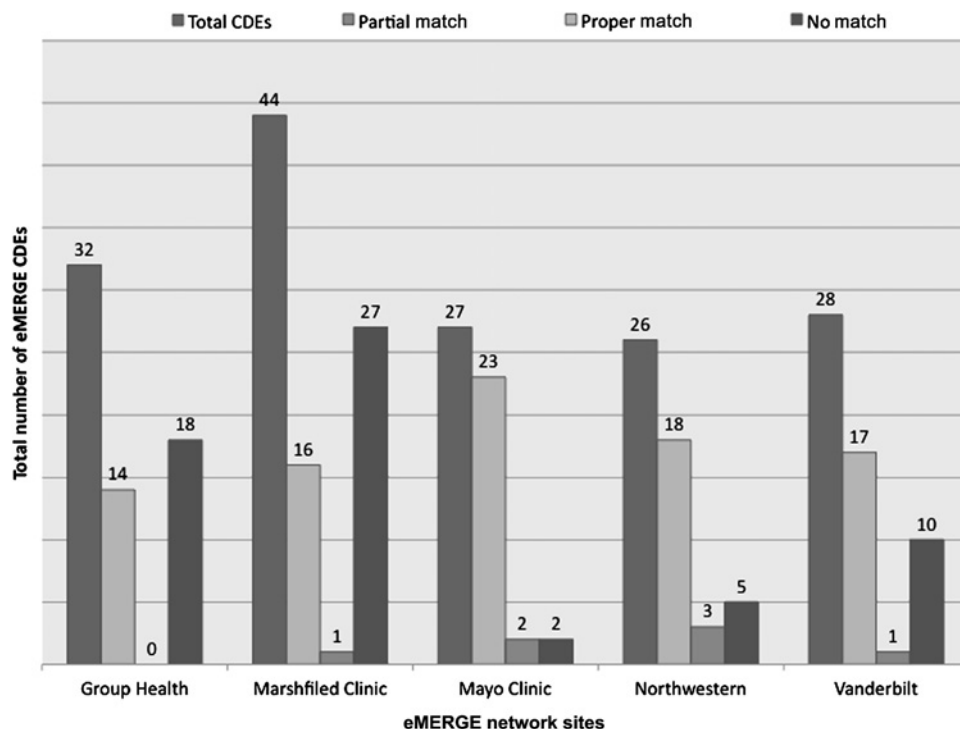
Cancer Data Standards Registry and Repository

Figure 2 shows the number of DEs for each eMERGE site that were mapped with the caDSR (November 19, 2009 snapshot of the caDSR release). While the number of proper matches was significant for all the sites, at the same time, the number of DEs that could not be mapped to the caDSR was also high. Examples of DEs that were not mapped included decade of birth of a subject or age of first statin use by a subject. Furthermore, some DEs had partial matches, as in the case of ECG impression text (referring to cardiologist-generated free-text impression from first normal ECG), which was mapped to the caDSR CDE electrocardiogram impression finding (caDSR ID=2008423) which had enumerated permissible values such as abnormal and borderline instead of ‘free-text’.

NCI thesaurus

Figure 3 shows the shows the number of DEs for each eMERGE site that were mapped with the NCI-T (November 19, 2009 NCI-T 09.09c release). We noticed that the number of DEs that could not be mapped to NCI-T was particularly high for Marshfield Clinic (studying cataract and low HDL). One of the main reasons for this was because many DE variables require post-coordination of terminology concepts for appropriate semantic representation, and NCI-T does not support post-coordination. Examples include median age at HDL-cholesterol measurements and median body mass index at HDL-cholesterol measurements for a particular study subject. Furthermore, DEs such as history of cerebrovascular disease were only partially mapped to NCI-T concepts, such as cerebrovascular disorder (NCI-T code=C2938).

Figure 2 eMERGE data elements mapped to caDSR (November 19, 2009 caDSR release). CDEs, common data elements.



SDTM terminology

Figure 4 shows the number of DEs from each individual eMERGE site that were mapped to the SDTM-T (released on May 1, 2009 as part of the SDTM Implementation Guide 3.1.1). We found that the SDTM-T was very well curated with no redundant/duplicate DEs, but lacked coverage for several eMERGE phenotype DE variables. Examples include subject enrollment age or self-reported smoking status. Furthermore, we considered the partial matches to be semantically inconsistent with the intended meaning of the DE (and hence did not report them in figure 4). For instance, when age at first diagnosis of cataract is matched partially against the SDTM-T entity, cataract, we considered the match to be semantically inconsistent.

SNOMED CT

Figure 5 shows the results for eMERGE DE mapping to pre-coordinated concepts in SNOMED CT. As illustrated above, while SNOMED CT provides a comprehensive coverage of the clinical and biomedical domain, the number of DEs that could not be mapped was higher because we were only considering pre-coordinated concepts for matching. Furthermore, in the case of partial matches, we observed that, even though the DE variable could be matched, the permissible values could not. For example, the DE variable, education level, was partially mapped to the SNOMED CT concept, education and/or schooling finding (SNOMED CT ID=365458002), since the DE permissible values such as 8th grade, associate degree, or doctoral degree

Figure 3 eMERGE data elements mapped to NCI thesaurus (November 11, 2009 NCI-T 09.09c release). CDEs, common data elements.

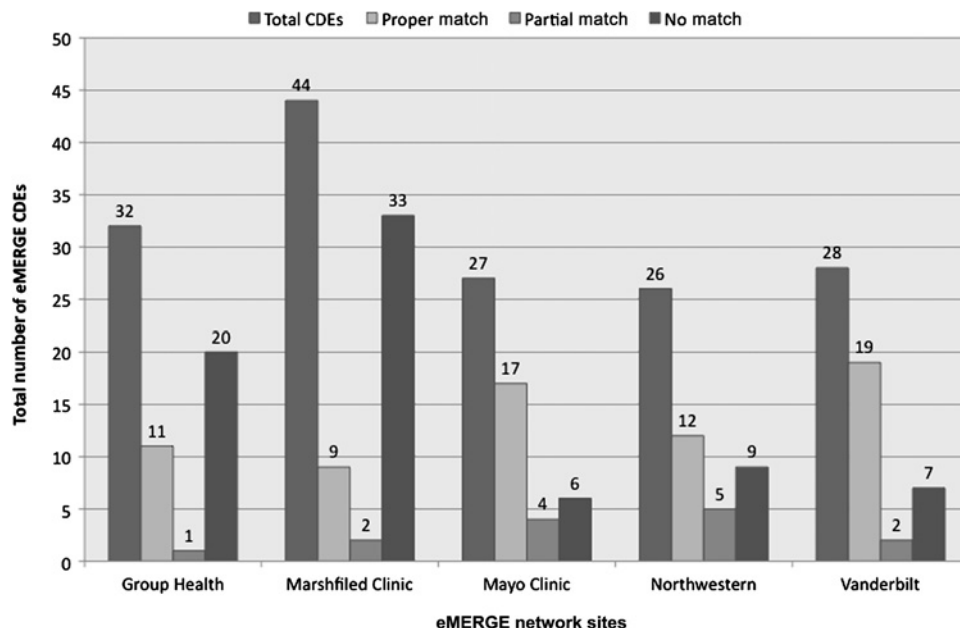
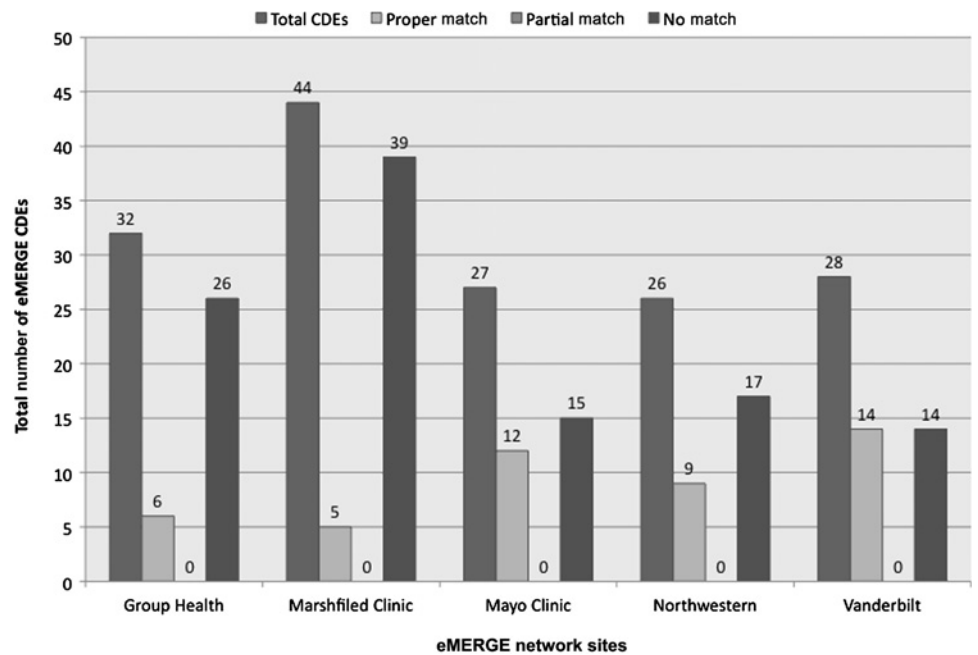


Figure 4 eMERGE data elements mapped to SDTM terminology (May 1, 2009 SDTM Implementation Guide 3.1.1 release). CDEs, common data elements.



could not be mapped to appropriate SNOMED CT concepts. We discuss results for mapping based on post-coordination using caDSR and SNOMED CT in the next section.

Mapping via post-coordination caDSR

Figure 6 shows the number of DEs for each eMERGE site that were mapped to the caDSR via post-coordination (April 20, 2010 snapshot of the caDSR database) that were not initially mapped (figure 3). As evident, the total number of eMERGE DEs mapped to the caDSR increased significantly. This was primarily due to the creation of new CDEs in the caDSR for all those DEs that were unmapped previously (figure 2). (Fifty-four new caDSR CDEs and nine new NCI-T concepts were created by NCI curators in collaboration with eMERGE investigators.) Examples of such CDEs included age of first statin use by a subject (caDSR

ID=3008893) with age as the object class and first HMG-CoA reductase inhibitor use as property concepts.

SNOMED CT

Figure 7 shows the number of DEs for each eMERGE site that were mapped to the SNOMED concepts via post-coordination that were not initially mapped via pre-coordination (figure 5). Since eleMAP does not support post-coordination at present, our process involved manually determining the best possible post-coordinated concept using the CliniClue SNOMED browser (<http://www.cliniclue.com>). Continuing with the above example, we can represent age of first statin use as follows:
 363819003|drug therapy observable|:
 {24645002|occurrence|=255216001|first|},
 {127489000|has active ingredient|=6302009|HMG-CoA reductase inhibitor|},

Figure 5 eMERGE data elements mapped to pre-coordinated SNOMED CT concepts (January 31, 2010 SNOMED International Release). CDEs, common data elements.

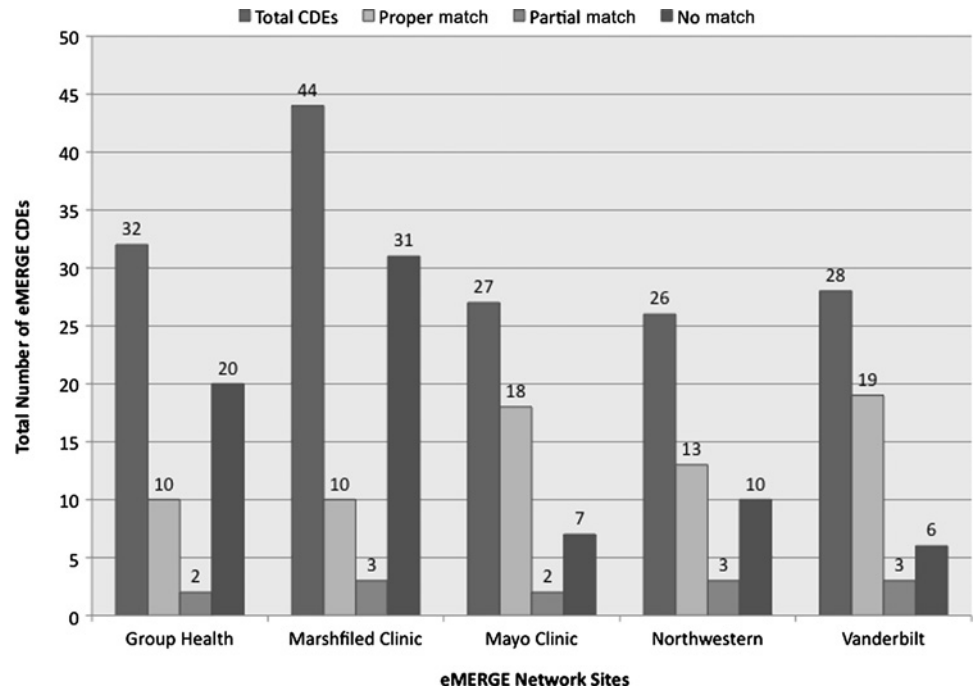
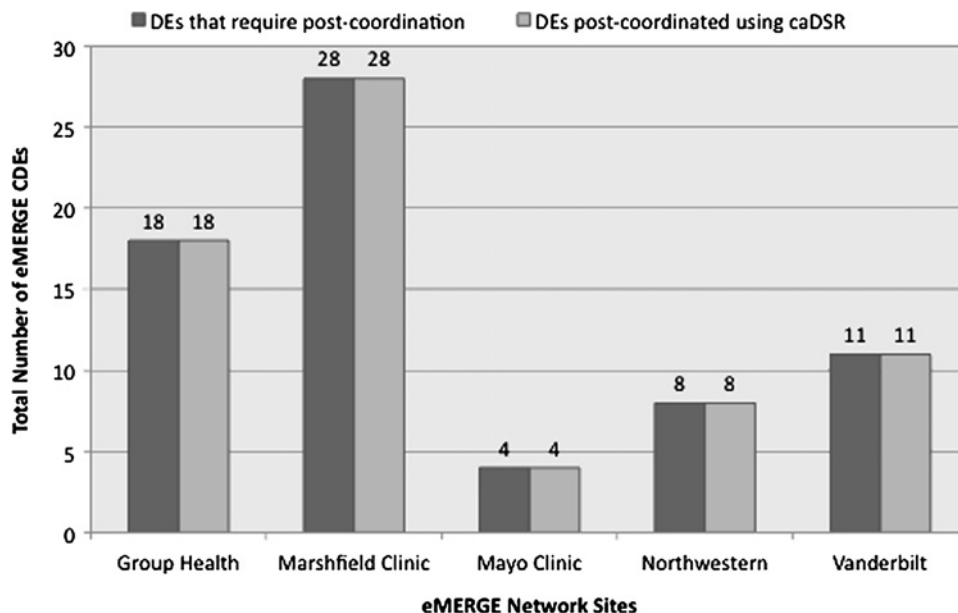


Figure 6 eMERGE data elements mapped to caDSR (April 20, 2010 caDSR release). CDEs, common data elements.



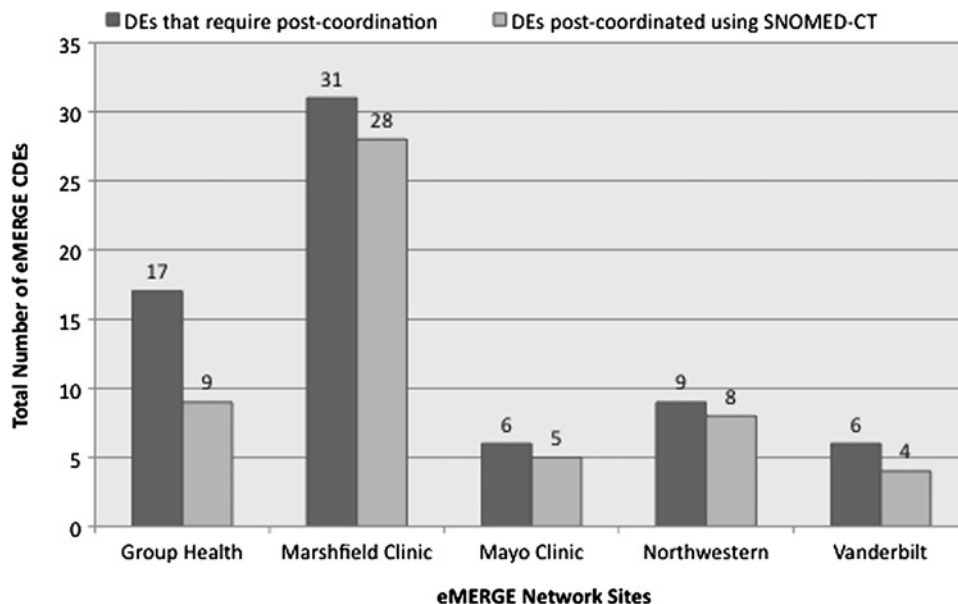
As is evident from figure 7, several DEs could not be mapped to post-coordinated SNOMED CT concepts. There are two main reasons for this: first, several atomic concepts such as *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edition, or decade does not exist in SNOMED CT. Consequently, representing DEs such as whether or not the subject meets dementia DSM III criteria and decade of birth of a subject could not be represented using SNOMED CT. Second, some DEs such as age at first qualifying for ICD-9 dementia code require definition of an 'attribute' (based on the compositional grammar; box 1) for representing the information about ICD-9 dementia codes. However, such attributes are not pre-defined (and approved) within SNOMED CT releases. Note that, unlike NCI-T, at the time of our study, there was no official means of requesting addition of concepts to SNOMED CT (US extension), although it is possible to make such requests at present.

**SUMMARY
Significance**

A major focus of clinical research informatics is to enable application of information technology for efficient translation and application of research findings to patient care and public health settings. Consequently, standardized representation of clinical data is an integral aspect of facilitating the speed and quality of clinical as well as translational research. However, given the vastness and richness, along with the complexity, of the domains from which clinical research data are generated, it is beyond any doubt that the potential to implement data standards into the clinical research domain is a daunting challenge.

In this work, our goal was to map and harmonize several phenotype data dictionaries from the eMERGE Network to standardized metadata and terminological resources. We adopted simple lexical matching techniques for querying four public,

Figure 7 eMERGE data elements mapped to post-coordinated SNOMED CT concepts (January 31, 2010 SNOMED International Release). CDEs, common data elements.



open-access resources, namely, caDSR, NCI-T, SDTM and SNOMED CT, which were implemented as part of the eleMAP toolkit for the mapping process. The mapped data dictionaries were reviewed by the eMERGE investigators, and have been made publicly available (refer to an eMERGE data dictionary for peripheral arterial disease, available as an online data supplement). Furthermore, to our knowledge, this work is the first to extensively use resources such as the caDSR to deposit phenotype datasets to NIH's dbGaP (Database for Genotypes and Phenotypes) repository. This is an important milestone that this is an important milestone because open-access repositories, such as dbGaP, while provide a consistent architecture for sharing genomics and phenotype data, mapping of DEs deposited to dbGaP to standardized terminologies and metadata resources will enable better indexing, querying, and visualization of information for researchers. Our work has also led to synergistic collaboration with PhenX (Consensus Measures for Phenotypes and Exposures²⁷), which provides investigators with high-priority, well-established, low-burden standard phenotypic measures for large-scale genomic research studies. In particular, both the teams jointly investigated caDSR for standardized collection as well as representation of phenotypic data in a dataset.²⁸

Discussion

As part of this investigation, we realized that, while simple lexical matching techniques for DE mapping works for DEs that are more 'commonly' used in clinical studies (eg, race, gender), the approach becomes rather ineffective for DEs that are either uncommon (eg, subject's decade of birth) or represent complex semantic information (eg, subject's estimated SE for baseline HDL). Furthermore, partial matching, which is highly pertinent for guiding the mapping process, often yielded results not directly useful in an automated environment. For example, there was no exact or normalized match for decade of birth in caDSR (until the April 20, 2010 release of caDSR), and mapping to person birth date or year of birth is semantically incorrect.

We also encountered that, in many cases, mapping to pre-coordinated terminology concepts resulted in confusing matches. For example, in NCI-T, left ventricular hypertrophy is represented by two different concept codes: C71076 and C50631. The contributing source of the former is CDISC and has ventricular hypertrophy as the parent concept, whereas the latter has FDA as the contributing source and cardiovascular system finding as the parent concept. They both have the semantic type of finding, although C71076 also has a semantic type of laboratory or test result. Additionally, in some cases, the permissible values can be represented by a higher level of granularity (eg, subject's unknown smoking status can be represented by NCI-T code C67151 (unknown if ever smoked)), although this does not universally hold (eg, subject's unknown race can be represented by NCI-T code C17998 (unknown)). As a consequence, because of the presence of such idiosyncratic issues in existing terminology and metadata standards, careful human intervention and curation is required for appropriate representation of DE semantics during the mapping process.

As illustrated above, post-coordination is the mechanism of supporting construction of detailed concepts using a set of pre-defined concepts in a terminology. Post-coordination is important because it takes away the need to define a pre-coordinated set of all possible concepts. However, in several circumstances, it is not considered best practice to define a very specific post-coordinated concept expression that cannot be reused and applied in other contexts. For instance, to model adjusted

baseline HDL-cholesterol measurement, we collaborated with the caDSR curators to define a new CDE (caDSR ID=3008953) that represents the HDL measurement adjusted for age (59 years), body mass index (BMI=29), and estrogen levels (no estrogen). These requirements associated with the baseline measure are added as textual comments in the CDE description, which, arguably, is not semantically computable. This begs the question whether CDEs are adequate for representation of such information, and one should consider more expressive representations such as HL7 Detailed Clinical Models (http://wiki.hl7.org/index.php?title=Detailed_Clinical_Models) or CEN/ISO EN13606 (<http://www.en13606.org/>), which we plan to explore in future iterations of eleMAP.

Limitations and future work

In terms of limitations of our current work, the validity of mappings was evaluated by only two reviewers (JP and JW). An independent evaluation would be required to confirm our results. We also did not apply sophisticated text analytics for finding the most appropriate mapping of DE variable and permissible values in our initial implementation of eleMAP. Although this was practical and useful, we intend to investigate more sophisticated approaches for pattern recognition and matching techniques.²⁹ Furthermore, the current release of eleMAP only allows users to harmonize their data dictionaries to standardized metadata and terminology resources—the users are responsible for mapping their local 'instance data' based on the mapped DEs, a task that can be tedious and laborious. Consequently, we are currently working on expanding the eleMAP platform to enable users to upload and automatically remap their local instance data to standards using the mapped data dictionaries.

Note that our work also brings us to the issue of finding equivalences between pre-coordinated and post-coordinated concepts. At least within the context of SNOMED CT, several researchers have studied this problem^{30 31} and proposed extensions to the underlying description logics for appropriate representation of post-coordinated concepts.^{32 33} In particular, Cornet³² proposed an extension to the SNOMED CT model based on explicit representation of the domain and range of relationship types and on the use of universal restrictions. This extension would enable validation of the definition of a post-coordinated concept (to see if it makes sense clinically or not) and finding equivalences, as well as providing more generic support for post-coordination. Although in this work we did not explore pre- and post-coordinated concept equivalency, in future we plan to study and implement such a feature in the eleMAP tool. Specifically, we are interested in developing techniques for detecting semantically inconsistent (and hence clinically nonsensical) post-coordinated concepts and DEs.

Conclusion

The aim of our study was to consider representation and mapping of phenotype DEs used in several studies in the eMERGE Network using standardized metadata and terminology resources. We extracted the DEs from the individual study-specific data dictionaries and mapped them to existing semantic standards for clinical and translational research. We also developed open-source software that could assist users in semi-automatic DE mapping and harmonization.

Acknowledgments We thank Diane Reeves, Luke Rasmussen, Jennifer Pacheco, and several members of the PhenX Project for numerous fruitful discussions leading to this work.

Funding The eMERGE Network was initiated and funded by NHGRI, in conjunction with additional funding from NIGMS through the following grants: U01-HG-004610 (Group Health Cooperative); U01-HG-004608 (Marshfield Clinic); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University, also serving as the Administrative Coordinating Center). JP's work is also funded in part by the Mayo Clinic Early Career Development Award.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Friemer N**, Sabatti C. The human phenome project. *Nature Genetics* 2003;**34**: 15–21.
2. **Electronic Medical Records and Genomics (eMERGE) Network**. <http://www.gwas.net> (accessed 15 Nov 2010).
3. **caDSR: Cancer Data Standards Registry and Repository**. <http://ncicb.nci.nih.gov/core/caDSR> (accessed 6 Mar 2010).
4. **CDISC Study Data Tabulation Model**. <http://www.cdisc.org/sdtm> (accessed 21 Dec 2010).
5. **Noy N**, de Coronado S, Solbrig H, *et al*. Representing the NCI Thesaurus in OWL DL: modeling tools that help modeling languages. *Appl Ontol* 2008;**3**:173–90.
6. **SNOMED-CT**. Systematized Nomenclature of Medicine-Clinical Terms. <http://www.ihtsdo.org/snomed-ct> (accessed 26 Feb 2011).
7. **Pathak J**, Wang J, Kashyap S, *et al*. eleMAP: an online tool for harmonizing data elements using standardized metadata registries and biomedical vocabularies. *Am Med Inform Assoc* 2010:1214.
8. **Richesson R**, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc* 2007;**14**:687–96.
9. **Noy N**. Semantic integration: a survey of ontology-based approaches. *ACM SIGMOD Record* 2004;**33**:65–70.
10. **Shvaiko P**, Euzenat J. A survey of schema-based matching approaches. *J Data Semantics IV*, LNCS 3730: 2005:146–71.
11. **Euzenat J**, Shvaiko P. *Ontology Matching*. Berlin, Germany: Springer, 2007.
12. **Ghazvinian A**, Noy NF, Musen MA, *et al*. Creating mappings for ontologies in biomedicine: simple methods work. *Am Med Inform Assoc* 2009:198–202.
13. **Ghazvinian A**, *et al*. What four million mappings can tell you about two hundred ontologies. 8th International Semantic Web Conference, 25–29 October, 2009. 2009:229–42.
14. **Mougin F**, Burgun A, Bodenrieder O. Mapping data elements to terminological resources for integrating biomedical data sources. *BMC Bioinform* 2006;**7**(Suppl 3): S6.
15. **Doan A**, Madhavan J, Dhamankar P, *et al*. Learning to match ontologies on the semantic web. *VLDB J* 2003;**12**:303–19.
16. **Eckert K**, Meilicke C, Stuckenschmidt H. Improving Ontology Matching Using Meta-level Learning, in 6th European Semantic Web Conference. Berlin, Germany: Springer-Verlag, 2009:158–72.
17. **Health IT**. Standards Committee Vocabulary Task Force. <http://healthit.hhs.gov/portal/server.pt?open=512&mode=2&objID=3004&PageID=20394> (accessed 16 Dec 2010).
18. **Team CRT**. HITSP/IS158: Clinical Research Interoperability Specification, 2010. http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=1&PrefixNumeric=158 (accessed 2010).
19. **Baader F**, Calvanese D, McGuinness D, *et al*. *The Description Logic Handbook*. Cambridge, UK: Cambridge University Press, 2003.
20. **Mailman M**, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet* 2007;**39**:1181–6.
21. **Enterprise Vocabulary Services**. <https://cabig.nci.nih.gov/concepts/EVS/> (accessed 9 Apr 2010).
22. **McGuinness DL**, Harmelen Fv. OWL Web Ontology Language Overview. 2004. <http://www.w3.org/TR/owl-features/>.
23. **Sauperl A**. Precoordination or not? A new view of the old question. *J Doc* 2009;**65**:817–33.
24. **Miller U**, Teitelbau R. Pre-coordination and post-coordination: past and future. *Knowl Organ* 2002;**29**:87–93.
25. **Noy N**, Shah NH, Whetzel PL, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**(Suppl 2):1–4.
26. **Rosenbloom S**, Miller RA, Johnson KB, *et al*. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc* 2006;**13**:277–88.
27. **Stover P**, Miller RA, Johnson KB, *et al*. PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol* 2010;**21**:136–40.
28. **Pathak J**, Wang J, Kashyap S, *et al*. Evaluating Phenotyping Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network. San Francisco, CA: AMIA Clinical Research Informatics (CRI) Summit, 7–11 March, 2011.
29. **Duda R**, Hart E, Stork D. *Pattern Classification*. New York, USA: Wiley, 2000.
30. **Andrews J**, Patrick TB, Richesson RL, *et al*. Comparing heterogeneous SNOMED CT coding of clinical research concepts by examining normalized expressions. *J Biomed Inform* 2008;**41**:1062–9.
31. **Green J**, Wilcke JR, Abbott J, *et al*. Development and evaluation of methods for structured recording of heart murmur findings using SNOMED CT post-coordination. *J Am Med Inform Assoc* 2006;**13**:321–33.
32. **Cornet R**. Definitions and Qualifiers in SNOMED CT. *Methods of Information Med* 2009;**48**:178–93.
33. **Rector A**, Brandt S. Why do it the hard way? The Case for an Expressive Description Logic for SNOMED. *J Am Med Inform Assoc* 2008;**15**:744–51.