Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects

Jyotishman Pathak, PhD1, Helen Pan, PhD3, Janey Wang, MS2, Sudha Kashyap2, Peter A. Schad, PhD3, Carol M. Hamilton, PhD3, Daniel R. Masys, MD2, Christopher G. Chute, MD, DrPH1

1Mayo Clinic, Rochester, MN; 2Vanderbilt University, Nashville, TN; 3RTI International, RTP, NC

Abstract

Combining genome-wide association studies (GWAS) data with clinical information from the electronic medical record (EMR) provide unprecedented opportunities to identify genetic variants that influence susceptibility to common, complex diseases. While mining the vastness of EMR greatly expands the potential for conducting GWAS, non-standardized representation and wide variability of clinical data and phenotypes pose a major challenge to data integration and analysis. To address this requirement, we present experiences and methods developed to map phenotypic data elements from eMERGE (Electronic Medical Record and Genomics) to PhenX (Consensus Measures for Phenotypes and Exposures) and NCI’s Cancer Data Standards Registry and Repository (caDSR). Our results suggest that adopting multiple standards and biomedical terminologies will expose studies to a broader user community and enhance interoperability with a wider range of studies, in turn promoting cross-study pooling of data to detect both more subtle and more complex genotype-phenotype associations.

Introduction

Systematic study of clinical phenotypes is important to better understanding the genetic basis of human diseases and more effective gene-based disease management [1]. While the recent advances in genotyping technologies to systematically ascertain large numbers of sequence variants (e.g., single nucleotide polymorphisms) for the complete genome of an individual has fueled numerous genotype-phenotype association studies [2, 3], our ability to fully understand the genetic basis of common diseases is significantly hindered by the inability to precisely specify the phenotypes (i.e., the outward physical manifestation of the genotypes). In particular, identifying and extracting phenotypes at large varies greatly between different medical specialties and institutions, and lacks the systematization and throughput compared to large-scale genotyping efforts. This makes it difficult to compare or combine GWAS studies even though several risk factors and phenotypes are common across multiple conditions (e.g., subject’s smoking behavior) [4].

To address this growing requirement, the U.S. National Human Genome Research Institute (NHGRI; part of National Institutes of Health) initiated two separate projects, called eMERGE (Electronic Medical Records and Genomics [5]) and PhenX (Consensus Measures for Phenotypes and Exposures [6]), in 2007. The overarching goal for eMERGE was to correlate whole genome scans with phenotype data extracted from the electronic medical record (EMR) systems. PhenX, on the other hand, provides investigators with high-priority, well-established, low-burden standard measures to collect phenotypic and environmental data for large-scale genomic studies. Despite the retrospective nature of the data elements collected in eMERGE from Electronic Medical Record (EMR) systems and prospective nature of data elements and measures in PhenX for GWAS, an integral part of both efforts is to standardize the collection as well as representation of phenotypic data in a dataset. However, in practice, data elements representative of the clinical data stored in the EMRs or research databases across different medical institutions are developed independent of each other, without any common data structure or representational format. Furthermore, the GWAS measures, while developed using a consensus-based process, evolve without any consideration of how clinical phenotypes are stored and represented. Arguably, to facilitate integration and analysis of data, it is vital for such activities to provide appropriate mappings of phenotype data elements to controlled biomedical vocabularies and terminological resources.

Toward this end, in this study we mapped EMR-derived phenotype data elements from eMERGE to the newly developed standardized phenotypic and environmental measures from PhenX, as well as widely used metadata repository for phenotypic data elements, NCI’s Cancer Data Standards Registry and Repository (caDSR [7]), to assess the common and diverse phenotype data elements among the EMR derived data and other data standards. Data elements that can be mapped to these resources present opportunities to cross-study analysis. On the other hand, data elements that cannot be mapped complements areas not currently present in the caDSR (diseases other than cancer), and further underlines the importance of using standard measures, as recommended by the PhenX, in prospective studies for meta-analysis across studies.

Background

eMERGE: Electronic Medical Records and Genomics Network

The eMERGE Network [5] 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 different 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, such as Dementia and Type 2 Diabetes. At the crux of eMERGE, is development of phenotype extraction algorithms that can be executed on institutional EMR systems. However, due to the lack of standardization of EMR data across institutions, one of the goals of eMERGE is to use phenotypic data elements that are harmonized to standardized metadata resources to facilitate consistent and interoperable representation of healthcare information.

PhenX: Consensus Measures for Phenotypes and Exposures

PhenX [4] addresses the need for standard measures in GWAS and other large-scale genomic research efforts. The goal of PhenX is to identify high-priority, well established, and broadly applicable measures for 21 research domains, such as cardiovascular and cancer. PhenX measures are selected by Working Groups (WG) of domain experts using a consensus process that includes input from the scientific community. The selected measures are then made freely available to the scientific community via the PhenX Toolkit (http://phenxtoolkit.org). Each PhenX measure refers broadly to a standardized way of capturing data on a certain characteristic of, or relating to a study subject. A PhenX Protocol is a standard procedure recommended by a Working Group for investigators to collect and record a PhenX Measure. The PhenX Toolkit is a resource for investigators who want access to high quality; standard measures and is valuable for all epidemiological studies.

Common Data Elements and NCI’s caDSR

The NCI Cancer Data Standards Registry and Repository (caDSR [7]) defines a comprehensive set of standardized metadata descriptors for cancer research data for use in information collection and analysis. It provides a database and a set of Application Programming Interfaces (APIs) to create, edit, deploy, and find common data elements (CDEs). It 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. Various NCI offices and partner organizations have developed the content of the caDSR by registration of data elements based on data standards, data collection forms, databases, clinical applications, data exchange formats, UML models, and vocabularies. Consequently, for enabling interoperability across phenotypic data elements derived within the eMERGE and PhenX projects, we leveraged caDSR CDEs for representing the data elements across both projects. In this report, we outline our methods in undertaking this task, and share our experiences on the utility of standards-based common metadata for the clinical research community. It is to be noted that independently, eMERGE and PhenX have been using caDSR to standardize their data elements, either by mapping to existing CDEs or curating new CDEs where applicable since the inception of both projects.

Materials

For this study, we used the eleMAP toolkit Version 1.0 (http://www.gwas.net/eleMAP) developed within the eMERGE network for data element harmonization [8]. eleMAP provides an uniform and intuitive interface for mapping data elements from several eMERGE studies to the caDSR and various biomedical vocabularies in the NCBO BioPortal [9]. We also used the PhenX toolkit Version 3.5 (http://www.phenxtoolkit.org) release with caDSR/CDE browser Version 3.2.05 Build 1 [4]. The toolkit presents a brief description of each phenotypic measure, its purpose, the rationale for its inclusion, the standard protocols for collecting the data, and relevant references.

Methods

Standardizing eMERGE Data Elements using caDSR

This step involved each individual eMERGE site first preparing a data dictionary for the phenotype data of interest using the using their “local” (i.e., institutional) terminology. Normalization (e.g., removing underscores, spaces) of the data elements was done to bring more uniformity. Expectedly, some of the data elements, such as Subject Gender, were common for all the studies, whereas others, such as Age of First Cataract Surgery, were specific to a particular study. Furthermore, the instances or value sets were either enumerated (Subject Gender can be Male, Female, or Unknown) or non-enumerated (Glucose Measurement is a continuous variable). Table 1 shows examples of data elements relevant for Type 2 Diabetes (proposed by Northwestern University within eMERGE). Furthermore, these data elements were categorized into different groups (e.g., Body Measures, Cardiovascular Disorders). Our overall goal was to map the data elements and permissible values to the caDSR metadata.

| Data Element Variable | Description | Type     | Unit | Permissible Values |
|-----------------------|-------------|----------|------|--------------------|
| Enrollment Age        | Age at enrollment | Integer  |      |                    |
| Sex                   | Subject’s gender | Encoded  |      | M=Male; F=Female; U=Unknown |
| Glucose Measurement   | Subject’s random glucose value | Decimal | mg/ DL | |

Table 1 Example eMERGE Data Elements for Type 2 Diabetes

For this process, we leveraged the eleMAP toolkit that provides an intuitive text search-based functionality for finding the relevant caDSR CDEs. In particular, it first attempts to find an exact string match for the data element variable. If no match is found, an approximate search is done by normalizing the original search string (e.g.,
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 a data element has an enumerated list of permissible values, the above process is repeated to find corresponding terms for the permissible values. As an example, Sex from Table 2 is mapped to the caDSR CDE Subject Gender (caDSR Public ID=2200604). If none of the existing CDEs in caDSR are appropriate, then new CDEs were curated (n=54) in collaboration with the caDSR curators.

**PhenX Measures Registered in caDSR**

We registered each PhenX measure and protocol in caDSR with a caDSR CDE name and public ID. Similar, to eMERGE, where applicable, we search for existing caDSR CDEs that match the PhenX measure to re-use them, and when no relevant CDEs are discovered, we curate new CDEs.

**Mapping eMERGE data elements to PhenX measures**

For the mapping process, we evaluated two different approaches:

- We used a lexical matching-based technique for finding similar data element variables and permissible values.
- We used the caDSR CDE identifier to find which data elements were mapped equivalently.

Our rationale for applying the second approach was straightforward: if a given eMERGE and PhenX data element were mapped (equivalence relation) to the same caDSR CDE, then by transitivity, the data elements represent the same semantics. For example, an eMERGE data element Height (of type Body Measures) and PhenX measure Standing Height (of type Anthropometrics) mapped to the same caDSR CDE Person Height Value (caDSR CDE ID=2179643). Consequently, both the eMERGE and PhenX data element are assigned an equivalence relationship.

Our rationale for the first approach, arguably a primitive one, is based on empirical evidence from prior research where simple string-based lexical matching techniques have outperformed advanced algorithms in terms of both precision and recall [10]. For example, eMERGE data element Gender would lexically match to the PhenX measure of Gender. As usually done for such a method, the mappings were performed in close collaboration among the authors, as well as, domain experts were consulted to address doubts and confusion via e-mails or telephone conferences.

We further distinguished between the following cases:

- **Equivalent**: the source and target data elements represent the exact same semantics and the informational entity. Example includes Height (eMERGE; source) and Standing Height (PhenX; target) as illustrated earlier.
- **Broader**: the semantics of the source data element is more granular and specific with respect to the target data element. For example, the data element Blood Pressure (PhenX; target) is broader than Diastolic Blood Pressure (eMERGE; source).
- **Narrower**: the inverse of broader. For instance, Fasting Plasma Glucose for Diabetes Screening (PhenX; target) is narrower than Fasting Lab Indicator (eMERGE; source).
- **No match**: there exists no semantically related target data element that can be mapped to a given source data element. As an example, Standard Error in Total Cholesterol Measurement (eMERGE; source) has no correspondence to any PhenX measure.
- **Out of scope**: the data element is either too abstract, or included for housekeeping purposes, or serves a purpose that is not related to the domain of interest. For instance, the PhenX measure Macaroni Cheese Frequency (target) is considered out of scope for eMERGE since it is not relevant to any of the phenotypes investigated by the project.

We describe our results and evaluation of applying both these approaches in the next section.

**Results**

For eMERGE, authors JP, JW and SK identified the unique eMERGE data elements (n=143) for the all the categories (n=12) corresponding to 13 different phenotypes studied by the eMERGE network. Four categories were assigned per author based on their familiarity with the category and the domain, and both the lexical matching-based and caDSR CDE identifier-based approaches outlined above were applied to find relevant correspondences.

**Lexical matching-based mapping**

As stated earlier, for the lexical matching-based technique, a conservative approach for finding the most appropriate data element was adopted. In particular, the PhenX toolkit was used to search for the relevant data elements, both at the variable and the instance level, and appropriate relationships (equivalent, broader, narrower, no match, or out of scope) were assigned. Figure 1 presents the results for this procedure indicating the total number of eMERGE data elements that were equivalent (and similarly other relationships) to PhenX data elements.

**caDSR Common Data Element-based mapping**

For the caDSR CDE-based approach, simplistically the caDSR CDEs that were mapped to eMERGE and PhenX data elements, respectively, were enumerated and the
intersection set (for equivalency) was identified per eMERGE variable category.

Discussion

Assessing eMERGE Data Element against PhenX and caDSR. As evident from Figure 1, while the lexical matching technique found few equivalences (8%) between the eMERGE and PhenX data elements, and majority had broader (41%), narrower (4%) or no (36%) relationships. These outcomes are consistent with the fact that the eMERGE studies are primarily focused on EMR-derived phenotyping, and hence, the phenotype specific data elements are representative of data stored in EMR systems that can range from very abstract (e.g., Cancer Indicator) to extremely granular (e.g., Ankle-Brachial Index after a Treadmill Test). PhenX measures, on the other hand, were developed primarily for investigators who are either planning a future study or expanding an existing one with the expectation that the measures, when readily available, can be used as part of standard protocols for collecting subject related data. Furthermore, PhenX also focused on environmental exposures (e.g., History of Daycare Attendance) that were out of scope for eMERGE. As a consequence, either many eMERGE data elements had a broader relationship to PhenX measures, or had no match. Interestingly enough, for the data elements that were equivalent, the corresponding mapped caDSR CDEs were not the same. (We discuss this issue later in this section.)

For the caDSR CDE-based mapping approach, the goal was to determine CDEs common and mapped to both eMERGE and PhenX data elements. We identified that a majority (97%) of caDSR CDEs did not match, or were not reused across both projects. One of the reasons for such a large non-overlap of data elements is due to non-overlap between the phenotypes and domain of study between both the projects. For example, several PhenX measures were modeled for cancer, reproductive health and speech and hearing—areas that eMERGE did not address. The second major reason for lack of overlap is more technical, and is associated with coverage and curation aspects of caDSR. We discuss this issue next in this section.

caDSR Coverage and Curation. In total, PhenX measures from 21 research domains have been registered as 352 CDEs in caDSR. Of these, 31 existing CDEs were re-used, and 321 newly created. The existing CDEs that PhenX measures map to are most commonly used data elements from Demographics, Anthropometrics, Alcohol and Tobacco Use, and Assays. The only exception is the “Perceived Stress Scale Questionnaire” (public ID: 2199495) in the Psychosocial domain. The large number of newly created CDEs fall in non-cancer research domains which include other disease areas (e.g., Speech and Hearing, Skin, Bone, Muscle and Joint.), environmental factors (Nutrition, Environmental Exposure, Physical Activity and Physical Fitness), and social domains (Social Environment, Psychosocial). This set of 321 newly created CDEs is a significant addition to the caDSR.

In our study, several caDSR CDEs did not match for the eMERGE and PhenX data elements. We see two main reasons for this: (1) mapping to granular, context-specific CDEs in the caDSR, and (2) presence of duplicate (or semantically similar) CDEs in the caDSR. For the first issue, several eMERGE data elements were mapped to phenotype specific caDSR CDEs (e.g.,

![Figure 1 Results for lexical matching-based mapping](image-url)
Dementia Cognitive Abilities Screening Instrument Count) that were not relevant for PhenX. Similarly, several PhenX data elements mapped to the caDSR CDEs (e.g., Paternal Grandfather’s Birthplace) were out of eMERGE’s scope. This aspect, while leading to lesser degree of overlap between the data elements for eMERGE and PhenX, illustrates the fact that the domains for these projects are non-overlapping. As more phenotypes are studied in eMERGE, in future we expect the degree of data element overlap with PhenX to significant increase. The second issue is more involved and technical. In its current incarnation, the caDSR provides a database and a set of APIs for creating, editing, sharing and using CDEs to facilitate interoperability. However, due to the limitations of the ISO/IEC 11179 model Version 2 used in the existing caDSR implementation as well as API and caDSR CDE browser limitations, not only is it difficult for end-users to query for the relevant CDEs, but it is also difficult to identify CDEs that are semantically similar, and hence, can be re-used. Consequently, often CDEs with overlapping semantics get curated, and users are presented with several similar CDEs for a given search query. For instance, at the time of writing this manuscript, a string search for Gender using the caDSR CDE browser, 67 different CDEs are returned as the query result, and the user is left with the exercise for selecting the most appropriate one, thereby leading to inconsistent CDE reuse and mapping. Continuing the above example, the data element Sex in eMERGE was mapped to the caDSR CDE Person Gender (caDSR Public ID=2200604), whereas PhenX mapped it to the caDSR CDE Gender Code (caDSR Public ID=2179640). It is abundantly clear, even from this simple example, that significant improvements with respect to CDE curation, software implementation and modeling, as well as education and training is required to ensure appropriate re-use of CDEs for data interoperability.

Limitations and Future Work. While caDSR is a very useful resource for data elements in individual studies to share with the research community, it has some limitations as described above. Adopting diverse set of metadata standards and terminologies will expose studies to a broader user community to enhance interoperability with a wider range of potential studies and promote cross-study pooling of data to detect both more subtle and complex genotype-phenotype associations. Consequently, both eMERGE and PhenX are investigating using CHI standards including, LOINC and SNOMED-CT, for future cross-study analysis.

Significance. In addition to the collaboration with eMERGE on the phenotypic data extracted from EMR, PhenX is collaborating with other projects including dbGaP (http://www.ncbi.nlm.nih.gov/gap) to develop a consistent rule set for mapping PhenX measures to dbGaP study variables. This will enable PhenX measures to be included in dbGaP, thereby facilitating sharing and access of variables from different studies for cross-study analysis. Through this study of mapping eMERGE data elements and PhenX measures, our outcomes can serve as a gateway to link mapped eMERGE EMR variables to other widely visible and diverse resources.

Conclusion. Wide-spread adoption and use of standard measures within clinical research will greatly facilitate cross-study analysis. Increased statistical power from cross-study analysis makes it possible to detect more subtle and more complex gene associations including gene-gene and gene-environment interactions. This study demonstrates the value of using a standardized metadata resource for exposing studies to a broader community, as well as, outlines several limitations of existing metadata resources.

Acknowledgment. This work is funded in part by the eMERGE (U01-HG-04599 and U01-HG-04603) and PhenX (U01-HG-004597-01) grants from NHGRI.

References

1. Friemer, N. and C. Sabatti, The Human Phenome Project. Nature Genetics, 2003. 34(1): p. 15-21.
2. Pearson, T. and T. Manolio, How to Interpret a Genome-wide Association Study. Journal of the American Medical Association, 2008. 299(11): p. 1335-1344.
3. Ku, C., et al., The discovery of human genetic variations and their use as disease markers: past, present and future. Journal of Human Genetics, 2010. 55(1): p. 403-415.
4. Stover, P., et al., PhenX: a toolkit for interdisciplinary genetics research. Current Opinion in Lipidology, 2010. 21(2): p. 136-140.
5. Electronic Medical Records and Genomics (eMERGE) Network. [cited 2010 August 2, 2010]; Available from: http://www.gwas.net
6. Consensus Measures for Phenotypes and Exposures (PhenX). [cited 2010 August 2, 2010]; Available from: http://www.phenx.org
7. caDSR: Cancer Data Standards Registry and Repository. March 6th, 2010; Available from: http://ncicb.nci.nih.gov/core/cADSRS
8. Pathak, J., et al., eleMAP: An Online Tool for Harmonizing Data Elements using Standardized Metadata Registries and Biomedical Vocabularies, in American Medical Informatics Association (AMIA) Annual Symposium, 2010.
9. Noy, N., et al., BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research, 2009. 37(Suppl 2): p. 1-4.
10. Ghazvinian, A., et al., Creating Mappings for Ontologies in Biomedicine: Simple Methods Work, in American Medical Informatics Association (AMIA) Annual Symposium. 2009. p. 198-202.