Demographics Identification: Variable Extraction Resource (DIVER)

Conference Paper · September 2012

DOI: 10.1109/HISB.2012.17

CITATIONS
5

READS
120

5 authors, including:

Alexander Hsieh
Columbia University
5 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Son Doan
Kaiser Permanente, San Diego, United States
60 PUBLICATIONS **1,451** CITATIONS
SEE PROFILE

Ko-Wei Lin
University of California, San Diego
31 PUBLICATIONS **242** CITATIONS
SEE PROFILE

Hyeoneui Kim
Seoul National University
88 PUBLICATIONS **1,765** CITATIONS
SEE PROFILE

Demographics Identification: Variable Extraction Resource (DIVER)

Alexander Hsieh, Son Doan, Michael Conway, Ko-Wei Lin, Hyeoneui Kim

Division of Biomedical Informatics, Department of Medicine
University of California, San Diego
La Jolla, USA
alhsieh@ucsd.edu

***Abstract* – Lack of standardization in representing phenotype data generated in different studies is a major barrier to data reuse for cross study analyses. To address this issue, we developed DIVER, a tool that identifies and standardizes demographic variables in dbGaP, based on simple natural language processing and standardized terminology mapping. In its evaluation using variables (N=3,565) from a range of pulmonary studies in dbGaP, DIVER proved to be an effective approach to standardizing dbGaP variables by successfully identifying demographic variables with high rates of recall and precision (98% and 94%, respectively). In addition, DIVER correctly modeled 79% of the identified demographic variables at the core semantic level. Examination of variables that DIVER could not handle shed light on where our tool needs enhancement so it can further improve its semantic modeling accuracy. DIVER is an important component of a system for phenotype discovery in dbGaP studies.**

Keywords – dbGaP; data standardization; phenotype variables; data reuse

I. INTRODUCTION

With the advancement of Genome Wide Association Studies (GWAS), abundant phenotype and genotype data that can be used and reused to identify genetic variants associated with health outcomes has become readily available. The database of Genotypes and Phenotypes (dbGaP), for example, developed by the National Heart Lung Blood Institute (NHLBI), contains over 2,300 data sets that include more than 130,000 phenotype variables collected from more than 300 studies [1], [2]. In spite of the huge potential for leading to high impact discoveries in understanding health outcomes at the genomics level, mining such data resources is often challenging due to the lack of standardization in the way that phenotype data are represented [3].

In this paper, we present the process of algorithmically identifying and standardizing demographic variables in dbGaP. This work was done as part of the Phenotype Finder in Data Resource (PFINDR) program funded by NHLBI. One of the main purposes of our project, Phenotype Discoverer (PhD) is to standardize phenotype variables in a way that supports an accurate and complete search in dbGaP [4].

Our initial goal was to make phenotype variables collected from pulmonary studies searchable. We plan to generalize our approach to standardizing the pulmonary variables in this phase to apply to the phenotype variables collected from other disease category studies, in subsequent phases.

The purpose of this article is to report the development of methods and describe evaluation results of our tool – Demographics Identification Variable Extraction Resource (DIVER). DIVER was designed to identify four types of demographic variables (i.e., age, race, ethnicity, and gender) and to standardize them by attaching relevant metadata information. We defined demographic variables broadly, extending our definition to include any variables related to age, gender, race, and ethnicity. Therefore, variables related to health history or findings such as "age when first diagnosed with breast cancer" and "age stopped smoking," were also captured within the scope of this work. We targeted the demographic variables first, as they are relatively simple and less variant than other types of phenotype variables, but among the most frequently used by researchers [5].

We also report cross mapping results of the demographic variables identified by DIVER to those in another standardized phenotype variable repositories, consensus measures for Phenotypes and eXposures (PhenX) [6], [7] and Cancer Data Standards Registry and Repository (caDSR) [8], [9]. The purpose of this mapping was to assess the feasibility of integrating and improving the interoperability of demographic variables in dbGaP with those in PhenX and caDSR.

TABLE I. DEMOGRAPHIC VARIABLES in dbGaP

Variable ID	Name	Description
phv00122459	HEAGEINF	Infant age heart condition noted (in months)
phv00066445	coc_dep_ons	Age onset of DSM4 cocaine dependence
phv00122968	AMINDIAN	Race-Native American (0=no,1=yes)
phv00022889	hisp	Is participant Hispanic or Latino?

II. BACKGROUND

A. Challenges in Reusing dbGaP Data

Advances in high-throughput technologies in genomics, imaging, and proteomics have led to an abundance of data. Identified as one of the top priority areas of research at recent NIH workshops and the "Big Data" announcement by several federal funding agencies, mining of such data is considered a critical and urgent research area that leads to new and better understanding of human diseases [10]. In order to properly mine the data,

we need to first standardize and integrate them in a way that enables comprehensive and accurate search and retrieval.

However, data are often collected without standardization, thus making it challenging to reuse them. For example, dbGaP provides an advanced search interface where users can perform focused searches by specifying search fields (e.g., variable name, variable description, document name, etc). As illustrated in **Table 1**, many variables in dbGaP are named without following a specific naming convention and are often labeled with abbreviated codes that are somewhat difficult for users to decipher. Consequently, searches against variable names do not yield reliable results. To resolve this querying obstacle, dbGaP allows users to search against full text resources using keywords. However, the full text search usually returns a large number of false positives [3], [11].

B. Related Studies

Non-standardized phenotype representation has been a barrier to the use and reuse of data generated and collected in different studies [12]. Many nationally funded projects have attempted to address this issue by cataloguing variables, in a standardized way, and registering them to public data repositories [5], [8], [13], [14]. This approach allows users generate and collect data using the standardized variables, while facilitating later data reuse without additional burden of data standardization.

In particular, PhenX aims to standardize key measures in GWAS and other large-scale genomic research, a goal closely related to our own PhD project [4], [7]. Using the PhenX toolkit, researchers have access to standard means of capturing data for well-defined and frequently used measures, in 21 research domains [5–7]. Efforts have been made to cross map phenotype variables from 16 dbGaP studies to PhenX variables [11]. This manual approach produced high quality mappings, but the need for adopting algorithmic approaches to identify similarities and differences among the variables also became apparent [11].

Another standardized variable repository, the NCI caDSR, defines and represents the data elements used in cancer research based on the ISO/IEC 11179 metadata standards [9], [15]. Establishing cross mappings among different standardized data repositories is important because it allows researchers to conduct valuable cross study analyses. As such, the data elements (i.e., variables) in PhenX and the eMERGE (Electronic Medical Record and Genomics) Network have been mapped to the Common Data Elements (CDE) in caDSR [11], [16], [17].

The eMERGE Network is a national consortium formed to develop, disseminate, and apply approaches to GWAS that use institutional EMR (Electronic Medical Records) as a phenotype data source [13], [18]. eMERGE addresses issues associated with the lack of standardization in EMR data across institutions by standardizing and harmonizing these data elements via metadata annotation and standardized concept mapping [17], [18]. To facilitate this process, eMERGE created eleMAP, a web environment where researchers can search, browse, and download harmonized phenotype variables along with their associated metadata [19]. Researchers can also harmonize their local phenotype data dictionaries to existing metadata and terminology standards such as caDSR, NCI Thesaurus (NCIT), and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) using eleMAP [19].

C. Core Constructs of DIVER

PhD and eMERGE take similar approaches to standardizing and harmonizing phenotype variables; standardized terminology mapping followed by metadata annotation. However, unlike eMERGE, where variables are mostly standardized and harmonized while submitted to eleMAP, in PFINDR, the task at hand is to standardize and harmonize the vast amount of phenotype variables already stored in dbGaP (more than 130,000). Therefore, an algorithmic means of processing those variables is crucial to facilitating variable standardization efficiently.

As previously described, another major challenge associated with standardizing the phenotype variables in dbGaP is that the variable names are often cryptic and do not convey much comprehensible meaning. Therefore, using MetaMap [20], DIVER processes variable descriptions rather than variable names, to capture core concepts that sufficiently convey variable meaning and to subsequently map them to a standardized terminology system.

We chose to use MetaMap on the grounds that it is a general purpose, highly configurable, and freely available tool maintained by the National Library of Medicine (NLM) that uses sophisticated and computationally expensive natural language parsing methods. As described in [21], using MetaMap in real-time applications can be challenging, due to the tool's relatively low processing speed compared to primarily statistically-based tools like MGREP . However, the coverage of MetaMap is better than some its statistically oriented competitors (e.g. MGREP) [22] Note that, for our PhD project, coverage and accuracy of the concept identifier used is more important than execution speed, since our tool standardizes variables once before making them available for quick retrieval.

DIVER also uses simple rule-based NLP algorithms to formalize the semantics of phenotype variables based on the outputs MetaMap generates with variable descriptions. Many studies have shown that NLP can effectively determine semantic categories and relations in the biomedical domain [23–25]. Model systems, such as ontologies, lexicons, and syntactic structures, serve as references for formalizing the syntactic and semantic structures of text, and thus are crucial to these studies.

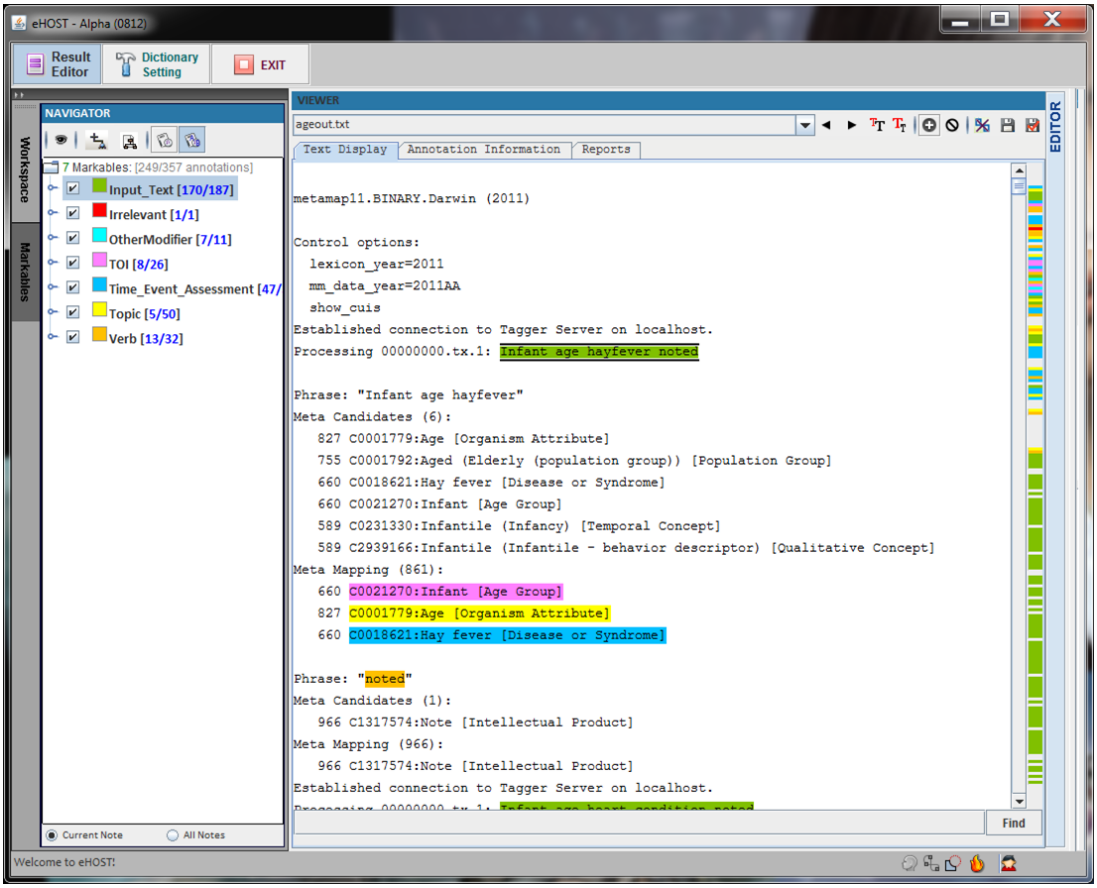

FIGURE 1. eHOST WORKSPACE FOR AANNOTATIONS

Likewise, the information models we developed for the 4 types of demographic variables played a critical role in our formalization of phenotype variable descriptions with DIVER. The process of developing the information models is described in the methods section.

D. Prior Explorative Study

In a prior explorative study, we tested the feasibility of identifying age variables by applying UMLS (Unified Medical Language System) [26] mapping and task-specific post-processing rules to dbGaP variable descriptions [27]. When tested with 200 variable descriptions, this approach successfully identified age variables with a high accuracy rate.

In this study we further enhanced the age variable identification function and expanded our previous approach to other demographic variables such as race, ethnicity and gender. We built DIVER in a way such that it streamlines the steps involved in identifying and standardizing demographic variables and lends itself to be readily adopted by other data repositories in need of standardization, to facilitate data integration and sharing

III. METHODS AND PROCEDURES

A. Developing Information Models

Based on the findings from the previous study [27], we identified an information model for age variables, which consists of four major information classes – theme (topic – i.e., age), target (or subject) of information, event related to the age assessment, and a linkage term that further specifies the time point of the event. We developed the models for other demographic variables using the age information model as a straw man model.

To develop the information models, we first extracted variable descriptions of 50 age variables, 22 race variables, 20 gender variables, and 5 ethnicity variables from the data dictionaries of 5 pulmonary studies in dbGaP. These variable descriptions were then processed with MetaMap to identify core concepts in the descriptions and their corresponding UMLS Concept Unique Identifier (CUI). The MetaMap text output was then imported into a text annotation tool called eHOST [28]. The eHOST workspace used by the reviewers is presented in **Figure 1**.

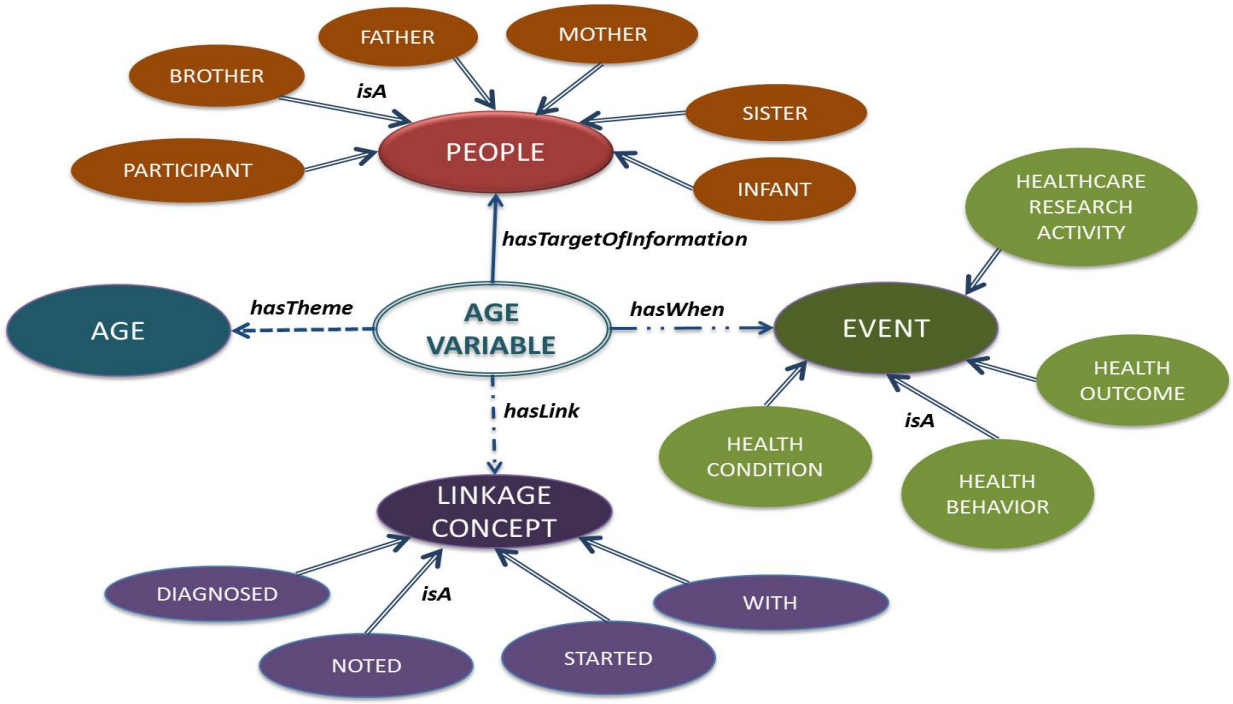

FIGURE 2. AGE INFORMATION MODEL

Two reviewers (AH, HK) collaboratively annotated each variable description by assigning information classes to the relevant key concepts identified by MetaMap. The reviewers were allowed to include new information model classes when necessary. The reviewers also noted whether MetaMap correctly mapped the key concepts assigned with an information class to the UMLS meta-thesaurus.

The annotated information was then exported as an XML document. The information model classes used in each type of demographic variable were collated and analyzed by the two reviewers. By establishing semantic relations among the information classes, four information models for the four demographic variables were constructed. The information model for the 'age' variable is presented in **Figure 2** as an example.

B. Developing Rules for Information Model Fitting

The same two reviewers (AH, HK) collaboratively analyzed the annotated MetaMap outputs to identify any patterns between the properties of the concepts and their assigned information model classes. These patterns served as the basis for the rules we created to (1) assign information classes to the concepts in the variable descriptions and to (2) determine if the variable description fell into one of the 4 target demographic variables.

For example, age-related variables can have different targets of information, such as study subject or family members of the study subject. We noted that when the target of information for a particular demographic variable is anybody other than the study subject, MetaMap recognizes the subject-specific term and maps it to concepts whose semantic type is either "population group" or "family group."

Also, when the age variable indicates the age at a specific time point other than current age, a term that further specifies the time point co-occurs, and usually refers to a health condition, health behavior, or research activity. These patterns were written in IF-THEN-ELSE format rules (see **Figure 3**) and tested with the 97 demographic variables used for developing the rules.

B. Developing a Metadata Annotation Schema

The final outputs of the DIVER process are standardized demographic variables fully annotated with pre-defined metadata. We developed the metadata schema for DIVER by benchmarking relevant metadata schemas of the three related initiatives that shared a similar purpose (i.e., eMERGE-eleMAP, caDSR, and PhenX) [5], [8], [19]. The DIVER metadata schema is presented in **Figure 4** with an age variable as an example.

```
IF Theme[i] is in {Age, Gender, Race, Ethnicity}
AND
SemanticType[i] = "age group" OR "family group"
THEN    TOI_PCN = CandidatePreferredName[i]
        AND TOI_CUI = CandidateCUI[i]
ELSE    //this is the default if no other TOI assigned
        TOI_PCN = "study subject"
        AND TOI_CUI = "C0681850"
```

FIGURE 3. RULE EXAMPLE

C. Implementing DIVER

The DIVER pipeline was written in Python. DIVER was implemented as a series of variable description processing steps. Each step is further detailed below.

Step 1. Variable description extraction: The data dictionaries in dbGaP were in XML format. For this first step of extracting variable descriptions and associated identifiers (i.e., variable id and study id), we wrote an xml parser and text extractor in Python.

Step 2. Variable description preprocessing: Because MetaMap was unable to process certain lexical variations, we added a simple lexical preprocessing step to this DIVER process. We identified a list of normalization tasks, based on findings from the annotation exercise previously described in section A, and wrote a Python script to complete each task. A few outstanding preprocessing tasks included removing any unnecessary information, such as miscellaneous punctuations (e.g., hyphens, underscores) and indexing related to data collection (e.g., Q10, EX7), and replacing certain shorthand expressions (e.g., mom to mother) that MetaMap failed to correctly recognize. This task list was continuously augmented throughout the evaluation phases as necessary whenever we encountered additional issues that needed to be resolved.

Step 3. MetaMap and output parser: We incorporated MetaMap into the DIVER pipeline. In this step, MetaMap took as an input the preprocessed variable descriptions and produced as an output an XML file, from which only relevant information, such as input text, mapped term, final mapping results (i.e., concept preferred name, CUI, and semantic type), was further extracted and passed to the next semantic analysis step. We wrote a Python script to process MetaMap outputs.

Step 4. Semantic mapping: We next implemented a set of rules that evaluates whether a variable description processed through MetaMap fitted one of the information models. As previously described, this rule engine formalizes each variable description by assigning information classes (see **Figure 2**) to the key concepts identified by MetaMap. Designation of the information classes was determined mainly through string matching and evaluation of semantic types.

The output of this semantic mapping step was an XML file containing variable id, variable description, type of demographic variable, and the UMLS CUI, as well as preferred concept names for each information class relevant to the variable.

Step 5: Metadata annotation: The last step of the DIVER process was to annotate the identified demographic variables with the pre-defined metadata items (see **Figure 4**). Our goal is to have these metadata added as a part of dbGaP, against which user queries could be run to improve the accuracy of the retrieved information.

sdGaP ID	sdv000000.v1
Variable ID	phv00122015.v1
Var ID	phv00122015
Version	v1
Variable Name Original	SAGE
Variable Name Standardized	Subject_Age_at_Study
Variable Description	Subject age at time of study
Variabel Category	age
Mapping	theme::subject age [C2348575] event::study [C2600334] linkageConcept::at
Data Type	integer
Unit of Measurement	year
Max	62
Min	18

FIGURE 4. DIVER METADATA SCHEMA

D. Evaluating DIVER Performance

Preliminary evaluation: We tested the DIVER pipeline with 2,454 variables collected from 5 pulmonary studies and 3 non-pulmonary studies in dbGaP. We included non-pulmonary studies to ensure the generalizability of applying these rules to non-pulmonary studies, even though the rules were developed primarily based on the demographic variables from pulmonary studies. Our rationale for this was that demographic variables should share the same characteristics across studies with different disease categories. The same two reviewers (AH, HK) independently reviewed the output and compared results to ensure reliability in the evaluation process.

The review was done at 2 levels. The first level review was to determine whether DIVER correctly picked up demographic variables and/or dropped non-demographic variables. The second level review was to determine if DIVER correctly assigned information model classes to the key concepts extracted from the variable descriptions of the demographic variables that it had identified. Four grading options were used: C for correct, A for added incorrectly (i.e., false positive), M for missed incorrectly (i.e., false negative), and W for when the information class was relevant but assigned to a wrong concept. The first level review was done only using C, A, and M grades.

Based on this testing, the rules and the pre-processing functions were further expanded to incorporate more diverse use cases. Also, many minor implementation errors were corrected.

Final evaluation: A subject matter expert (KL) manually reviewed the study descriptions of the 300 studies registered to dbGaP and identified 26 pulmonary studies. Excluding 10 studies whose data dictionaries were unavailable at the time of this evaluation, we retrieved the data dictionaries of the remaining 16 studies and ran them through DIVER.

The same two reviewers who performed the preliminary evaluation reviewed the DIVER outputs for the final evaluation. The same two-level, 4-option grading scheme was employed.

E. Mapping to PhenX and caDSR

The demographic variables identified from the evaluation set were mapped to PhenX using the PhenX tool kit [5] and to caDSR using the web-based Common Data Element (CDE) browser [8]. One of the authors (AH) manually mapped the dbGaP demographic variables to the PhenX measures and CDE. A matching level was scored using 4 grades: E for "exact" matches, B for "broad" matches (i.e., the mapped item has more general meaning than the dbGaP variable), N for "narrow" matches (i.e., the mapped item has more specific meaning than the dbGaP variable), O for "other" matches (i.e., the mapped item has the same theme but different qualifiers than the dbGaP variable).

We standardized demographic variables using 3 major information classes; theme, target of information, and event referring to the time point of assessment. Variable compatibility with these information classes played an important role in determining matching level. To first be considered as any type of match, two variables must have the same theme class. Variables that are "exact" matches also share both target of information and event, in addition to theme. Match designations "broad" and "narrow" specify that theme and target of information classes are comparable between the two. As an example, a dbGaP age variable that has a specific event modifier mapped to the age variable, but no event modifier in caDSR or PhenX, is considered a "broad" match. Variables that do not share information besides theme we marked as having matching level "other". Simply having a common theme is not enough to justify aggregation and comparison, for cross-study analysis, between different types of targets of information. A second reviewer (HK) reviewed the mapping results to ensure accuracy. Any disagreements were resolved through open discussion.

IV. RESULTS

The DIVER process is summarized in the Data Flow Diagram (DFD) in **Figure 5** with an age variable as an example. The DIVER tester was implemented as a web-based tool and is available at http://pfindr-data.ucsd.edu/diver/.

A. DIVER Performance

A total of 3,569 variable descriptions were extracted from the 16 pulmonary studies. Frequency distributions of the target demographic variables as determined by the reviewers are presented in Table II. About 10% of the total variable descriptions processed by DIVER were duplicates (i.e., variables that have identical variable descriptions). In addition, some of the 16 studies had already been previously used for the development and preliminary evaluation. Therefore, the frequencies were calculated separately, and only with the variables used in the final evaluation (i.e., test set), after removing duplicates.

TABLE II. FREQUENCY DISTRIBUTIONS OF VARIABLE DESCRIPTIONS

Variable Types	With Duplicates		Without Duplicates	
	All	***Test Set***	***All***	***Test Set***
Demographics total	282	255	253	229
Age	228	214	217	205
Ethnicity	4	3	3	2
Gender	25	16	14	8
Race	25	19	18	14
Non-Demographics	3,283	2,781	2,933	2,482
Total	3,565	3,036	3,186	2,711

DIVER showed a high rate of precision, recall, and accuracy when assessed at the demographic category level. In this level of assessment we only evaluated whether DIVER correctly identified and categorized input variables into the 4 types of demographic variables (see **Table 2**). The proportion of duplicates was relatively small and the accuracy scores did not change significantly when calculated at the unique variable level – 99.44% for all variables and 99.37% for unique variables. Therefore, we only present here the precision and recall scores calculated with the data before removing duplicates.

We then assessed the DIVER performance at the semantic level by accounting for its ability to correctly identify and assign relevant information classes to the demographic variables (see **Table 3**). The percent-correct scores were then calculated separately for each style of demographic variables. In total, DIVER identified 79% of the demographic variables in the dataset correctly at the semantic level.

TABLE III. DIVER PERFORMANCE

Measurement		All	Test Set
Demographics Category Level	Accuracy[a]	99.36%	99.34%
	Recall	98.58%	98.81%
	Precision	94.83%	94.68%
	F-Measure	0.9667	0.9670
Semantic Level (percent correct)	Theme/TOI[b]	98.35% (N=61)	97.67% (N=43)
	Theme/TOI[b]/ Event	57.14% (N=35)	56.67% (N=30)
	Theme/TOI[b]/ Event/Link	76.88% (N=186)	78.77% (N=179)
	Combined	79.08% (N=282)	79.37% (N=252)

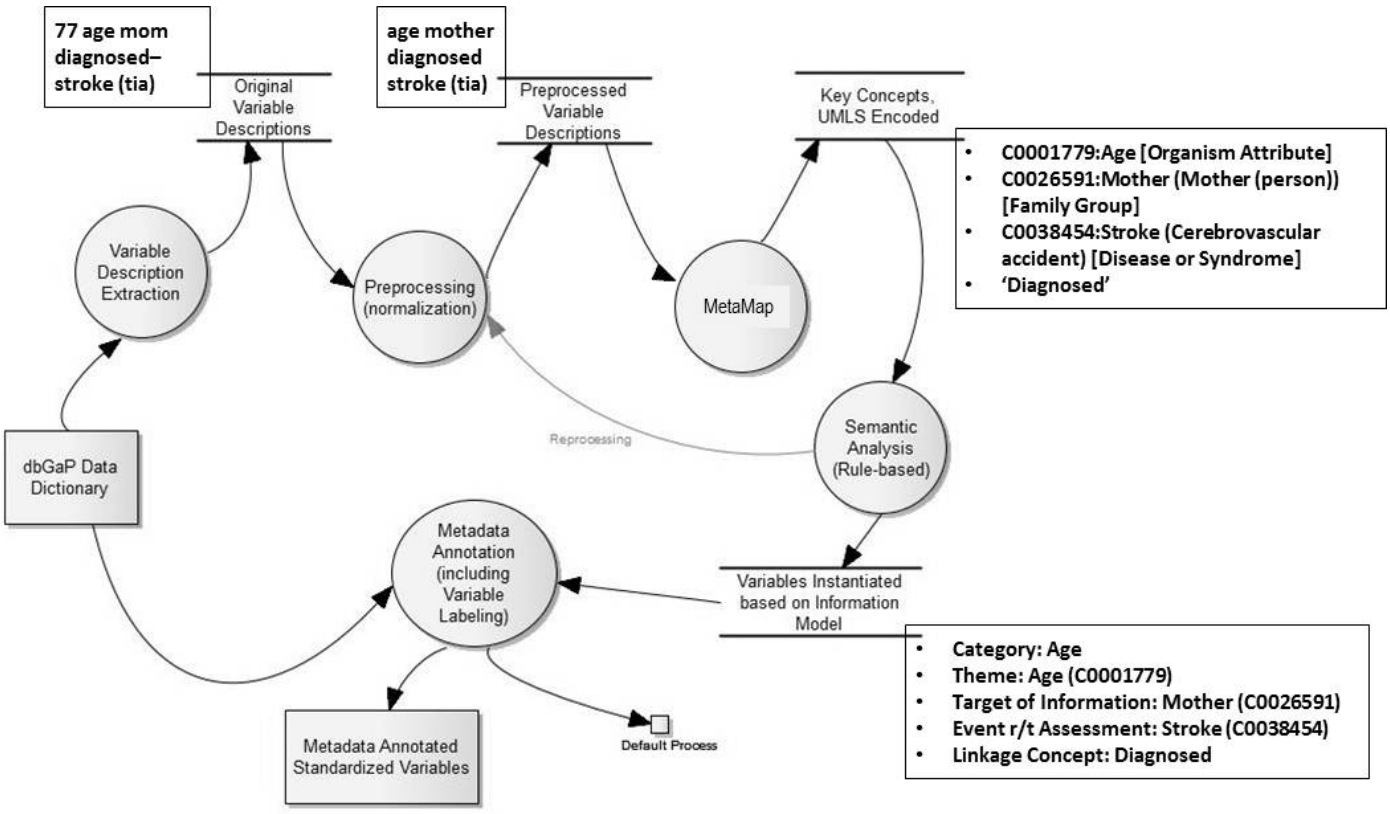

FIGURE 5. DIVER DATA FLOW

B. Cross mapping to PhenX and caDSR

The 253 unique demographic variable descriptions were standardized into 174 unique demographic variables and mapped to the PhenX Measures and the caDSR CDE. The mapping results are presented in **Table 4**. The number of variables in each demographic variable category included in this mapping is provided in parentheses. Mapping to PhenX and to caDSR yielded 13 (7%) and 23 (13%) exact matches, respectively. In both mappings, the majority of the mappings were deemed to fall in "other matches."

TABLE IV. MAPPING RESULTS

Matching Level / Target System		Age (163)	Gendr (1)	Race (8)	Ethnct (2)	Total (174)
Exact	*PhenX*	9	1	1	2	13
	caDSR	17	1	3	2	23
Broad	*PhenX*	2	0	7	0	9
	caDSR	6	0	4	0	10
Narrow	*PhenX*	0	0	0	0	0
	caDSR	3	0	1	0	4
Other	*PhenX*	152	0	0	0	152
	caDSR	137	0	0	0	137

V. DISCUSSION

This study showed that DIVER, built with a readily available open source NLP tool, simple information models, and standardized terminology mapping, can identify and standardize demographic variables in dbGaP with a high rate of accuracy.

Simple approaches to identifying and modeling demographic variables, such as keyword based string matching, may show reasonably good performance in identifying the theme concepts (i.e., Age, Race, Sex, and Ethnicity) and subject of information concepts, but struggle when handling inputs more diverse. DIVER uses a string matching algorithm for its preprocessing step to normalize frequently used lexical variants, and for its modeling step to identify themes. However, correctly interpreting ambiguous (or polysemous) terms (e.g., "White" as color vs. "White" as racial group) requires a more sophisticated semantic modeling step, much like the one that DIVER provides. In addition, certain modifier concepts (i.e., "event" concepts in this study) can take on a wide range of diseases/conditions or behaviors. Utilizing semantic types is therefore a sensible approach to recognizing such concepts, as complete identification of all possible "event" concepts upfront remains a challenge.

Through the DIVER process, idiosyncratic demographic variables are formalized based on the information models and their key concepts are annotated with UMLS CUI. The process of phenotype variable standardization is critical to facilitating data interoperability and reuse. In the PhD system, phenotype variable search will be done against standardized concepts rather than raw strings to improve the completeness of the results by expanding the search terms to synonyms and child concepts.

A. Failure analysis

Analyses of the 21% tested variables that DIVER failed to correctly standardize revealed some errors and limitations in DIVER, to be addressed in future enhancements of this tool. We describe below the 6 outstanding causes of failure.

MetaMap parsing error: MetaMap struggled to correctly tokenize specific concepts as units in the descriptions. It would often split apart words of a phrase (e.g., "drink of alcohol" became "drink" and "alcohol") or include unnecessary terms in a conceptual token (e.g., "infant high bp" as one unit instead of "infant" and "high bp" individually). This incorrect partitioning led to incorrect concept mapping, which caused the rule-based semantic analysis step to fail. This type of error affected mostly the identification of an event concept.

MetaMap UMLS mapping failure: MetaMap failed to correctly map the terms from variable descriptions to UMLS. Besides errors in parsing, lack of content coverage in UMLS contributed to the number of incorrect mappings (e.g., "smoking pipe" as in "age when first smoked pipe"). MetaMap failed to recognize cases where lay terms were used to describe synonymous, known medical procedures (e.g., "tonsils removed" to tonsillectomy (C0040423), "irregular heart beat" to arrhythmia (C0003811)). However, there were also certain mapping errors not easily explained. For example, MetaMap failed to correctly map "smoking cigarettes", as in "age stopped smoking cigarettes", while correctly mapping "smoking cigar" (C0453996), as in "age stopped smoking cigar".

MetaMap limitations in handling negation: MetaMap did not recognize "other than" as a negation term, which led to incorrect DIVER outputs. Metamap uses an implementation of the NegEx negation detection algorithm [29]. NegEx locates "negation trigger terms" (i.e. terms indicating that a concept is negated) using simple regular expressions. The fact that the NegEx algorithm was originally designed for discharge summaries, rather than short, often ungrammatical variable descriptions could be a reason that it failed to recognize some of the negation patterns common in dbGaP variable descriptions. For example "age when last used drugs other than marijuana, cocaine, or opiates" was modeled to mean "age when last used marijuana, cocaine, or opiates".

Instead of using the default MetaMap implementation of NegEx, we could develop an independent NegEx-style negation module, with additional rules and trigger terms appropriate for dbGaP, or try a statistical NLP system for detecting negations based on Conditional Random Fields, which covers medical and genomics text [29]. We are also considering adding rules for identifying and promoting more general concepts for the terms following "other than" as an option. In the above example, the second approach would model the input description as "age when last used drugs".

Limitations in the semantic analysis rules: The specification of "event" concepts focuses on diseases, conditions, and health behaviors. Therefore event concepts like "implantation of cardiac pacemaker" that specify the time point of age measurement were not correctly modeled. We will need to include procedure concepts in the list of event types in future versions of the rule engine. Generic procedures also often specify devices or anatomical concepts to help clarify or convey relevant meaningful information. To account for this, we will add rules that recognize device or anatomical site concepts as valid modeling components when they appear following a procedure concept.

Incorporating concepts with the semantic type *Finding*: To improve the sensitivity, DIVER did not utilize the semantic type *Finding*, as it covers a wide range of concepts. This led to us missing a few relevant event terms (e.g., "baseline state" (C1290922)) and, more importantly, dropping key theme concepts such as "Race – other" (C0425379) and "death age" (C0742983).

Resolution of some of the limitations and errors described here may require improving NLP algorithms both in and out of MetaMap, an effort by no means insubstantial. Therefore, we would need to assess the cost-benefit ratio involving this step once we examine also the other types of variables used to search phenotypes in dbGaP, and not only the demographics ones. Our goal is not in formalizing variable descriptions with 100 % accuracy and sophistication. For example, we doubt modeling "age of first use of substance other than cocaine, marijuana, opioid" fully at the level of specific kind of substances negated would provide much added value to researchers. A simpler version – "age of first use of substance" – would be enough to convey proper meaning to researchers. Along these lines, we are preparing an extensive user requirement analysis. Practicality issues will be discussed with dbGaP users during the requirement analysis.

B. Limitations

DIVER was designed to standardize variables solely based on their descriptions. Considering permissible values can also provide critical information about their respective variables, this could potentially be a limitation that would need to be addressed via future enhancements. Upon manual review, however, we determined that the

values usually did not alter the overall meaning of demographic variables we targeted, and often supplemented the description in elucidating variable meaning. DIVER attaches value-related metadata such as format, allowed values, maximum and minimum, and value type to support cross study analyses. Making value-related information readily available to researchers is important as it helps them determine whether similar variables from different studies are directly comparable or require some pre-processing, across domains.

Our findings regarding the cross mapping of dbGaP demographic variables to caDSR and PhenX produced similar results as described in [11], [17]. We did not find many exact matches in both resources. We also noticed that certain dbGaP variables were mapped to more than one caDSR CDE due to the duplicates in caDSR.

We suspect that matching levels adopted in [11], such as "comparable" and "related", will be more informative to dbGaP users considering the target of mapping is a phenotype variable. However, assessing matching level using the grading option presented in [11] would require extensive review by domain experts and is planned for a future study.

VI. Conclusion

This study demonstrated that DIVER, a tool developed based on standardized terminology mapping and simple NLP, can successfully identify and standardize demographic variables in dbGaP. The DIVER approach can be extended to include the other phenotype variables in dbGaP, processing of which would be a cumbersome and labor-intensive task if attempted manually. Complete standardization of the phenotype variables in dbGaP would create new opportunities for cross study analyses and support research initiatives in a previously uncharted 'sea' of data. DIVER is the first of a series of components that are being developed to facilitate phenotype discovery in dbGaP studies.

Acknowledgements

We thank Dr. Lucila Ohno-Machado for her leadership and invaluable input on this project. We also thank Ustun Yildiz and Vinay Venkatesh for their technical expertise. This project was supported in part by the grants UH2HL108785 and U54HL108460 (NIH/NHLBI).

References

[1] M. D. Mailman et al., "The NCBI dbGaP database of genotypes and phenotypes.," *Nature genetics*, vol. 39, no. 10, pp. 1181-6, Oct. 2007.

[2] "The database of Genotypes and Phenotypes." [Online]. Available: http://www.ncbi.nlm.nih.gov/gap/. [Accessed: 10-Mar-2012].

[3] K.-W. Lin et al., "Testing the adequacy of a public GWAS database as a cohort discovery tool." .

[4] "Phenotype Discoverer (PhD) for dbGaP." [Online]. Available: http://pfindr.net/. [Accessed: 03-Jul-2012].

[5] "PhenX: consensus measures for Phenotypes and eXposures." [Online]. Available: https://www.phenxtoolkit.org/. [Accessed: 25-Jun-2012].

[6] C. M. Hamilton et al., "The PhenX Toolkit: get the most from your measures.," *American journal of epidemiology*, vol. 174, no. 3, pp. 253-60, Aug. 2011.

[7] P. J. Stover, W. R. Harlan, J. A. Hammond, T. Hendershot, and C. M. Hamilton, "PhenX: a toolkit for interdisciplinary genetics research.," *Current opinion in lipidology*, vol. 21, no. 2, pp. 136-40, Apr. 2010.

[8] "Common Data Element Browser." [Online]. Available: https://cdebrowser.nci.nih.gov/CDEBrowser/. [Accessed: 05-Jul-2012].

[9] P. A. Covitz et al., "caCORE: a common infrastructure for cancer informatics.," *Bioinformatics (Oxford, England)*, vol. 19, no. 18, pp. 2404-12, Dec. 2003.

[10] C. Ober et al., "Getting from genes to function in lung disease: a National Heart, Lung, and Blood Institute workshop report.," *American journal of respiratory and critical care medicine*, vol. 182, no. 6, pp. 732-7, Sep. 2010.

[11] H. Pan et al., "Using PhenX measures to identify opportunities for cross-study analysis.," *Human mutation*, vol. 33, no. 5, pp. 849-57, May 2012.

[12] L. Ohno-Machado et al., "iDASH: integrating data for analysis, anonymization, and sharing.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 19, no. 2, pp. 196-201.

[13] "The eMERGE Network." [Online]. Available: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page. [Accessed: 10-May-2012].

[14] "PROMISE: Dynamic Tools to Measure Health Outcomes from the Patient Perspective." [Online]. Available: http://www.nihpromis.org/. [Accessed: 07-Jul-2012].

[15] ISO/IEC, "Information technology - Metadata registries (MDR) - Part 3: Registry metamodel and basic attributes."

[16] J. Pathak et al., "Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects.," *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, vol. 2011, pp. 41-5, Jan. 2011.

[17] J. Pathak et al., "Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, no. 4, pp. 376-86.

[18] C. A. McCarty et al., "The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies.," *BMC medical genomics*, vol. 4, p. 13, Jan. 2011.

[19] "eleMAP." [Online]. Available: https://victr.vanderbilt.edu/eleMAP/. [Accessed: 20-May-2012].

[20] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 17, no. 3, pp. 229-36.

[21] N. H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang, and M. A. Musen, "Comparison of concept recognizers for building the Open Biomedical Annotator.," *BMC bioinformatics*, vol. 10 Suppl 9, no. Suppl 9, p. S14, Jan. 2009.

[22] N. Bhatia, N. Shah, D. Rubin, A. Chiang, and M. Musen, "Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap," in *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2009.

[23] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.," *Bioinformatics (Oxford, England)*, vol. 17 Suppl 1, pp. S74-82, Jan. 2001.

[24] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.," *Journal of biomedical informatics*, vol. 36, no. 6, pp. 462-77, Dec. 2003.

[25] R. N. A and P. V. A, "From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions." .

[26] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett, "The Unified Medical Language System: an informatics research collaboration.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 5, no. 1, pp. 1-11.

[27] A. Hsieh, M. Conway, and H. Kim, "Identifying Age Variables in dbGaP using Natural Language Processing." 2012.

[28] "eHOST: Extensible Human Oracle Suite of tools." [Online]. Available: http://code.google.com/p/ehost/. [Accessed: 11-Apr-2012].

[29] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries.," *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301-10, Oct. 2001.