INFORMATICS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Using PhenX Measures to Identify Opportunities for Cross-Study Analysis

Huaqin Pan,[1]* Kimberly A. Tryka,[2] Daniel J. Vreeman,[3,4] Wayne Huggins,[1] Michael J. Phillips,[1] Jayashri P. Mehta,[2†] Jacqueline H. Phillips,[3] Clement J. McDonald,[5] Heather A. Junkins,[6] Erin M. Ramos,[6] and Carol M. Hamilton[1]

[1]RTI International, Research Triangle Park, North Carolina; [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland; [3]Regenstrief Institute, Inc., Indianapolis, Indiana; [4]Indiana University School of Medicine, Indianapolis, Indiana; [5]Lister Hill National Center for Biomedical Communication, National Library of Medicine, National Institutes of Health, Bethesda Maryland; [6]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

For the Deep Phenotyping Special Issue

Received 3 November 2011; accepted revised manuscript 28 February 2012.
Published online 13 March 2012 in Wiley Online Library (www.wiley.com/humanmutation).DOI: 10.1002/humu.22074

ABSTRACT: The PhenX Toolkit provides researchers with recommended, well-established, low-burden measures suitable for human subject research. The database of Genotypes and Phenotypes (dbGaP) is the data repository for a variety of studies funded by the National Institutes of Health, including genome-wide association studies. The dbGaP requires that investigators provide a data dictionary of study variables as part of the data submission process. Thus, dbGaP is a unique resource that can help investigators identify studies that share the same or similar variables. As a proof of concept, variables from 16 studies deposited in dbGaP were mapped to PhenX measures. Soon, investigators will be able to search dbGaP using PhenX variable identifiers and find comparable and related variables in these 16 studies. To enhance effective data exchange, PhenX measures, protocols, and variables were modeled in Logical Observation Identifiers Names and Codes (LOINC®). PhenX domains and measures are also represented in the Cancer Data Standards Registry and Repository (caDSR). Associating PhenX measures with existing standards (LOINC® and caDSR) and mapping to dbGaP study variables extends the utility of these measures by revealing new opportunities for cross-study analysis.

Hum Mutat 33:849–857, 2012. Published 2012 Wiley Periodicals, Inc.*

KEY WORDS: phenotype; environmental exposure; epidemiologic methods; GWAS

Introduction

The influx of genome-wide association studies (GWAS) has led to the identification of many genetic variants associated with disease outcomes. More than 1,000 publications are currently included in the Catalog of Published GWAS [Hindorff et al., 2009]. Despite the vast potential for cross-study comparisons, the lack of standard phenotypic and environmental measurements has limited the ability to combine data from GWAS and other large-scale genomic and epidemiologic studies [Hindorff et al., 2009; Manolio, 2009; Thorisson et al., 2009]. Standard measures are critical for combining data from seemingly disparate studies with similar underlying risk factors, increasing statistical power so that relatively modest or more complex associations can be identified and initial findings from GWAS can be replicated [Burton et al., 2009; Fortier et al., 2010; García-Closas and Lubin, 1999; Khoury et al., 2009]. However, in most longitudinal clinical studies, each investigator develops a set of clinical variables that are not same across other studies.

In response to a clear need for standard measures of phenotypes and exposures, PhenX (consensus measures for Phenotypes and eXposures) engaged 21 working groups of experts to identify high-quality, relatively low-burden, well-established measures of phenotypes and exposures. These measures were vetted by the scientific community prior to inclusion in the PhenX Toolkit (https://www.phenxtoolkit.org). The PhenX Toolkit provides researchers with a source of standard measures suitable for a variety of study designs in population-based research. Because the PhenX Toolkit provides a variety of high-quality measures, investigators can come to the Toolkit and select measures to expand their study, especially to add measures that are beyond the primary research focus of the study.

The nomenclature for the PhenX Toolkit was defined by the PhenX Steering Committee and is shown in Table 1. Currently, the PhenX Toolkit includes 295 measures spanning 21 research domains [Hamilton et al., 2011; Hendershot et al., 2011]. A measure usually comprises multiple variables or questions; so most measures correspond to many items in the other data sets described in this article.

Challenges in phenotype harmonization have been widely recognized, and efforts have been made in this emerging research field [Bennett et al., 2011; Fortier et al., 2010]. To help address these problems, all 295 PhenX measures have been mapped to multiple resources, including the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap/), Logical Observation Identifiers Names and Codes (LOINC®; http://loinc.org/), and the

Additional Supporting Information may be found in the online version of this article.

†Current address: Merck & Co., Inc., North Wales, Pennsylvania.

*Correspondence to: Huaqin Pan, RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709. E-mail: hpan@rti.org

Contract grant sponsor: This work was supported by the National Human Genome Research Institute and the American Recovery and Reinvestment Act through NHGRI U01 HG004597-01 to RTI International; by the National Library of Medicine through HHSN27620080000006C and National Center for Research Resources 3UL1RR025761-02S6 to Regenstrief Institute; and by the Intramural Research Program of the National Institutes of Health.

Table 1. PhenX Toolkit Nomenclature

PhenX Toolkit nomenclature	Definition	Example
PhenX domain	A PhenX domain is a field of research with a unifying theme and easily enumerated quantitative and qualitative measures (e.g., demographics, anthropometrics, organ systems, complex diseases, and lifestyle factors).	Alcohol, tobacco and other substances
PhenX measure	A PhenX measure refers broadly to a standardized way of capturing data on a certain characteristic of or relating to a study subject.	Alcohol—lifetime use
PhenX protocol	A PhenX protocol is a standard procedure recommended by a working group for investigators to collect and record a PhenX measure.	Alcohol Use Disorder and Associated Disabilities Interview Schedule, Fourth Edition Version (AUDADIS-IV)
PhenX variable	A PhenX variable is developed for each data element collected by the PhenX protocols. PhenX variables can be found in the PhenX data dictionary files at the "My Toolkit" page when the PhenX protocol is selected. For each PhenX variable, the data dictionary includes a unique variable name, variable ID, variable description, and other related attributes such as data types, units, and permitted values when applicable.	PX030101_lifetime_use—In your entire life, have you had at least one drink of any kind of alcohol, not counting small tastes or sips?
PhenX collection	A PhenX collection is a collection of measures with a shared characteristic, target population, or topic. The measures included in a collection may cut across research domains.	Alcohol use

Cancer Data Standards Registry and Repository (caDSR) of the cancer Biomedical Informatics Grid (caBIG; https://cabig.nci.nih.gov/). This article describes how PhenX measures were integrated into these standards and demonstrates the utility of this approach.

Database of Genotypes and Phenotypes

The dbGaP database, which was created by the National Center for Biotechnology Information, is a public repository for individual-level genotype, sequence, and phenotype data, and the associations between them [Mailman et al., 2007]. dbGaP currently contains more than 125,000 variables. Many of these variables may be similar enough to PhenX variables that they could be considered comparable or related for cross-study analysis. To help researchers interested in PhenX variables find similar variables in dbGaP, we developed a process for mapping dbGaP study variables to PhenX variables. As a proof of concept, variables from 16 completed studies deposited in dbGaP were mapped to PhenX measures. These results will be fully incorporated in dbGaP and will bring to light additional opportunities for cross-study analysis.

Investigators who submit data to dbGaP will be asked to identify PhenX variables as part of the data submission process. PhenX variables will then be highlighted as such in dbGaP. Because dbGaP was established before PhenX measures were developed, none of the studies currently in dbGaP used PhenX protocols. However, we know that there are many variables in dbGaP that are similar, or nearly identical, to PhenX variables and potentially could be combined with data collected using PhenX protocols as well as with each other. Although it is possible to run full-text searches within the dbGaP database to find data that are similar, experience tells us that the full-text searches for variables are likely to return large numbers of false positives. For example, a search on "education" will return more than 10,000 variables.

To make it easier for researchers to find non-PhenX variables that might be compared or combined with PhenX variables, scientists from PhenX and dbGaP investigated the feasibility of mapping dbGaP variables to PhenX variables. The first attempt began by examining four dbGaP studies, so that we could begin to develop the process and refine our ideas about what it means to map one variable to another. Each scientist was given all the variables for the four studies, including the variable description and a link to the variable report page on the dbGaP website. Using this information and all of the information available on the PhenX Toolkit, each scientist generated his or her own set of mappings for the dbGaP variables. Many factors such as measurement concept, protocol, code category (answer list), and measurement unit were discussed. Based on these discussions, the team decided on the following two levels of mapping:

- Comparable: The data collected in these variables are conceptually the same and should be able to be compared either directly or after a straightforward transformation/conversion. Examples of comparable variables are:
  - variables whose data were collected using the same protocol, or
  - variables that are recognized as producing the same data (e.g., age) or producing data that can be easily transformed (e.g., measured weight in kilograms and pounds, or "birth date" and "birth year"), although they do not share identical protocols.
- Related: The data in these variables cannot be directly compared, but could be compared after further manipulation. Examples of related variables are:
  - variables that are not collected using the same study protocols, but that measure similar properties (e.g., "measured weight" and "self-reported weight");
  - multiple PhenX variables that might need to be combined to reflect a single dbGaP variable (e.g., "weight" and "height" for "body mass index"); or
  - variables that have different qualifiers (e.g., "since last visit" vs. "have you ever," "regularly" vs. "at least once a week," or "hormone therapy" vs. "hormone therapy with a specific hormone name").

It is possible that a dbGaP variable neither corresponds nor is related to a PhenX variable or measure. Such a lack of correspondence or relation could be considered "not found," but this mapping level is not explicitly shown when looking at the dbGaP variable. Rather, variables that do not have a mapping level simply are not displayed.

A dbGaP variable can be mapped to multiple PhenX variables and/or measures. For example, the dbGaP variable phv00111936 (smok_evr: smoked more than 100 cigarettes or five packs in a lifetime) is mapped to four PhenX variables (as comparable to one variable and related to the other three) that are associated with three different PhenX Measures (see Fig. 1).

Once the mapping criteria had been agreed on, the remaining studies were mapped. Mapping was performed by at least two independent curators. Results were compared and, as before, discrepancies were resolved by consensus after discussion. Any new mapping criteria that were developed during this process were added to the guidelines for future use.

Figure 1. Screen shot of the report page for dbGaP variable phv00111936. In the column "Mapping," a green circle indicates that the mapping result is "comparable"; a half-filled yellow circle indicates that the mapping result is "related." The links below "PhenX Variable" take the user to a complete list of dbGaP variables with mapping to the PhenX variable. The "Measure" links take the user to the PhenX website's measures page.

In this report, we show the results of mapping 13 Gene Environment Association Studies (GENEVA) consortium studies and three electronic Medical Records and Genomics (eMERGE) network studies to PhenX. The GENEVA consortium (https://www.genevastudy.org) consists of 16 GWAS that aim to accelerate the understanding of genetic and environmental contributions to health and disease on a collection of mostly traditional epidemiologic cohorts [Cornelis et al., 2010]. The eMERGE network (https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page) is a national consortium formed to develop, disseminate, and apply approaches to research that combine DNA biorepositories with electronic medical record systems for large-scale, high-throughput genetic research [McCarty et al., 2011; Kho et al., 2011]. We used variable descriptions from GENEVA and eMERGE studies released in dbGaP at the time of this mapping. Table 2 lists the dbGaP studies, their dbGaP accession numbers, the total number of variables from each study, and the number of variables that mapped to a PhenX variable or measure. The percentage of variables mapped for a particular study ranges from 23% to 80%, but the effective mapping rate for all studies is somewhat higher than this because all of the studies contain variables that are not phenotype data, and therefore cannot be expected to have an analog in PhenX. These types of variables include administrative data such as IDs (e.g., for subjects, subjects' parents, locations of data collection, etc.), consent status, or information about instrumentation (e.g., sequencing platforms). Aside from administrative variables, there are some dbGaP variables that do not map to PhenX. In general, these variables reflect concepts that are study specific (e.g., "Are your ear lobes creased?" or "What is your US shoe size?").

Results of mapping the dbGaP studies to PhenX are summarized in the PhenX–dbGaP variables cross-reference table in Supp. Table S1. For these 16 dbGaP studies, the cross-reference table lists a total of 2,041 mappings, with 604 dbGaP variables mapped to 504 PhenX variables and 52 PhenX measures. The cross-reference table is available at the PhenX Toolkit website (https://www.phenxtoolkit.org). Examples of these mappings are illustrated in Table 3, in which individual PhenX variables are mapped to many variables from multiple studies, highlighting opportunities for cross-study analysis at the investigator's discretion. Note that "lipid_total_cholesterol" (a PhenX variable) maps to "dyslipidemia" (a condition). Although this mapping may be at first disconcerting, it is actually a good example of how mapping can identify data that is comparable or related, even though the reasons for collecting the data were different. "Lipid_total_cholesterol" is a variable associated with the PhenX lipid profile measure, and the data collected can be used to derive the condition "dyslipidemia." On the dbGaP website, mapping information for a variable is shown on that variable's report page. Figure 1 is a screenshot of the report

Table 2. List of dbGaP Studies Mapped to PhenX

Study name[a]	dbGaP accession[b]	Study variables		
		# Total	# Mapped	% Mapped
GENEVA—addiction	phs000092	140	90	64
GENEVA—birth weight	phs000096	226	130	57
GENEVA—blood clotting	phs000304	89	33	37
GENEVA—dental caries	phs000095	95	76	80
GENEVA—diabetes	phs000091	37	23	62
GENEVA—early onset stroke	phs000292	33	15	45
GENEVA—glaucoma	phs000308	59	32	54
GENEVA—lung cancer	phs000093	34	16	47
GENEVA—melanoma	phs000187	38	15	39
GENEVA—oral clefts	phs000094	39	22	56
GENEVA—preterm birth	phs000103	63	28	44
GENEVA—prostate cancer	phs000306	62	14	23
GENEVA—venous thrombosis	phs000289	53	35	66
eMERGE—cataract	phs000170	62	37	60
eMERGE—peripheral arterial disease	phs000203	44	22	50
eMERGE—electrocardiogram QRS	phs000188	36	16	44

[a]Names of studies have been shortened to reflect the title of the initiative they are part of, as well as the major research area to save space. Full names can be found at the dbGaP page for each study.
[b]To find the most recent version of these studies in dbGaP, use the following base URL and add the dbGaP accession to the end:
http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id =.
dbGaP, database of Genotypes and Phenotypes; GENEVA, Gene Environment Association Studies; eMERGE, electronic Medical Records and Genomics.

page for the dbGaP variable phv00111936 (smok_evr). The "Terms Linked to this Variable" section lists the PhenX variables mapped to smok_evr. The left column shows the level of mapping; a full green circle indicates comparable, and a half-filled yellow circle indicates related. The second column lists the name of the PhenX variable or measure that has been mapped to; these names are linked to a search page that displays all of the dbGaP variables that map to that PhenX variable (see Supp. Fig. S1). The third column gives a short definition of the mapped PhenX variable, whereas the measure column lists the PhenX measure associated with the variable. The names of the PhenX measures are links to the measure on the PhenX website.

Figure 2 shows the number of dbGaP mappings to PhenX as a function of the PhenX measure. Although dbGaP variables map to more than 100 different PhenX measures, only the 25 PhenX measures with the most mappings are shown here. When looking at this plot, you should keep in mind the following points:

- A single dbGaP variable can map multiple times onto a PhenX measure because there are multiple variables that are either comparable or related. For example, dbGaP variable phv00142512 (hgla: family history of glaucoma in first-degree relatives) is mapped to the PhenX measure personal and family history of eye disease and treatments six times, each time representing a different variable (one each asking about glaucoma for mother, father, sister, brother, daughter, and son).
- A single dbGaP variable can map to multiple PhenX measures because there are variables in each measure that are comparable or related. This was described earlier for the variable phv00111936 (smok_ever), for which a single dbGaP variable mapped into three different measures.
- Although in some cases there are PhenX measures that contain a single basic concept (e.g., gender, age, height, or weight), the same piece of information can be collected in other PhenX measures. This is a consequence of PhenX absorbing entire instruments to retain their coherence rather than just cherry picking particular questions from an instrument to include in PhenX. For example, the concept of "gender" is represented in PhenX as its own measure, gender. It is also present in the following PhenX measures: cancer–personal and family history,

Table 3. Examples of Four PhenX Variables Mapped to Variables from 16 Mapped dbGaP Studies

PhenX variable name	dbGaP variables	Level
PX030601_Cigarette_smoking_current	Cigarette smoke[b]	Related
	packyrs_ca[g]	Related
	currsmoke[g]	Related
	m_Smoker[a]	Related
	packyr[f]	Related
	Smoking_Status[j]	Related
	Smoking_Status[k]	Related
PX021502_Self_reported_weight_Lbs	Weight[b]	Comparable
	Wt_Kg[c]	Comparable
	wt[d]	Comparable
	weight[h]	Comparable
	WEIGHT[l]	Comparable
	m_WtM_OGTT[a]	Comparable
	apptwt_kgs[i]	Comparable
	weight[f]	Comparable
	weight[j]	Comparable
	Weight[k]	Comparable
PX040201_lipid_total_cholesterol	Counts_Total_Cholesterol_Measurement[j]	Comparable
	chol[d]	Related
	dyslipidemia[k]	Related
PX030301_Alcohol_30Day_Frequency	alcohol_use_past_month[j]	Comparable
	alco[f]	Related
	alcohol[d]	Related
	m_Drinker[a]	Related
	m_Alc_Drnks[a]	Related

[a]GENEVA—birth weight. [b]GENEVA—blood clotting. [c]GENEVA—dental caries. [d]GENEVA—diabetes. [e]GENEVA—early onset stroke. [f]GENEVA—glaucoma. [g]GENEVA—prostate cancer. [h]GENEVA—addiction. [i]GENEVA—venous thrombosis. [j]eMERGE—cataract. [k]eMERGE—peripheral arterial disease. [l]eMERGE—electrocardiogram QRS.
dbGaP, database of Genotypes and Phenotypes; GENEVA, Gene Environment Association Studies; eMERGE, electronic Medical Records and Genomics.

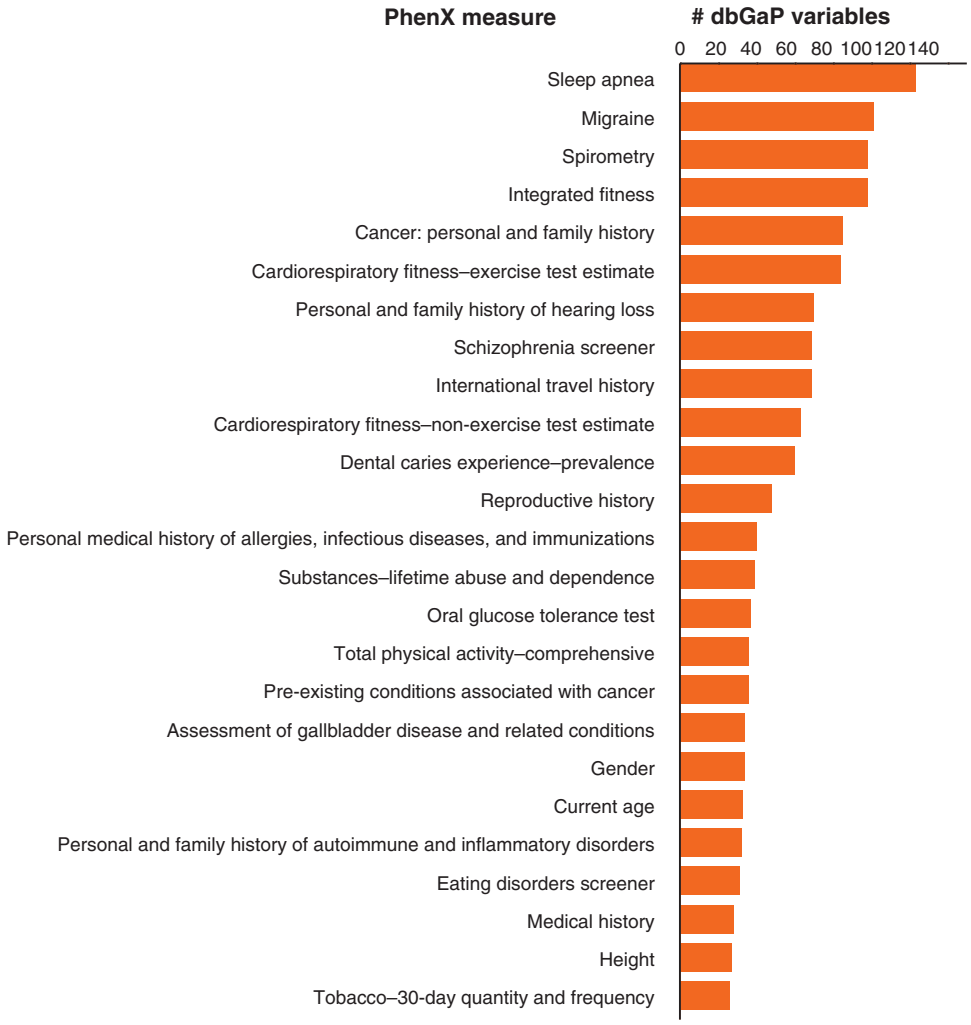

Figure 2. A plot that shows the number of dbGaP mappings to PhenX as a function of the PhenX measure.

sleep apnea, spirometry, schizophrenia screener, migraine, cardiorespiratory fitness–exercise test estimate, integrated fitness, and international travel history among others.

The final point explains why the measures that have the most dbGaP mappings to PhenX are those like "sleep apnea" and "migraine," rather than "gender" or "current age." A measure like "sleep apnea" or "migraine" contains an extensive protocol that collects a large number of discrete variables including age, gender, height, and weight in addition to the more specific data suggested by its name. Therefore, these measures will have many dbGaP variables mapped to them from a single study, whereas the measure "gender" may have only one variable mapped to it from each study.

For the pilot study described, only a handful of dbGaP studies comprising a relatively small number of variables were selected, and it was relatively easy to identify all variables related to a given concept (e.g., diabetes, race, smoking). The manual approach, although somewhat laborious, resulted in thoughtful, consistent mappings. That said, scaling up will present challenges, and using natural language processing (NLP) algorithms to identify similarities and differences among the variables may be helpful in this regard. For example, NLP has been used successfully to identify cataract cases from electronic health records [Peissig et al., 2012]. Perhaps in the future, NLP can be used to augment and extend the described approach.

Logical Observation Identifiers Names and Codes

LOINC® (http://loinc.org/) is a vocabulary standard for identifying laboratory tests, clinical measurements and reports, survey instruments, and other kinds of clinical observations. By providing universal identifiers for a wide range of measurements and observations, LOINC® enables exchange and aggregation of electronic health data from independent systems for many purposes [Vreeman et al., 2010b; McDonald et al., 2003]. LOINC® has been widely adopted in the private and public sectors, both within the United States and by users in more than 140 countries worldwide. Notably, the Health Information Technology (HIT) Standards Committee of the Federal Office of the National Coordinator for Health Information Technology recently adopted LOINC® as the coding system for transmitting results of laboratory and other tests, assessment instruments, and many other clinical variables [Health IT Standards Committee, 2011]. LOINC® has now incorporated all of the PhenX content, enabling results of PhenX measures from independent systems to be shared using the same exchange, storage, and processing infrastructure that health information systems use

61450-3 Each time you drank milk, how much did you usually drink in the past 30 days PhenX

STATUS

Trial – caution, may change.

BASIC ATTRIBUTES

Class/Type: PHENX/Clinical

NORMATIVE ANSWER LIST:

SEQ#	Answer	Code	Answer ID
1	Less than 1 cup	1	LA13923-0
2	1 cup (8 ounces)	2	LA13924-8
3	More than 1 cup	3	LA13925-5

SURVEY QUESTION:

Text: Each time you drink milk, how much do you usually drink?

Source: PhenX.050201020100

PARTS

Part Type	Part No.	Part Name
Component	LP102639-4	Each time you drank milk, how much did you usually drink in the past 30D *[Each time you drank milk, how much did you usually drink in the past 30 days]*
Method	LP95333-8	PhenX

IMAGE

8 ounce glass of milk

Source: Consensus measures for Phenotypes and Exposures

Figure 3. An example of accessory (image) content for a PhenX variable as represented in LOINC®.

for sending a serum glucose test result or a chest X-ray report. Here, we describe the process of representing PhenX content in LOINC®, advantages to this linkage, and some of the lessons learned.

Each term in LOINC® provides a "fully specified" name using an established model that contains six main axes (Supp. Table S2) [McDonald et al., 2011]. The model produces names that are detailed enough to distinguish among similar clinical observations. As a collection, PhenX contains many kinds of measurements, from laboratory tests to anthropomorphic measures and validated questionnaires. LOINC® has developed a robust model for representing standardized assessment instruments, recognizing that they have psychometric properties that are essential for interpreting meaning [Vreeman et al., 2010a]. Thus, in addition to the structured name, LOINC® stores many other attributes about the individual variable, including the exact question text and source, example units of measure (for quantitative variables) and full answer lists (for categorical variables), references, descriptions, and external copyright information when applicable. LOINC® also creates terms for named collections of variables (called "panels" in LOINC®) and enumerates the child elements contained in that set into an explicit hierarchy.

Through iterative development, the LOINC® team incorporated the entire set of PhenX measures into LOINC®, either by creating new LOINC® terms or by linking the PhenX variables to existing LOINC® terms. We extracted content from the PhenX Toolkit for every variable in each measure and domain, starting with a small set of PhenX content that was first represented in LOINC® version 2.29 as a proof of concept. Some variables, such as head circumference and gestational age, were already present in LOINC®, but the majority of them were not. We modeled variables new to LOINC® according to the established naming conventions. From the protocol text, we extracted and stored the key accessory attributes (e.g., units of measure or the allowable answer choices). Many PhenX variables are defined or illustrated by graphics (e.g., line drawings or photographs) to show exactly how a measurement should be taken or how to answer that particular question. The LOINC® team created a mechanism for storing and displaying these graphics in the free desktop mapping program called the Regenstrief LOINC® Mapping Assistant (http://loinc.org/relma) and the online LOINC® search application (http://search.loinc.org). Figure 3 illustrates how the accessory content for a PhenX variable is represented in LOINC®, including the structured answer list, exact question text, and a

reference image. To capture the hierarchical arrangement of variables into collections, we created LOINC® panel terms at the level of each PhenX domain, measure, and protocol. These named panels include all of the corresponding PhenX child elements in a formal hierarchy linked to that panel. Over time, the LOINC® team added the remainder of the PhenX content (Supp. Table S3). LOINC® has now completed modeling of all PhenX variables from 295 measures in 21 research domains; 138 existing LOINC® terms were mapped to PhenX variables, and approximately 4,500 new LOINC® terms were added based on the PhenX content.

Incorporating the PhenX content into LOINC® has many advantages. Adding the PhenX measures to LOINC® enables the results to be shared using the same HIT infrastructure and standards that are now becoming widely adopted. In addition, the LOINC® model provides the same uniform computable representation of the PhenX content as other standard assessments and data sets contained in LOINC®, including many mandated by the Centers for Medicare & Medicaid Services (CMS) and provided by other National Institutes of Health institutes, such as the Patient Reported Outcomes Measurement Information System (PROMIS; http://www.nihpromis.org/) [Gershon et al., 2010; Riley et al., 2011] and Quality of Life Outcomes in Neurological Disorders (Neuro-QOL; http://www.neuroqol.org/). Having such a common representation that promotes sharing will accelerate genomic and other clinical research. Moreover, because of LOINC®'s broad adoption worldwide, representing the PhenX measures in LOINC® will widen the audience for PhenX measures.

The process of integrating the PhenX content into LOINC® elucidated several important lessons. Many of the PhenX measures selected instruments and protocols that were initially conceptualized as paper data collection forms. As the LOINC® team defined its terms and parsed this content into its data model, it revealed many of the same challenges that were encountered with coding other widely used survey instruments [Vreeman et al., 2010b]. For example, some protocols did not specify all of the variables needed to collect the data, or lacked sufficient detail to precisely define the observation. In other cases, the information model of the protocols differed substantially from the typical information model used to exchange data between clinical care systems with LOINC® and messaging standards like Health Level Seven International (HL7). The LOINC® team always found solutions to these problems through discussions with the PhenX team. One strategy was to turn a long list of "check all that apply (yes or no)" questions into a single variable with an answer choice list that could be repeated as many times as necessary. For example, a protocol requiring answers of yes or no to a long list of potential diseases could be transformed into an active diseases variable whose answer values could be the diseases present. This approach dramatically reduced the number of LOINC® observation codes necessary to cover all of the PhenX variables and was consistent with the prevailing health data exchange and storage conventions. We anticipate that these challenges will diminish as survey instrument developers become acquainted with the formality required for computer representation of instruments in LOINC®.

LOINC® was chosen as the vocabulary standard for several reasons. The goal was to represent PhenX content in a widely adopted vocabulary standard that would enable data aggregation using prevailing conventions (e.g., HL7 messaging). The value of LOINC® in this context is that it provides a set of universal identifiers and a uniform model of that instrument across any context. LOINC® is well suited for clinical observations and formal surveys and questionnaires, and it is the standard adopted by the HIT Standards Committee for laboratory and non-laboratory measurements and observations. When this pilot study was initiated, LOINC® already contained many similar complete packages of standardized assessments and data sets, including the CMS-required Minimum Data Set (https://www.cms.gov/MinimumDataSets20/), Outcome and Assessment Information Set (https://www.cms.gov/OASIS/), the new Continuity Assessment Record and Evaluation instrument, Patient Health Questionnaire, PROMIS [Gershon et al., 2010], and Neuro-QOL. Making PhenX content available in the same model and format will facilitate data interoperability and data exchange.

Common Data Elements in caDSR

The caDSR is a data standards repository in caBIG [caBIG Strategic Planning Workspace, 2007; Kakazu et al., 2004]. It is an open-source, open-access information network designed to enable secure data exchange throughout the cancer research community. The caDSR includes a catalog of common data elements (CDEs). Each CDE is a unique pairing of a data element concept, which represents the question metadata, and a value domain, which represents the answer metadata. One or more CDEs are either assigned or created for every PhenX protocol (it is possible for a PhenX measure to have multiple protocols). There are 353 PhenX protocols mapped with 379 CDEs; 343 of these CDEs were newly created for PhenX.

PhenX has reused existing CDEs when available. The need to create so many new CDEs is not surprising; the CDEs previously available were focused either on general demographic concepts such as gender, race, and age, or on specific concepts related to cancer, whereas the focus of PhenX is much broader. PhenX represents 21 research domains, most of which are outside the traditional cancer research domain; such domains include the psychiatric, psychosocial, and social environments domains. For example, two new CDEs were created for the protocols of the measure assay for chlamydia/gonorrhea: immunology *Chlamydia trachomatis* assay laboratory finding result (3151324) and immunology gonorrhea assay laboratory finding result (3153202). At the request of the caDSR administrator, the PhenX CDEs' workflow status was changed from "draft new" to "released" so that they would be available for reuse; they have been already used by other studies. In the caDSR, PhenX protocols are organized by research domain and can be located using the CDE Browser (https://cdebrowser.nci.nih.gov/CDEBrowser/), as shown in Figure 4.

Table 4 shows an excerpt from the cross-reference table that includes LOINC® codes and caCDR CDEs that are associated with each PhenX protocol. The comprehensive cross-reference table provided, in Supp. Table S4, is available on the Toolkit website and will serve as a valuable resource to investigators as well as bioinformaticists.

Discussion

Recognition and use of the PhenX Toolkit continues to increase as investigators begin to realize the importance of collecting data with standard instruments or tools. At of the end of January 2012, there were 259,077 visitors to the Toolkit website. Most Toolkit visitors are from the United States, the United Kingdom, and Australia, but there have also been visitors from 143 other countries. There are currently 637 registered users. Registered users have access to additional features such as the "My Toolkit" for collecting and saving selected measures. Joining the Toolkit network makes it possible for users of the network to contact each other. The idea is that the network can be used to facilitate collaboration at the study-design

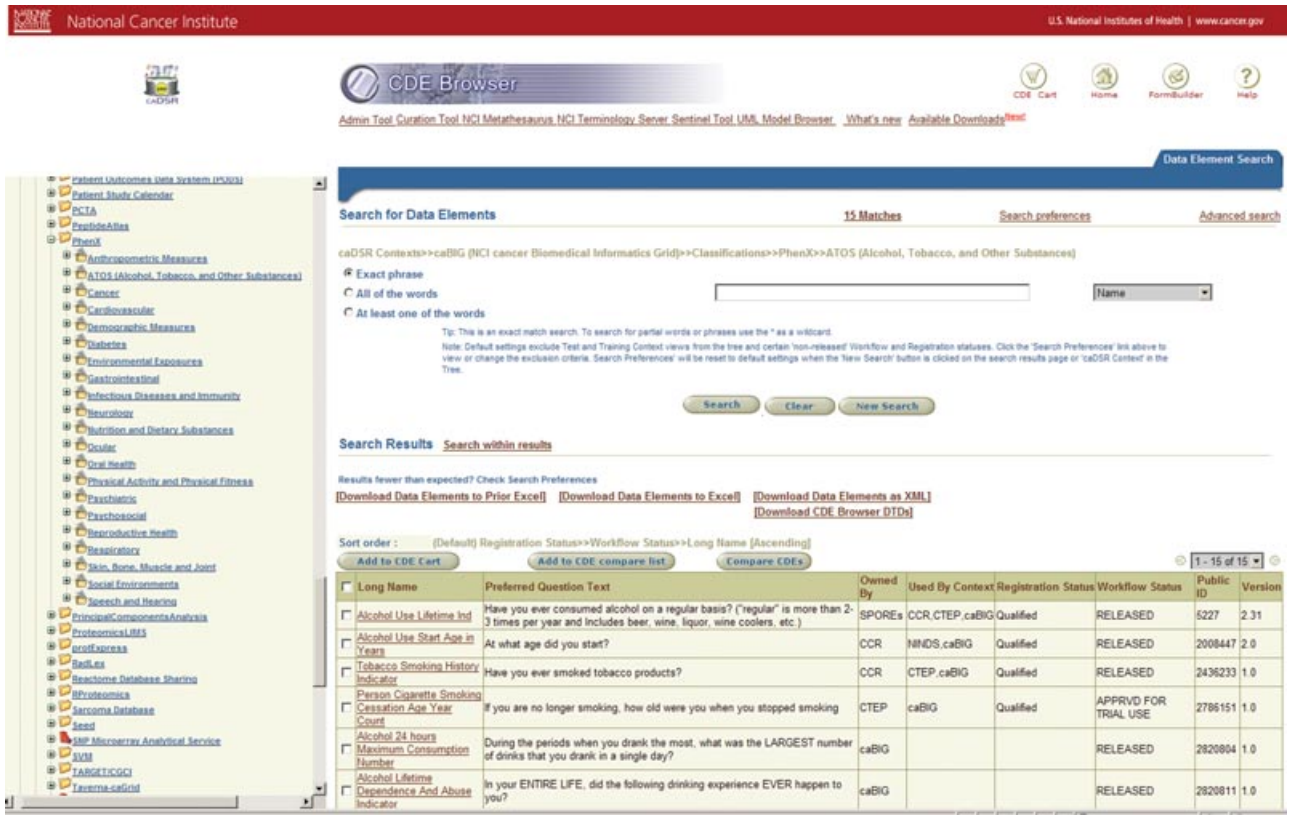

Figure 4. Common Data Elements (CDEs) for PhenX domains listed in CDE Browser.

Table 4. Representation of Five PhenX Protocols in LOINC® and caDSR CDE

PhenX protocol name	PhenX protocol ID	LOINC® identifier	caDSR CDE (public ID)
Current age	PX010101	Current age proto (62293-6)	Derived person age value (2423393)
Gender	PX010701	Gender (46098-0)	Gender code (2179640)
Tobacco—smoking status	PX030601	Tobac smoke status adoles proto (62553-3)	Adolescent tobacco smoking history indicator (2923486)
Lipid profile	PX040201	Lipid profile proto (62391-8)	Person high cholesterol indicator (2936262)
Personal history of type 1 and type 2 diabetes	PX140501	Pers hx type 1 and 2 diabetes proto (62797-6)	Person diabetes personal medical history assessment description text (3070673)

Each row of the table represents one PhenX protocol. Supp. Table S2 contains the complete list of LOINC® codes and CDEs that are equivalent PhenX measures and protocols. LOINC®, Logical Observation Identifiers Names and Codes; caDSR, Cancer Data Standards Registry and Repository; CDE, common data element.

phase as well as retrospective cross-study analyses. Early adopters of PhenX measures include PhenX RISING (Real world, Implementation, SharIng) project (https://www.phenx.org/Default.aspx?tabid=748), the National Eye Institute Glaucoma Human Genetics Collaboration consortium, and the Gulf Long-Term Follow-up Study (http://nihgulfstudy.org/). Additional information about early adopters is available on the Toolkit website.

dbGaP and PhenX will continue to collaborate and extend the relationship between the two resources. As noted previously, when new studies submit their data to dbGaP and identify their variables as PhenX, this information will be stored in the database. Other areas of development in dbGaP include adding the ability to filter search results to return variables submitted as, or mapped to, PhenX; mapping additional retrospective studies; and adding other languages/ontologies beneath the "Terms Linked to this Variable" heading on the variable report page (e.g., International Classification of Diseases-9 codes or Medical Subject Headings terms). These developments will expand the ability of investigators to identify variables of interest across dbGaP. This information can be used prospectively, at the study design stage, or retrospectively, to identify opportunities for cross-study analysis with or without the need for harmonization.

PhenX intends to keep its mappings to LOINC® and caBIG CDEs up to date when the Toolkit is expanded or updated, either by linking to concepts already extant in those resources or by creating new concepts within them (as described earlier). By maintaining collaborations and close connections to these resources (and potentially adding resources), PhenX will be able to expand and update the cross-reference table accordingly. The results presented here extend the utility of PhenX measures and add value to resources like dbGaP, LOINC®, and caDSR.

The goal of associating PhenX measures with existing standards is to make it easier for investigators to share data and to compare and combine study results. Integrating PhenX measures into existing standards (LOINC®, CDE) and mapping PhenX variables to dbGaP study variables extend the utility of PhenX measures and

reveal new opportunities for cross-study analysis. The primary limitation of data sharing is that study-specific measures are needed to support scientific inquiry. That is, deciding what measures are needed to effectively address a specific research question is inherent to study design. Striking a balance between the inclusion of study-specific measures and the inclusion of standard measures is necessary; both types of measures will affect the overall scientific impact of the study results. The work presented here will facilitate individual investigators to recognize and realize the potential of data sharing and cross-study analysis. Linking dbGaP, LOINC®, caDSR, and PhenX resources will help promote data sharing and thus will have a significant positive impact on biomedical research.

Acknowledgments

Disclosure Statement: The authors declare no conflict of interest.

References

Bennett SN, Caporaso N, Fitzpatrick AL, Agrawal A, Barnes K, Boyd HA, Cornelis MC, Hansel NN, Heiss G, Heit JA, Kang JH, Kittner SJ, and others. 2011. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. Genet Epidemiol 35:159–173.

Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. 2009. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 38:263–273.

caBIG Strategic Planning Workspace. 2007. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Stud Health Technol Inform 129:330–334.

Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF, Feenstra B, Feingold E, and others. 2010. The Gene, Environment Association Studies Consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 34:364–372.

Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, Deschênes M, Knoppers BM, Doiron D, Keers JC, Linksted P, Harris JR, and others. 2010. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol 39:1383–1393.

García-Closas M, Lubin JH. 1999. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. Am J Epidemiol 149:689–692.

Gershon RC, Rothrock N, Hanrahan R, Bass M, Cella D. 2010. The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. J Appl Meas 11:304–314.

Hamilton, CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, Hammond JA, Huggins W, Jackman D, Pan H, Nettles DS, Beaty TH, and others. 2011. The PhenX Toolkit: get the most from your measures. Am J Epidemiol 174:253–260.

Health IT Standards Committee. 2011. Recommendations to the Office of the National Coordinator for Health Information Technology (ONC) on the assignment of code sets to clinical concepts [data elements] for use in quality measures. [Letter]. Accessed at: http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_12811_955546_0_0_18/HITSC_CQMWG_VTF_Transmit_090911.pdf.

Hendershot TP, Pan H, Haines J, Harlan WR, Junkins HA, Ramos EM, Hamilton CM. 2011. Use the PhenX Toolkit to add standard measures to your study. Curr Protoc Hum Genet 71:1.21.1–1.21.18.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367.

Kakazu KK, Cheung LW, Lynne W. 2004. The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. Hawaii Med J 63:273–275.

Kho A, Pacheco J, Peissig P, Rasmussen L, Newton K, Weston N, Crane P, Pathak J, Chute C, Bielinski S, Kullo I, Li R, Manolio T, Chisholm R, Denny J. 2011. Electronic medical records for genetic research: results of the eMERGE Consortium. Sci Transl Med 3:79re1.

Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JP, Janssens AC, Ostell J, and others. 2009. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol 170:269–279.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, and others. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181–1186.

Manolio TA. 2009. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. Pharmacogenomics 10:235–241.

McCarty C, Chisholm R, Chute C, Kullo I, Jarvik G, Larson E, Li R, Masys D, Ritchie M, Roden D, Struewing J, Wolf W, eMERGE Team. 2011. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 4:13.

McDonald CJ, Huff SM, Mercer K, Hernandez J, Vreeman DJ, editors. 2011. Logical Observation Identifiers Names and Codes (LOINC®) users' guide. Indianapolis, Indiana: Regenstrief Institute.

McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P. 2003. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 49:624–633.

Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, Chen L, Denny JC, Wilke RA, Pathak J, Carrell D, Kho AN, Starren JB. 2012. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 19:225–234.

Riley WT, Pilkonis P, Cella D. 2011. Application of the National Institutes of Health Patient-Reported Outcome Measurement Information System (PROMIS) to mental health research. J Ment Health Policy Econ 14:201–208.

Thorisson GA, Muilu J, Brookes AJ. 2009. Genotype–phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet 10:9–18.

Vreeman DJ, McDonald CJ, Huff SM. 2010a. LOINC®: a universal catalogue of individual clinical observations and uniform representation of enumerated collections. Int J Funct Inf Pers Med 3:273–291.

Vreeman DJ, McDonald CJ, Huff SM. 2010b. Representing patient assessments in LOINC®. AMIA Annu Symp Proc 2010:832–836.