scientific data

OPEN

ARTICLE

Identifying Datasets for Cross-Study Analysis in dbGaP using PhenX

Huaqin Pan[1], Vesselina Bakalov[1], Lisa Cox[1], Michelle L. Engle[1], Stephen W. Erickson[1], Michael Feolo[2], Yuelong Guo[3], Wayne Huggins[1], Stephen Hwang[1], Masato Kimura[2], Michelle Krzyzanowski[1], Josh Levy[4], Michael Phillips[1], Ying Qin[1], David Williams[1], Erin M. Ramos[5] & Carol M. Hamilton[1]

Identifying relevant studies and harmonizing datasets are major hurdles for data reuse. Common Data Elements (CDEs) can help identify comparable study datasets and reduce the burden of retrospective data harmonization, but they have not been required, historically. The collaborative team at PhenX and dbGaP developed an approach to use PhenX variables as a set of CDEs to link phenotypic data and identify comparable studies in dbGaP. Variables were identified as either comparable or related, based on the data collection mode used to harmonize data across mapped datasets. We further added a CDE data field in the dbGaP data submission packet to indicate use of PhenX and annotate linkages in the future. Some 13,653 dbGaP variables from 521 studies were linked through PhenX variable mapping. These variable linkages have been made accessible for browsing and searching in the repository through dbGaP CDE-faceted search filter and the PhenX variable search tool. New features in dbGaP and PhenX enable investigators to identify variable linkages among dbGaP studies and reveal opportunities for cross-study analysis.

Introduction

Secondary analysis using multiple study datasets can validate findings from individual studies and generate more power to detect subtle and complex associations not possible by individual studies[1–7]. However, identifying the relevant datasets to include in a secondary analysis from publicly available data repositories can be challenging. One fundamental challenge in searching the metadata of archived biological datasets lies in the unstructured nature of the variable description. Additionally, variation in the semantic terms poses a barrier for data findability and reuse. For example, “drink regularly past month” and “alcohol consumption frequency last 30 days” can be recognized as comparable concepts by manual curation, but not by simple keyword search. Ultimately, heterogeneity in data collection, and especially the lack of standard measurement protocols used to collect data, restricts the ability to combine or harmonize data from multiple studies over time and therefore limits the overall impact of individual studies. Ideally, each funded study would have further scientific impact after its initial analysis through secondary analyses and meta-analyses designed to answer research questions requiring large patient populations.

The National Institutes of Health (NIH) Strategic Plan for Data Science (SPDS) promotes FAIR principles (i.e., **F**indable, **A**ccessible, **I**nteroperable, **R**eusable) to facilitate data sharing. Objective 2–3 of the SPDS, “Leverage Ongoing Initiatives to Better Integrate Clinical and Observational Data into Biomedical Data Science,” states that the NIH will “promote use of the NIH Common Data Elements Repository[8]”. Several NIH Common Data Element (CDE) resources have been established to address the data silo challenge[9–12]. PhenX (consensus measures for Phenotypes and eXposures), as one of the NIH CDE Repository projects, is a community-driven project to establish and promote the use of standard data collection protocols to improve the quality and consistency of data collection and to facilitate data sharing. The PhenX Toolkit shares measurement protocols that are

[1]RTI International, Research Triangle Park, NC, USA. [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [3]GeneCentric Therapeutics Inc., Durham, NC, USA. [4]Levy Informatics, Chapel Hill, NC, USA. [5]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. ✉e-mail: hpan@rti.org

Identical	Variables that are immediately ready for direct harmonization between datasets, without any transformation needed to combine the data. At the time of curation, this term was reserved for prospective, investigator self-identified use for future submissions to the dbGaP database.
Comparable	Two variables that were conceptually similar and that contain data that can be directly harmonized or compared after a simple logical or mathematical transformation.
Related	Bioassay variables and other instances when the methods for data collection may be distinct. This distinction is to alert investigators that they should review the methods carefully before proceeding with transformation of the data for harmonization.

Table 1. Definitions of mapping levels.

recommended by experts. A subset of PhenX protocols is included in the NIH CDE Repository. Each standard protocol can itself be a CDE, and the protocols often contain a set of variables, which can also be used as CDEs to harmonize data. Incorporating standard measurement protocols at the study design stage can reduce the need for data harmonization over time by addressing the challenge of data harmonization at its source[13–17]. Retrospective data harmonization has many challenges and entails a labor-intensive process to combine data collected using different data collection protocols[18–24]. Use of CDEs can help identify comparable study datasets from multiple data sources and repositories, enable cross-study analysis, and reduce the burden of retrospective data harmonization[25,26].

The PhenX Toolkit (https://www.phenxtoolkit.org) is a catalog of measurement protocols recommended for use by working groups of domain experts. In general, PhenX protocols are well-established, low-burden methods of collecting important data to assess phenotype and exposure data in studies involving human participants, including clinical, translational, genomic, and epidemiological studies. The protocols included in the PhenX Toolkit are relevant to a wide range of research domains[27–31]. Specialty collections in the PhenX Toolkit provide additional depth in specific research areas, such as Social Determinants of Health and COVID-19[32–36]. The database of Genotypes and Phenotypes (dbGaP) at the National Library of Medicine's National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/gap/) was created to provide a common location for storage of data from NIH-funded genomics-based and other studies. As a controlled access data repository that archives and distributes data from research studies investigating the interrelated nature of genotypes, phenotypes, and exposures[37,38] dbGaP has adapted to manage a wide and continually expanding variety of data types since its creation. It supports the NIH genomic data sharing policy[39] by allowing for the secondary use of data by investigators who have an approved research proposal consistent with the data use limitations for each dataset. However, most studies present in the dbGaP database were completed prior to the introduction and recommendation of CDEs, and before the first release of the PhenX Toolkit in 2009. The lack of CDEs and the heterogeneity of terminology among dbGaP studies increases the monetary and personnel burdens for investigators, who must manually review hundreds of data elements to identify studies which may be suitable for data harmonization and meta-analysis.

In this paper, we describe the process for mapping PhenX CDEs consisting of either protocols or individual variables to dbGaP study variables; the development of tools to help investigators find datasets using these linkages; and new options available for the identification of CDEs when submitting studies to dbGaP. PhenX protocols (e.g., "Tobacco - Age of Initiation of Use - Adolescent" [PX030701]) consist of a list of variables to collect data about phenotypes and exposures (e.g., PX030701_First_Cigarette_Smoking_Age). The dbGaP study datasets (e.g., Jackson Heart Study [JHS] Cohort, phs000286) consist of variables recording phenotype data (e.g., TOBA2, phv00128497). The PhenX – dbGaP variable mapping process compares the PhenX variables with the dbGaP variables and identifies all the dbGaP variables that can be mapped to relevant PhenX variables. The addition of PhenX CDE linkages at the variable level enables investigators who visit dbGaP to identify linked variables across studies which were not retrievable using keyword search, thus enhancing the findability of linked datasets.

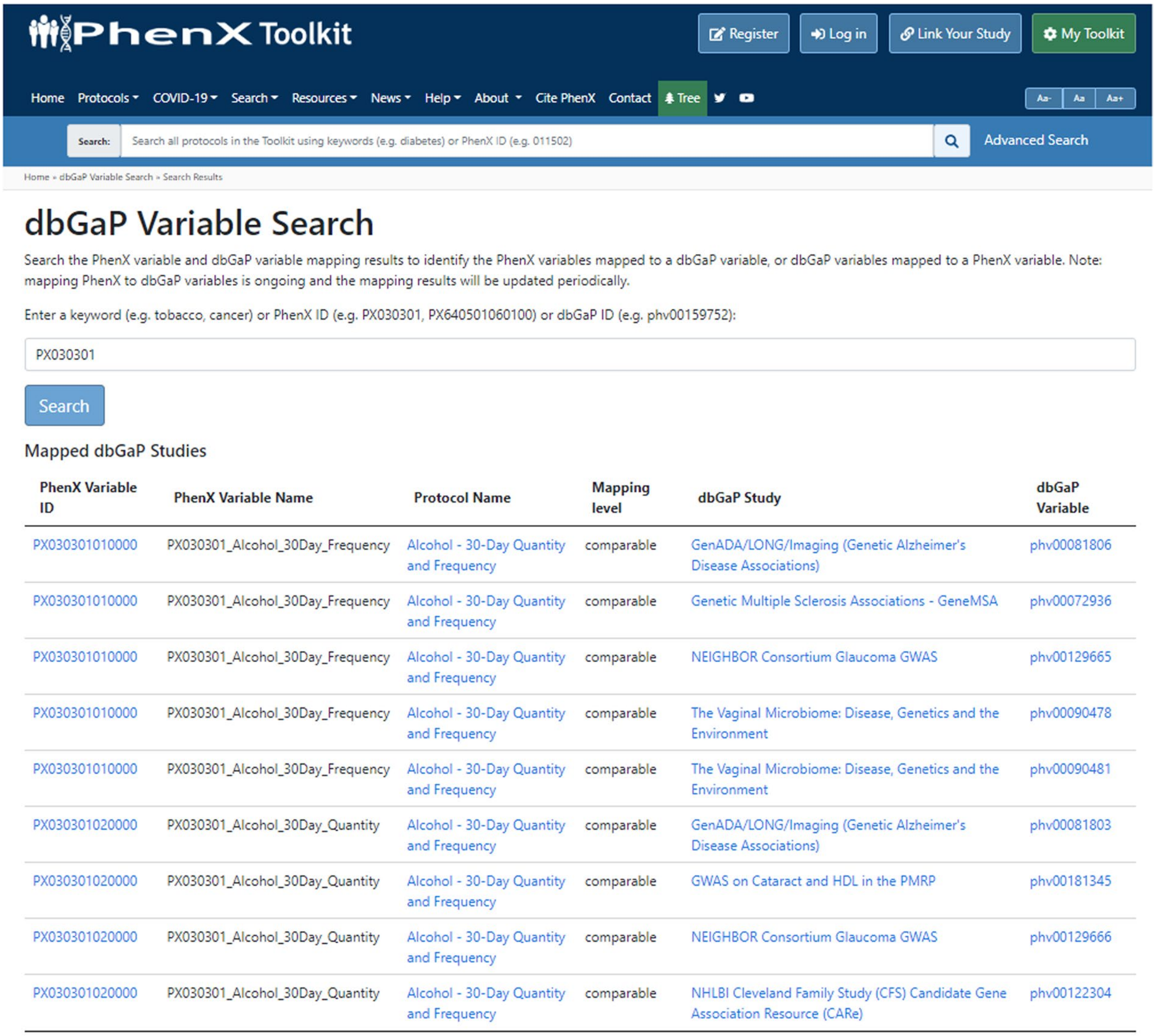
Results

Linking studies via PhenX-dbGaP variable mapping.

In this paper, we report the mapping results from the February 28, 2017, data freeze, which includes the June 8, 2016, release of dbGaP and the August 30, 2016, release of the PhenX Toolkit. The results of this mapping include 13,653 dbGaP linked variables that come from 521 dbGaP studies using PhenX variables (see "Mapping PhenX variables to dbGaP study variables" in Methods for more detail). These variable mappings include "many-to-many" relationships because both PhenX measurement protocols and dbGaP studies have redundant variables (Table 2). For example, the PhenX variables collecting information on "History of Cancer" are present in "Personal History" for multiple disease and condition measurement protocols. Similarly, some dbGaP variables in longitudinal studies were measured at multiple visits.

Identifying opportunities for cross-study analysis.

Scenario 1. PhenX users collecting alcohol use data identify dbGaP studies linked by variable mappings. First, the user can use the PhenX "dbGaP Variable Search" under "Search" (https://www.phenxtoolkit.org/vsearch) to search keywords or PhenX or dbGaP identifiers. For example, searching PhenX measurement protocol ID "PX030301" (which represents the protocol titled "Alcohol - 30-Day Quantity and Frequency") returns nine dbGaP variables from six dbGaP studies with links to dbGaP web pages (Fig. 1). The PhenX variable ID links to the measurement protocol page providing complete variable information, and the dbGaP results link to the dbGaP study and variable pages, respectively. Additionally, users browsing the PhenX Toolkit can visit the individual PhenX measurement protocol page, for



dbGaP Variable Search

Search the PhenX variable and dbGaP variable mapping results to identify the PhenX variables mapped to a dbGaP variable, or dbGaP variables mapped to a PhenX variable. Note: mapping PhenX to dbGaP variables is ongoing and the mapping results will be updated periodically.

Enter a keyword (e.g. tobacco, cancer) or PhenX ID (e.g. PX030301, PX640501060100) or dbGaP ID (e.g. phv00159752):

PX030301

Search

Mapped dbGaP Studies

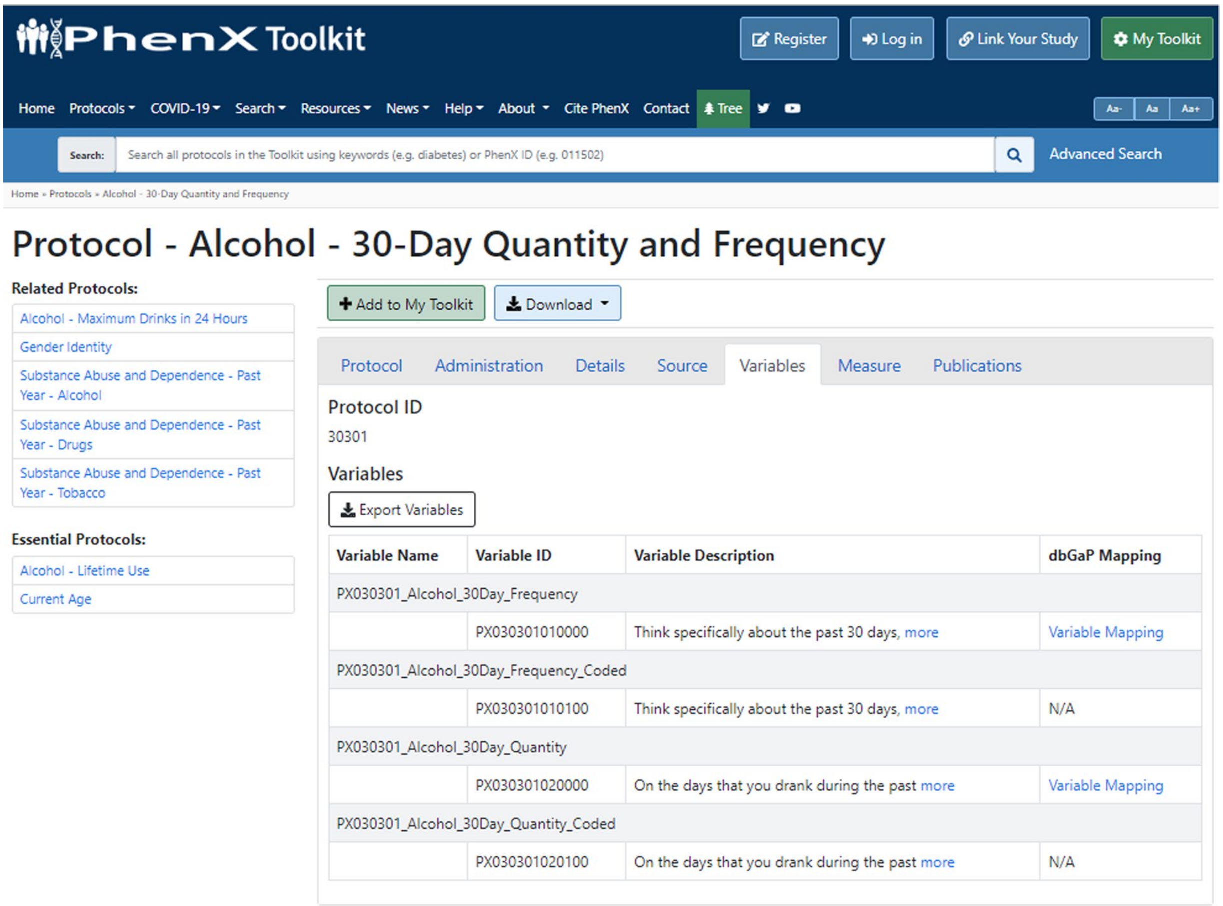
PhenX Variable ID	PhenX Variable Name	Protocol Name	Mapping level	dbGaP Study	dbGaP Variable
PX030301010000	PX030301_Alcohol_30Day_Frequency	Alcohol - 30-Day Quantity and Frequency	comparable	GenADA/LONG/Imaging (Genetic Alzheimer's Disease Associations)	phv00081806
PX030301010000	PX030301_Alcohol_30Day_Frequency	Alcohol - 30-Day Quantity and Frequency	comparable	Genetic Multiple Sclerosis Associations - GeneMSA	phv00072936
PX030301010000	PX030301_Alcohol_30Day_Frequency	Alcohol - 30-Day Quantity and Frequency	comparable	NEIGHBOR Consortium Glaucoma GWAS	phv00129665
PX030301010000	PX030301_Alcohol_30Day_Frequency	Alcohol - 30-Day Quantity and Frequency	comparable	The Vaginal Microbiome: Disease, Genetics and the Environment	phv00090478
PX030301010000	PX030301_Alcohol_30Day_Frequency	Alcohol - 30-Day Quantity and Frequency	comparable	The Vaginal Microbiome: Disease, Genetics and the Environment	phv00090481
PX030301020000	PX030301_Alcohol_30Day_Quantity	Alcohol - 30-Day Quantity and Frequency	comparable	GenADA/LONG/Imaging (Genetic Alzheimer's Disease Associations)	phv00081803
PX030301020000	PX030301_Alcohol_30Day_Quantity	Alcohol - 30-Day Quantity and Frequency	comparable	GWAS on Cataract and HDL in the PMRP	phv00181345
PX030301020000	PX030301_Alcohol_30Day_Quantity	Alcohol - 30-Day Quantity and Frequency	comparable	NEIGHBOR Consortium Glaucoma GWAS	phv00129666
PX030301020000	PX030301_Alcohol_30Day_Quantity	Alcohol - 30-Day Quantity and Frequency	comparable	NHLBI Cleveland Family Study (CFS) Candidate Gene Association Resource (CARe)	phv00122304

Fig. 1 PhenX "dbGaP Variable Search" Tool. This tool (https://www.phenxtoolkit.org/vsearch) allows users to search keyword, PhenX identifiers, or dbGaP identifiers to find the PhenX-dbGaP variable mappings. For example, searching PhenX ID "PX030301" returns 9 dbGaP variables from 6 dbGaP studies with links to dbGaP.

example, "Alcohol - 30-Day Quantity and Frequency" (https://www.phenxtoolkit.org/protocols/view/30301). Navigating to the "Variable" tab shows available mappings to dbGaP variables listed for each PhenX variable in the measurement protocol (Fig. 2).

Scenario 2. A dbGaP user is interested in all study data sets collecting "Age when first smoke cigarettes." Using the dbGaP SOLR-faceted "Advanced Search" tool (https://www.ncbi.nlm.nih.gov/gap/advanced_search/), users can enter search keywords "age" AND "first" AND "smoke" into the search field. This returns 35 variables in four studies annotated with the "Common Data Elements" (Fig. 3). Clicking the "Variables" tab (to the right of the default "Studies" tab), reveals four pages of results, which the user can navigate between using the black arrows at the top of the search results page. This includes the variable TOBA2 on page 3 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/variable.cgi?study_id=phs000286.v6.p2&phv=128497).

Scenario 3. A user annotates CDE source during dbGaP data submission. In addition to the retrospective mapping of PhenX CDEs to dbGaP studies already included in the database, we modified the dbGaP study data submission process to capture PhenX CDE use for new studies being submitted to dbGaP. Two additional columns were added to the dbGaP data dictionary to record CDE linkage: VARIABLE_SOURCE and SOURCE_VARIABLE_ID. Using these columns, a researcher depositing data in dbGaP can indicate that a submitted variable used a specific PhenX protocol (e.g., VARIABLE_SOURCE = "PhenX" and SOURCE_VARIABLE_ID = "PX130301"). These columns are intended to capture mappings that are tagged at the "identical" mapping level, which is reserved specifically for the prospective use of PhenX measurement protocols in data collection.



Protocol - Alcohol - 30-Day Quantity and Frequency

Related Protocols:

- Alcohol - Maximum Drinks in 24 Hours
- Gender Identity
- Substance Abuse and Dependence - Past Year - Alcohol
- Substance Abuse and Dependence - Past Year - Drugs
- Substance Abuse and Dependence - Past Year - Tobacco

Essential Protocols:

- Alcohol - Lifetime Use
- Current Age

Protocol ID

30301

Variables

Variable Name	Variable ID	Variable Description	dbGaP Mapping
PX030301_Alcohol_30Day_Frequency			
	PX030301010000	Think specifically about the past 30 days, more	Variable Mapping
PX030301_Alcohol_30Day_Frequency_Coded			
	PX030301010100	Think specifically about the past 30 days, more	N/A
PX030301_Alcohol_30Day_Quantity			
	PX030301020000	On the days that you drank during the past more	Variable Mapping
PX030301_Alcohol_30Day_Quantity_Coded			
	PX030301020100	On the days that you drank during the past more	N/A

Fig. 2 PhenX protocol page contains mappings to dbGaP variables. At the PhenX protocol page, Alcohol - 30-Day Quantity and Frequency" (https://www.phenxtoolkit.org/protocols/view/30301), navigating to the "Variable" tab shows available mappings to dbGaP variables listed for each PhenX variable in the measurement protocol.

The other two mapping levels ("comparable" and "related") are for retrospective mapping of variables by the PhenX curation teams (Table 1).

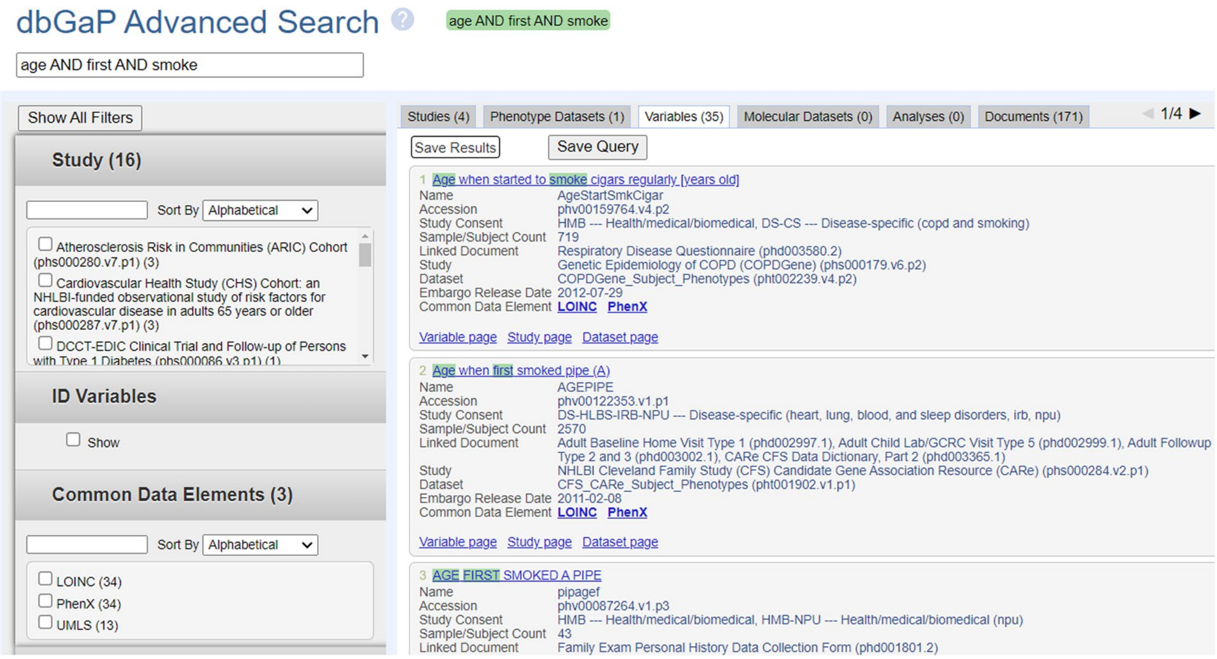
Discussion

Unique challenges exist in mapping PhenX variables to dbGaP study variables. The PhenX-dbGaP mapping compares the PhenX and dbGaP variable description fields, which are both stored as a string of unstructured terms, often covering multiple concepts. For example, the PhenX variable "In your entire life, have you had at least 1 drink of any kind of alcohol, not counting small tastes or sips?" (PX030101010000) is comparable to dbGaP variables such as "Has patient ever consumed at least one (alcoholic) drink/week for one year or more" (phv00034229) or "Was there ever a period when you drank alcoholic beverages regularly?" (phv00161469). A search using keywords representing a unique concept in the concise space of one or two words (e.g., diabetes or lung cancer) can yield good results, but using a lengthy string to impose specificity representing multiple concepts (alcohol, time frame, qualifiers) poses a challenge.

We tested available Natural Language Processing (NLP) tools during the mapping effort in 2016[40] and found that a common problem in the predicted result sets was the inclusion of many false positive results due to the unstructured multi-concept nature of dbGaP variable description. We determined that weeding out the false positives would be managed most effectively by a manual curation process. One possible method of applying NLP more effectively to map these data would be to generate enough relevant training data to improve performance sufficiently to reduce the amount of effort needed for manual curation.

Because manual curation is not scalable and because both the PhenX and dbGaP databases continue to grow, development of NLP tools that can be applied to the variable mapping effort is a compelling approach to consider. Recently, there have been significant advances in the development and application of NLP algorithms[41–48]. The performance of the NLP tools relies on the size, quality, and content coverage of a well-annotated and structured training set. The set of mapping results presented here provides an invaluable training and test dataset for the future application of an NLP tool to refine algorithms for recognizing multi-concept strings.

Variable mapping as annotation of CDEs increases findability in data repositories. There are many unidentified opportunities for secondary analysis of studies measuring phenotypic aspects of complex diseases in



dbGaP Advanced Search

age AND first AND smoke

age AND first AND smoke

Show All Filters

Study (16)

Sort By Alphabetical

□ Atherosclerosis Risk in Communities (ARIC) Cohort (phs000280.v7.p1) (3)

□ Cardiovascular Health Study (CHS) Cohort: an NHLBI-funded observational study of risk factors for cardiovascular disease in adults 65 years or older (phs000287.v7.p1) (3)

□ DCCT-EDIC Clinical Trial and Follow-up of Persons with Type 1 Diabetes (phs000086.v3.p1) (1)

ID Variables

□ Show

Common Data Elements (3)

Sort By Alphabetical

□ LOINC (34)
□ PhenX (34)
□ UMLS (13)

Studies (4) | Phenotype Datasets (1) | Variables (35) | Molecular Datasets (0) | Analyses (0) | Documents (171) | 1/4

Save Results | Save Query

1 Age when started to smoke cigars regularly [years old]
Name AgeStartSmkCigar
Accession phv00159764.v4.p2
Study Consent HMB --- Health/medical/biomedical, DS-CS --- Disease-specific (copd and smoking)
Sample/Subject Count 719
Linked Document Respiratory Disease Questionnaire (phd003580.2)
Study Genetic Epidemiology of COPD (COPDGene) (phs000179.v6.p2)
Dataset COPDGene_Subject_Phenotypes (pht002239.v4.p2)
Embargo Release Date 2012-07-29
Common Data Element LOINC PhenX

Variable page Study page Dataset page

2 Age when first smoked pipe (A)
Name AGEPIPE
Accession phv00122353.v1.p1
Study Consent DS-HLBS-IRB-NPU --- Disease-specific (heart, lung, blood, and sleep disorders, irb, npu)
Sample/Subject Count 2570
Linked Document Adult Baseline Home Visit Type 1 (phd002997.1), Adult Child Lab/GCRC Visit Type 5 (phd002999.1), Adult Followup Type 2 and 3 (phd003002.1), CARe CFS Data Dictionary, Part 2 (phd003365.1)
Study NHLBI Cleveland Family Study (CFS) Candidate Gene Association Resource (CARe) (phs000284.v2.p1)
Dataset CFS_CARe_Subject_Phenotypes (pht001902.v1.p1)
Embargo Release Date 2011-02-08
Common Data Element LOINC PhenX

Variable page Study page Dataset page

3 AGE FIRST SMOKED A PIPE
Name pipagef
Accession phv00087264.v1.p3
Study Consent HMB --- Health/medical/biomedical, HMB-NPU --- Health/medical/biomedical (npu)
Sample/Subject Count 43
Linked Document Family Exam Personal History Data Collection Form (phd001801.2)

Fig. 3 Find PhenX mappings in the dbGaP SOLR-faceted "Advanced Search" tool. dbGaP variables with PhenX mappings can be found in the SOLR-faceted "Advanced Search" tool, https://www.ncbi.nlm.nih.gov/gap/advanced_search/. Searching "age AND first AND smoke" returns 35 variables in the "Variables" tab where PhenX is listed under the "Common Data Elements" facet on the left.

Variables with different units that require a simple algorithmic transformation to convert	
Average number of packs smoked per day	How many cigarettes per day do/did you smoke?
Free-form numeric annual income number	Quantitative income category of "$35,000–$50,000" or "below poverty line"
Variables addressing similar concepts with different semantic terminology	
Have you had any of these clinician-diagnosed illnesses—Stroke?	Stroke ever diagnosed
	Subject's history of disease—stroke
	Hemorrhagic stroke
	Ischemic stroke
	CVA (Cerebrovascular accident)
	Cerebral stroke
	Cerebrovascular accident
	Cerebrovascular apoplexy
	Cerebrovascular stroke
	Brain Vascular Accident

Table 2. Examples of "comparable" mapping level with various scope and diversity.

publicly available data repositories. To take advantage of the large amount of data present and publicly available in repositories, NIH has been encouraging data reuse through multiple funding opportunities[49–52]. Previously, identifying datasets for inclusion in any secondary analysis, such as a meta-analysis, had been a laborious process that required hours of manual review to identify datasets with potential for harmonization due to heterogeneity of variable semantics and differences in data collection methods. The results reported here use PhenX variables as a set of CDEs to annotate data elements in dbGaP studies and, furthermore, provide an additional way of indexing dbGaP studies for more effective browsing and searching. Increasing the annotation of CDE data elements within the dbGaP repository will streamline the process of study identification and data harmonization. The linkage and annotation, together with the submission and search tools developed, addresses this barrier of missing linkage due to semantic variation and increases the searchability of CDEs in the dbGaP repository.

The dbGaP search at the variable level often returns a high proportion of false positive results (e.g., in matching a single keyword without the context, the chemical "lead" can't be distinguished from the "lead" in electrocardiograph) and/or a high proportion of false negative results by missing relevant results using alternative semantics (e.g., smoking vs cigarette). The newly developed CDE browsing facets (dbGaP) and search tool helps the user to retrieve studies with more specificity and sensitivity and decreases investigator burden by reducing the time required for manual review. dbGaP users can now identify relevant study datasets in dbGaP, which was not possible before. Additionally, when PhenX users browse the PhenX measurement protocols to decide whether to include them during study design, the availability of linked dbGaP studies enables them to identify publicly available datasets for future cross-study analysis. With insightful CDE inclusion during study design and CDE annotation in data repositories, individual studies can broaden their impact with lower costs through reuse of other study datasets in data repositories.

Rich metadata enhances interoperability. The FAIR principles advocate for standard metadata practices, such as including qualified references to other metadata, to increase interoperability[53]. The PhenX Toolkit is designed to improve consistency of data collections and, thus, supports data interoperability and reusability[53,54]. To date, the adoption of PhenX measurement protocols in prospective studies has been recommended in 523 NIH Funding Opportunity Announcements and 28 Notices. Using PhenX variables as CDEs to link retrospective studies in dbGaP, as reported in this paper, provides another important resource to identify studies with sufficient commonality to support cross-study analysis. Moving forward, to maximize the value of common measures used in various studies and to enhance interoperability, PhenX plans to expand this effort of annotating CDEs among studies to additional NIH Data Repositories and Data Commons that are willing to collaborate. Additionally, we propose to utilize Logical Observation Identifiers Names and Codes (LOINC) standards and adopt the FHIR (Fast Healthcare Interoperability Resources) specification to enhance PhenX measurement protocols' interoperability for integration into the NIH data ecosystem.

Methods

Mapping PhenX variables to dbGaP study variables. To help streamline the mapping process, we developed an R Shiny search tool to help organize and refine the dbGaP variable dataset for curation. The tool allowed the PhenX curation team to use Boolean logic (AND, OR, NOT) to combine search queries to scan variable descriptions for multiple concepts, and multiple Boolean statements could be combined to create more complex queries. This team of curators manually evaluated these initial keyword search results to decide whether each suggested pair of the PhenX-dbGaP variables were conceptually equivalent to qualify as "comparable" or "related" as a mapping. The determination was based upon a set of criteria measuring the potential for harmonization of the datasets collected, as described below and summarized in Table 1. Each set of potentially mapped PhenX-dbGaP variables was then reviewed by independent curators to ensure consistent and high-quality mapping. Any discrepancy was resolved in a group discussion. A "comparable" mapping level was defined as two variables that were conceptually similar and that contained data that can be directly harmonized or compared after a simple logical or mathematical transformation, including a categorical approximation or grouping, as listed in Table 2. A "related" mapping level was assigned to bioassay variables and other instances when the methods for data collection may be distinct. This distinction is to alert investigators that they should review the methods carefully before proceeding with transformation of the data for harmonization. For example, "Concentration of Immunoglobulin E" and "Immunoglobulin E (EDTA plasma)" bioassays often measure the same analyte, but the methodology may differ despite the dataset's appearing directly harmonizable in terms of measurements and units.

New dbGaP variable search tool released in PhenX Toolkit. To make PhenX-dbGaP mapping accessible, we developed a dbGaP Variable Search Tool at the PhenX Toolkit website (https://www.phenxtoolkit.org/vsearch). We loaded the mapping results into the PhenX Toolkit database and used PHP and JavaScript to enable the web interface. To find PhenX-dbGaP variable mappings, searches can be run using keywords, PhenX measurement protocol ID, PhenX variable ID, or dbGaP variable ID.

New search tools and features released in dbGaP. PhenX-dbGaP variable mappings are loaded by dbGaP and accessed through a SOLR-based faceted search (Fig. 3.) To facilitate searching of PhenX-dbGaP variable linkages, we additionally indexed PhenX variable ID, name, and descriptions for dbGaP studies and variables. With this development, searches can be executed based on PhenX properties (e.g., PhenX variable name, accession, keywords) or dbGaP properties (e.g., dbGaP variable name, accession, description). Although we implemented this strategy for PhenX variables specifically, we adopted it more generally to embrace CDE linkages from multiple initiatives such as Logical Observation Identifiers Names and Codes (LOINC) and Unified Medical Language System (UMLS). The CDE mappings that enable searching are also represented as links displayed on each dbGaP variable page. These CDE links can be authored by either the data submitter via a data dictionary or another party doing a retrospective mapping (e.g., the PhenX team).

In addition to facilitating direct CDE (e.g., PhenX) to dbGaP variable mapping, we modified dbGaP's study data submission process to include two additional columns to dbGaP's data dictionary: VARIABLE_SOURCE and SOURCE_VARIABLE_ID. Using these columns, a researcher depositing data in dbGaP can indicate that a submitted variable is mapped to a specific PhenX measurement protocol, such as VARIABLE_SOURCE = "PhenX" and SOURCE_VARIABLE_ID = "PX130301." These columns are intended to capture mappings that are tagged at the "identical" mapping level, which is reserved specifically for the prospective, submitter-identified use of PhenX measurement protocols in data collection. The other two mapping levels (comparable and related) are reserved for retrospective mapping of variables.

In summary, the use of CDEs can help identify comparable study datasets from multiple data sources and repositories, enable cross-study analysis, and reduce the burden of retrospective data harmonization. We presented an approach of using PhenX measurement protocols and variables to identify CDEs linking across the study-specific variables in dbGaP studies. Both dbGaP and PhenX have developed browse and search tools to access variables and studies linked by the mappings. Users of dbGaP and PhenX can browse and search variables of interest to find other studies that have collected data using the same variable. This development adds features in dbGaP and PhenX that help investigators identify opportunities for cross-study analysis and maximize research benefits beyond the original objective of a single study.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Code availability

Code sharing is not applicable to this article as the development distributed linked metadata (dbGaP variable ID linked to PhenX variable ID) in existing databases and features from PhenX and dbGaP.

Received: 8 September 2021; Accepted: 23 August 2022;
Published online: 01 September 2022

References

1. Nagel, M. *et al*. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet* **50**, 920–927, https://doi.org/10.1038/s41588-018-0151-7 (2018).
2. Popovic, M. *et al*. Genome-wide meta-analysis identifies novel loci associated with free triiodothyronine and thyroid-stimulating hormone. *J Endocrinol Invest* **42**, 1171–1180, https://doi.org/10.1007/s40618-019-01030-9 (2019).
3. Nalls, M. A. *et al*. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology* **18**, 1091–1102, https://doi.org/10.1016/s1474-4422(19)30320-5 (2019).
4. Winkler, T. W. *et al*. Genome-wide association meta-analysis for early age-related macular degeneration highlights novel loci and insights for advanced disease. *BMC Med Genomics* **13**, 120, https://doi.org/10.1186/s12920-020-00760-7 (2020).
5. Li, M. *et al*. Genome-wide meta-analysis identifies three novel susceptibility LOCI and reveals ethnic heterogeneity of genetic susceptibility for iga nephropathy. *J Am Soc Nephrol* **31**, 2949–2963, https://doi.org/10.1681/ASN.2019080799 (2020).
6. Kunkle, B. W. *et al*. Novel alzheimer disease risk loci and pathways in African American individuals using the african genome resources panel: A meta-analysis. *JAMA Neurol* **78**, 102–113, https://doi.org/10.1001/jamaneurol.2020.3536 (2021).
7. Di Narzo, A. *et al*. Meta-analysis of sample-level dbGaP data reveals novel shared genetic link between body height and Crohn's disease. *Hum Genet* **140**, 865–877, https://doi.org/10.1007/s00439-020-02250-3 (2021).
8. National Institutes of Health, Office of Data Science Strategy. *NIH strategic plan for data science*, https://datascience.nih.gov/strategicplan (2018).
9. Warzel, D. B. *et al*. Common data element (CDE) management and deployment in clinical trials. *AMIA Annu Symp Proc*, 1048 (2003).
10. Loring, D. W. *et al*. Common data elements in epilepsy research: development and implementation of the NINDS epilepsy CDE project. *Epilepsia* **52**, 1186–1191, https://doi.org/10.1111/j.1528-1167.2011.03018.x (2011).
11. Lawlor, M. W. *et al*. NINDS common data elements for congenital muscular dystrophy clinical research: A national institute for neurological disorders and stroke project. *J Neuromuscul Dis* **5**, 75–84, https://doi.org/10.3233/JND-170248 (2018).
12. National Institutes of Health, National Library of Medicine. *CDE repository*, https://cde.nlm.nih.gov/ (n.d.).
13. Voight, B. F. *et al*. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**, 579–589, https://doi.org/10.1038/ng.609 (2010).
14. Griffith, L. E. *et al*. Comparison of standardization methods for the harmonization of phenotype data: An application to cognitive measures. *Am J Epidemiol* **184**, 770–778, https://doi.org/10.1093/aje/kww098 (2016).
15. Spjuth, O. *et al*. Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *Eur J Hum Genet* **24**, 521–528, https://doi.org/10.1038/ejhg.2015.165 (2016).
16. Fortier, I. *et al*. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol* **46**, 103–105, https://doi.org/10.1093/ije/dyw075 (2017).
17. Johnson, S. B., Butow, P. N., Kerridge, I., Bell, M. L. & Tattersall, M. H. N. How well do current measures assess the impact of advance care planning on concordance between patient preferences for end-of-life care and the care received: A methodological review. *J Pain Symptom Manage* **55**, 480–495, https://doi.org/10.1016/j.jpainsymman.2017.09.008 (2018).
18. Bennett, S. N. *et al*. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genet Epidemiol* **35**, 159–173, https://doi.org/10.1002/gepi.20564 (2011).
19. Budin-Ljosne, I. *et al*. Data sharing in large research consortia: experiences and recommendations from ENGAGE. *Eur J Hum Genet* **22**, 317–321, https://doi.org/10.1038/ejhg.2013.131 (2014).
20. Yang, L., Chen, Y., Yu, C. & Shen, B. Biobanks and their clinical application and informatics challenges. *Adv Exp Med Biol* **939**, 241–257, https://doi.org/10.1007/978-981-10-1503-8_10 (2016).
21. Sollini, M., Cozzi, L., Antunovic, L., Chiti, A. & Kirienko, M. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. *Sci Rep* **7**, 358, https://doi.org/10.1038/s41598-017-00426-y (2017).
22. Basu, A. *et al*. Call for data standardization: Lessons learned and recommendations in an imaging study. *JCO Clin Cancer Inform* **3**, 1–11, https://doi.org/10.1200/CCI.19.00056 (2019).
23. Jovicich, J. *et al*. Harmonization of neuroimaging biomarkers for neurodegenerative diseases: A survey in the imaging community of perceived barriers and suggested actions. *Alzheimers Dement (Amst)* **11**, 69–73, https://doi.org/10.1016/j.dadm.2018.11.005 (2019).
24. Tratwal, J. *et al*. Reporting guidelines, review of methodological standards, and challenges toward harmonization in bone marrow adiposity research. Report of the methodologies Working Group of the International Bone Marrow Adiposity Society. *Front Endocrinol (Lausanne)* **11**, 65, https://doi.org/10.3389/fendo.2020.00065 (2020).
25. Simko, L. C. *et al*. Challenges to the standardization of trauma data collection in burn, traumatic brain injury, spinal cord injury, and other trauma populations: A call for common data elements for acute and longitudinal trauma databases. *Arch Phys Med Rehabil* **100**, 891–898, https://doi.org/10.1016/j.apmr.2018.10.004 (2019).
26. Meeuws, S. *et al*. Common data elements: Critical assessment of harmonization between current multi-center traumatic brain injury studies. *J Neurotrauma* **37**, 1283–1290, https://doi.org/10.1089/neu.2019.6867 (2020).
27. Hamilton, C. M. *et al*. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol* **174**, 253–260, https://doi.org/10.1093/aje/kwr193 (2011).
28. Stover, P. J., Harlan, W. R., Hammond, J. A., Hendershot, T. & Hamilton, C. M. PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol* **21**, 136–140, https://doi.org/10.1097/MOL.0b013e3283377395 (2010).
29. Maiese, D. R., Hendershot, T. P. & Strader, L. C. PhenX: Establishing a consensus process to select common measures for collaborative research. RTI Press publication no. MR-0027-1310. https://doi.org/10.3768/rtipress.2013.mr.0027.1310 (2013).
30. McCarty, C. A. *et al*. PhenX RISING: real world implementation and sharing of PhenX measures. *BMC Med Genomics* **7**, 16, https://doi.org/10.1186/1755-8794-7-16 (2014).

31. Hendershot, T. *et al*. Using the PhenX toolkit to add standard measures to a study. *Curr Protoc Hum Genet* **86**, 1 21 21–21 21 17, https://doi.org/10.1002/0471142905.hg0121s86 (2015).
32. Conway, K. P. *et al*. Data compatibility in the addiction sciences: an examination of measure commonality. *Drug Alcohol Depend* **141**, 153–158, https://doi.org/10.1016/j.drugalcdep.2014.04.029 (2014).
33. Barch, D. M. *et al*. Common measures for national institute of mental health funded research. *Biol Psychiatry* **79**, e91–96, https://doi.org/10.1016/j.biopsych.2015.07.006 (2016).
34. Eckman, J. R. *et al*. Standard measures for sickle cell disease research: the PhenX Toolkit sickle cell disease collections. *Blood Adv* **1**, 2703–2711, https://doi.org/10.1182/bloodadvances.2017010702 (2017).
35. Garcia-Cazarin, M. L., Mandal, R. J., Grana, R., Wanke, K. L. & Meissner, H. I. Host-agent-vector-environment measures for electronic cigarette research used in NIH grants. *Tob Control* **29**, s43–s49, https://doi.org/10.1136/tobaccocontrol-2017-054032 (2020).
36. Kaufman, A. R., Persoskie, A., Twesten, J. & Bromberg, J. A review of risk perception measurement in tobacco control research. *Tob Control* **29**, s50–s58, https://doi.org/10.1136/tobaccocontrol-2017-054005 (2020).
37. Mailman, M. D. *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39**, 1181–1186, https://doi.org/10.1038/ng1007-1181 (2007).
38. Tryka, K. A. *et al*. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res* **42**, D975–979, https://doi.org/10.1093/nar/gkt1211 (2014).
39. Paltoo, D. N. *et al*. Data use under the NIH GWAS data sharing policy and future directions. *Nat Genet* **46**, 934–938, https://doi.org/10.1038/ng.3062 (2014).
40. Doan, S. *et al*. PhenDisco: phenotype discovery system for the database of genotypes and phenotypes. *J Am Med Inform Assoc* **21**, 31–36, https://doi.org/10.1136/amiajnl-2013-001882 (2014).
41. Velupillai, S., Mowery, D., South, B. R., Kvist, M. & Dalianis, H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform* **10**, 183–193, https://doi.org/10.15265/IY-2015-009 (2015).
42. Neveol, A. & Zweigenbaum, P. Making sense of big textual data for health care: Findings from the section on clinical natural language processing. *Yearb Med Inform* **26**, 228–234, https://doi.org/10.15265/IY-2017-027 (2017).
43. Kreimeyer, K. *et al*. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform* **73**, 14–29, https://doi.org/10.1016/j.jbi.2017.07.012 (2017).
44. Jovanovic, J. & Bagheri, E. Semantic annotation in biomedicine: the current landscape. *J Biomed Semantics* **8**, 44, https://doi.org/10.1186/s13326-017-0153-x (2017).
45. Wang, Y. *et al*. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* **19**, 1, https://doi.org/10.1186/s12911-018-0723-6 (2019).
46. Sheikhalishahi, S. *et al*. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med Inform* **7**, e12239, https://doi.org/10.2196/12239 (2019).
47. Koleck, T. A., Dreisbach, C., Bourne, P. E. & Bakken, S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* **26**, 364–379, https://doi.org/10.1093/jamia/ocy173 (2019).
48. Wang, J. *et al*. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on Pubmed. *J Med Internet Res* **22**, e16816, https://doi.org/10.2196/16816 (2020).
49. National Institutes of Health (NIH). *Secondary analyses and archiving of social and behavioral datasets in aging (R03). Funding Opportunity Announcement (FOA) Number RFA-AG-12-005*, https://grants.nih.gov/grants/guide/rfa-files/rfa-ag-12-005.html (2011).
50. National Institutes of Health (NIH), U.S. Food and Drug Administration (FDA). *Secondary analyses of existing datasets of tobacco use and health (R21 Clinical trial not allowed). Funding Opportunity Announcement (FOA) Number RFA-OD-21-003*, https://grants.nih.gov/grants/guide/rfa-files/RFA-OD-21-003.html (2021).
51. National Institutes of Health (NIH). *Secondary analysis and integration of existing data to elucidate the genetic architecture of cancer risk and related outcomes (R01 clinical trial not allowed). Funding Opportunity Announcement (FOA) Number PAR-20-276*, https://grants.nih.gov/grants/guide/pa-files/PAR-20-276.html (2020).
52. National Institutes of Health (NIH). Secondary analysis of existing datasets in heart, lung, and blood diseases and sleep disorders (R21 Clinical Trial not allowed). Funding Opportunity Announcement (FOA) Number PAR-20-078 (2019).
53. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
54. Wilkinson, M. D. *et al*. Addendum: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **6**, 6, https://doi.org/10.1038/s41597-019-0009-6 (2019).

Acknowledgements

The authors acknowledge the contributions of the PhenX Steering Committee, Working Groups, liaisons from the National Institutes of Health, and RTI PhenX Team members. The dbGaP part of this research is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

C.M., H.P. (RTI), E.R. (NHGRI), and M.K., M.F. (dbGaP) developed the idea of mapping PhenX to dbGaP. H.P., J.L., S.H. (RTI) led the mapping team at multiple phases respectively; V.B., L.C., S.E., Y.G., W.H., S.H., M.K., M.P. (RTI) performed the mapping curation and quality check; Y.Q., S.H., D.W. (RTI) performed feature development and mapping data loading to PhenX; M.F., M.K. (dbGaP) performed feature development and mapping data loading to dbGaP, M.E., C.M. (RTI) and E.R. (NHGRI) performed critical review and revision to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.