Phenotype Discovery in NHLBI Genomic Studies

Final Report

Hyeoneui Kim, RN, PhD
Son Doan, PhD
Ko-Wei Lin, DVM, PhD
Michael Conway, PhD
Alexander Hsieh
Asher Garland
Seena Farzaneh
Neda Alipanah
Stephanie Fedujio Feupe
Hua Xu
Mindy Ross
Dexter Friedman
Rebecca Walker

and
Lucila Ohno-Machado, MD, PhD

1 Project Overview

We developed PhenDisco (Phenotype Discoverer) to facilitate phenotypic data search in dbGaP. The main functionalities that PhenDisco supports are summarized in **Table 1**. Note that these functionalities were identified and prioritized based on the user requirement analyses as well as the recommendations provided by the NIH program officers and Scientific Advisory Board (SAB).

Table 1. Main functionalities of PhenDisco

ID	Type	User/Functional Requirements	Source
USE CASE 1. Search Studies			
QI1	Query Input	Frequently used search terms are auto-completed as users type them in	User Group
QI2	Query Input	Spelling errors in frequently used search terms are auto-corrected as users type them in	User Group
QI4	Query Input	Search fields in the advanced search menu are presented with easier to understand names	User Group
QI5	Query Input	Additional information (e.g., definition) on the search fields in the advanced search menu is available	User Group
QI6	Query Input	Search fields in the advanced search menu are structured based on semantics not on alphabetic order	User Group
QI7	Query Input	Users can choose an option to expand search terms through synonyms with free text search menu	User Group
QI9	Query Input	Users can search for very specific populations groups (e.g., Amish, Gambian, African, etc), which are rolled up to more general race and ethnicity categories	SAB, NIH Officers
USE CASE 2. Retrieve Studies			
QO1	Query Output	Retrieved studies are displayed in the order of relevancy	User Group
QO2	Query Output	Users can see succinct metadata of each retrieved study in the result screen	User Group SAB
QO3	Query Output	The keywords or other terms in the study information or phenotype variables which are relevant to the search are highlighted	User Group
QO5	Query Output	The search output is organized in a way that supports quick browsing	NIH Officers
QO6	Query Output	If returned study is a sub study of a bigger study, then the search output shows the title of the bigger study as well	NIH Officers
QO7	Query Output	Users can sort the returned studies using various parameters (i.e., study level metadata)	User Group NIH Officers
QO8	Query Output	Users can select studies from the returned list and save for later review (i.e., PubMed style)	User Group NIH Officers
QO9	Query Output	Users can select study level metadata items to display in the output screen	User Group
QO10	Query Output	PhenDisco shows other relevant or similar studies (like Amazon)	User Group
QO11	Query Output	Retrieved studies are displayed in the order of relevancy	User Group

2 Core Development Activities and Achievements

2.1 Requirement Analysis

We interviewed 8 users to identify user requirements for PhenDisco. Four of them (A Zambon, N Heintzman, J Kim, C Woelk) were researchers at UCSD, whose research focuses were on genomics, bioinformatics, biomedical informatics, pharmacogenomics and proteomics. In addition, as suggested by SAB, we downloaded the list of the researchers who have requested the dbGaP data and recruited 4 additional users (E Smith, Zhao, O Harismandy, LM Meneades) who have interacted with dbGaP. Our initial goal was to interview 10 users. However, we observed that the input from the users converged into the common themes when we reached 6 interviews. Therefore, for this phase, after verifying the common themes with 2 additional interviews, we stopped recruiting additional interviewees. With the help from the users, we developed five search scenarios to guide the PhenDisco development process. An example search scenario narrative is presented in **Table 2**.

Table 2 an Example Search Scenario

Goal: To investigate the risk factors of stroke in the African American population	
Narratives	Research suggests that African Americans have higher mortality from stroke compared to Caucasians; in fact, they appear to be twice as likely. In addition, stroke tends to occur earlier in life and survivors are more likely to be disabled or have limitations to their daily functioning and activity. African American women especially have a lower survival rate if the stroke is ischemic (caused by a blood clot). The explanations for this are not clear; however, it is known that hypertension is very common in the African American population and 1 in 3 are affected. Other risk factors noted are sickle cell anemia, diabetes, obesity, and smoking. It has been documented that African Americans are less likely to receive tissue plasminogen activator (tPA), which is an approved treatment for stroke when compared to the white population, so socioeconomic and social factors may come into play. This is a topic that lends itself to exploration of genetic studies to understand if there are any underlying clues to the etiology of this phenomenon.

2.2 Variable Standardization

We have successfully applied the information model based variable standardization to demographic variables and other phenotypic variables. The PhenDisco standardization pipeline is comprised of preprocessing, meaningful concept identification and mapping, semantic role assignment, variable categorization, and same variable identification (**Figure 1**). The last 3 steps were developed based on heuristic rules.

We also produced standardized metadata about the studies stored in dbGaP via manual abstraction. Main disease topics were encoded with the UMLS (Unified Medical Language System) Metathesaurus, geographic information (i.e., study location) was encoded with ISO 3166-2 subdivision code, which encompasses state and country information. IRB approval requirement, and data use consent types were also added. There are several metadata items readily available in dbGaP such as study types, platform information, sample size, and sample demographics. We also incorporated these items in PhenDisco after further standardizing their values.

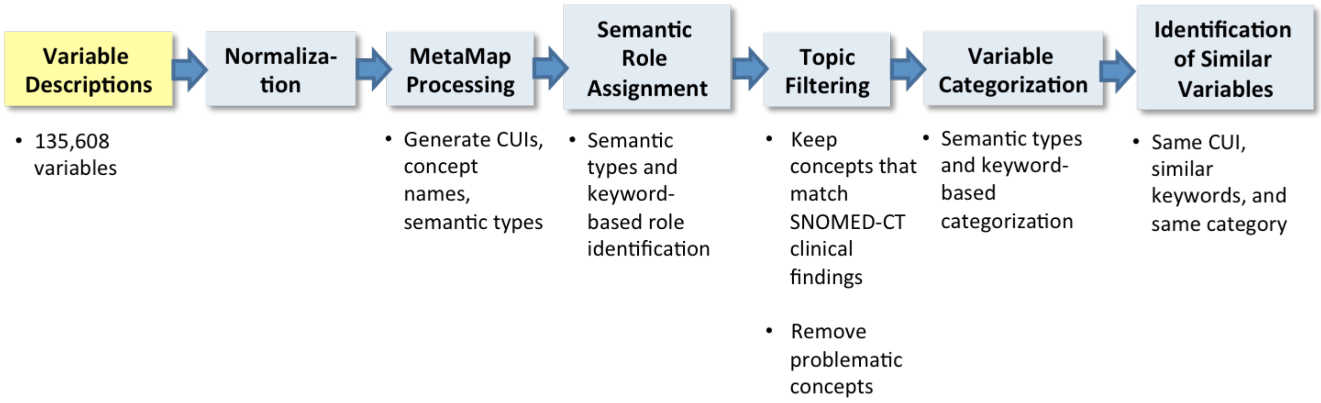

Variable Descriptions
Normaliza-tion
MetaMap Processing
Semantic Role Assignment
Topic Filtering
Variable Categorization
Identification of Similar Variables
• 135,608 variables
• Generate CUIs, concept names, semantic types
• Semantic types and keyword-based role identification
• Keep concepts that match SNOMED-CT clinical findings
• Remove problematic concepts
• Semantic types and keyword-based categorization
• Same CUI, similar keywords, and same category

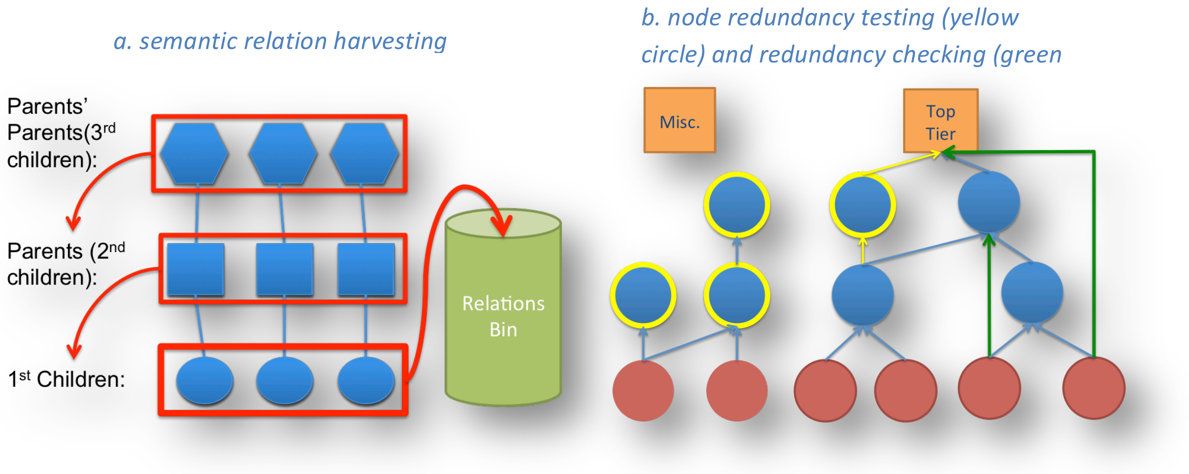
Figure 1 PhenDisco standardization pipeline

We conducted a large-scale evaluation of the performance of this pipeline. Five human reviewers reviewed the pipeline outputs of 2000 phenotypic variables of dbGaP and assessed the accuracy in concept identification and mapping, semantic role assignment, variable categorization, and similar variable identification. The full results will become available though a publication, which we are currently preparing. We noted that the MetaMap processing step still remained as a challenge, causing many incorrect concept identification and mapping which had cascade effects on the subsequent steps of categorization and similar variable identification. In addition, the PhenDisco standardization pipeline showed relatively poor performance with the lifestyle and environment related variables in general, indicating the need of new approach to process these types of variables.

2.3 Ontology Development

In order to formalize the semantic relations among phenotype variables, we created an ontology derived from standardized phenotype variables. We first processed the entire 130 thousands phenotype variables with the standardization pipeline, which identifies topic and subject of information terms and mapped them to the concepts in UMLS Metathesaurus. This process resulted in 5,096 unique CUIs (Concept Unique Identifiers). We also developed ontology building algorithms that performs following tasks with the unique concepts prepared from the PhenDisco standardization pipeline.

- Semantic relations harvesting: gathers concepts that are in the parent-children relationships with a given concept (Figure 2.a)
- Node legitimacy testing: removes concept the concept node that either has less than 2 children or is not directly related to a top tier concept (Figure 2.b)
- Redundancy elimination: removes redundant hierarchical relations



a. semantic relation harvesting

b. node redundancy testing (yellow circle) and redundancy checking (green

Figure 2 Ontology building

We tested our ontology building algorithms with 100 unique concepts as input. Our algorithms generated a concept hierarchy with 190 subsumptive relations. Upon manual review, we discovered 10 problematic relations. Nine of them were caused by the erroneous semantic relations in UMLS. The remaining one case was not exactly incorrect but the two concepts were deemed too distant. This was caused by the node legitimacy testing.

We were unable to build the complete concept hierarchy for PhenDisco. However, we developed a concept hierarchy with 2000 topic concepts, a metadata item defined at the study level. This hierarchy was used to implement search expansion function.

2.4 Ranking Algorithms

We explored the feasibility of building supervised ranking algorithms based on the search results generated for 124 search cases ranked by 3 human experts. The number of studies retrieved and ranked per each case varied from 3 to 140. On average, the human experts reviewed and ranked 30 studies per each case. A larger training data was required but expert ranking data was time consuming and tedious to generate. In addition, the wide variety in phenotype variables used warrants additional training by our algorithms to ensure accuracy in ranking as new studies are added to the system. In addition, human experts might have disagreement on ranking results. The models learned from divergent opinions can have wide confidence intervals and thus may be less reliable. Therefore, we implemented BM25F algorithm, which is an unsupervised model based on a variation of term-frequency, inversed frequency (TF-IDF) for ranking in PhenDisco.

2.5 User Interface Development

PhenDisco interface was designed based on the user requirement analysis. We intentionally kept the overall style of the site to be concordant with dbGaP to avoid confusing the existing dbGaP users and take advantage of their familiarity with that interface. Also, we replicated some of the original dbGaP designs which were well received by the users in PhenDisco. We completed the homepage screen with a basic search function; results display page with many added functions, and the advanced search page that supports a more precise search through a structured query menu.

PhenDisco supported basic search and advanced search function Some highlights of the core functionalities are listed below:

- **Auto-complete**: as users type in a search term, PhenDisco will show potential matching terms based on the term list we compiled from dbGaP and the GWAS catalogue (4) (Figure 3).

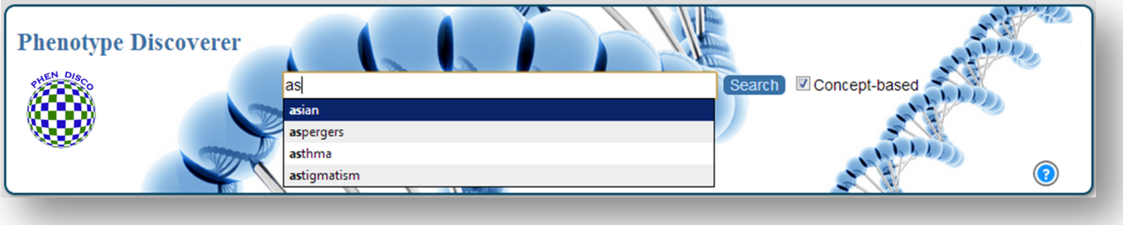
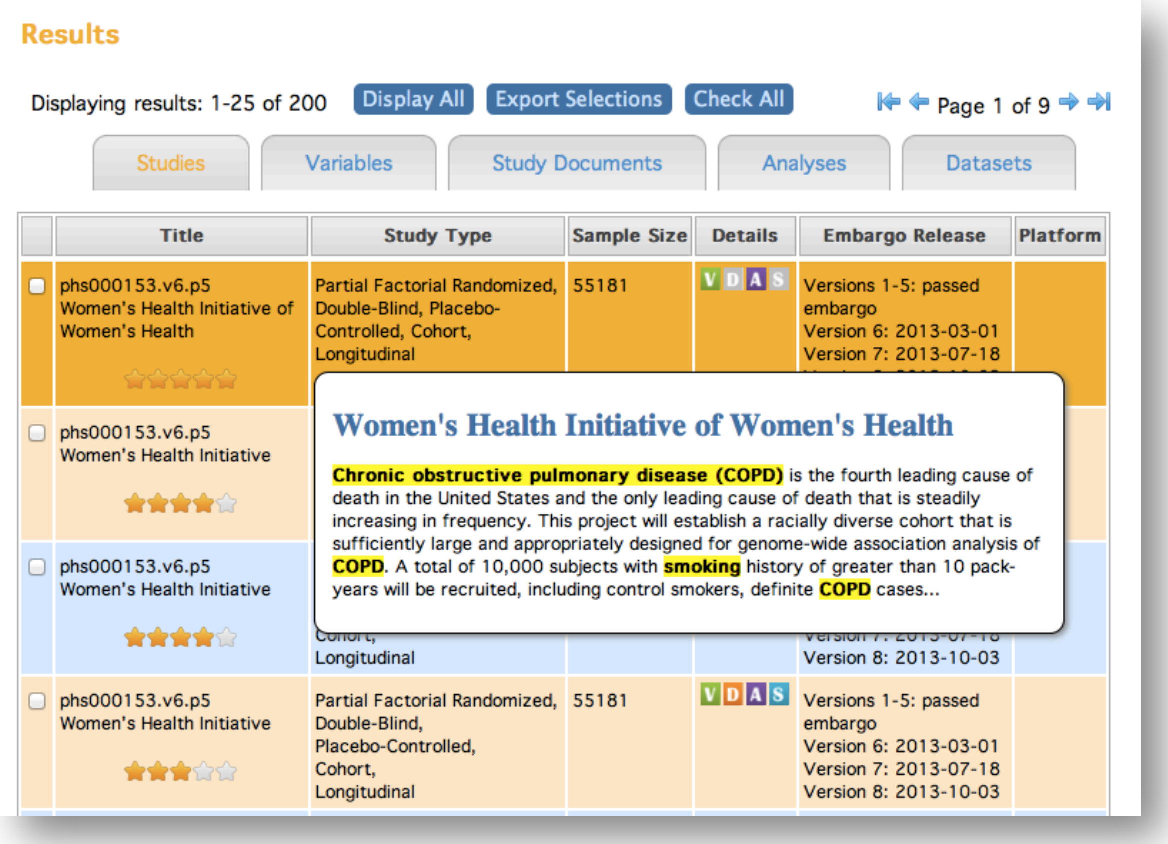


Figure 3 Auto-complete function

- **Search keywords highlighting**: to make it easier for users to decide the relevancy of the retrieved studies, PhenDisco highlights the relevant keywords in the study descriptions. This function will be applied to variable descriptions in the future release (Figure 4).



	Title	Study Type	Sample Size	Details	Embargo Release	Platform
□	phs000153.v6.p5 Women's Health Initiative of Women's Health ★★★★★	Partial Factorial Randomized, Double-Blind, Placebo-Controlled, Cohort, Longitudinal	55181	V D A S	Versions 1-5: passed embargo Version 6: 2013-03-01 Version 7: 2013-07-18	
□	phs000153.v6.p5 Women's Health Initiative ★★★★☆					
□	phs000153.v6.p5 Women's Health Initiative ★★★★☆	Cohort, Longitudinal			Version 7: 2013-07-18 Version 8: 2013-10-03	
□	phs000153.v6.p5 Women's Health Initiative ★★★☆☆	Partial Factorial Randomized, Double-Blind, Placebo-Controlled, Cohort, Longitudinal	55181	V D A S	Versions 1-5: passed embargo Version 6: 2013-03-01 Version 7: 2013-07-18 Version 8: 2013-10-03	

Women's Health Initiative of Women's Health

Chronic obstructive pulmonary disease (COPD) is the fourth leading cause of death in the United States and the only leading cause of death that is steadily increasing in frequency. This project will establish a racially diverse cohort that is sufficiently large and appropriately designed for genome-wide association analysis of COPD. A total of 10,000 subjects with smoking history of greater than 10 pack-years will be recruited, including control smokers, definite COPD cases...

Figure 4 Highlighting search keywords

-

- **Concept-based search**: by default, PhenDisco performs *concept-based* search, meaning that PhenDisco expands search terms to synonyms based on the concept mapping to the UMLS Metathesaurus. The concept based search will also incorporate hierarchical expansion once the dbGaP concept ontology is deployed. With hierarchical expansion, PhenDisco will retrieve the studies that contain more specific terms than the search terms, which will improve search recall. For example, with the search term "cardiovascular disease", PhenDisco will retrieve the studies containing "hypertension", "atherosclerosis", "congestive heart failure", and so on (Figure 5).

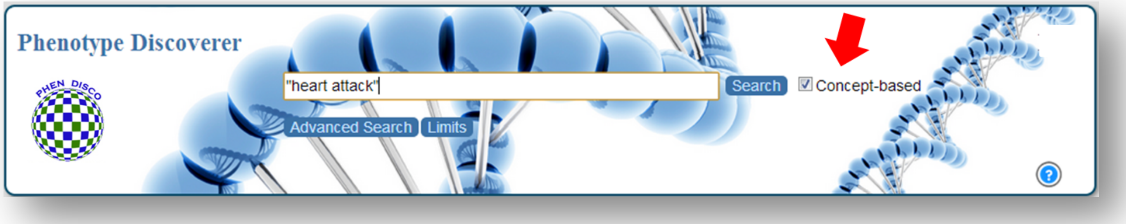
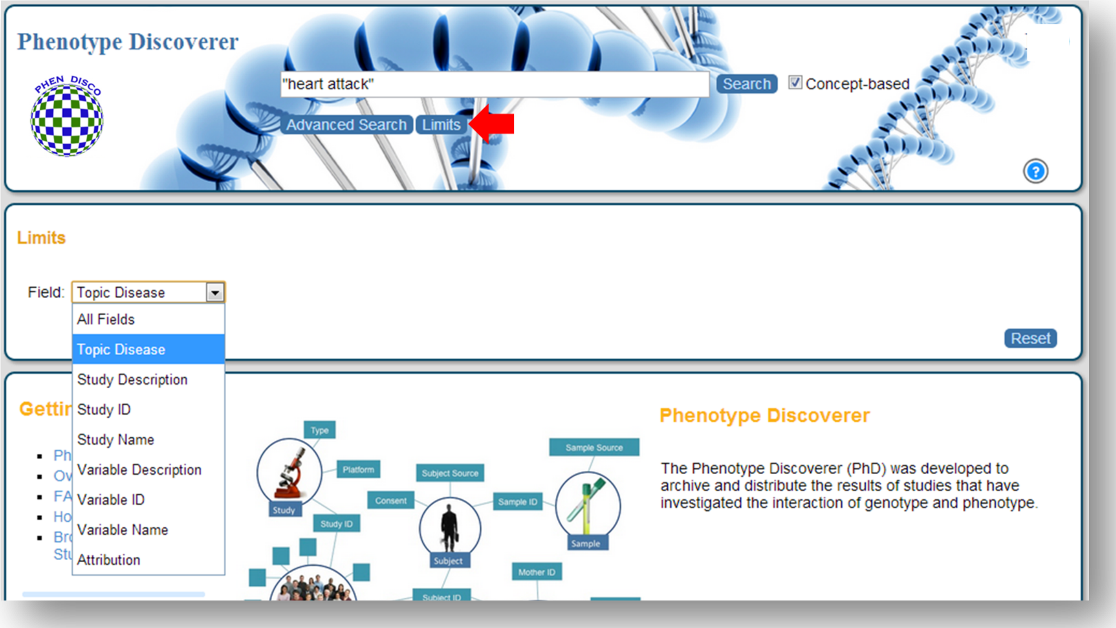


Figure 5 Concept-based search option

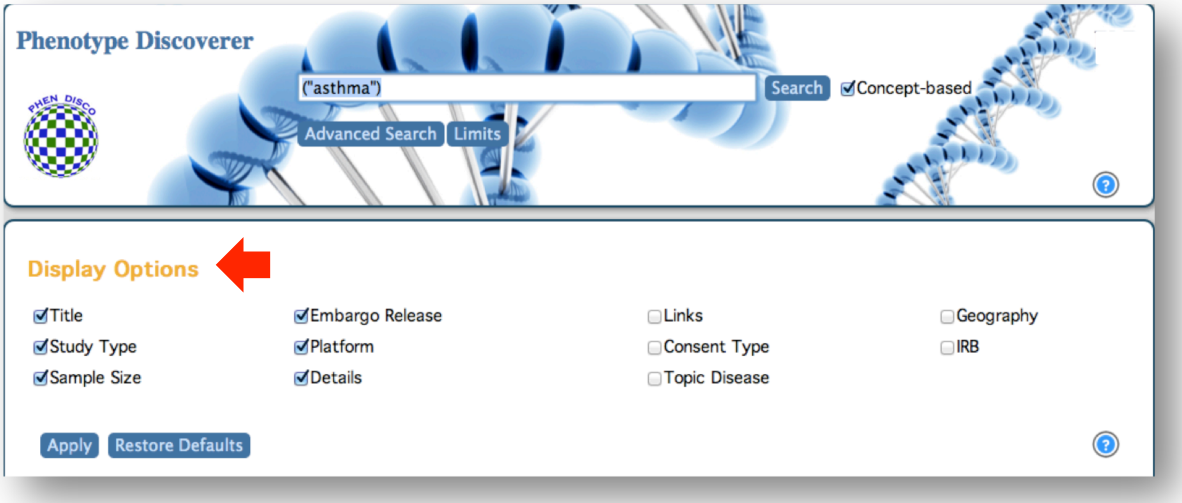
- **Focused search (limiting search space)**: The "limit" option of the dbGaP search menu is being replicated in PhenDisco. Currently, this focused search is supported for Topic Disease, Study ID, Study Name, Study Description, Variable ID, Variable Name, Variable Description, and Attribution. We will implement the full limit function in the next release of PhenDisco. We understand that dbGaP is currently revising the limit function by adding or removing certain limit options. We will keep checking dbGaP for changes, since we understand that our communication with the dbGaP may not be allowed at this point (Figure 6).



The Phenotype Discoverer (PhD) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.

Figure 6 Focused search using "Limits" option

- **Customization of the result display**: current PhenDisco offers a list of study level metadata items as displayable information. Users can select study level metadata items to display using the display option menu in the result screen (Figure 7).

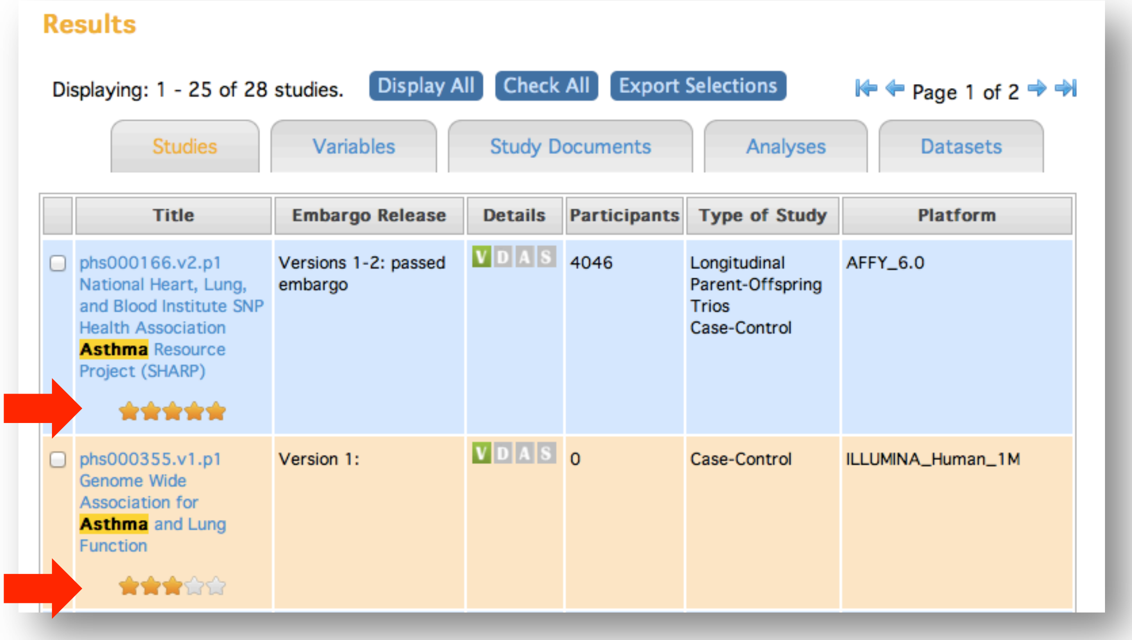


Display Options

☑Title	☑Embargo Release	□Links	□Geography
☑Study Type	☑Platform	□Consent Type	□IRB
☑Sample Size	☑Details	□Topic Disease	

Figure 7 Result display options

- **Ranking by relevancy**: current PhenDisco sorts the search results by the relevancy determined using the BM25F algorithm. The custom built ranking algorithms are currently under evaluation and will replace the BM25F algorithms in the next release of PhenDisco (Figure 8).



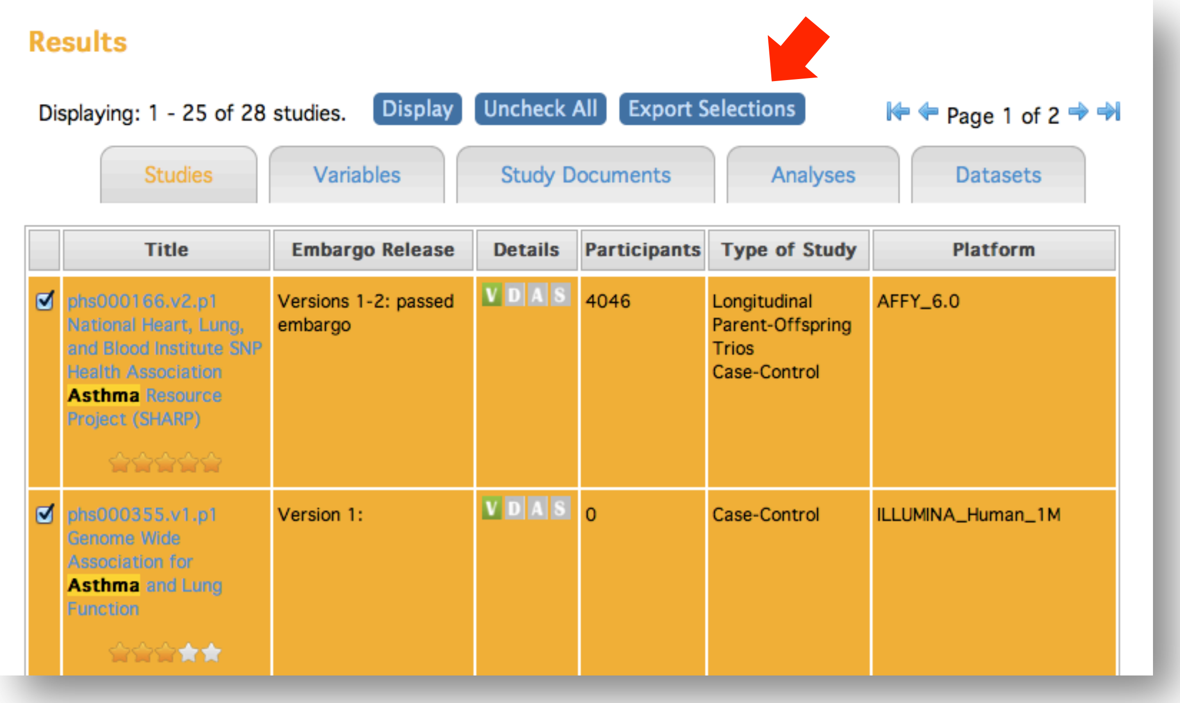
Results

Displaying: 1 - 25 of 28 studies.

	Title	Embargo Release	Details	Participants	Type of Study	Platform
□	phs000166.v2.p1 National Heart, Lung, and Blood Institute SNP Health Association **Asthma** Resource Project (SHARP) ★★★★★	Versions 1-2: passed embargo	V D A S	4046	Longitudinal Parent-Offspring Trios Case-Control	AFFY_6.0
□	phs000355.v1.p1 Genome Wide Association for **Asthma** and Lung Function ★★★☆☆	Version 1:	V D A S	0	Case-Control	ILLUMINA_Human_1M

Figure 8 Search results ranked by relevancy

- **Exporting the returned search results:** Users can export all or subset of the returned studies with their metadata to a CSV files (Figure 9).



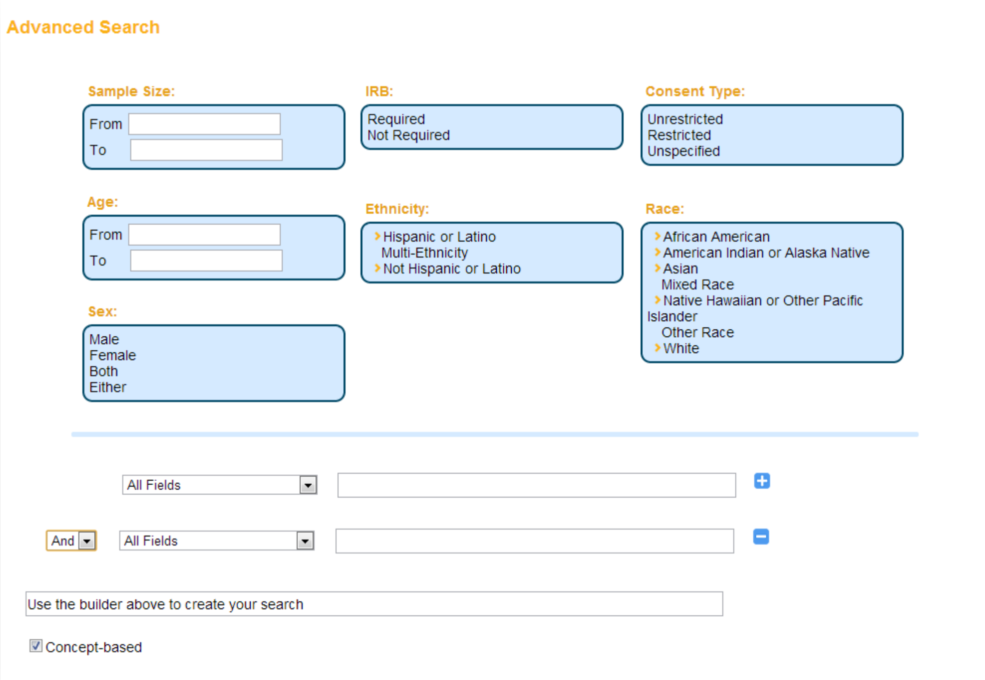
Results

Displaying: 1 - 25 of 28 studies.

	Title	Embargo Release	Details	Participants	Type of Study	Platform
☑	phs000166.v2.p1 National Heart, Lung, and Blood Institute SNP Health Association **Asthma** Resource Project (SHARP)	Versions 1-2: passed embargo	V D A S	4046	Longitudinal Parent-Offspring Trios Case-Control	AFFY_6.0
☑	phs000355.v1.p1 Genome Wide Association for **Asthma** and Lung Function	Version 1:	V D A S	0	Case-Control	ILLUMINA_Human_1M

Figure 9 Exporting selected records

- **Advanced search**: users can perform more precise search using the structured search interface. (Figure 10).



Advanced Search

Sample Size:
From
To

IRB:
Required
Not Required

Consent Type:
Unrestricted
Restricted
Unspecified

Age:
From
To

Ethnicity:
Hispanic or Latino
Multi-Ethnicity
Not Hispanic or Latino

Race:
African American
American Indian or Alaska Native
Asian
Mixed Race
Native Hawaiian or Other Pacific Islander
Other Race
White

Sex:
Male
Female
Both
Either

All Fields

And All Fields

Use the builder above to create your search

☑ Concept-based

Figure 10 Advanced search screen

3 PhenDisco

The entire source codes and the data generated from this project are available in GitHub (https://github.com/DBMI/pFINDR). The PhenDisco website (https://phendisco.ucsd.edu) is currently under major upgrade to meet UCSD's IT security and safety requirement. We closed PhenDisco in March of 2015. Given that PhenDisco operates on the old version of dbGaP data (July 2013), its utility to the real world users was questionable. As shown in the Usage Metrics table below, unlike the accesses occurred right after Phase I, the accesses occurred after Mar 2014 seem less meaningful as the majority of the visits were very shorts and the bounce rate is quite high (Table 3). This might mean that people were averted from further exploring the site knowing that the site was running on the old version of data. And it is understandable as the dbGaP site was up and running on the up to date data. The detailed page navigation information (Figure 11) confirms this usage pattern. PhenDisco received more visits from outside US.

Table 3 Usage metrics

Usage Metrics	Jun 2013 ~ Feb 2014	Mar 2014 ~ Apr 2015
Visits	668	1,139
Unique Visitors	204	762
Page-views	5,717	2,446
Pages/visit	9	2
Average visit duration	14 sec	2 sec
Bounce rate	32.78%	67.87%
% new visit	30.54%	66.02%

Table 4 Visits by countries

Countries	Jun 2013 ~ Feb 2014	Mar 2014 ~ Apr 2015
US	608	623
UK	8	17
China	5	7
Japan	5	15
Rumania	0	40
France	0	6
Other	42	491

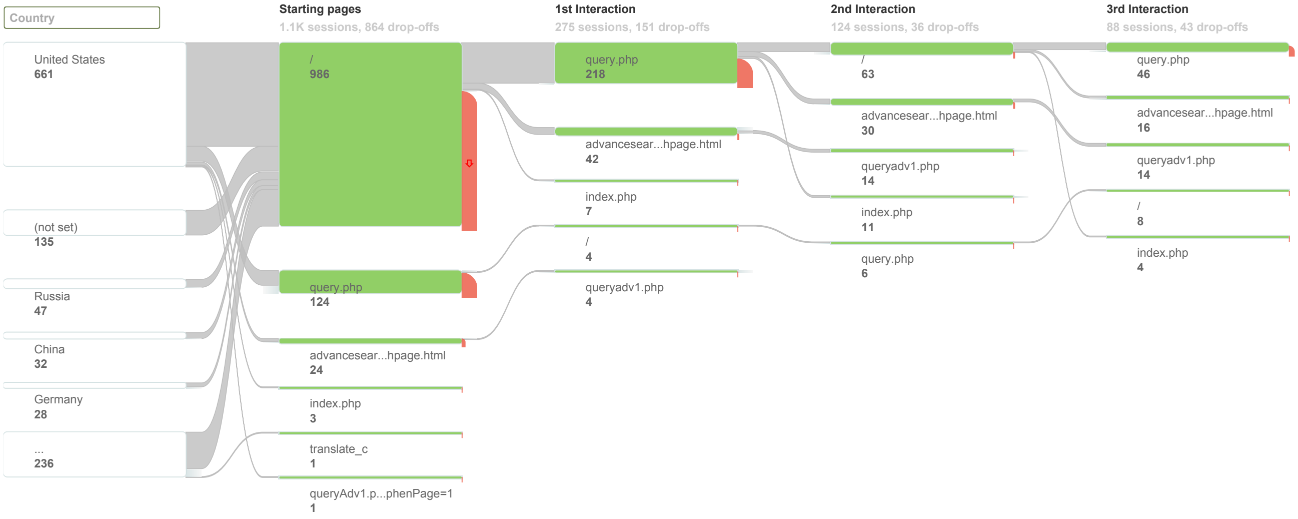


Figure 11 Site navigation log

Therefore, we made PhenDisco available only internally as a legacy system for future research purposes. However, the audiovisual recordings of the related presentations and demos will remain available to public:

- https://www.youtube.com/watch?v=8E48u6DL9tQ
- https://www.youtube.com/watch?v=pJrgxBzbz88
- http://www.slideshare.net/sondoan/phendisco-phenotype-discovery-system-for-the-database-of-genotypes-and-phenotypes-dbgap

4 Summary and Lessons Learned

Phase I of this project was dedicated to understanding the challenges in using dbGaP and identifying the user requirements. We also developed a *lite* NLP based standardization pipeline and applied it to the entire phenotype variables in dbGaP (up to July 2013 version). This "breadth-oriented" approach reached 70% of overall accuracy. In Phase II we focused on refining the standardization pipeline and developing the additional functionalities (i.e., hierarchical expansion of search and similar variable identification), which we started in the end of Phase I. These were among the high priority functionalities that the users identified during the requirement analysis step. We completed the algorithms that automatically update the concept hierarchy as new phenotype variables become available to support hierarchical expansion of search concepts. The first version of algorithms for similar variable identification was developed.

After the completion of this project, we continued investigating the new approaches to improving the standardization pipeline. As a test case, we took the *standardization of lifestyle variables*, with which the PhenDisco standardization pipeline struggled the most. In this study we adopted a machine-learning approach (neural network) and focused first on identifying lifestyle variables. Our neural network based classification algorithms showed a promising performance of 89% accuracy in identifying life style related variables. The initial data preparation work was accepted by 2016 Nursing Informatics Congress and is scheduled for presentation on June 29. The algorithm development and classification results were submitted to 2016 Fall AMIA symposium as a paper and is currently under review. This follow up study suggests that the text processing order of the PhenDisco pipeline might need to change. The PhenDisco standardization pipeline first identifies the key concepts and their semantic roles in the context (i.e., variable description narratives). Based on the key concepts and the semantic roles, the pipeline then classifies the variable into variable categories. The first standardization step often suffers with errors, which propagate to the second classification step. That is the classification of the variables highly depends on the standardization step thus reached the similar level of accuracy of 70%. By reverting the order – i.e., categorizing the variables based on machine-learning algorithms then identifying the key concepts and their semantic roles-could show higher accuracy as semantic ambiguities around the key concepts can be resolved. We are planning to experiment this approach in a future study.

The usage statistics show that PhenDisco was rarely used in 2015. Even though PhenDisco received more than 1000 visits, majority of them left the site immediately. This might caused by the disclaimer we put in the home page that says the system was running on the 2013 version of data. We did not advertise the system actively in Phase II due to the reduced duration and scope of the project. However we might have continued to attract visitors to the site by presenting the related works at various scholarly venues. This indicates the high level of interests in research communities on the resources that dbGaP provides and the related standardization efforts. One alternative approach that might have improved the use of PhenDisco is to focus on maintaining the data resource up to date with the basic level of standardization than completing the algorithms for advanced functionalities.   Nonetheless, this project provided an invaluable opportunity to understand and address the challenges around data reuse and standardization, which now became even more critical with the BD2K initiatives.

5 Related Publications and Presentations

Publications

[1] Hsieh, A, Doan S, Conway, M, Lin, KW, Kim H. Demographics Identification: Variable Extraction Resource (DIVER), Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference, pp.40-49, 27-28 Sept. 2012 doi: 10.1109/HISB.2012.17

[2] Lin KW, Hsieh A, Farzaneh S, Doan S, Kim H. Standardizing Phenotype Variables in the Database of Genotypes and phenotypes (dbGaP) based on Information Models. AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:110. eCollection 2013. PubMed PMID: 24303316.

[3] Alipanah N, Lin KW, Venkatesh V, Farzaneh S, Kim HE. Phenotype Information Retrieval for Existing GWAS Studies. AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:4-8. eCollection 2013. PubMed PMID: 24303228; PubMed Central PMCID: PMC3845737.

[4] Lin KW, Tharp M, Conway M, Hsieh A, Ross M, Kim J, Kim HE. Feasibility of using Clinical Element Models (CEM) to standardize phenotype variables in the database of genotypes and phenotypes (dbGaP). PLoS One. 2013 Sep 18;8(9):e76384. doi: 10.1371/journal.pone.0076384. eCollection 2013. PubMed PMID: 24058713; PubMed Central PMCID: PMC3776754.

[5] Doan S, Lin KW, Conway M, Ohno-Machado L, Hsieh A, Feupe SF, Garland A, Ross M, Jiang X, Farzaneh S, Walker R, Alipanah N, Xu H, Kim H. PhenDisco: Phenotype Discovery System for the Database of Genotypes and Phenotypes (dbGaP). *J Am Med Inform Assoc*. 2013 Sep 3. doi: 10.1136/amiajnl-2013-001882. [Epub ahead of print] PubMed PMID: 23989082.

Presentations

[1] Lin KW, Doan S, Hsieh A, Kim H. Identification of similar variables in the database of genotypes and phenotypes (dbGaP). 2015 *AMIA Summits Transl Sci*, San Francisco, CA. Mar 23-27

[2] Richardson A, Fireman E, Kim H. Extracting a concept hierarchy from UMLS to support hierarchical expansion in data base search. 2015 *AMIA Summits Transl Sci*, San Francisco, CA. Mar 23-27

[3] Walker R, Kim H, Feudjio SF, Farzaneh S, Ross M, Doan S, Ohno-Machado L, Lin K. User requirement analysis for the database of genotypes and phenotypes (dbGaP): a multidimensional approach for query tool design. 2014 2015 *AMIA Summits Transl Sci*, San Francisco, CA. Apr 07-11

[4] Kim H, Doan S, Ohno-Machado L. PhenDisco (Phenotype Discoverer): a new information retrieval system for the database of Genotypes and Phenotypes (dbGaP). 2013 The 3rd Annual Translational Bioinformatics Conference. Seoul Korea Oct 2-4

[5] Lin KW, Hsieh A, Farzaneh S, Doan S, Kim H. Standardizing Phenotype Variables in the Database of Genotypes and phenotypes (dbGaP) based on Information Models. 2013 *AMIA Summits Transl,* San Francisco, CA. Mar 18-22

[6] Feupe SF, Walker R, Kim H. Assessment of Race and Ethnicity Variables in dbGaP: Inconsistencies and the Impact on Search in the database. 2013 *AMIA fall symposium.* Washington DC. Nov 16-20

[7] Lin KW, Kim H. Building a Domain Analysis Model for the Data Stored in the Databse of Genotypes and Phenotypes (dbGaP). 2013 *AMIA fall symposium.* Washington DC. Nov 16-20

[8] Doan S, Lin KW, Walker R, Farzaneh S, Alipanah N, Kim H. A Rule-Based Natural Language Processing System in Tagging and Categorizing Phenotype Variables in NCBI's database of Genotypes and Phenotypes (dbGaP). 2013 *AMIA fall symposium.* Washington DC. Nov 16-20

[9] Hsieh A, Conway M, Kim H. Identifying and standardizing age variables using NLP. 2012 *AMIA fall symposium*, Chicago, Nov 2-7 (nominated for distinguished poster award)

[10] Alipanah N, Kim H, L Ohno-Machado: Building an Ontology of Phenotypes for Existing GWAS Studies. Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 *IEEE Second International Conference*, pp. 111.

[11] Lin K, Tharp M, Conway M, Hsieh A, Ross M, Kim J, Kim H. Feasibility of using Clinical Element Models (CEM) to standardize phenotype variables in the databases of Genotype and Phenotype (dbGaP). Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 *IEEE Second International Conference*, pp. 123. La Jolla, CA. Oct 27-28

[12] Lin KW, Ross M, Chapman WW, Conway MA, Ohno-Machado L, Finn F, Kim H. Testing the adequacy of a public GWAS database as a cohort discovery tool. 2012 *American Thoracic Society International Conference*. San Francisco, May 18-23

[13] Ross M, Kim J, Lin KW, Ohno-Machado L, Kim H. Data Locked Inside Databases: A Text Classification in the Database of Genotypes and Phenotypes (dbGaP) to Address Challenges in Retrieving Clinical Information from Public Data Repositories. 2012 *American Thoracic Society International Conference*. San Francisco, May 18-23

[14] PhenDisco (Phenotype Discoverer): a new information retrieval system for the database of Genotypes and Phenotypes (dbGaP). 2013, The 3rd Annual Translational Bioinformatics Conference. Seoul Korea, October 2-4

[15] Phenotype Finder IN Data Resources (pFINDR) and Demo. iDASH Third Annual All Hands Symposium, UCSD. September 16-17

[16] Kim H, Doan S. PhenDisco: phenotype discovery system for the database of genotypes and phenotypes. Informatics Journal Club Webinar. September 5, 2013.