DCC-harmonized phenotypes for the scientific community

Updated 10/20/2021

The DCC has undertaken two projects related to study phenotypes in TOPMed. Please see the sections below for more information.

Information about these projects is available in a published manuscript [1]. If you use the datasets described on this page, please cite the following paper:

Stilp AM, Emery LS, Broome JG, Buth EJ, Khan AT, Laurie CA, Wang FF, Wong Q, Chen D, D'Augustine CM, Heard-Costa NL, Hohensee CR, Johnson WC, Juarez LD, Liu J, Mutalik KM, Raffield LM, Wiggins KL, de Vries PS, Kelly TN, Kooperberg C, Natarajan P, Peloso GM, Peyser PA, Reiner AP, Arnett DK, Aslibekyan S, Barnes KC, Bielak LF, Bis JC, Cade BE, Chen MH, Correa A, Cupples LA, de Andrade M, Ellinor PT, Fornage M, Franceschini N, Gan W, Ganesh SK, Graffelman J, Grove ML, Guo X, Hawley NL, Hsu WL, Jackson RD, Jaquish CE, Johnson AD, Kardia SLR, Kelly S, Lee J, Mathias RA, McGarvey ST, Mitchell BD, Montasser ME, Morrison AC, North KE, Nouraie SM, Oelsner EC, Pankratz N, Rich SS, Rotter JI, Smith JA, Taylor KD, Vasan RS, Weeks DE, Weiss ST, Wilson CG, Yanek LR, Psaty BM, Heckbert SR, Laurie CC. A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. Am J Epidemiol. 2021 Oct 1;190(10):1977-1992. doi: 10.1093/aje/kwab115. PMID: 33861317; PMCID: PMC8485147.

DCC phenotype harmonization project

The TOPMed DCC has harmonized over 100 phenotype variables related to heart, lung, blood, and sleep domains. The main goal of the DCC harmonization project is to provide harmonized phenotypes that are well-documented, reproducible, and as homogeneous across studies as possible. In harmonized datasets and documentation, the DCC typically uses “phenotype” to refer to the observable characteristic (e.g., diastolic blood pressure) and “variable” to refer to the specific data vector values for a given phenotype (e.g., bp_diastolic_1). To enable reproducibility, all study data were acquired from dbGaP.

Datasets and documentation of the harmonized variables have been submitted to two NIH-designated repositories: dbGaP [2] and BioData Catalyst [3].

Full documentation for each harmonized variable is also provided in a GitHub repository [4]. The documentation for each harmonized variable includes the identifiers of the original dbGaP study variables used in harmonization as well as the code that was used to transform them into the harmonized variable. This repository also includes a reproducible example that instructs users how to use the documentation to reproduce a simulated harmonized variable.

The phenotype tagging project

In addition to the phenotype harmonization project, the DCC has undertaken a related project to label over 16,000 dbGaP study variables with 65 phenotype concepts from heart, lung, blood, and sleep domains. We refer to this process as “variable tagging.” These labels enable researchers to more easily identify variables of interest that can be used in future harmonization efforts. The results of the tagging project are available in the dbGaP user interface.

The list of tags and instructions for identifying phenotype tags can be found on the DCC phenotype tagging details page [5].

Links

[1] https://pubmed.ncbi.nlm.nih.gov/33861317/ [2] https://www.ncbi.nlm.nih.gov/gap/ [3] https://biodatacatalyst.nhlbi.nih.gov/ [4] https://github.com/UW-GAC/topmed-dcc-harmonized-phenotypes [5] https://topmed.nhlbi.nih.gov/dcc-phenotype-tagging-details