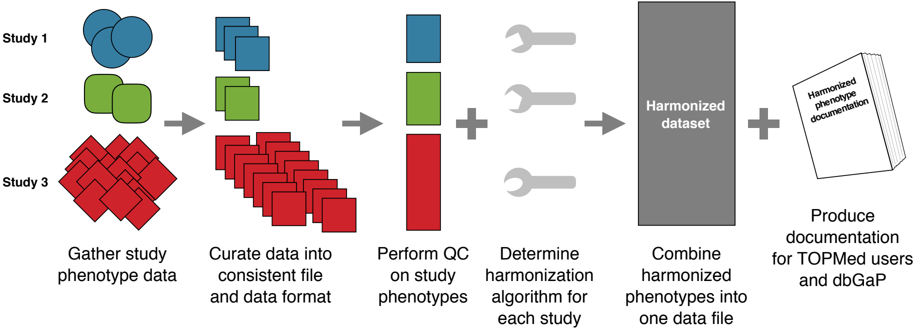
TOPMed phenotype harmonization

Adrienne Stilp
Leslie Emery
Susan Heckbert
Cathy Laurie

TOPMed Data Coordinating Center
University of Washington

March 11, 2017

General harmonization steps



Gather study phenotype data

Curate data into consistent file and data format

Perform QC on study phenotypes

Determine harmonization algorithm for each study

Combine harmonized phenotypes into one data file

Produce documentation for TOPMed users and dbGaP

Maelstrom essentials for phenotype harmonization

1. Collaborative framework
2. Expert input
3. Valid data input
4. Valid data output
5. Rigorous documentation
6. Respect for stakeholders

Fortier et al., 2016, Int J. Epidemiol., doi 10.1093/ije/dyw075

Collaboration for phenotype harmonization in TOPMed

- Phenotype harmonization committee
  - Have experience with previous harmonization efforts
  - Provide advice about general approach
- Working group members
  - Have the necessary domain expertise
  - Determine source phenotypes and harmonization algorithm to use
  - Assist with data QC
- Study liaisons
  - Have expertise about specific study phenotypes
  - Assist with data QC
- DCC analysts
  - Implement harmonization algorithm
  - Perform data QC
  - Save final harmonized phenotype

How is the DCC implementing phenotype harmonization?

1. Acquire study phenotype data from dbGaP
2. Construct a relational database to hold source and harmonized phenotype data
3. Compute harmonized phenotypes using a systematic workflow
4. Automatically produce datasets and documentation for harmonized phenotypes using information stored in the database

Why dbGaP?

- Stable repository for genotype and phenotype data
- Already contains most phenotype data for TOPMed studies
- Curated and in a consistent format
- Provides mechanism to track data provenance via study and variable accession numbers
- Tracks participants' consent

Overall, dbGaP allows for better reproducibility of harmonized phenotypes.

The DCC's relational database for phenotypes

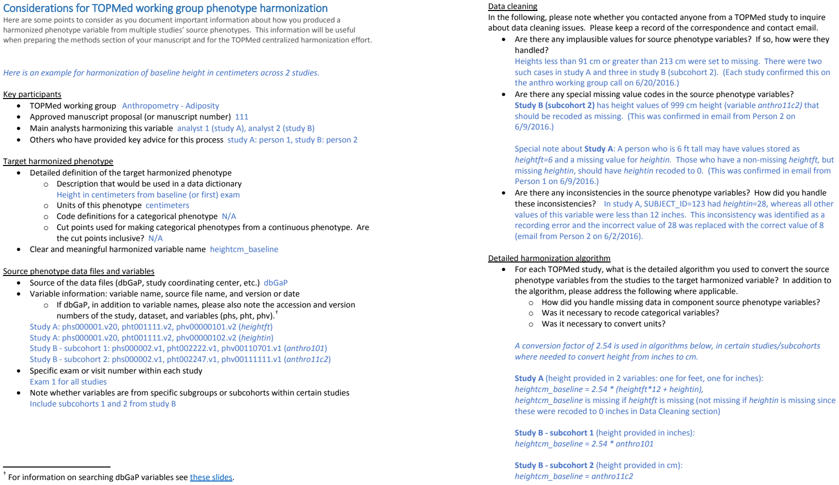
- Source phenotypes:
  - Consistent, uniform data access for DCC analysts
  - Can be queried to search for available source phenotypes
  - Store all versions downloaded from dbGaP
- Harmonized phenotypes:
  - Linked to component source phenotypes and algorithms
  - Can be queried to determine number of TOPMed subjects with a specific phenotype
    - Requires the phenotype of interested to have already been harmonized!
  - Easy to update when component source data are updated
- Custom R package interface for DCC analysts

DCC's workflow for phenotype harmonization

- Receive algorithm instructions and source phenotype identifiers from working group members
- Work in "harmonization units"
  - Study or subcohort within a study
  - Look up source phenotypes to include
  - Define harmonization algorithm to use
- Perform QC on both source study data and harmonized data
- Store harmonized phenotype and metadata in the database
- Include age at measurement for each harmonized phenotype

We also provide harmonization guidelines to working groups that may choose to do harmonization on their own.

Harmonization guidelines for working group members



Considerations for TOPMed working group phenotype harmonization

Here are some points to consider as you document important information about how you produced a harmonized phenotype variable from multiple studies' source phenotypes. This information will be useful when preparing the methods section of your manuscript and for the TOPMed centralized harmonization effort.

Here is an example for harmonization of baseline height in centimeters across 2 studies.

Key participants

- TOPMed working group Anthropometry - Adiposity
- Approved manuscript proposal (or manuscript number) 111
- Main analysts harmonizing this variable analyst 1 (study A), analyst 2 (study B)
- Others who have provided key advice for this process study A: person 1, study B: person 2

Target harmonized phenotype

- Detailed definition of the target harmonized phenotype
  - Description that would be used in a data dictionary
    Height in centimeters from baseline (or first) exam
  - Units of this phenotype centimeters
  - Code definitions for a categorical phenotype N/A
  - Cut points used for making categorical phenotypes from a continuous phenotype. Are the cut points inclusive? N/A
- Clear and meaningful harmonized variable name heightcm_baseline

Source phenotype data files and variables

- Source of the data files (dbGaP, study coordinating center, etc.) dbGaP
- Variable information: variable name, source file name, and version or date
  - If dbGaP, in addition to variable names, please also note the accession and version numbers of the study, dataset, and variables (phs, pht, phv).[†]

  Study A: phs000001.v20, pht001111.v2, phv00000101.v2 (*heightft*)
  Study A: phs000001.v20, pht001111.v2, phv00000102.v2 (*heightin*)
  Study B - subcohort 1: phs000002.v1, pht002222.v1, phv00110701.v1 (*anthro101*)
  Study B - subcohort 2: phs000002.v1, pht002247.v1, phv00111111.v1 (*anthro11c2*)
- Specific exam or visit number within each study
  Exam 1 for all studies
- Note whether variables are from specific subgroups or subcohorts within certain studies
  Include subcohorts 1 and 2 from study B

Data cleaning

In the following, please note whether you contacted anyone from a TOPMed study to inquire about data cleaning issues. Please keep a record of the correspondence and contact email.

- Are there any implausible values for source phenotype variables? If so, how were they handled?
  Heights less than 91 cm or greater than 213 cm were set to missing. There were two such cases in study A and three in study B (subcohort 2). (Each study confirmed this on the anthro working group call on 6/20/2016.)
- Are there any special missing value codes in the source phenotype variables?
  Study B (subcohort 2) has height values of 999 cm height (variable *anthro11c2)* that should be recoded as missing. (This was confirmed in email from Person 2 on 6/9/2016.)

  Special note about **Study A**: A person who is 6 ft tall may have values stored as *heightft=6* and a missing value for *heightin*. Those who have a non-missing *heightft,* but missing *heightin*, should have *heightin* recoded to 0. (This was confirmed in email from Person 1 on 6/9/2016.)
- Are there any inconsistencies in the source phenotype variables? How did you handle these inconsistencies? In study A, SUBJECT_ID=123 had *heightin*=28, whereas all other values of this variable were less than 12 inches. This inconsistency was identified as a recording error and the incorrect value of 28 was replaced with the correct value of 8 (email from Person 2 on 6/2/2016).

Detailed harmonization algorithm

- For each TOPMed study, what is the detailed algorithm you used to convert the source phenotype variables from the studies to the target harmonized variable? In addition to the algorithm, please address the following where applicable.
  - How did you handle missing data in component source phenotype variables?
  - Was it necessary to recode categorical variables?
  - Was it necessary to convert units?

  A conversion factor of 2.54 is used in algorithms below, in certain studies/subcohorts where needed to convert height from inches to cm.

  Study A (height provided in 2 variables: one for feet, one for inches):
  *heightcm_baseline = 2.54 * (heightft*12 + heightin),*
  heightcm_baseline is missing if *heightft* is missing (not missing if *heightin* is missing since these were recoded to 0 inches in Data Cleaning section)

  Study B - subcohort 1 (height provided in inches):
  *heightcm_baseline = 2.54 * anthro101*

  Study B - subcohort 2 (height provided in cm):
  heightcm_baseline = anthro11c2

[†] For information on searching dbGaP variables see these slides.

QC of source data

Potential issues:

- Biologically invalid values?
- Internal inconsistencies in the study data?
- Missing data?

What to do?

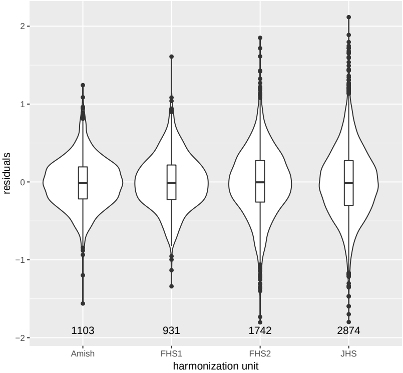
- Which measurement (if any) is correct?
- Should subjects with discrepant data be excluded?
- Often requires talking to study liaisons

No blanket procedure can be applied to all phenotypes! To solve QC issues, we usually need to refine the harmonization algorithm.

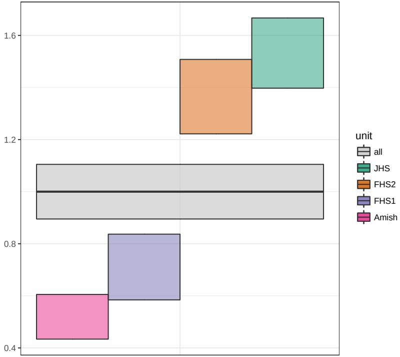
Example of QC in harmonized Red Blood Cell Count

We fit a mixed model accounting for age, sex, ancestry, harmonization unit and kinship.

RBC ~ sex + age + unit + PC1 +...+ PC5 + (1|kinship)
Wald test of mean variation among units: p−value=1e−4



Unit−specific variances relative to pooled variance
Likelihood test p−value = 7e−55



If one unit is very different, need further exploration!

Harmonized phenotype datasets and documentation

- Automatically produced from database
- Documentation includes:
  - dbGaP accession numbers of source phenotypes
  - algorithms used to produce a harmonized phenotype
- Will be provided to TOPMed investigators
- Eventually will be posted on dbGaP
- Allows reproduction of harmonized phenotypes by the community
- Will include LOINC codes for harmonized phenotypes

Example harmonized dataset and data dictionary

- Dataset:

```
topmed_subject_id study_subject_id age_at_hemoglobin_mcnc_bld_1 hemoglobin_mcnc_bld_1
         10000742             2361                           22                  13.9
         10004478            27345                           64                  12.6
         10011485            20420                           34                  15.3
         10014577            23615                           28                  15.3
         10020979            15756                           56                  14.5

age_at_hematocrit_vfr_bld_1 hematocrit_vfr_bld_1 age_at_platelet_ncnc_bld_1 platelet_ncnc_bld_1
                         22                 35.8                         22                 185
                         64                 40.4                         64                 248
                         34                 36.7
                         28                 38.5                         28                 156
                         56                 36.2
```

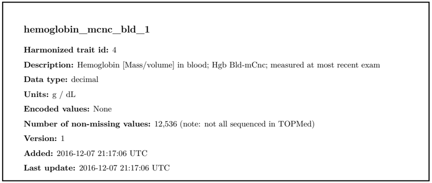
- Data dictionary:

```
VARNAME                       VARDESC                                     TYPE     UNITS             LOINC_code
topmed_subject_id             TOPMed-assigned subjectID
study_subject_id              subjectID assigned by each study
age_at_hemoglobin_mcnc_bld_1  age at measurement of hemoglobin_mcnc_bld_1 decimal  years
hemoglobin_mcnc_bld_1         Hemoglobin [Mass/volume] in blood           decimal  g / dL           718-7
age_at_hematocrit_vfr_bld_1   age at measurement of hematocrit_vfr_bld_1  decimal  years
hematocrit_vfr_bld_1          Hematocrit [Volume fraction] of blood       decimal  %                20570-8
age_at_platelet_ncnc_bld_1    age at measurement of platelet_ncnc_bld_1   decimal  years
platelet_ncnc_bld_1           Platelets [#/volume] in Blood               integer  thousands / uL   26515-7
```

*The phenotypes shown here are randomly-generated fake data instead of actual harmonized phenotypes.

More comprehensive documentation

Harmonized phenotype metadata



hemoglobin_mcnc_bld_1

Harmonized trait id: 4

Description: Hemoglobin [Mass/volume] in blood; Hgb Bld-mCnc; measured at most recent exam

Data type: decimal

Units: g / dL

Encoded values: None

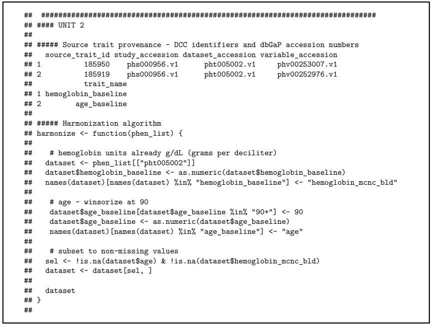
Number of non-missing values: 12,536 (note: not all sequenced in TOPMed)

Version: 1

Added: 2016-12-07 21:17:06 UTC

Last update: 2016-12-07 21:17:06 UTC

Component phenotypes and algorithm (by unit)



```
##  ##############################################################################
## #### UNIT 2
##
## ##### Source trait provenance - DCC identifiers and dbGaP accession numbers
##   source_trait_id study_accession dataset_accession variable_accession
## 1          185950    phs000956.v1      pht005002.v1     phv00253007.v1
## 2          185919    phs000956.v1      pht005002.v1     phv00252976.v1
##            trait_name
## 1 hemoglobin_baseline
## 2        age_baseline
##
## ##### Harmonization algorithm
## harmonize <- function(phen_list) {
##
##    # hemoglobin units already g/dL (grams per deciliter)
##   dataset <- phen_list[["pht005002"]]
##   dataset$hemoglobin_baseline <- as.numeric(dataset$hemoglobin_baseline)
##   names(dataset)[names(dataset) %in% "hemoglobin_baseline"] <- "hemoglobin_mcnc_bld"
##
##    # age - winsorize at 90
##    dataset$age_baseline[dataset$age_baseline %in% "90+"] <- 90
##    dataset$age_baseline <- as.numeric(dataset$age_baseline)
##    names(dataset)[names(dataset) %in% "age_baseline"] <- "age"
##
##    # subset to non-missing values
##   sel <- !is.na(dataset$age) & !is.na(dataset$hemoglobin_mcnc_bld)
##   dataset <- dataset[sel, ]
##
##   dataset
## }
##
```

Maelstrom essentials in TOPMed phenotype harmonization

1. Collaborative framework
   - Harmonization committee
   - Phenotype working groups
   - Study liaisons
2. Expert input
   - Working Groups and Study Liaisons
3. Valid data input
   - QC of source phenotypes
4. Valid data output
   - QC of harmonized phenotypes
5. Rigorous documentation
   - study designs and sample ascertainment
   - provenance of source and harmonized phenotypes
   - harmonization algorithms
   - QC results
6. Respect for stakeholders
   - participant consent
   - data use limitations
   - interests of study investigators and funding organizations