

Phenotype Harmonization in Association Studies

SISG 2019 Module 17

Leslie Emery

based on slides from Adrienne Stilp

July 24, 2019

What is phenotype harmonization?

Phenotype data harmonization in TOPMed

General steps for harmonization

QC of study source phenotypes

QC of harmonized phenotypes

DCC harmonization for TOPMed

Resources

Phenotype harmonization is a specific case of data harmonization

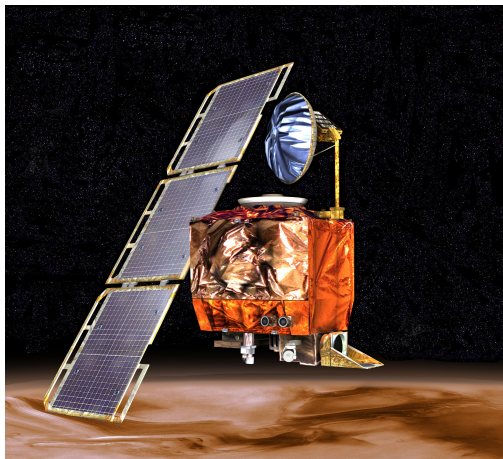
Data harmonization refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies. This process is becoming more and more significant in demography and sociology research, since the needs of data harmonization is rapidly growing as the volume and the need to share existing data explodes.

- Data Sharing for Demographic Research, University of Michigan

Do we really need to worry about harmonizing data?

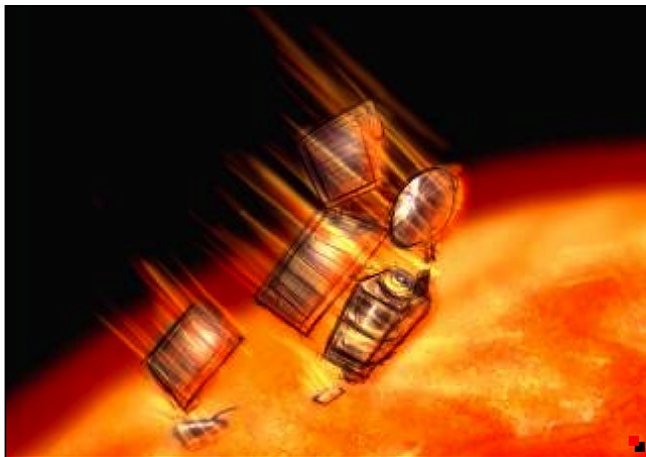
There is **much more** to data harmonization than just renaming variables and combining files together!

The Mars Climate Orbiter



NASA/JPL/Corby Waste

Lack of data harmonization can be disastrous!



More details from Wikipedia

What is phenotype harmonization?

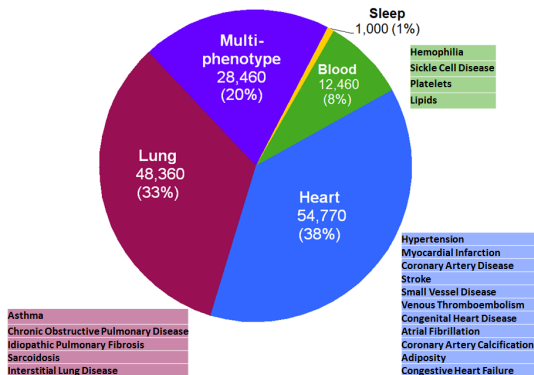
Phenotype harmonization is the process by which source phenotype variables from different studies are transformed so that they can be analyzed together.

Phenotype harmonization is a major undertaking in large projects that combine pre-existing studies. Examples include:

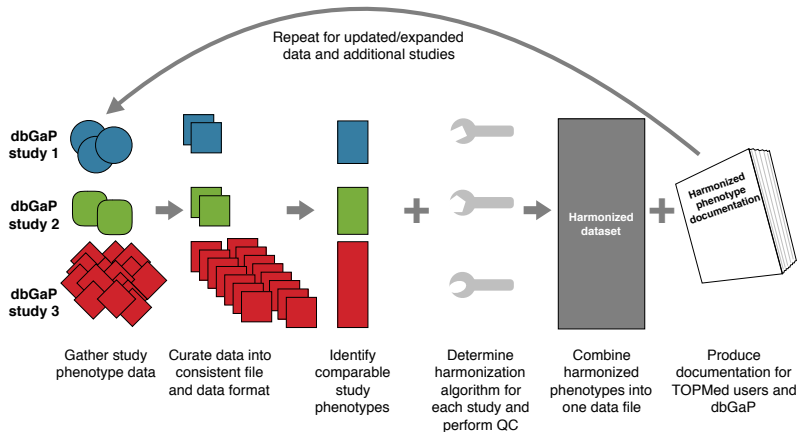
- ▶ NHLBI GO Exome Sequencing Project
- ▶ CHARGE Consortium (Cohorts for Heart & Aging Research in Genomic Epidemiology)
- ▶ TOPMed (Trans Omics for Precision Medicine)

Source phenotypes in TOPMed

- ▶ ~ 144,000 participants
- ▶ > 80 different studies
- ▶ 30 studies with source phenotype data via dbGaP



Harmonizing phenotypes from many studies in TOPMed



Working with phenotype data has unique challenges

- ▶ Genotypes
 - ▶ Big but **homogeneous**
 - ▶ Similar across studies – automated processing
- ▶ Phenotypes
 - ▶ Small but **heterogeneous**
 - ▶ Every study collects data differently – manual effort required
- ▶ Too much noise can cause a loss of power and mask true associations.

What needs to be done in phenotype harmonization?

1. Define the target harmonized phenotype
2. Decide which studies can be included
3. Process source phenotype data by study
 - ▶ Perform QC
 - ▶ Determine harmonization algorithm
 - ▶ Once per study
4. Estimate quality of harmonized phenotype output
 - ▶ More QC
 - ▶ May need to repeat previous steps
5. Document and disseminate harmonized phenotypes

Documentation

Your harmonized phenotype data should be reproducible.

- ▶ Accurate reporting in papers
- ▶ Able to add new studies in the future

What do you need?

- ▶ Definition of the harmonized phenotype
- ▶ Which component source phenotypes were used
 - ▶ Source file?
 - ▶ Version?
- ▶ What algorithms were used
 - ▶ Ideally, the exact code you used
- ▶ How QC issues were addressed

QC of study source phenotypes

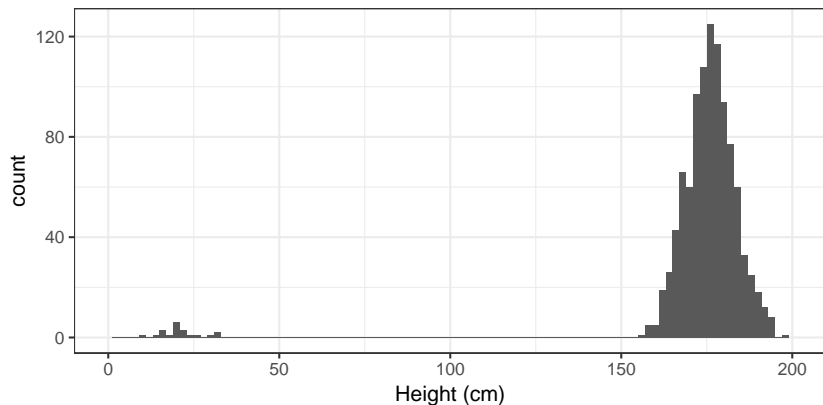
Potential QC issues:

- ▶ Biologically invalid values
- ▶ Extreme phenotypes
- ▶ Missing data
- ▶ Internal inconsistencies

And a lot of others you can't predict!

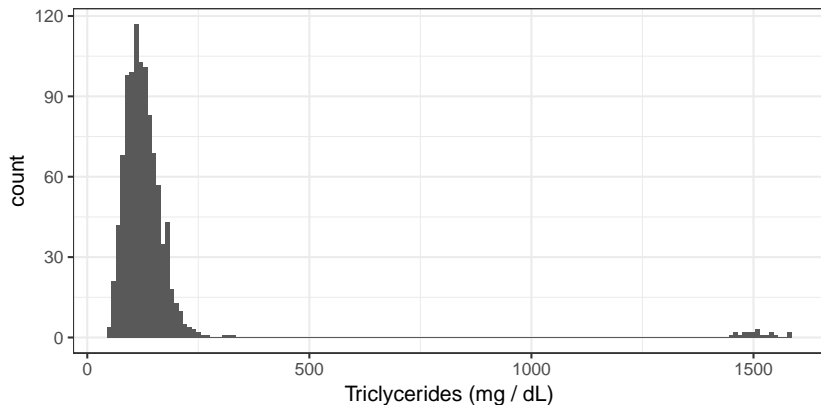
Are these values biologically invalid?

Example: Implausibly small height measurements



Are these extreme phenotypes true?

Example: Extreme triglycerides levels



Are any source phenotypes missing values?

Example: Missing data in some components for diabetes

	subject_id	diabetes_self_report	diabetes_meds
1	a	1	1
2	b	0	.
3	c	1	1
4	d	1	.
5	e	0	0
6	f	1	1
7	g	0	0
8	h	1	1
9	i	1	1
10	j	1	1

Are there any internal inconsistencies in the data?

Example: Self-reported vs. MD-diagnosed diabetes

	subject_id	self_report	md_diagnosis	
1	a	0	0	
2	b	0	0	
3	c	0	0	
4	d	1	1	
5	e	0	0	
6	f	1	0	# discrepant
7	g	0	0	
8	h	1	1	
9	i	0	0	
10	j	0	1	# discrepant

How do we fix problems?

- ▶ Which measurement (if any) is correct?
- ▶ Should you exclude subjects with discrepant data?
- ▶ Should outliers be excluded?
 - ▶ Measurement issue?
 - ▶ Real values indicative of rare variants with high effects (e.g., LOF)?

No blanket answer for all phenotypes!

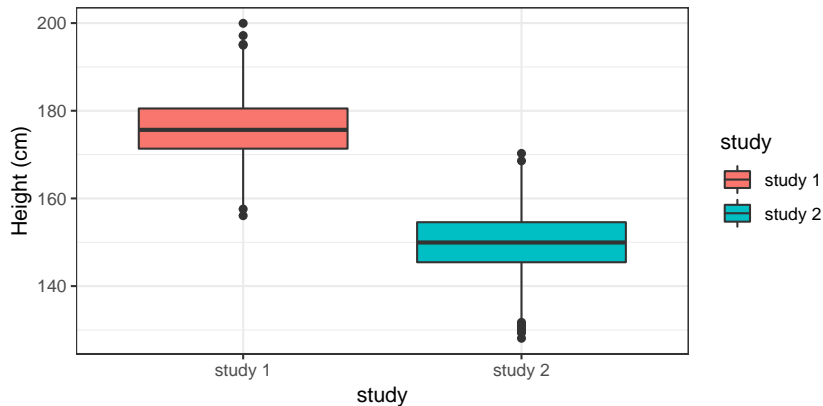
- ▶ Involve both study members and domain experts
- ▶ Clearly specify how you decide to handle these QC issues

QC of harmonized phenotypes

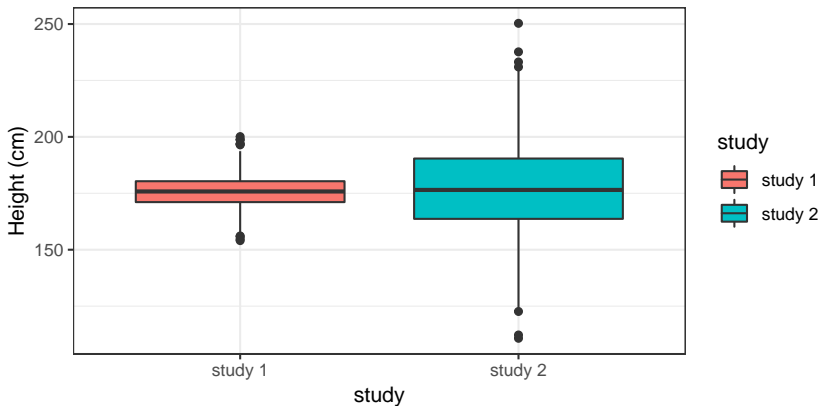
Are some studies very different than others?

- ▶ Quantitative phenotypes:
 - ▶ Mean
 - ▶ Standard deviation
 - ▶ General distribution
- ▶ Categorical phenotypes:
 - ▶ Frequency
- ▶ May need to look at batch effects from other variables, e.g.
 - ▶ Assay or device used?
 - ▶ Questionnaire version?
- ▶ Look at residual variance after fitting a null model
 - ▶ For WGS with related subjects, fit a mixed model
 - ▶ Fixed effects: age, sex, study
 - ▶ Random effects: genetic relatedness matrix

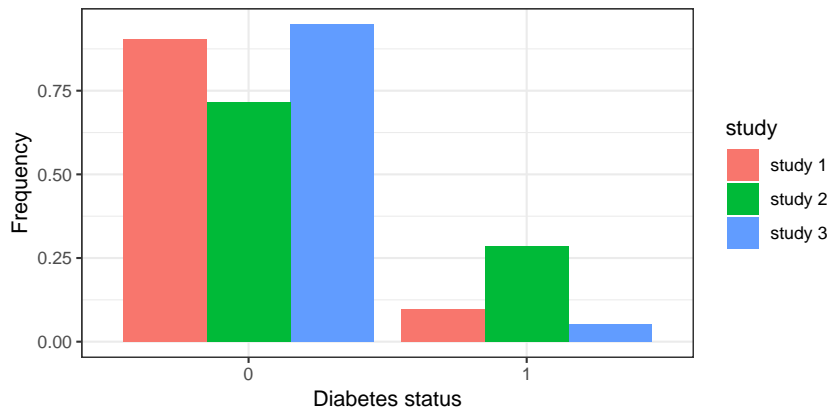
Do the studies have similar means?



Do the studies have similar standard deviations?



Do the studies have different frequencies?



What do you do if you find a difference between studies?

- ▶ Is there a valid reason for the difference?
 - ▶ Expected differences due to study design?
 - ▶ e.g. Higher prevalence of disease in a study targeting cases
 - ▶ Different distributions due to ancestry?
- ▶ Is there an error in the harmonization algorithm?
- ▶ Do you need to treat this study's data differently?
- ▶ Is the study too different to be included?
- ▶ Do you need to adjust for the difference in analysis?

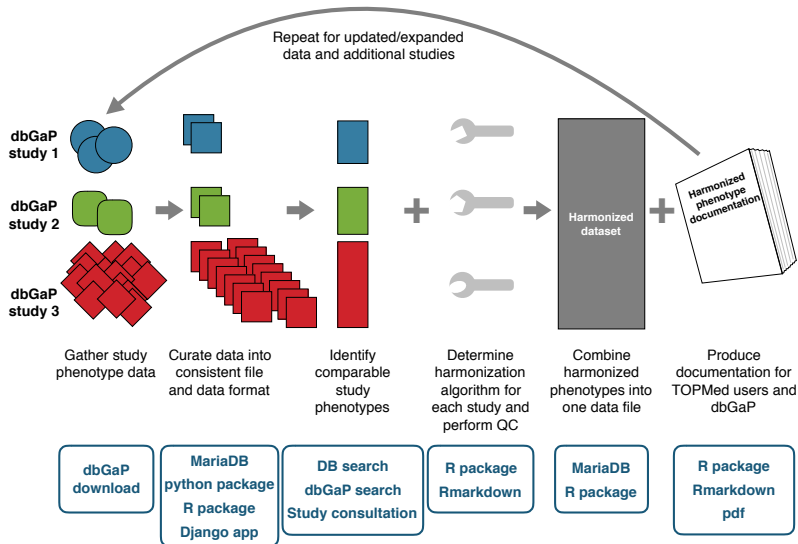
Again, no blanket answer to these questions!

- ▶ Need to involve both study members and domain experts

DCC harmonization for TOPMed

- ▶ Acquire study data from dbGaP
 - ▶ Provides a bookkeeping trail for documentation
 - ▶ Available to the general scientific community
- ▶ Store data in a relational database
 - ▶ Both study phenotypes and harmonized phenotypes
 - ▶ Includes everything needed to recreate a harmonized phenotype
 - ▶ Metadata
 - ▶ Component phenotypes and versions
 - ▶ Algorithms
 - ▶ Allows automated production of datasets and documentation

Harmonization workflow in TOPMed



TOPMed phenotype harmonization progress

Demographic:	Lipids:	Blood cell count:	Inflammation:	
annotated_sex_1	fasting_lipids_1	basophil_ncnc_bld_1	tnfa_1	mmp9_1
ethnicity_1	hdl_1	eosinophil_ncnc_bld_1	tnfa_r1_1	mpo_1
geographic_site_1	ldl_1	hematocrit_vfr_bld_1	tnfr2_1	opg_1
hispanic_subgroup_1	lipid_lowering_medication_1	hemoglobin_mcnc_bld_1	cd40_1	pselectin_1
race_1	total_cholesterol_1	lymphocyte_ncnc_bld_1	crp_1	
subcohort_1	triglycerides_1	mch_entmass_rbc_1	eselectin_1	
		mchc_mcnc_rbc_1	icam1_1	
		mcv_entvol_rbc_1	il1_beta_1	
		monocyte_ncnc_bld_1	il6_1	
		neutrophil_ncnc_bld_1	il10_1	
		platelet_ncnc_bld_1	il18_1	
		pmv_entvol_bld_1	isoprostane_8_epi_pgf2a_1	
		rbc_ncnc_bld_1	lppla2_act_1	
		rdw_ratio_rbc_1	lppla2_mass_1	
		wbc_ncnc_bld_1	mcp1_1	
Common covariates:	VTE:			Atherosclerosis:
bmi_baseline_1	vte_case_status_1			cac_score_1
current_smoker_baseline_1	vte_followup_start_age_1			cac_volume_1
ever_smoker_baseline_1	vte_prior_history_1			carotid_plaque_1
height_baseline_1				carotid_stenosis_1
weight_baseline_1				cimt_1
				cimt_2
	Blood pressure:			
	antihypertensive_meds_1			
	bp_diastolic_1			
	bp_systolic_1			

- ▶ 6 research scientists
- ▶ > 3 years
- ▶ 2 - 17 studies included in each phenotype

Guidelines for Phenotype Harmonization

- ▶ Always use subject ids in phenotype files
 - ▶ Don't use sample ids (they may change if you detect a mixup)
- ▶ Decide who will do the harmonization
 - ▶ You or the studies?
- ▶ Provide clear instructions to the harmonizers
 - ▶ Description of target phenotype
 - ▶ Clear algorithm definition
 - ▶ How to handle missing data and QC issues
- ▶ Perform sanity checks on the files you receive
- ▶ Document, document, document!

Helpful references

- ▶ Bennett, SN et al. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genet Epidemiol.* 2011 Apr; 35(3): 159-73
- ▶ Doiron, D et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol.* 2013 Nov 21; 10(1): 12
- ▶ Fortier, I et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol.* 2017 Feb 1; 46(1): 103-106