Phenotype and Exposure Data Harmonization

Leslie Lange

June 5-6, 2012

Background

Recent genetic data has tended to some level of "harmonization"

- Relatively limited number of platforms so far (e.g. genome-wide association; "exome chip") limits heterogeneity
- Variable names (e.g. SNPs) often have standardized names (rs numbers, chromosomal positions)

Phenotype and exposure data

- Data collection individual to each study
  - Questionnaires and data collection forms
  - Variable names
  - Measurement units
  - Biomarker assays
- Some studies began many years ago ("era" effects)

Overall goal

- Maximize the sample sizes of phenotype and environmental exposure data for samples with existing genetic data to increase statistical power to detect associations
- Facilitate identification of variables needed by investigators
- Reduce duplication of data harmonization efforts

Maximizing phenotype and exposure data for samples with genetic data

- Maximize utility of existing phenotype and exposure data:
  - Perform harmonization of a panel of phenotypes and exposures to produce a set of composite phenotypes across studies
  - Ensure all potential existing phenotypes and exposures are incorporated into dbGaP
- Obtain new phenotypes and exposures on existing study participants with genetic data
- For new projects, encourage use of a set of standardized phenotype and exposure measurements

Phenotypes in dbGaP

- Often *many* variables for a given phenotype when a basic search is done
  - Multiple visits
  - Sub-cohorts (e.g. Framingham)
  - Different definitions (e.g. self-report; biomarker-defined; etc.)
- Variables and/or definitions may have different key words to indicate a common phenotype (e.g. hypertension; high blood pressure; HTN)
- Varying levels of documentation submitted to dbGaP
- Additional documentation for phenotype details not always readily available

NHLBI HeartGO

- ~55,000 phenotype and exposure variables
- In 6 studies:
  - ARIC: 7,209
  - CARDIA: 9,332
  - CHS: 11,791
  - FHS: 20,585
  - JHS: 4,429
  - MESA: 2,068

→"Harmonized" phenotype and exposure data set of ~140 variables (e.g. BMI_baseline, current_smoker_baseline, former_smoker_baseline)

Phenotype harmonization

Multi-step, iterative process:

- Obtain input from phenotype-specific “working group” and disease/trait experts
- Scan variables in all projects and identify a first set of all variables related to trait(s) of interest
- Additional input from working group, trait experts, and cohort representatives, narrow down list
- Checking: sample size, measurement units, distribution, assay, visit, etc.

Phenotype	Study	Data Set	Variable	Description
Asthma	FHS	CARE_EX1_2S_V1_0108	B128	D203-006-ASTHMA
Asthma	FHS	CARE_EX1_3S_V1_0108	C93	WHEEZING OR ASTHMA IN INTERIM
Asthma	FHS	CARE_EX1_3S_V1_0108	C94	WHEEZING OF LONG DURATION
Asthma	FHS	CARE_EX1_3S_V1_0108	C95	SEASONAL WHEEZING
Asthma	FHS	CARE_EX1_3S_V1_0108	C96	WHEEZING WITH RESPIRATORY INFECTIONS
Asthma	FHS	CARE_EX1_3S_V1_0108	C374	CDI: ASTHMA
Asthma	FHS	CARE_EX1_4S_0108	D117	WHEEZING OR ASTHMA
Asthma	FHS	CARE_EX1_4S_0108	D118	WHEEZING-LONG DURATION
Asthma	FHS	CARE_EX1_4S_0108	D119	WHEEZING-SEASONAL
Asthma	FHS	CARE_EX1_4S_0108	D120	WHEEZING-WITH RESPIRATORY INFECTIONS
Asthma	FHS	CARE_EX1_4S_0108	D360	CDI-ASTHMA
Asthma	FHS	CARE_EX1_5S_0108	E339	WHEEZING OR ASTHMA
Asthma	FHS	CARE_EX1_5S_0108	E340	TYPE OF WHEEZING OR ASTHMA
Asthma	FHS	CARE_EX1_5S_0108	E646	CDI-ASTHMA
Asthma	FHS	CARE_EX1_6S_0108	F309	WHEEZING OR ASTHMA
Asthma	FHS	CARE_EX1_6S_0108	F641	CDI-ASTHMA
Asthma	FHS	CARE_EX1_7S_0108	G127	ASTHMA IN INTERIM
Asthma	FHS	CARE_EX1_7S_0108	G128	WHEEZING IN CHEST FOR LAST 12 MONTHS
Asthma	FHS	CARE_EX1_7S_0108	G418	CDI - ASTHMA
Asthma	FHS	CARE_EX1_7S_0108	G670	RESP-EVER HAD ASTHMA
Asthma	FHS	CARE_EX1_7S_0108	G671	RESP-IN 12 MOS, HAD ASTHMA ATTACK
Asthma	FHS	CARE_EX1_7S_0108	G672	RESP-CURRENTLY TAKING MEDS FOR ASTHMA
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A407	CDI - ASTHMA
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A539	EVER HAD ASTHMA
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A540	ASTHMA - STILL HAVE IT
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A541	ASTHMA - DIAGNOSED BY DOCTOR
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A542	ASTHMA - WHAT AGE DID IT START
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A543	ASTHMA - WHAT AGE DID IT STOP
Asthma	FHS	CARE_EX3_1S_V4_0108	G3A544	ASTHMA - RECEIVED MEDICAL TREATMENT
Asthma	FHS	CARE_RESP1_6S_0108	RQ014	EVER HAD ASTHMA
Asthma	FHS	CARE_RESP1_6S_0108	RQ015	MONTHS
Asthma	FHS	CARE_RESP1_6S_0108	RQ016	ASTHMA
Asthma	FHS	CARE_SLEEP1_1998S_0108	ASTH1215	pt had attack of asthma in last 12 months?
Asthma	FHS	CARE_SLEEP1_1998S_0108	ASTHMA15	MD said pt had asthma?
Asthma	FHS	CARE_SLEEP1_1998S_0108	ISTRD1	inhaled steroids for asthma
Asthma	FHS	CARE_SLEEP1_1998S_V1_0108	ASTH1215	pt had attack of asthma in last 12 months?
Asthma	FHS	CARE_SLEEP1_1998S_V1_0108	ASTHMA15	MD said pt had asthma?
Asthma	FHS	CARE_SLEEP1_1998S_V1_0108	ISTRD1	inhaled steroids for asthma
Asthma	FHS	CARE_SLEEP1_2003S_0108	ISTRD2	INHALED STEROIDS FOR ASTHMA
Asthma	FHS	CARE_SLEEP1_2003S_0108	OAIA2	(LEUKOTRIENE RECEPTOR ANTAGONISTS AND INHIBITORS OF LIPO-OXYGENASE)
Asthma	FHS	CARE_SLEEP1_2003S_0108	hi201d	have asthma?
Asthma	FHS	CARE_SLEEP1_2003S_0108	hi201e	Current asthma: do you still have asthma?

Data Set	Variable	Definition
DERV1C1	HYPERT05	HYPERTENSION, DEFINITION 5
DERV2C1	HYPERT25	V2 hypertension, definition 5
DERV3C1	HYPERT35	V3 HYPERTENSION, DEF. 5
DERV4C1	HYPERT45	Hypertension, Definition 5
DERV1C1	HYPERT06	HYPERTENSION, DEFINITION 6
DERV1C1	HYPERT04	HYPERTENSION, DEFINITION 4
DERV2C1	HYPERT26	V2 hypertension, definition 6
DERV2C1	HYPERT24	V2 hypertension, definition 4
DERV3C1	HYPERT36	V3 HYPERTENSION, DEF. 6
DERV3C1	HYPERT34	V3 HYPERTENSION, DEF. 4
DERV4C1	HYPERT46	Hypertension, Definition 6
DERV4C1	HYPERT44	Hypertension, Definition 4
A4F08	A08HBP	HIGH BLOOD PRESSURE
B2F08	B08HBP	HIGH BLOOD PRESSURE
C1F08	C08HBP	HIGH BLOOD PRESSURE
D1F08A	D08HBP	HIGH BLOOD PRESSURE?
DFLWUP1	FY096HBP	HIGH BLOOD PRESSURE? - MON 96
DFLWUP1	FY108HBP	HIGH BLOOD PRESSURE? - MON 108
E1F08	E08HBP	HIGH BLOOD PRESSURE?
F1F08	F08HBP	HIGH BLOOD PRESSURE?
FAMILY15_LAD_LONG	htn	HTN: abnormal bp (sys GE 140 or dia GE 90) or meds
FAMILY15_LAD_LONG	htndx	HTN: self report of MD dx of HTN
FAMILY15_LAD_LONG	htnx	HTN: self report of MD dx of HTN or sys GE 140 or dia GE 90 or meds
BASEBOTH	HYPER	CALCULATED HTN STATUS
YR10	HYPER	CALCULATED HTN STATUS
YR3	HYPER	CALCULATED HTN STATUS
YR4	HYPER	CALCULATED HTN STATUS
CARE_EX1_1S_V3_0108	A70	HISTORY OF HYPERTENSION
CARE_EX1_2S_V1_0108	B373	HYPERTENSION-ON TREAT OR ELEVATED BP
CARE_EX1_3S_V1_0108	C332	HYPERTENSION
DERIVE05	HTN017	Hypertension Status Per JNC7
MESA_EXAM_1	HIGHBP1	HYPERTENSION: SELF-REPORT
MESA_EXAM_1	HTN1C	Hypertension by JNC VI (1997) criteria
MESA_EXAM_2	HTN2C	Hypertension by JNC VI (1997) criteria,
MESA_EXAM_3	HTN3C	Hypertension by JNC VI (1997) criteria,
MESA_EXAM_4	HTN4C	Hypertension by JNC VI (1997) criteria,

Phenotype	Data Set	Variable	Definition
HTN med	DERV1C1	HYPTMD01	HYPERTENSION LOWERING MED. USE, DEF. 1
HTN med	UC480602	HYPTMDCODE01	HYPERTENSION LOWERING MEDICATION WITHIN PAST 2 WEEKS (V1)
HTN med	DERV2C1	HYPTMD21	Hypertension Meds (Self reported)
HTN med	DERV3C1	HYPTMD31	V3 HYPERTENSION MEDICATIONS, DEF. 1
HTN med	DERV4C1	HYPTMD41	V4 Hypert Med in Past 2 Wks: Self-rptd
HTN med	GW000605A	HYPTMDCODE41	HYPERTENSION LOWERING MEDICATION WITHIN PAST 2 WEEKS (V4)
HTN med	B2F08	B08BPMED	EVER TAKEN MEDS FOR HBP
HTN med	B2F09MHB	B09HBNM	NAME OF HBP MED
HTN med	B2F09MHB	B09HBNOW	TAKING HBP MED NOW?
HTN med	C1F08	C08HBNOW	CURRENTLY TAKING HBP MEDICATION
HTN med	C1F09MHB	C09HBNM	NAME OF HBP MEDICATION
HTN med	D1F08A	D08HBNOW	CURRENTLY TAKING MEDS FOR HBP
HTN med	D1F9MHBA	D09HBNM	NAME OF HBP MEDICATION
HTN med	E1F08	E08HBNOW	CURRENTLY TAKING MEDS FOR HBP
HTN med	E1F09MHB	E09HBNM	NAME OF HBP MEDICATION
HTN med	F1F08	F08HBNOW	CURRENTLY TAKING MEDS FOR HBP
HTN med	CARE_EX1_3S_V1_0108	C11	CALCIUM CHANNEL BLOCKERS
HTN med	CARE_EX1_3S_V1_0108	C12	BETA BLOCKERS
HTN med	CARE_EX1_3S_V1_0108	C13	ANTI-ARRHYTHMICS
HTN med	CARE_EX1_3S_V1_0108	C14	PERIPHERAL VASODILATORS
HTN med	CARE_EX1_3S_V1_0108	C16	OTHER HYPERTENSIVE DRUGS
HTN med	CARE_EX1_3S_V1_0108	C17	DIURETICS
HTN med	CARE_EX1_3S_V1_0108	C19	POTASSIUM SPARING DIURETICS
HTN med	CARE_EX1_3S_V1_0108	C20	RESERPINE DERIVATIVES
HTN med	CARE_EX1_3S_V1_0108	C21	ALDOMET
HTN med	CARE_EX1_3S_V1_0108	C22	CLONIDINE
HTN med	CARE_EX1_3S_V1_0108	C23	WYTENSIN
HTN med	CARE_EX1_3S_V1_0108	C24	GANGLIONIC BLOCKERS
HTN med	CARE_EX1_3S_V1_0108	C25	RENIN ANGIOTENSIN DRUGS
HTN med	DERIVE05	BPM01	Antihypertensive Medication
HTN med	MSRA	MSRA30A	30a: Past 2 wks took high blood pressure medications?
HTN med	MESA_EXAM_1	A2A1C	Angiotensin type 2 antagonists
HTN med	MESA_EXAM_1	A2AD1C	Combinations of angiotensin II antagonis
HTN med	MESA_EXAM_1	ACE1C	ACE Inhibitors without diuretics
HTN med	MESA_EXAM_1	ACED1C	ACE Inhibitors with diuretics
HTN med	MESA_EXAM_1	ALPHA1C	Alpha-blockers without diuretics
HTN med	MESA_EXAM_1	ALPHAD1C	Alpha-blockers with diuretics
HTN med	MESA_EXAM_1	BETA1C	Beta-blockers without diuretics
HTN med	MESA_EXAM_1	BETAD1C	Beta-blockers with diuretics
HTN med	MESA_EXAM_1	CCB1C	Any calcium-channel blocker = CCIR or CC
HTN med	MESA_EXAM_1	DIUR1C	Any diuretic
HTN med	MESA_EXAM_1	HTNMED1C	Hypertension Medication

Challenges of retrospective harmonization

- Time consuming
- Differing levels of ability to “harmonize”
- Inconsistent measuring units and/or definitions
  - Sometimes even within study
  - Sometimes not enough documentation to figure it out
- Impact of medications (“era” effects – can have profound impact)

Phenotypes in dbGaP

- Data submitted often limited to the “primary” study variables
  - Ancillary studies often have important phenotype or exposure information, but may not be involved in or aware of genetic efforts
  - Additional visits for existing cohorts may have variables of interest

Recommendations
Retrospective phenotype harmonization

- Develop panel of “harmonized” phenotypes and exposures of important measures

  - common variable names

  - common units of measure and definitions (to the extent possible)
- Need to ensure there is adequate variable definitions for every variable in dbGaP
- Year of visit and/or sub-cohort information should be clearly documented for every variable
- Studies should flag/note variables with “special issues”

Recommendations
Retrospective phenotype harmonization (contd.)

Identify point person or committee:

- Respond to questions from studies about this process
- Ensure that studies provide information to dbGaP in a standardized way
- Obtain input from phenotypic experts and cohorts to identify specific composite variable definitions

Recommendations
Additional phenotypes from existing samples

Survey of existing studies with genetic data in dbGaP for additional phenotypes and/or exposure data available from

- Ancillary studies
- Additional visits for longitudinal studies

Pros/Cons:

- Pros: Relatively cheap and fast
- Cons: Many variables only on subsets; not standardized

Recommendations
Prospective data collection

Collect new phenotypes and/or exposures for participants in existing studies

Pros:

- Input on panel of measures to collect
- Could be standardized with respect to definitions, units, assays, variable names, etc.

Cons:

- Often requires additional visit(s)
- Requires resources (funding support)
- Consideration for burden on participants

Leverage existing harmonization and standardization efforts

- Existing and previous consortia

Harmonization:

- GENEVA
- CARe
- CHARGE
- NHLBI PFINDR

Standardization of variables:

- PhenX Toolkit (https://www.phenxtoolkit.org/)
- NIA