LABOR AND POPULATION

Toward Development of a Synthetic Cohort of Phenotypic and Genotypic Data

Social Science and Genetics Association Consortium

October 29, 2011

Jinkook Lee, Erik Meijer, & Bas Weerman

Scientific opportunity for a synthetic cohort with harmonized data is ripe

- **The large collection of cohorts studies that have data on exposures throughout the life course have added genetic information.**
- **Open sharing policies through dbGap have created databases of extensive genotype data linked to phenotypes, but phenotypes are incompletely harmonized.**
- **Through measurement calibration, harmonized phenotype data can be developed, then linked to genotype data**

Tasks to be accomplished for the creation of a synthetic cohort

- **Catalog the existing cohorts with genetic information**
- **Develop analytical framework**
  - **Measurement calibration to create crosswalks between phenotype measures**
  - **Sampling, population representativeness**
- **Develop infrastructure for data sharing**

Catalog HRS family of surveys:
Mega meta data repository: https://metadata.rand.org

RAND Survey Meta Dat...

https://mmicdata.rand.org/megametadata/

Welcome to the Meta Data Repository | Register | Login

RAND SURVEY META DATA REPOSITORY

search the repository

or try advanced search

Home | Search Studies | Browse Studies | Obtain Study Data | Contextual Statistics | Papers | Help

Welcome to the Survey Meta Data Repository

Search the Meta Data Repository

search | search tips

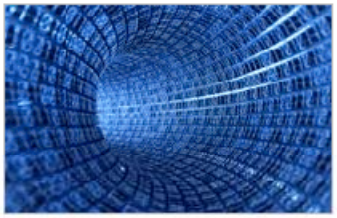
or try advanced search, including by topic

OR...

- Browse Survey Meta Data
- View Identically Defined Variables for Cross-Country Studies
- Explore General Statistics by Country and Year

About the Repository

The Survey Meta Data Repository is a collection of HRS-family survey data. It includes a digital library of survey questions, a search engine for finding comparable questions across the surveys, and a set of identically defined variables for cross-country analysis.



Studies in the Repository

RAND harmonized datasets for cross country analysis:

- RAND Health and Retirement Study (RAND HRS)
- RAND Harmonized English Longitudinal Study of Ageing (RH ELSA)
- RAND Harmonized Korean Longitudinal Study of Aging (RH KLoSA)
- RAND Harmonized Survey of Health, Ageing, and Retirement in Europe (RH SHARE)

External studies currently available in the repository:

- Health and Retirement Study (HRS)
- English Longitudinal Study of Ageing (ELSA)
- Survey of Health, Ageing, and Retirement in Europe (SHARE)
- Korean Longitudinal Study of Aging (KLoSA)
- Indonesia Family Life Survey (IFLS)
- China Health, Aging, and Retirement Longitudinal Study (CHARLS)
- Irish Longitudinal Study on Ageing (TILDA)
- Longitudinal Aging Study in India (LASI)
- Japanese Study on Aging and Retirement (J-STAR)
- Mexican Health and Ageing Study (MHAS)
- Study on Global Ageing and Adult Health (SAGE)

This project is funded by National Institute on Aging, National Institutes of Health (RC2 AG036619-01 and R01 AG030153)

This page last updated on May 10, 2011, 2:45pm PST · privacy policy · use and disclaimer

Measurement calibration

- **Different surveys often measure the same/similar concepts, but slightly differently; for combining surveys we'd need harmonized data.**
- **Sometimes, harmonized variables can be logically deduced from original variables; often not.**

Harmonized phenotype measures

Demographics	Number of household respondents; education: years of education; education: categorical summary; current marital status: with partnership, without partnership; marital history: number of marriage, never married, number of times divorced, number of times widowed; length of current marriage; length of longest marriage; religion (not available for ELSA); parental mortality: mother's current age or age at death; parental mortality: father's current age or age at death
Health	Self-report of health; whether health limits work; activities of daily living (ADLs): some difficulty; instrumental activities of daily living (IADLs): some difficulties; other functional limitations: raw recode; ADL summary: sum ADLs where respondent reports any difficulty; IADL summary: sum IADLs where respondent reports any difficulty; other summary indices: mobility, large muscle, gross fine motor activities; mental health (CESD score); doctor diagnosed health problems: ever have condition; doctor diagnosed health problems: memory-related disease; health behaviors: physical activity or exercise; health behaviors: drinking; health behaviors: smoking (cigarettes)
Financial and Housing Wealth	Net value of business; value of primary residence; value of all mortgage (primary residence); net value of primary residence; net value of real estate; net value of cars; net value of stocks, mutual funds, and investment funds; value of checking, savings, or money market accounts; net value of bonds and bond funds; net value of non-housing financial wealth; total family wealth (respondent & spouse)
Income	Individual earnings; income from employer pension or annuity; individual income from public pension; individual unemployment benefits or workers compensation (not available for ELSA); family capital income; family government transfer income; total family income (respondent & spouse)
Family structure	Number of people living in household; number of children; number of living siblings; number of living parents
Employment history	Currently working for pay; whether self-employed; labor force status; hours of work per week at current job; weeks worked per year at job; wage rate; level of physical effort at current job; years of tenure on current job; occupation code for current job; month and year last job ended

Measurement calibration

- **If we have a survey that administers the multiple versions to the same respondents, we can use modeling to construct a statistical "crosswalk" harmonize the measures by using either predicted or imputed values from the model.**
- **Differences may exist in response scales; further analysis, using anchoring vignettes.**
- **E.g., subjective well-being**

Sampling, population representativeness

- **The combination of separate datasets is not a representative sample of a well-defined and meaningful population.**
- **What assumptions are necessary to be able to claim that results are relevant to more general populations?**
- **This will typically be assumptions about conditional distributions being "universal"; the specific assumptions will be model-dependent.**

Infrastructure for data sharing

- Central secure server
- All data de-identified
- Linkage to genetic data based on identifiers
- Data storage encrypted
- Restricted data access

LABOR AND POPULATION