The (Un)intended Consequences of Comparative Data Sharing

Peter Granda
University of Michigan

CSDI Workshop March 27-29, 2014
Bethesda, MD

ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
50 years

Good Guidelines Never Go Begging

- Data Harmonization
  - Common theme in social science cross-cultural survey research but relatively new in the other sciences
  - Increased emphasis placed on data sharing has encouraged researchers in all fields to consider new approaches to facilitate collaborations
  - What role has this CCSG guideline played in this new environment?

Recent Examples of Harmonization with Health Data

- One closely related to the survey focus of the guideline: Profiles of public health agencies in the U.S.
  - Sponsored by the Robert Wood Johnson Foundation
- One more complex related to developing best practices to study the development and progression of chronic diseases
  - International effort; grant sponsored

Data Harmonization

Synoptic Analysis of ASTHO, NACCHO, NALBOH Surveys

DATA HARMONIZATION: UPDATES

The Data Harmonization Project seeks to promote PHSSR through facilitating the use of various data resources, particularly data from the ASTHO, NACCHO and NALBOH Profile studies. Funded by the Robert Wood Johnson Foundation, the Data Harmonization Project coordinates a variety of the mechanics of the three associations' separate Profile surveys of boards of health and health departments. The three Profile surveys provide the primary data on public health agencies, and the differing methods and methodologies employed to gather these data previously have hampered efforts to link the Profile surveys for research on the public health system.

To date, harmonization efforts have achieved a number of notable successes:

- All three associations conducted a census of their respective constituent agencies over a single 15-month period in 2010-11, allowing researchers to use data that draws from public health agencies over the same time period.
- The surveys utilized the same geographic and demographic questions, and efforts were made to link local boards of health and local health departments where possible. This can be difficult - a district health department services multiple counties and may work with multiple county-level boards of health, or a single regional board of health may oversee multiple local health departments. The 2010-11 NALBOH and NACCHO surveys clarify these connections between local health agencies and boards.
- The surveys used the same basic survey administration methodology and the same vendor to supply the online survey portals.
- The three associations began, and continue, to align their surveys' questions through uniform wording, uniform definitions of survey terms, and documentation of instances where terms or wording varies.
- In 2012, data on state health initiatives, local board governance issues, and local health department services will be available for each county. The Robert Wood Johnson Foundation provides it free to the public through its archiving agreement with the University of Michigan's Inter-University Consortium for Political and Social Research.

Current Information

ASTHO, NACCHO and NALBOH currently provide information on their websites about their survey materials and reports. ASTHO maintains a Data & Analysis page where users can download its questionnaire and codebook. NACCHO recently published the results of its 2010 National Profile of Local Health Departments. Lastly, users can visit the NALBOH publications page to

RESOURCES

Funding Opportunities
Frontiers in PHSSR
Reference Library
Data Harmonization
PHSSR Key Publications
PHSSR Quick Links
Technical Assistance

Harmonization Collaborators

- Support specific methods and development of software for data harmonization and integration

p3g PUBLIC POPULATION PROJECT IN GENOMICS AND SOCIETY

clsa élcv Canadian Longitudinal Study on Aging / Étude longitudinale canadienne sur le vieillissement

ICPSR A PARTNER IN SOCIAL SCIENCE RESEARCH 50 years

mælstrom

Maelstrom Research fosters the development of methods and software to support data harmonization and integration. Our international team provides open access to tools and metadata describing studies and harmonization programs as well as expertise to research projects making use of harmonized data.

Resources

Data harmonization methods, open-source software, federated analysis solutions, and publications.

Partnerships

Partnerships to develop new tools and provide infrastructure support to networks or individual studies.

Metadata

Repository of information facilitating the conception and achievement of data harmonization programs.

News

New publication: Data harmonization and federated analysis of population-based studies: the BioSHaRE project http://t.co/rofTgEqCly — *3 months 1 week* ago

The longitudinal study catalogue includes new features: description of study subpopulations and data collection events. — *4 months 1 week* ago

Mica 8.2 is now available. Release notes here: http://t.co/tpjk9FJKHD — *4 months 1 week* ago

Opal 1.14.0 is now available. Release notes here: http://t.co/difTOJQjzn — *10 months 2 weeks* ago

Opal 1.13.0 is now available. Release notes here: http://t.co/MUe1sSlzdY — *11 months 3 weeks* ago

Data Harmonization Research and Tools

Created in 2012, Maelstrom Research emerged from the activities of the Public Population Project in Genomics and Society, (P³G), the Biobank Standardization and Harmonization for Research Excellence in the European Union (BioSHaRE-EU), and the Canadian Longitudinal Study on Aging (CLSA).

Our goals are to foster a better usage of individual study data and encourage collaborative research through improved documentation, interoperability and integration of datasets. We conduct applied and methodological research, and develop methods and open-source software, principally to support retrospective (i.e. pre-existing) data harmonization and integration. These methods and software allow studies and harmonization projects to more efficiently:

- Collect and document data
- Evaluate the potential to harmonize and share existing data
- Process data from multiple studies under a common (i.e. harmonized) format
- Share harmonized data through secure central pooling or federated database systems

Harmonization Methods

The Maelstrom international team conducts methodological research to improve the reliability and validity of harmonized data and facilitate the data harmonization process as a whole.

Best practices methodology

Because of the complexity and heterogeneity of existing research data, a rigorous methodology must support retrospective harmonization to enable valid comparison and/or integration of information across studies. Given the increasing number of international collaborative research projects achieving secondary analyses of data, the need for best practices and guidelines facilitating data harmonization becomes all the more pressing. Such best practices and guidelines help ensure the validity of research making use of harmonized data.

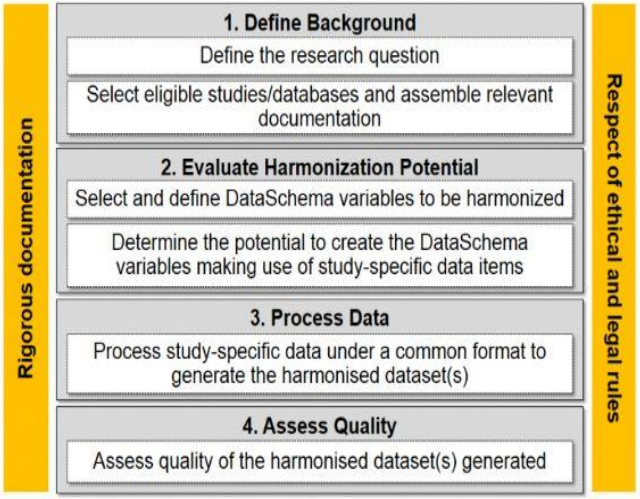
The general framework under which Maelstrom conducts retrospective harmonization projects involves the following steps (see figure):

1. Defining the background
   - Defining the research question(s) and scientific objectives and develop a protocol reflecting potential and limitations of the research project;
   - Selecting participating studies/databases and developing proper underpinning knowledge on each study/database;
2. Evaluating the harmonization potential
   - Identifying and defining the DataSchema: the variables that will serve as reference -or target- for the harmonization of study-specific data items;
   - Determining the potential for each study to construct each of the DataSchema variables and ascertaining the study-specific data items required to create these variables.;
3. Processing data
   - Processing study/database-specific data under a common format (DataSchema format) and generating the harmonized dataset (or independent study-specific harmonized datasets) to be used for data analysis;
4. Assessing quality
   - Assessing quality and homogeneity of the harmonized dataset(s) generated

To ensure reproducibility of the process, each step must be appropriately documented. Naturally, confidentiality and privacy of study participants' information must also be ensured throughout the process.

Maelstrom's harmonization methods are founded in the DataSHaPER approach, which supports the identification and definition of variables serving as reference for harmonization (i.e. the DataSchema), and the evaluation of compatibility of information collected across multiple studies.

For each of retrospective harmonization step outlined above, open-source web-based software is being developed to facilitate and streamline the entire process (for more information see the Software section). A more detailed retrospective harmonization best practices and guidelines document will be made available shortly.



Steps of a typical harmonization project

Expert advisory groups

DataSchema Catalogue

Domain Coverage

The DataSchema Catalogue section provides an overview of variables targeted for harmonization by different harmonization initiatives. A DataSchema is a list of core variables that serves as reference for the data harmonization and integration process by defining common format variables to be derived by participating studies. A DataSchema not only delineates *what* data a study needs to collect but also *how* it needs to collect it in order to generate harmonized data.

Variables found in existing DataSchemas can be re-used in new harmonization initiatives. The DataSchema Catalogue can thereby help new users identify and define variables targeted for harmonization to be used in their own data harmonization initiative. This then reduces the time required to generate new DataSchemas and facilitates collaboration between new and existing harmonization initiatives.

Name	Description
Asthma Respiratory DataSchema	This DataSchema provides a structure supporting data harmonization across respiratory cohorts. ***Under development***
Cognitive Measures DataSchema	This DataSchema assesses the quality of different statistical approaches used to integrate cognitive variables in individual participant data meta-analysis. ***Finalized***
Functional Somatic Symptoms DataSchema	This DataSchema supports the emergence of a series of collaborative research projects on somatic symptoms in Europe. ***Under development***
Healthy Obese Project DataSchema	This DataSchema assesses the prevalence of metabolically healthy obese and explore determinants, metabolic profiles and clinical outcomes of healthy obesity. ***Active***
Pediatric Multiple Sclerosis DataSchema	This DataSchema studies the relationships between genes and environmental and biological risk factors associated with Multiple Sclerosis in children.***Under development***
Physical Activity in Healthy Aging DataSchema	This DataSchema supports the validation of the DataSHaPER approach and investigates the relationship between physical activity and healthy aging. ***Finalized***
Retrospective Generic DataSchema	This DataSchema was created to estimate the harmonization potential across 53 large population-based cohorts across the world. ***Finalized***

Cognitive Measures DataSchema

Objectives

The Cognitive Measures DataSchema was developed as an output of a research project led Dr L. Griffith (Canada) and E. van den Heuvel (The Netherlands). The project applied different statistical methods to combine constructs such as cognition across four Canadian cohorts, and evaluated the impact of these methods on quality of combined data. Participating cohorts included: the Canadian Longitudinal Study on Aging (CLSA); the Canadian Community Health Survey (CCHS); the Canadian Study of Health and Aging (CSHA); and the Quebec Longitudinal Study on Nutrition and Successful Aging (NuAge)

Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis

Griffith L, van den Heuvel E, Fortier I, Hofer S, Raina P, Sohel N, Payette H, Wolfson C, Belleville S. Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis (March 2013) AHRQ Publication No. 13-EHC040-EF

Pubmed ID: 23617017

View article

Population Targeted

Elderly

Criteria used in the selection of DataSchema variables

- The variable is relevant to answer the specific research questions addressed;
- The variables is commonly measured across cohorts.

Dataset Type:

Harmonization dataset

Studies

No studies found.

Status

Finalized

Contact

Dr. Lauren Griffith (McMaster University)

Procedures used to develop harmonization guidelines for this research community

- Literature review/environmental scan including CCSG guideline
- Phone survey of existing methods and tools
- Series of meetings with experts
- Case studies testing proposed harmonization guidelines

Goals and Limitations of Harmonization

- Importance of meta-analyses involving large amounts of individual participant data
- Importance of harmonizing collection and management of bio-samples across biobanking studies
- OUTPUT = 'RETROSPECTIVE" HARMONIZATION

Why Output Harmonization?

- Implementing compatible protocols often not possible or desirable
- Health science investigators believe that evidence obtained through using different designs and measures is superior to repeating identical protocols
- Study-specific data collection methods sometimes required for technical and scientific reasons
- Impossible to predict harmonization potential at start of project

But still only one factor in the total harmonization process

Harmonization is a generic term for procedures used predominantly in official statistics that aim at achieving, or at least improving, the comparability of different surveys and measures collected. The term is closely related to that of standardization (see Sample Design and Questionnaire Design). Harmonizing procedures may be applied in any part of the survey life cycle, such as study design, choice of indicators, question wording, translation, adaptation, questionnaire designs, sampling, field work, data coding, data editing, or documentation.

THE FUTURE OF HARMONIZATION

ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
50 years