Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling

Dany Doiron[1], Parminder Raina[2], Vincent Ferretti[3], François L'Heureux[1] and Isabel Fortier[1,4]

1) Public Population Project in Genomics (P^3G), Montreal, Quebec, Canada
2) Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada
3) Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada
4) Research Institute – McGill University Health Centre, Montreal, Quebec, Canada

Correspondence: Public Population Project in Genomics, 2155 Guy Street, 4th floor, Montreal, QC, Canada, H3H 2R9
E-mail: ddoiron@p3g.org

INTRODUCTION

The quality and breadth of data and samples collected by cohort studies around the world are providing increasing opportunities to advance knowledge in chronic disease aetiology. However, very few individual cohorts provide the large sample sizes and accompanying statistical power required to investigate relatively rare diseases or complex gene-gene and gene-environment interactions [1-3]. Even large cohorts such as UK Biobank [4] and Kadoorie Study [5] will take at least a decade to generate sufficient numbers of incident cases of diseases such as rheumatoid arthritis or bladder cancer [2] and investigators making use of the data generated by these cohorts will face important statistical power limitations when exploring the interactions between genetic and environmental risk factors [2,6]. To increase the sample sizes available for statistical analyses, harmonization and pooling of information collected by different studies is increasingly being employed [2-3,7]. By making use of existing data sources, this approach can allow researchers to address important health issues in a relatively shorter and more cost-effective manner than through the creation of new large research infrastructures.

Over the past decade, an increasing number of organizations have provided networking opportunities, achieved collaborative research, and developed resources to support harmonization and pooling of data between studies and biobanks. The International Society for Biological and Environmental Repositories (ISBER; www.isber.org), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI; www.bbmri.eu), European Network of Genomic and Genetic Epidemiology (ENGAGE; www.euengage.org), Promoting Harmonisation of Epidemiological Biobanks in Europe (PHOEBE; www.phoebe-eu.org), Cohorts of Norway (CONOR) [8]; and the Public Population Project in Genomics (P^3G; www.p3g.org) are examples of organizations that have been pivotal in setting the groundwork to foster optimal use, exchange and combination of data and biological samples. However, the vast array of information collected by individual studies, the heterogeneity of study designs, collection tools and procedures, and the variability of ethical, legal, social and cultural contexts still pose major challenges and limit the potential to harmonize and pool data. Addressing these challenges requires that networks interested in collaborative research and data pooling have access to efficient data harmonization tools and methods. To date, no organization is offering an integrated suite of resources to support comprehensive and systematic procedures for data harmonization. A rigorous methodology supported by a suite of software applications dedicated to data documentation, harmonization, processing and pooling is required to complement the work done to date by international organizations working towards promoting collaborative research. The overarching objective of this paper is to present a framework for a harmonization platform aiming to foster development of such a suite of software and methods.

APPROACHES TO DATA HARMONIZATION

The goal of data harmonization is to achieve, or improve, compatibility of data collected from similar but independent sources in order to enable pooling or sharing of information [9]. Data harmonization can be achieved through different approaches, each with their own shortcomings and advantages [9-11]. Approaches to harmonization can be characterized as being either (a) prospective or retrospective, and (b) stringent or flexible.

When harmonization takes place prospectively (i.e. prior to data collection), a group of studies may decide to make use of identical data collection tools and procedures [12-13], referred to here as *stringent-prospective* harmonization. Under this approach, once consensus on the specific measures and standard operating procedures is attained amongst participating study investigators, and used to collect data, harmonization is relatively straightforward and data collected can be considered as standardized across different studies. However, as stringent-prospective harmonization allows no flexibility, it is not possible to adapt data collection tools and procedures to the specific cultural or scientific contexts of individual studies. Although this approach provides standardized and thus compatible data, imposing identical measures and procedures across a large number of studies is very challenging.

As an alternative to *stringent* harmonization, *flexible* harmonization has also been proposed by Fortier et al. [11]. Under *flexible-prospective* harmonization, investigators of a network will agree on common variables but allow some flexibility for individual studies to adapt data collection to their specific needs. However, to allow pooling, a rigorous assessment of heterogeneity is required and the information each study conveys needs to be deemed similar enough for the specific scientific purpose of the research program. The EPIC study (European Prospective Investigation into Cancer and Nutrition) [14] and Canadian Partnership for Tomorrow Project [15] are two examples of *flexible-prospective* harmonization initiatives. Coordination amongst participating data collection centres has been central to both of these longitudinal projects in order to ensure data compatibility. However, a certain level of flexibility in centre-specific assessment methods is allowed in order to better measure regional/cultural variations and to account for centre-specific scientific foci. Since both prospective approaches (stringent and flexible) require the implementation of identical or compatible procedures in emerging studies and the collection of new data, they necessitate a substantial amount of time and resources to generate results.

In addition to prospective approaches, consortia may wish to make use of *retrospective harmonization* to support pooling of existing data. Since very few established studies have used identical collection methods and procedures, retrospective harmonisation, by design, has to be flexible. Under this approach, information conveyed by each study has to be systematically assessed to determine the level of compatibility between studies [16]. This methodology typically involves defining a set of target variables to be pooled and determining the potential for each study to generate each target variable. The DataSHaPER [17] and the Integrated Public Use Microdata Series (IPUMS)-International [18] projects are examples of research initiatives in the health and social sciences that have made use of systematic retrospective harmonization approaches to assess compatibility of existing data. As demonstrated by these initiatives, the retrospective harmonization process demands time and access to appropriate expertise and adequate methodologies. Compared to prospective harmonization, the quantity of valid data that can be harmonized is limited and is directly linked to the heterogeneity of studies and collection tools. However, given the potential to make use of previously collected data, the main advantage of retrospective harmonization is that it can be achieved with relatively modest time and costs.

HARMONIZATION PLATFORM

Research funders are increasingly emphasizing the importance of data sharing and secondary analysis of publicly funded datasets [19-21] as a means to maximize the research potential of existing resources and provide greater returns on research investments. Moreover, many major funding bodies have recently agreed upon common principles and goals to increase the accessibility of health research data [22]. Ensuring that data is made widely available to the research community will stimulate the development of harmonization programs and hopefully promote the development of high impact and cost-effective research strategies.

To support the increasing need for data pooling, access to user-friendly software and standard harmonization procedures are required. However, no central resource is currently available to provide access to a repository of software, guidelines and technical support facilitating the documentation, harmonization, processing and pooling of data. The need for such a resource has shaped the proposal for a platform to support data harmonization efforts. Creation of this platform is facilitated by existing harmonization tools developed over the past half a decade under the umbrella of the Public Population Project in Genomics (P^3G) and its partner projects [16,23-25]. Initial development of the platform is supported by P^3G, BioSHaRE-EU (Biobank Standardization and Harmonization for Research Excellence in the European Union), the Canadian Longitudinal Study on Aging [26], and the Canadian Partnership for Tomorrow project [15].

A five-step methodology is proposed by the platform to provide a solid framework for *retrospective* harmonization. As a first step of this methodology, participating study attributes and type of information collected (e.g. data, samples) have to be described. Information such as study designs, sampling protocols, and data access policies must be formally documented in order to evaluate sources of study heterogeneity and feasibility of harmonization. All relevant information describing data elements and collection modes such as data dictionaries, questionnaires and standard operating procedures must also be catalogued to assess the compatibility of data collected by individual studies. Steps 2, 3, and 4 aim to respectively, identify variables that will serve as reference for harmonization, evaluate the potential for each study to generate these variables, and process individual study data under a common format. In step 2, target variables to be harmonized between studies are selected and defined. Step 3 involves the development of rules which determine the specific information each study needs to collect to generate target variables. These rules are then applied to conduct a systematic evaluation of the potential for each participating study to generate the target variables. In step 4, processing algorithms are developed and used to process study data into the common target variable format. Processed data is then aggregated in a derived database of harmonized variables. In the fifth and last step, pooled data is disseminated to investigators and appropriate statistical analyses are conducted. The whole process must necessarily be conducted in the utmost respect for ethical considerations related to the use of data from each participating study. Encrypting the data by changing participant identifiers or sharing aggregated data only are examples of methods used to ensure respect of privacy and confidentiality of individual-level data.

For each step outlined, open-source web-based software are being developed to facilitate and streamline the entire harmonization process. Construction of these software is facilitated by the availability of tools such as: the P^{3}G Catalogues, which allow documentation of studies; the DataSHaPER, which supports identification of variables serving as reference for harmonization and evaluation of the compatibility of information collected by multiple studies; and Opal and Mica, two software applications which support processing of data under a common format and facilitate the management and dissemination of harmonized databases. The suite of software is mainly written in Java and makes use of well established open source frameworks. In order to encourage uptake among potential users, all platform software will be licensed under the open source GPL3 licence and made freely accessible to the scientific community. Opal and Mica can already be downloaded from the OBiBa website (**O**pen Source Software for **Bi**o**Ba**nks; www.obiba.org). A website hosting the new integrated suite of software and guiding users through each step of the harmonization process will be available in the near future.

CONCLUSION

While a considerable number of harmonization initiatives have been established over the past decade, relatively little attention has been attributed to methodologies and tools used to achieve data harmonization and pooling in the literature. The harmonization platform initiative aims to offer a structured resource that supports achievement and documentation of rigorous data harmonization programs. As a methodological resource for study consortia and networks, the platform will create opportunities for collaborative use of research infrastructures and will promote innovative research within the epidemiological, public health, genomics, and social sciences communities. By facilitating data harmonization and pooling, we hope to help enhance the capacity of individual studies to improve the health and well-being of the population.

ACKNOWLEDGEMENTS

This work was supported by Genome Canada, Genome Quebec, and the Canadian Partnership Against Cancer. Additional funding came from the Tier 1 Canada Research Chair in GeroScience (P. R.), and the Raymond and Margaret Labarge Chair in Research and Knowledge Application for Optimal Aging (P. R.). V. F. is recipient of an Investigator Award from the Ontario Institute for Cancer Research, through the generous support of the government of Ontario.

REFERENCES

1. Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet* 2005. **6** (4): 287-98.
2. Burton PR, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009. **38** (1): 263-73.
3. Khoury MJ. The case for a global human genome epidemiology initiative. *Nat Genet* 2004; **36** (10): 1027-8.
4. Peakman TC, Elliott P. The UK Biobank sample handling and storage validation studies. *Int J Epidemiol* 2008. **37** (Suppl 1): i2-6.
5. Chen Z, et al. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol* 2005. **34** (6): 1243-9.
6. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* 2006; **7** (10): 812-820.
7. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol* 2009; **24** (12): 727-31.
8. Næss Ø, et al. Cohort Profile: Cohort of Norway (CONOR). *Int J Epidemiol* 2008; **37** (3): 481-485.
9. Granda P, Blasczyk E. *Data harmonization, guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Centre, Institute for Social Reseach, University of Michigan, 2010.
10. Granda P, Wolf C, Hadorn R. Harmonizing survey data. In: Harkness Ja, et al (eds). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, New Jersey: John Wiley & Sons, 2010: 315-332.
11. Fortier I, et al. Invited Commentary: Consolidating Data Harmonization – How to Obtain Quality and Applicability? *Am J Epidemiol* 2011; **174** (3): 261-264.
12. Craig CL, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003; **35** (8): 1381-95.
13. Cook DG, Shaper AG, Macfarlane PW. Using the WHO (Rose) Angina Questionnaire in Cardiovascular Epidemiology. *Int J Epidemiol* 1989; **18** (3): 607-613.
14. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 1997; **26** (suppl 1): S6-14.
15. Borugian MJ, et al. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. *CMAJ* 2010; **182**:1197-1201.

16. Fortier I, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010. **39** (5): 1383-1393.
17. Fortie, I, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol* 2011; **40** (5): 1314-1328.
18. Esteve A, Sobek M. Challenges and methods of international census harmonization. Historical Methods; 2003. **36** (2): 66-79.
19. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. *Bull WHO* 2010; **88**: 462-466.
20. Schofield PN, et al, Sustaining the Data and Bioresource Commons. *Science* 2010. **330** (6004): 592-593.
21. Piwowar HA, et al. Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med* 2008; **5** (9): e183.
22. Walport M, Brest P. Sharing research data to improve public health. *Lancet* 2011; **377** (9765): 537-539.
23. Knoppers BM, et al. Population Genomics: The Public Population Project in Genomics (P3G): a proof of concept? *Eur J Hum Genet* 2008; **16** (6): 664-665.
24. Wolfson M, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* 2010; **39** (5): 1372-1382.
25. OBiBa. *Open Source Software for Biobanks*. 2010 [cited 2010 July 30]; Available from: http://www.obiba.org/.
26. Raina P, et al. The Canadian Longitudinal Study on Aging (CLSA). *Can J Aging* 2009; **28** (3): 221-229.