

# Canadian Institutes of Health Research–Institute of Aging: Profile

---

## Harmonizing Data for Collaborative Research on Aging: Why Should We Foster Such an Agenda?

Isabel Fortier,<sup>1,2</sup> Dany Doiron,<sup>2</sup> Christina Wolfson,<sup>3</sup> and Parminder Raina<sup>4</sup>

<sup>1</sup> Research Institute of the McGill University Health Centre

<sup>2</sup> Public Population Project in Genomics

<sup>3</sup> McGill University

<sup>4</sup> McMaster University

Manuscript received: / manuscrit reçu : 05/12/11

Manuscript accepted: / manuscrit accepté : 11/12/11

Correspondence and requests for offprints should be sent to: / La correspondance et les demandes de tirés-à-part doivent être adressées à:

Isabel Fortier, Ph.D.

Research Institute – McGill University Health Centre

Allan Memorial Institute

1025 Pine Avenue West

Room P2.028

Montreal, QC H3A 1A1

(isabel.fortier@mail.mcgill.ca)

---

### Introduction

Despite longstanding awareness that the aging process is inextricably linked to the multifaceted changes occurring throughout an individual's lifetime (from the level of the cell, to individual psychological and behavioral factors, and to broad social contexts), a clear picture of all relevant interactions and their combined effects has not yet emerged. Furthermore, impacts of the complex interactions between biological, psychological, environmental, and social factors can take years to manifest, bringing additional challenges to the development of studies investigating this multifaceted picture. Despite these challenges, recent advances in science and technology (e.g., capacity to generate genotypes at very low cost, and the development of specialist measures adapted to elderly populations) now let us look forward to promising new avenues for research. However, to advance our understanding of the causal pathways leading to both adverse events and favorable outcomes for today's and tomorrow's seniors, it is essential to invest in infrastructures that enable the

ongoing collection of a wide range of information and are constructed to support the next generation of research potentials and requirements. The Canadian Longitudinal Study on Aging (CLSA) (Raina et al., 2009) is an example of a study that will provide such infrastructure in Canada.

The CLSA cohort will include 50,000 participants (45–85 years of age) to be followed over 20 years. In addition, the CLSA is planning to collaborate with a wide variety of national and international data collection efforts such as the Canadian Multicenter Osteoporosis Study ([www.camos.org](http://www.camos.org)), CARTaGENE ([www.cartagene.qc.ca](http://www.cartagene.qc.ca)), the EPIC Elderly Study (<http://epic.iarc.fr/research/elder.php>), and the Health and Retirement Study (<http://hrsonline.isr.umich.edu>) to conduct collaborative research addressing etiological and comparative policy analyses relevant to the aging population. The adoption of different designs and scientific targets in these cohorts offers unique opportunities to enable investigators representing different research infrastructures to learn from each other's experiences. However,

important scientific and policy advances will be achieved only if valid comparison or integration of study-specific data is feasible across cohorts and databases. Fortunately, the scientific promises of comparative and harmonized research are well recognized (National Research Council, 2001). An increasing number of countries are developing initiatives that support the creation of harmonized or compatible datasets to capture the multifaceted lives of older individuals and their families (Lee, 2007). Through various initiatives, Canada is also increasing efforts to foster harmonization of the rich existing and emerging national and international infrastructures that will help us to advance the science of aging.

### **Building Cohort Infrastructures to Support Research**

Cohort studies continue to be invaluable resources for the scientific community in a range of research fields. The information gathered in cohorts is critical for us to leverage Canadian health and social science, support training of the next generation of researchers, and ensure advancement of our understanding of the causal pathways of a broad variety of health and social outcomes. The impact of these studies, however, depends on the quality and breadth of information collected and generated. Cohort investigators must then face important financial, scientific, and technical challenges to ensure collection of comprehensive information on a range of diverse health outcomes as well as on risk and prognostic factors. In addition, the investigators must support regular follow-up of participant health and exposure profiles over extended periods.

Although substantial resources are needed for the development of such infrastructures, individual studies will often not have the statistical power or specific data items required to explore interactions and combined effects of the numerous factors affecting healthy aging. *Data linkage* and *data harmonization* are two complementary, but distinct, processes that may enhance the value of a given cohort or database. Data linkage can be described as “the bringing together from two or more different sources, data that relate to the same individual, family, place or event” (Holman et al., 2008, p. 767). In order to enrich databases with information not originally collected, many cohorts will link individual-level data to other data sources on the participant’s health (e.g., hospitalization databases or cancer registries), social environment (e.g., social deprivation indicators generated using census data), or physical environment (e.g., traffic and pollutant levels). Whereas linkage is used by cohort investigators to facilitate the combination of a variety of information on the same individuals, data harmonization, in contrast, essentially aims to achieve or improve comparability of similar measures collected

by separate studies or databases for different individuals (Granda & Blasczyk, 2010).

### **Harmonization as Support for Collaborative Research**

#### *What Is Data Harmonization?*

The simplest form of data harmonization can be conducted when different cohorts use identical (or standard) measures and data collection protocols (Hamilton et al., 2011). To implement this approach, investigators of emerging studies must agree on a common set of compatible questionnaires, measures, and standard operating procedures to collect data. Such an approach can be referred to as prospective harmonization. Essentially, applying compatible procedures at different collection sites leads to a high degree of homogeneity and reduces the manipulation and processing of data required to generate study-specific information under a common format and achieve harmonized data analysis. Under certain conditions, however, even when different measures and procedures are employed by cohort investigators, similar data items can be processed to allow valid harmonized data analysis. This flexible approach to harmonization is generally used to support comparison and/or integration of information provided by existing cohorts and is referred to as retrospective harmonization. As cohorts generally make use of different questionnaires, measures, and standard operating procedures to collect data, retrospective harmonization necessarily requires a rigorous assessment of the compatibility of information collected in individual studies (Esteve & Sobek, 2003; Fortier et al., 2010). In addition, comprehensive procedures must be undertaken to process individual cohort data under a common format and to ensure quality and validity of the harmonized database created. Even if it is technically challenging, retrospective harmonization is particularly valuable to optimize the utility of existing cohorts or databases.

#### *Why Harmonize Data?*

Data harmonization is essential to enable cross-national and -provincial comparative research and help inform policy decisions addressing the social and economic challenges of an aging population (National Research Council, 2001). Having access to compatible data allows researchers to properly explore similarities and differences across time and place and, for example, to investigate the impact of specific policy interventions on issues such as the cost and efficiency of health care delivery, work and retirement, and the health and well-being of aging populations. Harmonized data can thereby allow countries and provinces to learn from each other’s experience and better determine the impacts of specific policies or programs. Prospectively,

the development and implementation of international standardized instruments or classifications by organizations such as the International Labor Organization (ILO), the United Nations Educational, Scientific and Cultural Organization (UNESCO), the Organization for Economic Cooperation and Development (OECD), and the World Health Organization (WHO) have facilitated research involving cross-national comparisons. However, the scope of constructs covered by such standardized instruments is limited, and their successful implementation varies across countries. Retrospectively, initiatives such as the Integrated Public Use Microdata Series International (IPUMS-International) have been successful in documenting the compatibility of existing international data in order to support achievement of cross-national comparative analyses.

Furthermore, an increasing number of investigators in the health and social sciences are employing data harmonization to realize the many benefits that pooled data analysis offers. Ensuring data compatibility through harmonization provides the ability to integrate (or pool) health outcomes, life habits/behaviors, and other relevant data across cohorts. Such pooling results in larger sample sizes for data analysis. From the standpoint of disease aetiology, pooled datasets can provide the very large sample sizes required to investigate the interplay between genetic, lifestyle, environmental, and social factors, and to consider relatively rare health outcomes and risk factors. In this regard, the need for additional statistical power has often led investigators to employ data harmonization and pooling to achieve research initiatives on, for example, environmental exposure (Cardis et al., 2007), gene-environment interactions (Riboli & Kaaks, 1997), and aging (Anstey et al., 2010; Cooper et al., 2011).

Ensuring harmonization across partner studies can increase the use and extend the scientific impact of individual cohorts. When compared to building new studies involving thousands of participants, having access to harmonized data can permit generation of novel research findings relatively rapidly and at lower cost. Ultimately, harmonization facilitates the emergence of collaborative research initiatives and thereby minimizes the duplication of research efforts.

#### *What To Keep in Mind when Harmonizing Data*

The scientific impact of any harmonization program depends on the quality of the information collected by investigators of the partner cohorts. In addition, the potential to integrate or compare data across cohorts is related to the heterogeneity of the designs and methods used in each study. Naturally, the success of harmonization initiatives also depends on the ability to access the data and samples collected. This access, in turn,

depends on the limitations and restrictions imposed on data usage by cohort investigators (the use of the data may be restricted to studies examining the same conditions as the original study – neurological diseases, depression, or mobility, for instance).

Despite clear advantages to harmonization, some barriers pose important challenges. For prospective harmonization, these challenges are principally related to the difficulty of the scientific community agreeing upon, and implementing, standardized data collection and procedures. Indeed, in epidemiological and population-based research, the notion of repeating identical studies may not be viewed as providing evidence as strong as that obtained by conducting studies on the same topic using different designs. The principle here is that replication of results in different settings using different methodologies provides stronger support for findings.

Nevertheless, the prospect of facilitating comparison and integration of data has generated increasing interest in prospective harmonization. The Canadian Partnership for Tomorrow Project (CPTP), a consortium aiming to investigate risk factors of cancer and other chronic diseases, is a good example of one initiative that has incorporated flexible prospective harmonization in its design. In CPTP, investigators from each of five regional cohorts in Canada (Atlantic Path, CARTa-GENE, Ontario Health Study, The Tomorrow Project, and BC Generations Project) agreed on a core set of information to be collected by all data collection sites as part of their study protocols (Borugian et al., 2010). Moreover, even when harmonization of a wide range of information is not appropriate, study investigators will often consider using common measures or recognized standards to support collection of specific data. The International Physical Activity Questionnaire (IPAQ) (Craig et al., 2003) and the WHO (Rose) Angina Questionnaire (Cook, Shaper, & MacFarlane, 1989) are examples of standards that have been widely used.

Although the use of distinct methods and measures is essential to answer the specific scientific objectives foreseen by cohort investigators, the potential to synthesize information depends on heterogeneity across existing studies. This heterogeneity is related to a range of study-specific factors including: (a) the study design, time period, and duration of the follow-up; (b) the type of information and samples collected; (c) the specific tools, instruments, and standard operating procedures used to collect or generate data; and (d) the data coding and data management systems employed. Retrospective harmonization thus requires access to extensive documentation such as study protocols, questionnaires, standard operating procedures, data dictionaries, and instrument calibration procedures. Access to such documentation is essential to allow

proper evaluation of data compatibility across studies. These documents also support the extensive technical work required to develop and apply the algorithms used to process study-specific data under a common format; estimate the level of heterogeneity of the data processed for each cohort; and estimate impact of potential bias.

Cohort governance structures, rules for data and sample access, and components of the consent forms completed by the participants also influence the feasibility of harmonization programs. For example, with respect to data usage policies, harmonization will be appropriate only if participant consent in each cohort permits the planned harmonized analyses. For some cohorts, access to data by, or transfer of data to, a third party or a central infrastructure will be limited or prohibited. In such situations, harmonization will be possible, but analyses may have to be restricted to comparison across individual studies. Therefore, proper evaluation of data usage policies and intellectual property conditions, as well as achievement of all procedures required for access to information, are intrinsic components of any harmonization program.

#### *What Resources Will Facilitate Harmonization and Collaborative Research?*

Working towards the development of Canadian standards for the collection of specific measures adapted to research in aging would certainly improve data harmonization potential and support the emergence of collaborative research initiatives.

Development of novel harmonization methods and resources is also required as an important step towards responding to the increasing needs of the Canadian research community. One essential resource would be a user-friendly web-based catalogue providing access to standard descriptions of Canadian study infrastructures and including information on (a) study design; (b) specific data and samples collected; and (c) potential for data and sample access. International initiatives such as the Public Population Project in Genomics (P<sup>3</sup>G; [www.p3gobservatory.org](http://www.p3gobservatory.org)), National Archive of Computerized Data on Aging (<http://www.icpsr.umich.edu/icpsrweb/NACDA>), Biobanking and Biomolecular Resources Research Infrastructure (<http://www.bbMRI.eu/>), and Integrative Analysis of Longitudinal Studies on Aging (<http://www.ialsa.org>) have already begun to develop such catalogues. These catalogues can be accessed by a broad range of researchers to (a) evaluate the interest in using individual cohort data to achieve research program objectives; (b) identify studies that could be part of specialized harmonization programs; and (c) access relevant questionnaires or standard operating procedures.

The construction of a catalogue describing Quebec-based studies has been recently funded by the Quebec Ministry of Economic Development, Innovation and Export Trade, but it would certainly be of interest to extend it to the other Canadian provinces. The Canadian Institutes of Health and Research (CIHR), along with other federal funding councils, clearly see the importance of data harmonization in Canada. They convened a meeting of harmonization experts from across the world in March 2011 to assess what will be required to support further development or establishment of harmonization resources in Canada. An important element of the conclusion reached at the CIHR meeting was that there is also an urgent need to invest in harmonization platform(s) that develop rigorous methods, tools, and software accessible to the scientific community and serving to support and facilitate harmonization across national and international databases.

## Conclusion

Harmonization is increasingly viewed by the research community as a very promising avenue to support advancement in health and social research. Good harmonization strategies can accomplish a number of objectives including: (a) the generation of comparable data across studies, across jurisdictions (provincial, national or international), and/or across measures repeated through time; (b) the augmentation of the scientific impact of individual cohorts and the optimization of the return on investments; (c) the emergence of collaborative research programs minimizing duplication of research efforts; and (d) specifically for retrospective harmonization, the generation of research projects relatively rapidly and at low cost, by making use of existing data. Fostering harmonization efforts is necessarily challenging, but provides a unique opportunity to increase the development of Canadian and international collaborative research that will result in improvements to the health and well-being of today's and tomorrow's seniors.

## References

- Anstey, K.J., Byles, J.E., Luszcz, M.A., Mitchell, P., Steel, D., Booth, H., et al. (2010). Cohort Profile: The Dynamic Analyses to Optimize Ageing (DYNOPTA) Project. *International Journal of Epidemiology*, 39(1), 44–51.
- Borugian, M.J., Robson, P., Fortier, I., Parker, L., McLaughlin, J., Knoppers, B.M., et al. (2010). The Canadian Partnership for Tomorrow project: Building a pan-Canadian research platform for disease prevention. *CMAJ: Canadian Medical Association Journal [journal de l'Association medicale canadienne]*, 182(11), 1197–1201.
- Cardis, E., Richardson, L., Deltour, I., Armstrong, B., Feychting, M., Johansen, C., et al. (2007). The INTERPHONE study: Design, epidemiological methods, and description of

- the study population. *European Journal of Epidemiology*, 22(9), 647–664.
- Cook, D.G., Shaper, A.G., & MacFarlane, P.W. (1989). Using the WHO (Rose) angina questionnaire in cardiovascular epidemiology. *International Journal of Epidemiology*, 18(3), 607–613.
- Cooper, R., Hardy, R., Aihie Sayer, A., Ben-Shlomo, Y., Birnie, K., Cooper, C., et al. (2011). Age and gender differences in physical capability levels from mid-life onwards: The harmonisation and meta-analysis of data from eight UK cohort studies. [doi:10.1371/journal.pone.0027899]. *PLoS ONE*, 6(11), e27899.
- Craig, C.L., Marshall, A.L., Sjostrom, M., Bauman, A.E., Booth, M.L., Ainsworth, B.E., et al. (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine and Science in Sports and Exercise*, 35(8), 1381–1395.
- Esteve, A., & Sobek, M. (2003). Challenges and methods of international census harmonization. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 36(2), 66–79.
- Fortier, I., Burton, P.R., Robson, P.J., Ferretti, V., Little, J., L'Heureux, F., et al. (2010). Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. *International Journal of Epidemiology*, 39(5), 1383–1393.
- Granda, P., & Blasczyk, E. (2010). *Data harmonization guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.
- Hamilton, C.M., Strader, L.C., Pratt, J.G., Maiese, D., Hendershot, T., Kwok, R.K., et al. (2011). The PhenX toolkit: Get the most from your measures. *American Journal of Epidemiology*, 174(3), 253–260.
- Holman, C.D., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., et al. (2008). A decade of data linkage in Western Australia: Strategic design, applications and benefits of the WA data linkage system. *Australian Health Review*, 32(4), 766–777.
- Lee, J. (2007). *Harmonization of aging surveys and cross-national studies of aging executive summary*. Meeting Report, Chiang Mai, Thailand. National Institute on Aging, Behavioral and Social Research Program.
- National Research Council. (2001). *Preparing for an aging world: The case for cross-national research*. Washington, DC.
- Raina, P.S., Wolfson, C., Kirkland, S.A., Griffith, L.E., Oremus, M., Patterson, C., et al. (2009). The Canadian Longitudinal Study on Aging (CLSA). *Canadian Journal on Aging [Revue canadienne du vieillissement]*, 28(Special Issue 03), 221–229.
- Riboli, E., & Kaaks, R. (1997). The EPIC Project: Rationale and study design. European Prospective Investigation into Cancer and Nutrition. *International Journal of Epidemiology*, 26(Suppl. 1), S6.