International Journal of Epidemiology, 2017, 103–115
doi: 10.1093/ije/dyw075
Advance Access Publication Date: 5 June 2016
Original article

Methodology

Maelstrom Research guidelines for rigorous retrospective data harmonization

Isabel Fortier,[1]* Parminder Raina,[2] Edwin R Van den Heuvel,[3] Lauren E Griffith,[2] Camille Craig,[1] Matilda Saliba,[1] Dany Doiron,[1] Ronald P Stolk,[4] Bartha M Knoppers,[5] Vincent Ferretti,[6] Peter Granda[7] and Paul Burton[8]

[1]Research Institute of the McGill University Health Centre, Montreal, QC, Canada, [2]McMaster University, Department of Clinical Epidemiology and Biostatistics, Hamilton, ON, Canada, [3]Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven, The Netherlands, [4]University Medical Center Groningen, Department of Epidemiology, Groningen, Groningen, The Netherlands, [5]McGill University, Centre of Genomics and Policy, Montreal, Montrreal, QC, Canada, [6]Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada, [7]University of Michigan, Inter-university Consortium for Political and Social Research (ICPSR), Ann Arbor, MI, USA and [8]University of Bristol, D2K Research Group, School of Social and Community Medicine, Bristol, UK

*Corresponding author. Research Institute of McGill University Health Centre, 2155 Guy Street, office 460, Montreal, QC, Canada. E-mail: ifortier@maelstrom-research.org

Accepted 16 March 2016

Abstract

Background: It is widely accepted and acknowledged that data harmonization is crucial: in its absence, the co-analysis of major tranches of high quality extant data is liable to inefficiency or error. However, despite its widespread practice, no formalized/systematic guidelines exist to ensure high quality retrospective data harmonization.

Methods: To better understand real-world harmonization practices and facilitate development of formal guidelines, three interrelated initiatives were undertaken between 2006 and 2015. They included a phone survey with 34 major international research initiatives, a series of workshops with experts, and case studies applying the proposed guidelines.

Results: A wide range of projects use retrospective harmonization to support their research activities but even when appropriate approaches are used, the terminologies, procedures, technologies and methods adopted vary markedly. The generic guidelines outlined in this article delineate the essentials required and describe an interdependent step-by-step approach to harmonization: 0) define the research question, objectives and protocol; 1) assemble pre-existing knowledge and select studies; 2) define targeted variables and evaluate harmonization potential; 3) process data; 4) estimate quality of the harmonized dataset(s) generated; and 5) disseminate and preserve final harmonization products.

Conclusions: This manuscript provides guidelines aiming to encourage rigorous and effective approaches to harmonization which are comprehensively and transparently documented and straightforward to interpret and implement. This can be seen as a key step

towards implementing guiding principles analogous to those that are well recognised as being essential in securing the foundational underpinning of systematic reviews and the meta-analysis of clinical trials.

Key words: Data harmonization, data integration, data processing, individual participant data, retrospective harmonization, meta-analysis

Key Messages

- A wide variety of initiatives use retrospective data harmonization as a keystone of their research work.
- Even when appropriate approaches are used, the terminologies, procedures, technologies and methods used vary markedly across initiatives.
- Building on the combined findings of a phone survey, expert workshops and case studies, we have developed, and here report, formal guidelines for retrospective harmonization comprising a series of essentials and interactive steps.
- The guidelines aim to encourage rigorous and effective approaches to harmonization, which are comprehensively and transparently documented and straightforward to interpret and implement.

Introduction

Collaborative research programmes co-analysing individual participant data across studies are central to contemporary health science. The rationales underpinning such an approach include ensuring: sufficient statistical power; more refined subgroup analysis; increased exposure heterogeneity; enhanced generalizability and a capacity to undertake comparison, cross validation or replication across datasets.[1–3] Integrative agendas also help maximizing the use of available data resources and increase cost-efficiency of research programmes.[1,4]

Co-analysis of data across multiple studies can be achieved in several ways, including: study-specific data analysis (independent analysis-by-study followed by meta-analysis of study-level estimates); pooled data analysis (data transferred to a central server and analysed as a collective whole); and federated data analysis (centralized analysis, but the individual-level participant data remain on local servers).[5,6] However, to ensure content equivalence across studies and minimize measurement/assessment error that can cause bias or impair statistical power,[7] all such approaches require use of harmonized data. Essentially, data harmonization achieves or improves comparability (inferential equivalence) of similar measures collected by separate studies.[8]

The use of compatible protocols to prospectively collect common measures undoubtedly facilitates harmonization.[9] However, implementation of a prospective approach is not always possible or suitable. Repeating identical protocols is not necessarily viewed as providing evidence as strong as that obtained by exploring the same topic but using different designs and measures. In addition, investigators often need, for technical or scientific reasons, to use study-specific data collection devices. Finally, it is almost impossible to foresee all future harmonization requirements when implementing a new study. Retrospective harmonization (i.e. harmonization after data collection) is thus often the only option to permit data integration.[10] Retrospective approaches have supported numerous, relatively small[11–15] as well as very large research programmes.[16–21] For instance, international human immunodeficiency virus (HIV) research networks[22–24] that integrate existing HIV-related data are crucial to support current and upcoming research needs and develop appropriate health policies in the field. However, the increasing number of such programmes stresses an imperative to ensure quality, reproducibility and transparency of the results produced.

In systematic reviews and meta-analyses, the validity of a review depends on the use of a rigorous and transparent methodology.[25] Whereas traditional or narrative reviews are useful when conducted properly, it is recognized that they can sometimes be of poor quality, biased or lead to inappropriate recommendations.[25] In the past decades, guidelines for the conduct and reporting of systematic reviews and meta-analyses have therefore been articulated and consistently updated by consensus of experts.[26–28] Such guidelines identify and provide a rationale for the steps required to conduct a rigorous review and are considered compulsory in preparing formal review articles. Ensuring the reproducibility and validity of harmonized data also demands rigorous procedures, which must be transparent if they are to be accepted as valid. However,

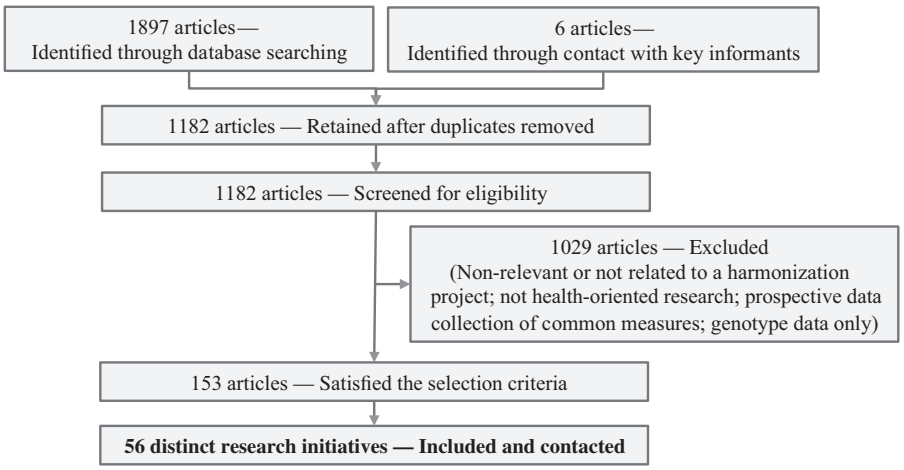

Figure 1. Flow chart describing selection of harmonization initiatives from literature search and references from key informants.

because no systematic guidelines are currently available, most investigators harmonizing data 'learn the hard way': repeatedly encountering significant pitfalls. Reports on retrospective harmonization procedures applied by research networks have been published,[13,20,29] and recently Rolland *et al.* described the process used at the Fred Hutchinson Cancer Research Center.[30] Although the paper provides a useful high -level overview of an approach comparable to the one we foster, it does not address the details of the component elements we developed in the past decade to formally underpin a generic harmonization guidelines applicable across disciplines. Nevertheless, Rolland's paper concurs with us that many researchers fail to reliably record basic information about the procedures used, decisions made and challenges encountered during the harmonization process and stresses the need to promote the creation of common and rigorous approaches to harmonization. Such guidance is essential for investigators new to the field to get to know issues to be addressed, and for groups reporting on their experience to identify the critical information to be made available if others are to properly estimate the quality of their work and learn from the successes and pitfalls they encountered.

The present paper provides an overview of the profile of key international initiatives and the approaches they use to harmonize data. It also details the guidelines developed in the past decade by Maelstrom Research and its partners, through a series of iterative reviews, consensus meetings and piloting within different harmonization programmes. The underlying goals of the guidelines are to foster a generic, but systematic, approach to retrospective data harmonization, and provide methodological guidance for investigators achieving harmonization and integration of pre-existing data. Detailed information and procedures are provided in supplementary materials, available as Supplementary data at *IJE* online.

Methods

The guidelines proposed are the results of three integrated activities carried out from 2006 to 2015. These comprise: a phone survey with major international initiatives to gather a clear overview of the current retrospective harmonization practices; formal workshops with experts to build the guidelines and overview its iterations; and a series of case studies to evaluate and pilot different iterations of the guidelines.

Exploring current practices

A literature search supplemented by references from key informants helped to identify initiatives having retrospectively harmonized individual participant's data across epidemiological studies (Figure 1). Research initiatives were selected instead of specific papers because most challenges faced and methods used ought logically to remain comparable across a given project, even if several publications are generated.

A literature search was undertaken in Medline®, EMBASE®, PsycINFO® databases and Google search engine using a range of keywords including 'harmonization, pooled analysis, multiple studies consortium and meta-analysis'. The search was supplemented by a review of the articles cited in the selected papers and references from key informants. Articles identified were defined as eligible if they were published from January 2000 to March 2014 and reported results from initiatives having: achieved retrospective harmonization and integration of individual participant data; integrated data from at least two epidemiological studies; and analysed data on risk factors and health outcomes.

A total of 1182 articles were retrieved after removal of duplicates. Screening of titles and abstracts led to the

identification of 153 articles satisfying all inclusion criteria. From those articles, 56 distinct initiatives conducting retrospective harmonization were identified and included in the survey. For each initiative, a key respondent was contacted by e-mail and, if not answering, re-contacted at least once by e-mail and once by phone to ask for participation. A semi-structured questionnaire was addressed to respondents agreeing to participate (lead investigators or a member of the research team responsible for data harmonization). The questionnaire addressed the aims, characteristics and infrastructure of the project, steps and methods applied to conduct the harmonization process, tools used and challenges faced. Descriptive analyses were conducted to explore the responses and compare characteristics of the participating and non-participating initiatives.

Developing and piloting of the guidelines

A series of international workshops were organized to gather input from experts and examine different iterations of the guidelines. More than 100 investigators from a variety of backgrounds (epidemiologists, computer scientists, statisticians, ethicists, data librarians, etc.), research interests (research on ageing, twins, cancer, diabetes, etc.) and over 15 countries provided input. Using an iterative review and consensus approach, a subgroup of core investigators brought together the results gathered through these meetings, established guiding principles and developed the Maelstrom Research guidelines. Iterative versions of the guidelines were produced and tested within a series of harmonization projects: Promoting Harmonisation of Epidemiological Biobanks in Europe;[31] Public Population Project in Genomics and Society;[32,33] Canadian Partnership for Tomorrow Project;[34] and Biobank Standardisation and Harmonisation for Research Excellence in the European Union.[35] More recently, the Biobanking and Biomolecular Resources Research Infrastructure–Large Prospective Cohorts[36] and the InterConnect project[37] also applied the guidelines proposed.

Results

Among the 56 study representatives contacted, 34 (60.7%) responded to the survey, two (3.6%) declined participation and 20 (35.7%) did not reply after three contacts. General characteristics of the 34 participating initiatives are presented in Table 1. A majority of the initiatives ($N = 25$; 73.5%) consisted of large consortia or collaborative networks addressing various research questions or generating harmonized datasets to serve longer-term goals; and 19 (55.9%) harmonized data only from studies of similar designs (e.g. all cohorts). Projects integrating data from multiple countries represented 76.5% ($N = 26$) of the initiatives. The number of individual studies within each initiative varied from 2 to 121, half of the initiatives ($N = 18$; 52.9%) harmonizing data from more than 10 studies. As for the total number of participants, 13 initiatives (38.2%) integrated data from more than 100 000 individuals, 15 (44.1%) from 10 000 to 100 000 individuals and six (17.6%) from less than 10 000 individuals. No differences were observed when the research areas, harmonization approaches or specific characteristics of the participating initiatives were compared with the non-participating initiatives (results not shown).

Infrastructures used to host and integrate data varied across initiatives. For the majority ($N = 26$; 76.5%), study-specific data were sent to a central location to permit integration and analysis. However, five (14.7%) initiatives included studies restricting data transfer, so data remained on study-specific servers. Three (8.8%) projects included some studies for which data were sent centrally and others in which they were hosted locally. When data were hosted locally, the harmonization process was generally rendered possible by sending the studies ready-to-use scripts to generate the harmonized variables and undertake a statistical analysis. Results generated were then combined using meta-analysis. However, two projects used a federated approach to remotely harmonize and analyse data hosted locally. Harmonization and processing were mainly achieved with regular statistical software ($N = 31$; 91.2%), except for three initiatives that used specialized software developed to support harmonization. As for data processing, algorithmic transformations (e.g. recoding of categories) was applied by all initiatives, and statistical models (e.g. regression analysis with standardized methods) were used by more than half ($N = 23$; 67.6%).

Respondents were asked to delineate the specific procedures or steps undertaken to generate the harmonized data requested. Sound procedures were generally described; however, the terminologies, sequence and technical and methodological approaches to these procedures varied considerably. Most of the procedures mentioned were related to defining the research questions, identifying and selecting the participating studies (generally not through a systematic approach), identifying the targeted variables to be generated and processing data into the harmonized variables. These procedures were reported by at least 75% of the respondents. On the other hand, few reported steps related to validation of the harmonized data ($N = 4$; 11.8%), documentation of the harmonization process ($N = 5$; 14.7%) and dissemination of the harmonized data outputs ($N = 2$; 5.9%).

A consensus approach was used to assemble information about pitfalls faced during the harmonization process

Table 1. General characteristics of the harmonization initiatives surveyed

Initiative (ref)	Countries	Number of studies	Study designs	Main topics
AirPROM[39a]	International	4	Cohort; Registry	Asthma and chronic pulmonary obstructive diseases
APCSC[40]	International	44	Cohort	Cardiovascular risk factors and stroke, coronary heart disease and total cardiovascular diseases
BioSHaRE[41]	International	8	Cohort; Cross-sectional	Metabolic risk factors and obesity
CHANCES[42a]	International	15	Cohort; Repeated cross-sectional	Cardiovascular diseases, diabetes mellitus, cancer, fractures and cognitive impairment
CLESA[11]	International	6	Cohort	Predictors of institutionalization, hospitalization and mortality
CLOSER[43a]	UK	9	Cohort; Panel	Broad topics (interdisciplinary research across longitudinal studies)
COSMIC[44]	International	19	Cohort	Cognitive measures and dementia
DYNOPTA[45]	Australia	9	Cohort	Cognitive measures, dementia and functional disabilities
ENGAGE[46]	International	36	Cohort; Cross-sectional	Cardiometabolic traits
ENRIECO[20]	International	19	Cohort	Environmental risk factors in pregnancy and early childhood
EPIC[47]	International	23	Cohort	Cancer and chronic diseases
EPOSA[13]	International	5	Cohort	Osteoarthritis
ERFC[17]	International	121	Cohort	Cardiovascular risk factors
EURALIM[21]	International	7	Cross-sectional	Diet and cardiovascular risk factors
GENEVA[29]	International	16	Observational study not specified; Clinical trial/intervention trial	Genetic and environmental risk factors for health and disease
GenomEUtwin[48]	International	8	Registry	Genetic and environmental risk factors for health and disease
HALCyon[49]	UK	9	Cohort	Physical capabilities
IALSA[50]	International	60	Cohort	Cognitive and physical capabilities, health, personality and well-being
INHANCE[51]	International	35	Case-control	Head and neck cancer
IDEFICS[12]	International	7	Cohort; Cross-sectional	Childhood obesity
IPD Meta-Analysis[52]	Canada	3	Cohort; Cross-sectional	Cognitive measures
LASA and NLSAA[53]	International	2	Cohort	Methodological differences in the harmonization of two longitudinal studies
MAGGIC[54]	International	31	Observational study not specified; Clinical trial/intervention trial	Survival of patients with heart failure with preserved or reduced left ventricular ejection fraction
MeRGE[55]	International	30	Case-control; Nested case-control	Restrictive diastolic filling pattern and mortality in patients post-acute myocardial infarction and patient with chronic heart failure
MORGAM[56]	International	28	Cohort; Repeated cross-sectional	Cardiovascular risk factors and outcomes

(Continued)

Table 1. Continued

Initiative (ref)	Countries	Number of studies	Study designs	Main topics
PAGE[57]	USA	8	Cohort; Cross-sectional; Nested case-control; Clinical trial/intervention trial	Genetic and environmental risk factors for health and disease
PROG-IMT[58]	International	50	Cohort; clinical trial/intervention trial	Cardiovascular events and carotid intima-media thickness
PPPSDC[16]	International	28	Case-control	Diet and cancer
PPSRH[59]	International	12	Cross-sectional	Self-rated health
RELATE[60a]	International	14	Cross-sectional; Panel	Early life conditions and older adult health
THLS[61a]	Finland	3	Cohort; Cross-sectional	Harmonization of clinical data between three studies
TLCS and HPHS[62]	USA	2	Cohort	Personality and health
TSC[63]	International	11	Cohort	Hypothyroidism, coronary heart disease and mortality risk
xTEND[64]	Australia	2	Cohort	Health and well-being

[a] This information was gleaned from the initiative's website or sources other than published articles.

AirPROM, Airway Disease Predicting Outcomes through Patient Specific Computational Modeling; APCSC, Asia Pacific Cohort Studies Collaboration; BioSHaRE, Biobank Standardisation and Harmonisation for Research Excellence in the European Union; CHANCES, Consortium on Health and Ageing: Network of Cohorts in Europe and in the USA; CLESA, Comparison of Longitudinal European Studies on Aging; CLOSER, Cohort & Longitudinal Studies Enhancement Resources; COSMIC, Cohort Studies of Memory in an International Consortium; DYNOPTA, Dynamic Analyses to Optimise Ageing; ENGAGE, European Network for Genetic and Genomic Epidemiology; ENRIECO, European initiative Environmental Health Risks in European Birth Cohorts; EPIC, European Prospective Investigation into Cancer and Nutrition; EPOSA, European Project on Osteoarthritis; ERFC, Emerging Risk Factor Collaboration; EURALIM, EURope ALIMentation; GENEVA, Gene Environment Association Studies; GenomEUtwin, GenomEUtwin; HALCYon, Health Ageing across the Life Course; IALSA, Integrative Analysis of Longitudinal Studies on Aging; IDEFICS, Identification and prevention of Dietary and lifestyle-induced health Effects In Children and infants; INHANCE, International Head and Neck Cancer Epidemiology; IPD Meta-Analysis, Harmonization of Cognitive Measures In IPD meta-analysis; LASA and NLSAA, Longitudinal Aging Study Amsterdam and Nottingham Longitudinal Study of Activity and Ageing; MAGGIC, Meta-analysis Global Group in Chronic Heart Failure; MeRGE, Meta-analysis Research Group in Echocardiography; MORGAM, MOnica Risk, Genetics, Archiving and Monograph; PAGE, Population Architecture using Genetics and Epidemiology; PROG-IMT, PROGression of Carotid Intima Media Thickness study; PPPSDC, Pooling Project of Prospective Studies of Diet and Cancer; PPSRH, Pooling Project on Self-Rated Health; RELATE, Research on Early Life and Aging Trends and Effects; THLS, National Institute for Health and Welfare (THL) studies (FINRISK cohorts, Health 2000 cohort and Helsinki Birth Cohort Study); TLCS and HPHLS, Terman Life Cycle Study and Hawaii Personality and Health Longitudinal Study; TSC, Thyroid Studies Collaboration; xTEND, eXtending Treatments, Education, and Networks in Depression study.

(Box 1), establish guiding principles and develop the guidelines. The iterative process (informed by workshops and case studies) permitted to refine and formalize the guidelines. The only substantive structural change to the initial version proposed was the addition of specific steps relating to the validation, and dissemination and archiving of harmonized outputs. These steps were felt essential to emphasize the critical nature of these particular issues.

The guidelines proposed include a series of essentials compulsory to the success of data harmonization (Box 2) and espouse an iterative process composed of a series of closely related and interdependent steps. An overview of the steps is provided below, but a comprehensive and structured description is presented as supplementary material. The Supplementary Material (available as Supplementary data at *IJE* online) lists, for each step and sub-step, The specific: aim; rational; procedures to be applied to ensure systematic process; issues to consider; resources that can be useful to facilitate the process; outputs generated; and an illustrative example. A checklist helping investigators to oversee the harmonization process is provided in Table 2.

Iterative steps toward data harmonization (see also Supplementary Material

Step 0: *Define the questions, objectives and protocol: develop a protocol reflecting the potential and limitations of the project. To ensure feasibility and reproducibility and to guide rational decision making, the objectives and research protocol must be clearly defined.*

Step 1: *Assemble information and select studies.*

Step 1a: *Document individual study designs, methods and content: ensure appropriate knowledge and understanding of*

Box 1. Overview of the potential pitfalls in data harmonization identified by the respondents and experts

- ensuring timely access to data;
- handling dissimilar restrictions and procedures related to individual participant data access;
- managing diversity across the rules for authorship and recognition of input from study-specific investigators;
- mobilizing sufficient time and resources to conduct the harmonization project;
- gathering information and guidance on harmonization approaches, resources and techniques;
- obtaining comprehensive and coherent information on study-specific designs, standard operating procedures, data collection devices, data format and data content;
- understanding content and quality of study-specific data;
- defining the realistic, but scientifically acceptable, level of heterogeneity (or content equivalence) to be obtained;
- generating effective study-specific and harmonized datasets, infrastructures and computing capacities;
- processing data under a harmonized format taking into account diversity of: study designs and content, study population, synchronicity of measures (events measured at different point in time or at different intervals when repeated) etc;
- ensuring proper documentation of the process and decisions undertaken throughout harmonization to ensure transparency and reproducibility of the harmonized datasets;
- maintaining long-term capacities supporting dissemination of the harmonized datasets to users.

Box 2. Absolute essentials required to achieve any successful harmonization project

Collaborative framework: a collaborative environment needs to be implemented to ensure the success of any harmonization project. Investigators involved should be open to sharing information and knowledge, and investing time and resources to ensure the successful implementation of a data-sharing infrastructure and achievement of the harmonization process.

Expert input: adequate input and oversight by experts should be ensured. Expertise is often necessary in: the scientific domain of interest (to ensure harmonized variables permit addressing the scientific question with minimal bias); data harmonization methods (to support achievement of the harmonization procedures); and ethics and law (to address data access and integration issues).

Valid data input: study-specific data should only be harmonized and integrated if the original data items collected by each study are of acceptable quality.

Valid data output: transparency and rigour should be maintained throughout the harmonization process to ensure validity and reproducibility of the harmonization results and to guarantee quality of data output. The common variables generated necessarily need to be of acceptable quality.

Rigorous documentation: publication of results generated making use of harmonized data must provide the information required to estimate the quality of the process and presence of potential bias. This includes a description of the: criteria used to select studies; process achieved to select and define variables to be harmonized; procedures used to process data; and characteristics of the study-specific and harmonized dataset(s) (e.g. attribute of the populations).

Respect for stakeholders: all study-specific as well as network-specific ethical and legal components need to be respected. This includes respect of the rights, intellectual property interests and integrity of study participants, investigators and stakeholders.

each study. Data comparability can be affected by heterogeneity of study-, population-, procedural- and data-related characteristics. Information related to design, time frame and population background will, for example, be required to evaluate study eligibility. In addition, information related to the specific data collected and, where relevant, standard operating procedures used will be essential to evaluate harmonization potential and guide data processing.

Table 2. Checklist helping to review the harmonization process

Step	Item	Description
Step 0: define the questions and objectives	1	The research question is well defined in term of population, exposure, comparator, outcome and timing
	2	The protocol takes into account questions related to feasibility (e.g. data access, realistic time-lines) and provides information required to guide the harmonization process
Step 1: assemble information and select studies		
Step 1a: document individual study designs, methods and content	3	Study-specific information gathered allows understanding study designs, time-line, population characteristics, data contents, standard operating procedures and ethico-legal requirements to access data
Step 1b: select participant studies	4	Studies are selected based on explicit selection criteria
Step 2: define variables and evaluate harmonization potential		
Step 2a: select and define the core variables to be harmonized (DataSchema)	5	The DataSchema variables are selected based on their relevance in answering the research question addressed, likelihood to be generated across a number of studies and, where relevant, input from experts
	6	The DataSchema variables are clearly defined, including their specific nature, format and acceptable level of heterogeneity
Step 2b: determine the potential to generate the DataSchema variables making use of study-specific data items	7	The potential (or not) for each study to create the DataSchema variables is assessed and documented
Step 3: process data		
Step 3a: ensure access to adequate study-specific data items and establish the overall data processing infrastructure	8	If harmonization is possible, the study-specific data items required to generate the DataSchema variables are made available in a computing infrastructure allowing data processing
	9	Quality of study-specific data items is assessed and considered adequate
Step 3b: process study-specific data items under a common format to generate the harmonized dataset(s)	10	Data processing is achieved using appropriate statistical models or processing algorithms
	11	Harmonized data are generated and algorithms or models used to process data are documented
Step 4: estimate quality of the harmonized dataset(s) generated	12	Quality and consistency of the harmonized data are assessed. Where appropriate, statistical models are applied to evaluate heterogeneity and potential bias
Step 5: disseminate and preserve final harmonization products	13	Harmonized data are available to approved users
	14	All information required to understand harmonization procedures and to analyse the harmonized data are accessible

Step 1b: *Select participant studies: select studies based on explicit criteria. To ensure consistency, designs of the studies included in a harmonization project must be similar enough to be considered compatible.*

Step 2: *Define variables and evaluate harmonization potential.*

Step 2a: *Select and define the core variables to be harmonized: outline the set of outcome, exposure and confounding variables that will serve as reference- or target-for the harmonization of study-specific data items and will serve to answer the research questions addressed (i.e. the DataSchema).*[38] *The nature of the DataSchema variables should reflect a satisfactory balance between targeting very precise concepts (e.g. identical questions) that optimize homogeneity, and acceptance of a greater degree of heterogeneity permitting inclusion of a larger number of studies. Explicit delineation and documentation are essential to inform the scientific meaning of the DataSchema variables*

Table 3. Impact of the level of information that is available from each study on the harmonization process

Level of information available	Location of study-specific individual participant data	Achievement of data processing	Application of the processing models (see Box 3)
Individual participant data	Transferred on a central server or remain on individual study's servers	Generally achieved centrally	All models
Aggregated data (e.g. means and frequencies)	Remain on individual study's servers	Achieved by study-specific teams. Can be centralized if a federated infrastructure is implemented	Limited to some models
Final results of statistical analysis	Remain on individual study's servers	Achieved by study-specific teams	Limited to some models

and facilitate proper decision making throughout the harmonization process. For example, the definition of the variable 'participant weight' should include its units (kg) and a record of the decision to accept (or not) both measured and self-reported weights. In many settings it is also crucial to define temporal proximity with other information of interest (e.g. collections of weight and physical activity).

Step 2b: *Determine the potential to generate the core (DataSchema) variables making use of study-specific data items: determine whether each study can construct-or not-each of the DataSchema variables as defined. It is necessary to evaluate which studies can provide data that enable generation of each of the DataSchema variables and to qualitatively assess the level of similarity between the study-specific and DataSchema variables. For example, only studies that measure participant weights could be viewed as being able to create the DataSchema variable 'Measured participant weight'.*

Step 3: *Process data.*

Step 3a: *Ensure access to adequate study-specific data items and establish the overall data processing infrastructure: ensure accessibility to, and quality of, the study-specific data items required to create the harmonized dataset. To allow data processing, it is essential to ensure availability and quality of all relevant study-specific data items. It is also a prerequisite to implement a data-processing infrastructure adapted to the context of the project and level of access to information allowed (access to individual participants' data, or access restricted to aggregated data or study-level results of statistical analysis) (Table 3). The data processing infrastructure will comprise both the study-specific (input data) and harmonized data generated (output data).*

Step 3b: *Process study-specific data under a common format to generate the harmonized dataset(s): convert the heterogeneous study-specific data items to DataSchema variables. Data processing is achieved using algorithms recoding study-specific data or statistical models based on contemporaneous analysis (Box 3). The procedures adopted will depend on the nature and format of the variables and on the data-processing infrastructure implemented.*

Step 4: *Estimate quality of the harmonized dataset(s) generated: understand the characteristics and utility of the harmonized dataset(s) generated. In order to ensure statistical analyses are run on data of acceptable quality, quality control procedures must be undertaken. The procedures should include verification of the algorithms or statistical models applied, and generation of basic quality checks and descriptive statistics (to evaluate consistency of the harmonized data across studies and explore potential influence of bias). When possible, procedures should be applied to test harmonization assumptions and assess heterogeneity.*

Step 5: *Disseminate and preserve final harmonization products: implement a sustainable infrastructure to preserve and disseminate harmonized data. In order for investigators not directly involved in the harmonization process to understand the steps and decisions taken, access to appropriate documentation must also be provided. This should include variable-specific metadata (e.g. harmonization potential, algorithms or statistical model used to process data) and description of the harmonization procedures applied. Ideally, all data and metadata should be made available in standard formats.*

Discussion

Achieving retrospective harmonization is necessarily challenging. This is particularly true for multidisciplinary initiatives like the ALPHA network (Analysing the Longitudinal Population-based HIV/AIDS data in Africa), including researchers from a variety of disciplines aiming to answer a broad range of research questions. Data harmonization is time consuming, and demands significant

Box 3. Examples of data processing models

Algorithmic transformation: Continuous and categorical variables, or both, with different but combinable ranges or categories (e.g. education level, household income)

Simple calibration model: Continuous metrics with calibration model (e.g. weight in kilograms or pounds)

Standardization model: Continuous constructs measured using different scales, with no known calibration method or bridging items (e.g. two independent memory scales)

Latent variable model: Continuous constructs measured using different scales, with no known calibration method but with bridging items (e.g. two memory scales, with some common items)

Multiple imputation models: Continuous or categorical constructs measured using overlapping scales permitting imputation of missing values (e.g. two overlapping scales measuring activities of daily living)

technical and scientific investments. Adding to the hurdle, harmonization is context-specific and the process generally needs to be repeated if new scientific questions arise. Furthermore, investigators need to ensure that data is only claimed to be harmonized if the process generated common variables of acceptable quality. Fortunately, a number of factors can facilitate the process and increase cost-effectiveness. For example, working within networks open to collaboration will facilitate sharing of data, resources and knowledge (Step 0). The identification of studies of interest (Step 1) and evaluation of the harmonization potential (Step 2) are facilitated by the existence of central metadata catalogues providing comprehensive information on existing study designs and content. Catalogues can also provide information useful to guide the development of prospective data collections. Data processing (Step 3) and the dissemination and preservation of the harmonized datasets (Step 5) are facilitated by usage of specialized software offering a secure, scalable and cost-effective computing environment. Access to comprehensive documentation about past harmonization initiatives can inform investigators about suitable processing models (Step 3) and quality control procedures (Step 4) and simplify achievement of harmonization in new, but similar contexts. Finally (Step 5), providing timely, appropriately governed access to harmonized datasets[38] helps to ensure effective return on the investments made and can act as a springboard to a wide range of additional research activities.

It is acknowledged that harmonization is important, requires thorough preparatory work, and has many elements that must be worked through carefully and systematically. However, many of the key steps to harmonization appear self-evident and straightforward even if time consuming to carry out. As a result, harmonization is often seen as a task that can easily be undertaken even by an 'enthusiastic amateur'. This precisely reflects early perspectives on systematic reviews, meta-analyses and clinical trials before formal guidelines and protocols were accepted as the norm. Unfortunately, no matter that many harmonization efforts are of high quality, the lack of collectively agreed terminologies and guidelines or protocols - emphasizing both documentation and quality control - makes it almost impossible for others to learn from those with practical experience, or even to objectively decide whether a particular harmonization project has been done well. To descend into cliché: reinvention of the wheel is all too common and, more seriously, the invention of non-functional wheels (e.g. with a missing axle) is far from rare. Virtually nobody with knowledge and experience in contemporary health science would argue that it would be preferable to undertake a clinical trial, a systematic review or a meta-analysis-particularly a first foray into any of these activities without following accepted guidelines. This ensures that no critical steps are missed, everything that others might later view as crucial information is properly documented and appropriate quality assurance criteria are evaluated. It is a central message of the current paper that harmonization should be viewed in precisely the same way and is the reason why we outlined these guidelines. Building robustly on the more detailed thinking laid out in supplementary materials (available as Supplementary data at *IJE* online), these guidelines have been applied to, and developed across, a number of harmonization initiatives that we believe have been successful. With this as a starting point, we encourage the scientific community, journal editors and funding agencies to debate and refine these guidelines with the aim of collectively agreeing on a generic protocol for data harmonization. Once this has been agreed, the harmonization procedures adopted in preparing a set of observational epidemiological studies for joint analysis can be held up to scrutiny against agreed best practice. Only then will harmonization initiatives - like systematic reviews, meta-

analyses and clinical trials - be reliably undertaken in an effective manner, and will such initiatives be properly evaluated in judging grant applications, reviewing papers or interpreting the published literature.

Supplementary data

Supplementary data are available at *IJE* online.

Funding

This work was supported by funds from the Quebec 'Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie'; the Canadian Partnership against Cancer, the European Union's Seventh Framework HEALTH-F4-2010 grant 261433 (BioSHaRE.eu); the Canadian Longitudinal Study on Aging, funded by the Canadian Institutes of Health Research and the Canadian Foundation for Innovation; the Ontario Institute for Cancer Research through funding provided by the Government of Ontario, Canada; and the Genome Canada and Genome Quebec funding agencies. The D2K (Data to Knowledge) programme of methods research in infrastructural epidemiology at the University of Bristol is supported by joint awards from the MRC and Wellcome Trust underpinning the ALSPAC project and the Biomedical Resource of the 1958 Birth cohort; MRC funding for the Welsh and Scottish Farr Institutes; ESRC funding of the CLOSER initiative and the BBMRI-LPC (Biobanking and Biomolecular Resources Research Infrastructure-Large Prospective Cohorts EU FP7, I3 grant). This work is also supported by a Tier 1 Canada Research Chair in GeroScience; and the Raymond and Margaret Labarge Chair in Research and Knowledge Application for Optimal Aging.

Acknowledgments

We would like to thank all the Maelstrom Research members, who have made this work possible, and particularly Catherine Hilgers, François l'Heureux and Mylène Deschênes. We would also thank the experts who participated in the survey and harmonization workshops, as well as collaborators from the DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) programme, CPTP (Canadian Partnership for Tomorrow Project), BioSHaRE.EU (Biobank Standardisation and Harmonization for Research Excellence in the European Union), P^3G (Public Population Project in Genomics and Society), BBMRI-LPC (Resources Research Infrastructure – Large Prospective Cohorts) and the InterConnect project.

Conflict of interest: None.

References

1. Gallacher JE. The case for large scale fungible cohorts. *Eur J Public Health* 2007;**17:**548–49.
2. Gallacher J, Hofer SM. Generating large-scale longitudinal data resources for aging research. *J Gerontol B Psychol Sci Soc Sci* 2011;**66(Suppl 1):**i172–79.
3. Frank C, Battista R, Butler L *et al*. *Making an Impact: A Preferred Framework and Indicators to Measure Returns on Investment in Health Research*. Ottawa, ON: CanadianAcademy of Health, 2009.
4. Dalziel M, Roswell J, Tahmina TN, Xiao Zhao. Impact of government investments in research & innovation: review of academic investigations. *Optimum ONline* 2012;**42**(2).
5. Wolfson M, Wallace SE, Masca N *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:**1372–82.
6. Gaye A, Marcon Y, Isaeva J *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014;**43:**1929–44.
7. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int J Epidemiol* 2003;**32:**51–7.
8. Granda P, Blasczyk E. *Data Harmonization. Guidelines for Best Practice in Cross-Cultural Surveys*. 3rd edn. Ann Arbor, MI: Survey Research Centre, Institute for Social Reseach, University of Michigan, 2011.
9. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: consolidating data harmonization - how to obtain quality and applicability? *Am J Epidemiol* 2011;**61:** 264–66.
10. Hutchinson DM, Silins E, Mattick RP *et al*. How can data harmonisation benefit mental health research? An example of the Cannabis Cohorts Research Consortium. *Aust N Z J Psychiatry* 2015;**49:**317–23.
11. Minicuci N, Noale M, Bardage C *et al*. Cross-national determinants of quality of life from six longitudinal studies on aging: the CLESA project. *Aging Clin Exp Res* 2003;**15:** 187–202.
12. Pigeot I, Barba G, Chadjigeorgiou, *et al*. Prevalence and determinants of childhood overweight and obesity in European countries: pooled analysis of the existing surveys within the IDEFICS Consortium. *Int J Obes* 2009;**33:**1103–10.
13. Schaap LA, Peeters GM, Dennison EM *et al*. European Project on Osteoarthritis (EPOSA): Methodological challenges in harmonization of existing data from five European population-based cohorts on aging. *BMC Musculoskel Disord* 2011;**12:**272.
14. Horwood LJ, Fergusson DM, Coffey C *et al*. Cannabis and depression: an integrative data analysis of four Australasian cohorts. *Drug Alcohol Depend* 2012;**126:**369–78.
15. Silins E, Horwood LJ, Patton GC *et al*. Young adult sequelae of adolescent cannabis use: an integrative analysis. *Lancet Psychiatry* 2014;**1:**286–93.
16. Smith-Warner SA, Spiegelman D, Ritz J *et al*. Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *Am J Epidemiol* 2006;**163:**1053–64.
17. Emerging Risk Factor Collaboration, Danesh J, Erqou S *et al*. The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. *Eur J Epidemiol* 2007;**22:**839–69.
18. Gehring U, Casas M, Brunekreef B *et al*. Environmental exposure assessment in European birth cohorts: results from the ENRIECO project. *Environ Health* 2013;**12:**8.

19. Boffetta P, McLerran D, Chen Y *et al.* Body mass index and diabetes in Asia: a cross-sectional pooled analysis of 900 000 individuals in the Asia cohort consortium. *PLoS One* 2011;**6:**e19930.
20. Hohmann C, Govarts E, Bergström A *et al.* Joint data analyses of European birth cohorts: two different approaches. *WebmedCentral EPIDEMIOLOGY* 2012;**3:**WMC003869.
21. Morabia A, Northridge ME, Beer-Borst S *et al. ftES. Harmonising Local Health Survey Data: The EURALIM Experience*. London: Springer, 2003.
22. ALPHA Network. 2016 http://alpha.lshtm.ac.uk/ (27 August 2015, date last accessed).
23. Zaba B, Calvert C, Marston M *et al.* Effect of HIV infection on pregnancy-related mortality in sub-Saharan Africa: secondary analyses of pooled community-based data from the network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA). *Lancet* 2013;**381:**1763–71.
24. Leroy V, Newell ML, Dabis F *et al.* International multicentre pooled analysis of late postnatal mother-to-child transmission of HIV-1 infection. Ghent International Working Group on Mother-to-Child Transmission of HIV. *Lancet* 1998;**352:** 597–600.
25. Akobeng AK. Principles of evidence based medicine. *Arch Dis Child* 2005;**90:**837–40.
26. Moher D, Liberati A, Tetzlaff J *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;**151:**264–69.
27. Stroup DF, Berlin JA, Morton SC *et al.* Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;**283:**2008–12.
28. Moher D, Shamseer L, Clarke M *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015;**4:**1.
29. Bennett SN, Caporaso N, Fitzpatrick AL *et al.* Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genet Epidemiol* 2011;**35:** 159–73.
30. Rolland B, Reid S, Stelling D *et al.* Toward rigorous data harmonization in cancer epidemiology research: one approach. *Am J Epidemiol* 2015;**182:**1033–38.
31. Norwegian Institute of Public Health. *PHOEBE - Promoting Harmonisation of Epidemiological Biobanks in Europe 2009*. http://www.fhi.no/artikler/?id = 73793 (27 August 2015, date last accessed).
32. Fortier I, Doiron D, Little J *et al.* Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol* 2011;**40:**1314–28.
33. Public Population Project in Genomics and Society. *Public Population Project in Genomics and Society 2014*. http://www.p3g.org/ (27 August 2015,date last accessed).
34. Canadian Partnership for Tomorrow Project. *Canadian Partnership for Tomorrow Project 2015*. http://www.partnershipfortomorrow.ca/ (27 August 2015, date last accessed).
35. Biobank Standardisation and Harmonisation for Research Excellence in the European Union. *BioSHaRE 2013*. https://www.bioshare.eu/ (27 August 2015, date last accessed).
36. Biobanking and Biomolecular Resources Research Infrastructure – Large Prospective Cohorts. BBMRI-LPC 2015 [cited 2015 August 27]. Available from: http://www.bbmri-lpc.org/ (27 August 2015, date last accessed).
37. InterConnect. InterConnect: a global initiative on diabetes gene-environment interaction 2015 [cited 2015 August 27]. Available from: http://www.interconnect-diabetes.eu/ (27 August 2015, date last accessed).
38. Fortier I, Burton PR, Robson PJ, *et al.* Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;**39**(5):1383–1393.
39. AirPROM. 2013 [cited 2015 August 27]. Available from: http://www.europeanlung.org/en/projects-and-research/projects/airprom/home (27 August 2015, date last accessed).
40. Woodward M, Barzi F, Martiniuk A, *et al.* Cohort profile: the Asia Pacific Cohort Studies Collaboration. *Int J Epidemiol* 2006;**35**(6):1412–16.
41. Doiron D, Burton P, Marcon Y, *et al.* Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol* 2013;**10**(1):12.
42. Boffetta P, Bobak M, Borsch-Supan A, *et al.* The Consortium on Health and Ageing: Network of Cohorts in Europe and the United States (CHANCES) project–design, population and data harmonization of a large-scale, international study. *Eur J Epidemiol* 2014;**29**(12):929–36.
43. Cohort & Longitudinal Studies Enhancement Resources. 2012. http://www.closer.ac.uk /(27 August 2015, date last accessed).
44. Sachdev PS, Lipnicki DM, Kochan NA *et al.* COSMIC (Cohort Studies of Memory in an International Consortium): an international consortium to identify risk and protective factors and biomarkers of cognitive ageing and dementia in diverse ethnic and sociocultural groups. *BMC Neurol* 2013; **13:**165.
45. Anstey KJ, Byles JE, Luszcz MA *et al.* Cohort Profile: The Dynamic Analyses to Optimize Ageing (DYNOPTA) project. *Int J Epidemiol* 2010;**39:**44–51.
46. Fall T, Hägg S, Mägi R *et al.* the role of adiposity in cardiometabolic traits: a mendelian randomization analysis. *PLoS Med* 2013;**10:**e1001474.
47. Riboli E, Hunt KJ, Slimani N *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;**5:**1113–24.
48. Muilu J, Peltonen L, Litton JE. The federated database - a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe. *Eur J Hum Genet* 2007;**15:**718–23.
49. Cooper R, Hardy R, Aihie Sayer A *et al.* Age and gender differences in physical capability levels from mid-life onwards: the harmonisation and meta-analysis of data from eight UK cohort studies. *PLoS One* 2011;**6:**e27899.
50. Hofer SM, Piccinin AM. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychol Methods* 2009;**1:**15–164.
51. Hashibe M, Brennan P, Benhamou S *et al.* Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *J Natl Cancer Inst* 2007;**99:**777–89.

52. Griffith L, van den Heuvel E, Fortier I *et al*. *Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-analysis*. Rockville, MD: Agency for Healthcare Research and Quality, 2013.
53. Bath PA, Deeg D, Poppelaars J. The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. *Ageing & Society* 2010;**30:**1419–37.
54. Meta-analysis Global Group in Chronic Heart Failure (MAGGIC).The survival of patients with heart failure with preserved or reduced left ventricular ejection fraction: an individual patient data meta-analysis. *Eur Heart J* 2012;**33:**1750–57.
55. Whalley GA, Gamble GD, Dini FL *et al*. Individual patient meta-analyses of restrictive diastolic filling pattern and mortality in patients post acute myocardial infarction and in patients with chronic heart failure. *Int J Cardiol* 2007;**122:** 207–15.
56. Evans A, Salomaa V, Kulathinal S *et al*. MORGAM (an international pooling of cardiovascular cohorts). *Int J Epidemiol* 2005;**34:**21–7.
57. Matise TC, Ambite JL, Buyske S *et al*. The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *Am J Epidemiol* 2011;**174:**849–59.
58. Lorenz MW, Polak JF, Kavousi M *et al*. Carotid intima-media thickness progression to predict cardiovascular events in the general population (the PROG-IMT collaborative project): a meta-analysis of individual participant data. *Lancet* 2012;**379:** 2053–62.
59. French DJ, Browning C, Kendig H *et al*. A simple measure with complex determinants: investigation of the correlates of self-rated health in older men and women from three continents. *BMC Public Health* 2012;**12:**649.
60. McEniry M, Moen S, McDermott J. Methods report on the compilation of the RELATE cross national data set on older adults from 20 low, middle and high income countries. University of Michigan, 2013. http://www.disc.wisc.edu/codebooks/qm512001.pdf (27 August 2015, date last accessed).
61. Silander K, Eklund N, Laukkanen M. *Protocol for retrospective harmonization of phenotype data in epidemiological sample collections*. 2012 (unpublished).
62. Kern ML, Hampson SE, Goldberg LR, Friedman HS. Integrating prospective longitudinal data: modeling personality and health in the Terman Life Cycle and Hawaii Longitudinal Studies. *Dev Psychol* 2014;**50:**1390–406.
63. Rodondi N, den Elzen WP, Bauer D, *et al*. Subclinical hypothyroidism and the risk of coronary heart disease and mortality. *JAMA* 2010;**304:**1365–74.
64. Allen J, Inder KJ, Lewin TJ *et al*. Integrating and extending cohort studies: lessons from the eXtending Treatments, Education and Networks in Depression (xTEND) study. *BMC Med Res Methodol* 2013; **122**.