Journal of Developmental Origins of Health and Disease

www.cambridge.org/doh

Life course of retrospective harmonization initiatives: key elements to consider

Isabel Fortier[1], Tina W. Wey[1], Julie Bergeron[1], Angela Pinot de Moira[2], Anne-Marie Nybo-Andersen[3], Tom Bishop[4], Madeleine J. Murtagh[5], Milica Miočević[6], Morris A. Swertz[7], Esther van Enckevort[8], Yannick Marcon[9], Michaela. Th. Mayrhofer[10], Jos Pedro Ornelas[11], Sylvain Sebert[12], Ana Cristina Santos[13], Artur Rocha[11], Rebecca C. Wilson[14], Lauren E. Griffith[15] and Paul Burton[16]

[1]Research Institute of the McGill University Health Centre, Montreal, QC, Canada; [2]Section of Epidemiology, University of Copenhagen, Denmark; [3]Department of Public Health, University of Copenhagen, Denmark; [4]Epidemiology Unit, University of Cambridge, England, UK; [5]School of Social and Political Sciences, University of Glasgow, Scotland, UK; [6]Department of Psychology, McGill University, Montreal, QC, Canada; [7]University Medical Center Groningen, University of Groningen, Netherlands; [8]Department of Genetics, University Medical Center Groningen, University of Groningen, Netherlands; [9]Epigeny, St. Ouen, France; [10]BBMRI-ERIC, Graz, Austria; [11]INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal; [12]University of Oulu, Finland; [13]Department of Epidemiology, Institute of Public Health of the University of Porto, Portugal; [14]Department of Public Health, Policy and Systems, University of Liverpool, Liverpool, England, UK; [15]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada and [16]Population Health Sciences Institute, Newcastle University, Newcastle-upon-Tyne, England, UK

Abstract

Optimizing research on the developmental origins of health and disease (DOHaD) involves implementing initiatives maximizing the use of the available cohort study data; achieving sufficient statistical power to support subgroup analysis; and using participant data presenting adequate follow-up and exposure heterogeneity. It also involves being able to undertake comparison, cross-validation, or replication across data sets. To answer these requirements, cohort study data need to be findable, accessible, interoperable, and reusable (FAIR), and more particularly, it often needs to be harmonized. Harmonization is required to achieve or improve comparability of the putatively equivalent measures collected by different studies on different individuals. Although the characteristics of the research initiatives generating and using harmonized data vary extensively, all are confronted by similar issues. Having to collate, understand, process, host, and co-analyze data from individual cohort studies is particularly challenging. The scientific success and timely management of projects can be facilitated by an ensemble of factors. The current document provides an overview of the 'life course' of research projects requiring harmonization of existing data and highlights key elements to be considered from the inception to the end of the project.

Introduction

Longitudinal pregnancy and birth cohort studies are powerful resources for exploring the developmental origins of health and disease (DOHaD). They provide the opportunity to investigate how parental and environmental factors occurring during early life (preconception, pregnancy, infancy, and childhood) influence fetal and child growth, developmental trajectories, and long-term susceptibility to disease.[1–3] However, the ability to address these questions depends on the access to data sets with large sample sizes, varied and heterogeneous exposure information, and long-term repeated follow-up. To achieve some of these requirements, the scientific community has increasingly begun to take advantage of the opportunity to combine data from existing cohort studies. A prerequisite for co-analysis of individual participant data (IPD) across studies is that the data formats and meanings are comparable, requiring, where possible, to harmonize study-specific data, i.e., to transform collected data to a common format.[4–6] Indeed, the number of such harmonization initiatives has increased exponentially during the past two decades, also driven by the call from the scientific communities and funders to make data findable, accessible, interoperable, and reusable (FAIR).[7]

The types of retrospective harmonization initiatives focusing on DOHaD research vary extensively. Risk factors (e.g., genetic background, mother's stress, air pollution), outcomes (e.g., birth weight, cognitive development, cancer) and life stages of interest differ from one initiative to the other. While most of these focus on specific research questions, initiatives like

Original Article

Cite this article: Fortier I, Wey TW, Bergeron J, Pinot de Moira A, Nybo-Andersen A-M, Bishop T, Murtagh MJ, Miočević M, Swertz MA, van Enckevort E, Marcon Y, Mayrhofer MT, Ornelas JP, Sebert S, Santos AC, Rocha A, Wilson RC, Griffith LE, and Burton P. (2023) Life course of retrospective harmonization initiatives: key elements to consider. *Journal of Developmental Origins of Health and Disease* **14**: 190–198. doi: 10.1017/S2040174422000460

Received: 29 June 2021
Revised: 17 February 2022
Accepted: 6 July 2022
First published online: 12 August 2022

Keywords:
Data harmonization; data processing; longitudinal data; cohort studies; Developmental Origins of Health and Disease (DOHAD)

Address for correspondence: Isabel Fortier, Research Institute of the McGill University Health Center, Montreal, QC, Canada.
Email: isabel.fortier2@mcgill.ca

Lifecycle,[8] ENRIECO,[9] RECAP-Preterm,[10] ReACH,[11] and Global Pregnancy CoLab COLLECT database[12] aim to address a broad range of objectives. Each of these initiatives also differs in magnitude, with the number of participating cohort studies varying from two or three[13] to over 20.[9,14,15] Finally, various governance, data warehouse, and data-sharing infrastructures can be adopted, while different methodological and operational approaches are used to handle data access, cleaning, harmonization, documentation, and co-analysis.

Although each initiative is unique, all are confronted by similar issues. Having to collate, understand, process, host, and co-analyze IPD from individual cohort studies is challenging. While it is the case for DOHAD research, it is also true for any initiative harmonizing and co-analyzing existing data across individual studies. First, it is often difficult to access structured documentation or obtain comprehensive information from local study teams related to cohort designs, participant follow-ups, and specific data items or samples collected/available. This can lead to important challenges in selecting appropriate data sources and ensuring the optimal use of data.[16] Second, organizational, ethical, and legal requirements typically restrict access to individual participant data or allow access, but only under specific conditions, often differing from one study to another. Therefore, time required to understand these rules and achieve data access procedures can be significant. Third, because of the heterogeneity of the information collected and the number and timing of study-specific data collection events, comparison and/or integration of data across studies present major methodological challenges. Fourth, data harmonization and co-analysis require access to secure and potentially sophisticated data sharing frameworks, methodological expertise, and specialized tools (e.g., standards, software), fundamentals that are not always accessible. Finally, for large-scale harmonization initiatives, maintaining the personnel, infrastructure and documentation required to support optimal long-term use of the harmonized data can be difficult.

While achieving optimal harmonization is and will remain challenging, scientific success and timely management of the initiatives can be facilitated by an ensemble of factors. In the following paper, we aim to provide an overview of the logistics and key elements to be considered from the inception to the end of collaborative epidemiologic projects requiring harmonizing existing data.

Methods

The paper content was generated using a consensus approach bringing together information from different sources. First, the experience of the authors in leading, collaborating in, or supporting over 50 harmonization initiatives from research networks in a broad range of research areas helped to build the core elements of the paper. Second, scans of the literature were used to identify additional challenges faced and solutions implemented by other harmonization initiatives. Third, the authors conducted a survey on a subset of 20 initiatives to answer specific questions raised and gather concrete examples of harmonization in practice. The survey included information about the challenges faced, the variables harmonized, the personnel and time required to achieve different tasks, and the data infrastructure implemented. Over 60 initiatives that, at various levels, informed the development of the paper are listed in Supplementary material S1. Consensus on paper content was achieved through a series of topic-specific meetings coordinated by Maelstrom Research,[17] ReACH (Research Advancement through Cohort Cataloguing and Harmonization),[11] EUCANconnect[18] and DataSHIELD[19] initiatives from 2018 to 2021. These meetings included cohort investigators and experts in various domains (e.g., epidemiologists, software architects, computer scientists, data analysts, statisticians, ethicists, lawyers, physicians, project coordinators, etc.).

Life course of harmonization initiatives

Harmonization initiatives can pursue divergent goals. Some have broad scientific objectives and engage numerous collaborators from various disciplines. Others are set up to answer very specific research questions and harmonize a limited number of variables across a limited number of studies. While each initiative is unique, investigators must generally develop and finance their research plan, implement a working environment adapted to their needs, and generate and preserved harmonized data to ultimately achieve statistical analysis. Figure 1 provides an overview of the conceptual workflow undertaken by harmonization initiatives. The workflow presented is complementary to the iterative harmonization steps proposed by the Maelstrom guidelines for retrospective harmonization.[4] To simplify reading, the workflow is described as linear. However, it needs to be adapted to the reality of each initiative, and a back-and-forth process is generally required to continually improve procedures and outputs generated based on the experience gained and results observed. An example of a specific harmonization initiative, the Prenatal Alcohol Exposure (PAE) project, is provided in Supplementary material S2.

Initiation

Conceive the project idea and proposal

The research questions addressed and research plan proposed by harmonization initiatives need to be Feasible, Interesting, Novel, Ethical, and Relevant (FINER).[20] As in any research project, this involves defining elements including the objectives to be pursued and research questions addressed; the specific exposures and outcomes required to answer the research questions; and the suitable population size and characteristics (e.g., mothers, children, age range, area of residence, primipara). In addition, the logistical, operational, technical, methodological, ethical, and legal elements specific to harmonization and co-analysis of IPD across independent cohort studies generally need to be outlined. Relevant elements to be considered depend on the objectives and scale of the initiative but can comprise defining the proposed governance model; the criteria used to select participating studies; the data infrastructure to be implemented; the operational and methodological approach to harmonization; and the statistical methods foreseen to validate and analyze harmonized data. Ideally, the protocol should also include a first evaluation of the harmonization potential across participating studies to outline the true potential of the project to answer the research questions addressed. Specialized catalogs documenting the design and content of mother-and-child studies are available to the research community to facilitate such evaluation.[8,10,11,21]

Apply for funding

Harmonization initiatives can require significant investment in time and expertise, both from the participating cohort studies and from the team coordinating the project, and the budget should reflect this reality. Costs relate to many factors including the scope of the initiative, the complexity of the governance and data infrastructure, the number of cohort studies involved, the quality of

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press

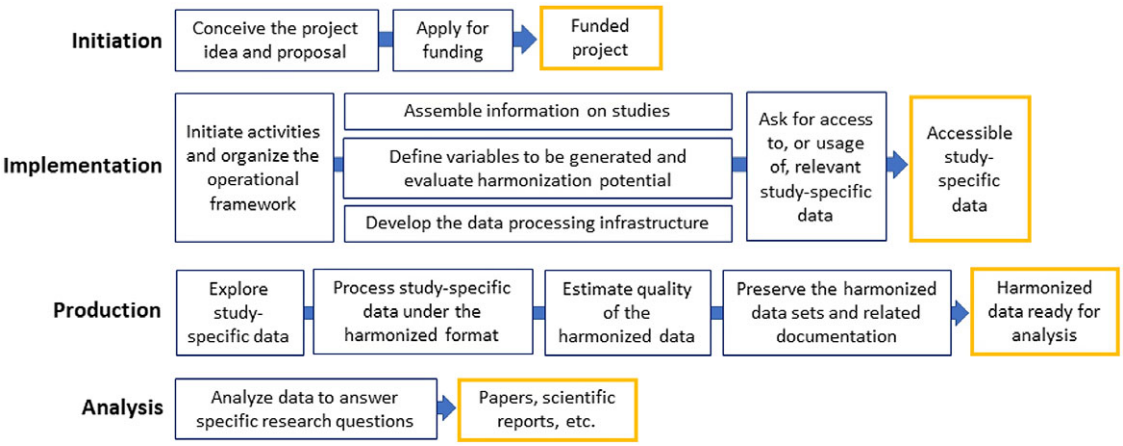

Fig. 1. Life course of a harmonization initiative.

study-specific data and metadata (information about data), and the number and type of harmonized variables to be generated across studies. Based on the survey results, relatively small-scale initiatives (e.g., aiming to harmonize 10–15 variables across four studies) can require 2–6 person-months (number of months, for the equivalent of one person working full time) to generate a validated harmonized data set, while large-scale initiatives (e.g., aiming to harmonize 150 variables across more than 10 studies) from 15 to over 80 person-months. Generally, most of the staff's time resources are dedicated to data inventory, cleaning, management, and processing.

For many initiatives, the time and costs required to obtain data should also be considered. Depending on the context, data access procedures (from submitting a demand for access to being ready to initiate harmonization) can take from a week to more than a year per study. In addition, study-specific data access fees might be applicable and may easily exceed 2,000€. Implementing a complex data infrastructure (e.g., distributed across several studies) could also be required by large initiatives. Setting up such an infrastructure demands time from technical experts (e.g., in data management and security) and can take several months. An overview of the timeline and costs of the PAE project is provided in Supplementary material S2. Finally, as in individual cohort studies, large-scale harmonization initiatives might also need long-term funding to ensure a sustainable platform and the maintenance and management of access to the harmonized data sets generated.

Implementation

Initiate activities and organize the operational framework

The success of large long-term harmonization initiatives often depends on building a collaborative, interdisciplinary team of experts and implementing flexible but efficient operational and governance models. Large initiatives bring together data users (investigators requesting harmonized data to achieve their research goals), data producers (stakeholders from participating cohort studies), and experts from specialized domains (e.g., longitudinal data analysts, ethicists, computer scientists, epidemiologists, clinicians). Team members generally come from different research groups, and each member brings its own professional background and level of expertise. To optimize projects operations and research outcomes, efforts might thus be required to build a unified approach and common understanding of various concepts.

Building consensus is not always necessary (e.g., in a small initiative with a narrow research question). However, to ensure efficient launch of activities, the team needs to rapidly delineate the practical requirements and operational details related to the research agenda, the data infrastructure to be implemented, and the data harmonization and analysis framework. Table 1 provides various examples of questions that could be addressed by the team.

Assemble information on studies

It is generally important to gather precise information about the characteristics of the studies actually enrolled so as to ensure the quality of the harmonized data set. Data comparability is affected by heterogeneity of the study-specific populations and data content. Access to comprehensive information on study designs, population characteristics, data collected, duration and timing of data collection events, and the standard operating procedures used can be required to confirm the eligibility of studies and estimate harmonization potential. Study-specific inclusion criteria are different for each research question addressed, but could include study-specific design (e.g., cohort studies), number of participants (e.g., at least 500 mothers recruited at baseline), sampling/recruitment frame (e.g., representative sample of pregnant women in a geographic area), years of recruitment (e.g., mothers recruited after 2010), number and frequency of data collection events (e.g., at least two data collection events during pregnancy), data/samples collected (e.g., smoking status, cord blood), specific time of collection (e.g., fasting glucose collected before 12 weeks of pregnancy), and potential to access IPD (e.g., IPD can be transferred to a central repository; or IPD can be analyzed remotely but cannot physically be shared/transferred or copied). Harmonization initiatives generally select studies before initiating the project. However, large-scale ones can address a broad range of research questions, each requiring inclusion of different subsets of studies presenting specific characteristics.

Define variables to be generated and evaluate harmonization potential

A DataSchema, or list of core variables (e.g., outcomes, risk factors) to be generated using study-specific data items, generally needs to be outlined. Selecting and defining these variables is probably the most scientifically challenging step of the harmonization process. It can require participation of researchers with specific domain expertise (e.g., nutrition, mental health), investigators or data

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press

Table 1. Examples of questions that could be addressed to help delineate analytical approach, practical requirements, and operations of a harmonization initiative

Research agenda
• What are the *specific* research questions?
• For each research question addressed, what are the population characteristics required to perform statistical analyses?
• For each research question addressed, what are the specific data items required to perform statistical analyses?
• What statistical models would be optimal to analyze data?
• When should statistical analysis be initiated (when will harmonized data need to be available)?
• What is the policy for authorship and recognition of partners' input?
Data infrastructure
• Is access to individual participant data (IPD) by users external to the cohort studies required to achieve analysis? If yes, is access to IPD acceptable for studies?
• Where will the study-specific and harmonized data sets be hosted (on the study-specific servers or transferred to a central server)?
• What will be the procedures required to access data?
• Which software will be used to support data management, harmonization, and analysis?
• What are the required characteristics of the data infrastructure to be implemented (e.g., security, access policy, servers' characteristics and capacity, backups, and persistent data storage requirements)?
Data harmonization and analysis framework
• Who will be responsible for making scientific decisions relating to the harmonization process and the specific variables to be generated?
• What will be the tools and standard operating procedures used to support data harmonization and quality control?
• Who will process study-specific data under the harmonized data format (the study-specific teams or centralized harmonization teams)?
• What will be the information generated to inform users about the data harmonization processes and outputs, and how will it be provided to users (e.g., decision rules, variables definitions, processing scripts, etc.)?

managers from member studies, and personnel with technical expertise in data harmonization. The information collected across cohort studies is generally not standardized, the wording of questions and measures used to evaluate the same constructs (e.g., level of physical activity, alcohol consumption) typically differ, and there is variation in the format, structure, and naming conventions of variables. In addition, the research questions addressed often require the analysis of longitudinal data (e.g., several data collection events during pregnancy or throughout the life of the child), but the collection events are likely to differ between and within studies in keyways that affect compatibility. As an example, Table 2 outlines information about mothers' binge drinking during pregnancy collected by five Canadian cohorts. Various DataSchema variables could be created using the data collected by these studies. These include, but are not limited to, a unique "binge drinking status during pregnancy" variable defined as binge drinking at least once during pregnancy (yes/no) or a "current binge drinking status" (yes/no) variable paired with the "time when binge drinking status was collected" (number of weeks of pregnancy). While there is rarely a unique or perfect solution, it is important to implement a rigorous and transparent decision-making process to select and define the DataSchema variables. The process should be guided by the scientific needs of the project, including specific requirements related to the statistical analysis planned.

Various elements can be used to define the DataSchema variables. These include the: nature of the variable (e.g., smoking status, highest completed level of education); value type (*e.g.*, integer, text, decimal); format, including the specific units (*e.g.*, kg) or list and description of the response options (e.g., 0 = Never; 1 = Almost never; 2 = Sometimes; 3 = Often; 4 = Very often); targeted individual or entity (e.g., the information is about the mother, father, neighborhood); targeted time period (e.g., first trimester, last 30 days, at birth); interdependence with other information needed to interpret the variable (e.g., birthweight and duration of pregnancy); acceptable sources of information (e.g., information obtained from questionnaire, registry, medical files); acceptable informants or who can provide the information (e.g., participant or proxy); acceptable time of collection (e.g., smoking status during first trimester of pregnancy can be collected at birth); acceptable question wording (e.g., binge drinking defined as 5 drinks or more); and acceptable procedures or devices used to generate the measure (e.g., weight needs to be measured, not self-reported by the participant).

Following the selection and definition of each DataSchema variable, it is possible to evaluate the potential (or not) for each study to generate it. According to the Maelstrom Research guidelines for retrospective data harmonization,[4] the harmonization potential is considered complete (fully achievable) if study-specific variables can directly generate the DataSchema variables (identical) or could be transformed to do so (compatible). The harmonization potential is however deemed impossible if study-specific variables cannot generate the DataSchema variables that have been defined (incompatible) or if the information is simply not collected (unavailable). It is also possible to define the harmonization potential as partial when it is possible to generate the variable but with an unavoidable loss of information. Evaluating the harmonization potential will often lead to adjustments in the initial DataSchema variable definition proposed (e.g., response options for binge drinking categories are adjusted to allow harmonization of more studies). Once finalized, the process will provide a clear overview of the harmonization potential across studies (which variables can be generated by which studies) and the study-specific data required to generate the DataSchema variables. Such processing and documentation can be generated using simple tools (e.g., Excel) or specialized resources.

Develop the data processing infrastructure

In parallel to documenting cohort studies and exploring harmonization potential, it is essential to determine the operating model and build the infrastructure required to host, manage, and analyze the data. For small-scale initiatives, the operating model may be simple and the data infrastructure rudimentary. For example, it could be limited to study-specific data sets uploaded on a server accessible by a single user who generates the DataSchema variables to answer a specific research question and never shares or reuses the harmonized data. However, a more sophisticated approach is often required.

Define the data harmonization and analysis operating models.

While various ethical, legal, methodological, and operating factors must be considered, data access and location are fundamental to inform the operating models to be implemented to harmonize and analyze data (Fig. 2). If transfer of study-specific IPD to external third parties is acceptable, data may be transferred to a central server and the harmonization process centralized.[22,23] But this is not always possible and may be unsuitable. Alternatively, study-specific data may remain on study-specific servers, and harmonized data generated by study-specific teams.[24] Each approach presents advantages and challenges (Table 3) and directly impacts operational decisions (e.g., number of servers required,

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press

Table 2. Example of information about frequency of binge drinking during pregnancy collected by five mother-and-child cohorts

	3D	AOF	APrON	Family	OBS
Data collection events	*Visit 1:* 8–14 weeks *Visit 2:* 20–24 weeks *Visit 3:* 32–35 weeks	*Visit 1:* 36 weeks	*Visit 1:* 1–13 weeks *Visit 2:* 13–28 weeks	*Visit 1:* 21–39 weeks	*Visit 1:* 12–16 weeks *Visit 2:* 28–32 weeks
Questions timeframe	*Visit 1:* Since becoming pregnant *Visits 2 & 3:* Since last visit	Since becoming pregnant	N/A	During pregnancy	*Visit 1:* Currently, during pregnancy *Visit 2:* Over the last 3 months
Definition of binge drinking	5 drinks or more on one occasion	5 drinks or more on one occasion	N/A	5 drinks or more in one day	4 drinks or more at same sitting or occasion
Format(s)	Continuous	Categorical	N/A	Categorical, Continuous	Categorical, Continuous
Response options or units	*Variable 1:* # days/week OR *Variable 2:* # days/month OR *Variable 3:* Total # days	*Variable 1:* Yes; No	N/A	*Variable 1:* Yes; No *Variable 2:* # times	*Visit 1, Variable 1*: One or more times a month; Less than once a month; None; Don't know; Prefer not to answer. *Variable 2:* #times/month *Visit 2, Variable 1:* 6 to 7 times a week; 4 to 5 times a week; 2 to 3 times a week; Once a week; 2 to 3 times a month; About once a month; 6 to 11 times a year; 1 to 5 times a year; Never; Don't know; Prefer not to answer

3D: 3D Study – Design, Develop, Discover; **AOF**: All Our Families; **APrON**: Alberta Pregnancy Outcomes and Nutrition; **FAMILY**: Family Atherosclerosis Monitoring in Early Life; **OBS**: Ontario Birth Study.

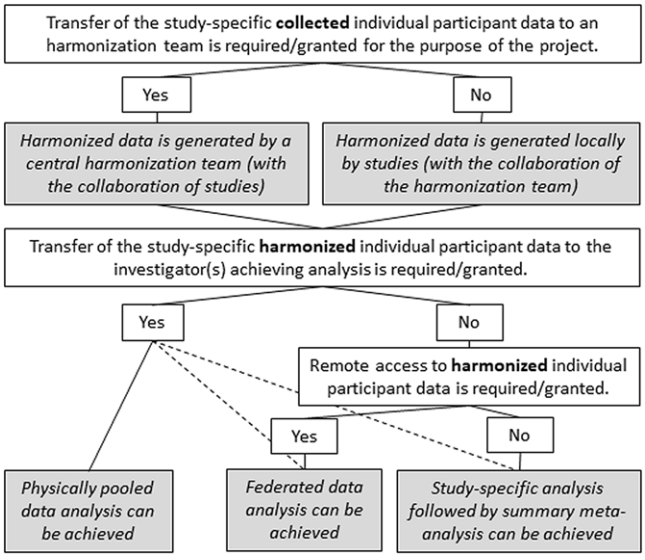

Fig. 2. Influence of the level of access to individual participant data on the data harmonization workflow.

distribution of personnel) and the data infrastructure required (e.g., type of access required, level of security, and computing capacities).

In turn, the possible operating models for statistical analysis are informed by the level of access to the harmonized IPD. Study-specific (analysis performed by studies followed by a meta-analysis of study-level estimates), pooled (data hosted on a central server and analyzed as a collective whole), or federated (centralized analysis, but the individual-level participant data remain on local servers) IPD analysis can be achieved (Fig. 2). Again, each approach presents advantages and challenges.[25–27] Study-specific IPD analyses followed by a meta-analysis of aggregate data (i.e., two-step IPD meta-analysis[28]) is often the approach selected.[29] The approach may reduce efforts to obtain and analyze data, as only aggregate data are required for combined analysis and as meta-analytical methods for aggregate data are well established. However, standardizing analyses among studies may require substantial effort, and statistical power and flexibility to explore interactive or heterogeneous effects (for example, across studies or subgroups) can be limited. In contrast, a pooled analysis approach (i.e., one-step IPD meta-analysis[28]) typically offers statistical power and flexibility, with the potential for greater insights into interactive or heterogeneous effects and interpretation of results (such as of pooled estimates).[14,30] However, it may necessitate high-performance processing environments to allow analysis of large amounts of data, and it often comes with substantial efforts to obtain access to IPD. The trade-offs between these first two approaches and strategies for choosing an approach have been discussed in detail elsewhere.[28,31] Finally, federated data analysis can represent a valid option.[26,32] The approach may support one- and two-stage meta-analyses, but it requires implementation of a distributed and interoperable data infrastructure supporting unified co-analysis of the harmonized data across studies. Additional information on these approaches is provided in Supplementary material S3.

Implement the data infrastructure. The data infrastructure provides the physical environment required to access, manage, process, document, and analyze data securely, but efficiently. As mentioned above, the infrastructure may be extremely simple, but large-scale initiatives often require implementation of complex computational environments. The nature of the infrastructure to be implemented is informed by factors including the type and volume of data needed (e.g., questionnaire data, genotypes, images), the statistical analyses foreseen, the location of study-specific and harmonized data, the type of access to IPD required by the various users, the hardware and software resources available to the initiative (and costs if needed to be acquired), the technical skills of the participating teams, the security requirements, and the need (or

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press

Table 3. Advantages and challenges related to processing collected data under the harmonized format centrally and by study-specific teams

Data harmonized	Advantages	Challenges
Centrally	• Facilitates implementation of standard data processing and decision-making across studies. • Facilitates standardization of quality control procedures. • Allows using diverse statistical models to generate harmonized data elements.	• Ensuring the harmonization team develops proper understanding of study-specific data sets and maintains close collaboration with studies. • Obtaining access to, and transfer of, data from each study. • Maintaining governance of the harmonized data generated respecting all study-specific requirements and policies.
By studies	• Facilitates data processing, as the study-specific teams are familiar with their data. • Simplifies procedures required to access study-specific data. • Ensures studies maintain full control of their data.	• Coordinating and ensuring quality and consistency of the data processing workflow achieved by the study-specific teams. • Funding at least one data manager per study and one coordinating the process across studies. • Possibilities to use statistical models to harmonize data elements are limited.

not) for long-term maintenance and potential scaling up of the infrastructure.

Data harmonization generally requires relatively limited computational power compared to statistical analysis. If statistical analysis is achieved on pooled data, the (internal or external) users analyzing data will require sufficient storage, memory, and processing power to deal with harmonized data from all studies. On the other hand, if analysis is performed by individual studies, computational requirements will be governed by the characteristics of each study data set. Obviously, all aspects of data security should be carefully considered. Proper access control to the data should be in place, and availability and integrity of the data should be ensured by backups, regular system maintenance, and proper monitoring. Where required, static data sets and backups should be encrypted, and there should be documented and auditable procedures for granting access to and/or transfer of data.

Ask for access to, or usage of, relevant study-specific data

Obtaining access or permission to use study-specific data is a prerequisite for initiating data processing. This is true even if data remain on local servers and are processed by the study-specific teams. The goals of ethical, bureaucratic, and technical procedures for data access governance are to protect the cohort study participants (ensuring study-consent stipulations are maintained), data producers (in some case intellectual property rights), and the study itself (to mitigate against reputational risk[33]). Data access committees are also responsible for maintaining adherence to supra-study regulations (e.g., the European Union's General Data Protection Regulation). Differences in regulatory environments and study-specific policies often mean that access procedures vary from study to study. Procedures may include submission of the project protocol, exchanges with members of scientific or data access committees, and completion of data transfer or privacy agreements. As harmonization initiatives need to access data from more than one study and few integrated (multistudy) access governance systems exist, significant delays are often encountered.[34] This is particularly true if independent applications for access need to be submitted to each study for each research question addressed. The data access process should thus be initiated as soon as possible and careful attention to the study-specific requirements and procedures is highly recommended. Providing a list of the exact variables required is often requested by the data access committees; to address the principle of data minimization, data access committees may check the variable list against the proposed research question for coherence. Preparation of this list can be informed by the result of the harmonization potential outlined above and needs to include all study-specific variables required to generate, understand, and validate the DataSchema variables and achieve the statistical analysis foreseen (e.g., all required confounders).

Production

Explore study-specific data

Once access is granted, data are generally prepared (preprocessing under a defined format might be required) and explored to ensure quality and deepen proper understanding of each study-specific data set. For example, the completeness, content, and format of the study-specific data can be verified. Issues observed at this stage often lead to adjustment of the harmonization potential estimated. Poor data quality may also lead to the exclusion of a data set.

If data is processed under the DataSchema format by the study teams, this step may be facilitated and limited to extracting the required study-specific data. However, if harmonization is achieved by a central team, the study-specific data must be rendered accessible to the team and generally explored with close communication with study teams. Ensuring consistent quality and validation procedures across study-specific data sets can be challenging and must be adapted to each project. Different standard operating procedures, methodological approaches, and tools have been proposed by the research community[35–37] to support quality assessment of study-specific data. An example of minimal procedures that can be used is outlined in Supplementary material S4.

Process study-specific data under the harmonized format

To enable analysis, it is necessary to convert the heterogeneous study-specific data items into the DataSchema variables format. Where appropriate (when harmonization is deemed possible), data processing is accomplished through algorithmic recoding or statistical modeling of study-specific data. The approach selected for each variable will depend on the scientific objectives of the project, the nature and format of the DataSchema variable, the study-specific data items available, the potential to access study-specific IPD, and whether the data processing is achieved centrally or by study-specific teams. Supplementary material S5 provides an overview of possible approaches (algorithms and statistical models) and considerations in their use. Figure 3 illustrates possible algorithmic processing applied to generate a variable on binge drinking status using the variables outlined in Table 2.

Establishing an efficient processing and quality assurance workflow and ensuring accuracy and consistency in decision making is challenging, especially for large-scale initiatives or if

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press

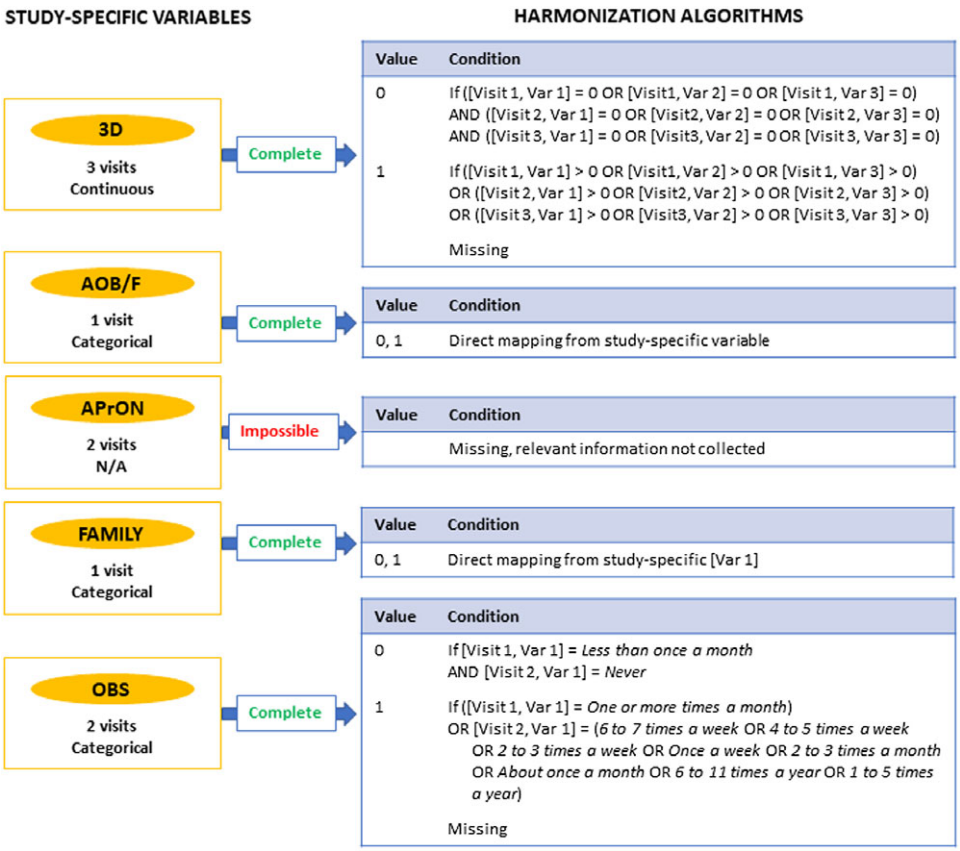

Fig. 3. Example of harmonization potentials and algorithms used to generate the variable "binge drinking during pregnancy". **Variable definition: Label:** Binge drinking during pregnancy (yes/no); **Definition:** Indicator of whether the mother ever binge drank at least once during pregnancy; **Value type:** Integer; **Format:** 0 = No, 1 = Yes; **Targeted individual:** Mother; **Targeted time period:** Throughout pregnancy; **Acceptable time of collection:** Can be collected in second or third trimester; **Acceptable question wording:** Binge drinking defined as five or more drinks *or* four or more drinks on one occasion *or* in one day. See Table 2 for information on the study-specific variables collected.

harmonization is achieved by different teams. Processing should be guided by the DataSchema variable definitions, and decision making (e.g., treatment of missing values) and quality assurance should be consistent across all data sets.

Estimate quality of the harmonized data

Once the harmonization process is finalized, it is often essential to explore the data sets generated to understand variable quality. This can include generating basic quality control checks (e.g., validating processing algorithms) and descriptive statistics (e.g., participant distributions, proportion of missing values) to evaluate the consistency across cohort studies (Supplementary material S6). When relevant, assessments of heterogeneity can be performed (e.g., testing for a statistical effect of study-specific question formats on the harmonized data generated). However, in practice it can be difficult to distinguish heterogeneity due to harmonization assumptions as opposed to population differences. A more comprehensive examination of relevant heterogeneity (and how to account for it) thus often needs to be performed at the stage of analysis.

Preserve the harmonized data sets and related documentation

Once validated and deemed of acceptable quality, the harmonized data set and its related documentation can be made available, and this, ideally in adherence to the FAIR[7] principles. Complying with the FAIR principles involves making data and metadata accessible to the scientific community, enabling long term access, ensuring their interoperability, and providing sufficient information to enable optimal use and reuse of the harmonized data. Documentation provided could include the harmonization protocol, selected information about cohort study designs and standard operating procedures, the DataSchema variable definitions, the harmonization potential across studies, the processing scripts or statistical models applied to generate harmonized data, and summary statistics on participant distributions or missing values. For large initiatives, creating a centralized metadata portal can provide user-friendly access to such information. However, maintenance of such portal, as well as long term preservation of the harmonized data sets in one or multiple secured data warehouses, can be challenging (e.g., to retain competent staff, maintain and when required scale up the infrastructure, etc.).

Analysis

Analyze data to answer specific research questions

Using harmonized data often involves working with an infrastructure where data are not available across all studies (missing values when harmonization is considered impossible), co-analyzing data available at different time points across studies, managing the heterogeneity of effects across studies, and using data that are not as precise as the study-specific data collected. Effectively, harmonizing heterogenous data often results in data reduction (e.g., transforming continuous variables into dichotomous) and subsequently to a potential lack of precision and reduction of power leading to underestimation of effects.[38]

Substantial time can be necessary to explore the harmonized data and conduct preliminary analysis.[39,40] It might be required to explore the impact of the harmonization potential of each DataSchema variable on the reduction in sample size and/or the diversity of variables included in the analysis. Harmonized data sets might have complex or limiting patterns of missing data that need to be examined; for example, it may be difficult to obtain the

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press

complete harmonized data across the same studies for the same DataSchema variables, leading to a trade-off between including more studies or more covariates in an analysis. Further exploration of the heterogeneity existing across studies and the potential effect of the harmonization process on the variables generated could also be suitable.[41,42] Various approaches are then available to analyze data and the analytical models determined by the research questions addressed, data infrastructure, and variable content.

Discussion

With the growing emphasis on a FAIR[7] approach to science, retrospective data harmonization is increasingly used to support research. However, to be FAIR, in addition to be accessible, data and associated documentation needs to be high quality. While the advantages of retrospective harmonization are significant, the limitations of harmonized data must also be recognized. Harmonizing existing data may not always generate as useful and high-quality data as hoped or expected. First, quality of the study-specific data is not always as good as anticipated. Second, defining DataSchema variables is a balancing act between generating very homogenous harmonized data and limiting the number of contributing cohort studies, or allowing more heterogeneity to retain more studies. Thus, there is often an unavoidable loss of precision in harmonized data generated. While a broad range of categories may be used by a given study to define, for example, the highest level of education, generating the variable across all studies could involve limiting the categories to "having completed secondary school or higher education (yes/no)". Third, as it is rarely possible to generate all DataSchema variables across all study-specific data sets, the harmonized data set will often only support subanalysis across selected variables and/or studies. Specialized statistical models working around missing values could help to overcome the problem but are not always applicable or suitable. Fourth, major complexities are introduced by the differing numbers and timing of data collection events across studies (before, during, and after pregnancy, as well as through the life of the participants). As the DataSchema variables defined can be time-dependent (i.e., need to be measured at a specific time point), this limits the harmonization potential. Fifth, factors related to the study-specific designs (e.g., population characteristics, sampling frames) can introduce biases.

Given these challenges, what motivates harmonization efforts? Might using published results to perform meta-analyses be a more sensible approach to synthesize information? It is easier and faster than selecting, exploring, harmonizing, integrating, documenting, and co-analyzing individual participant data across multiple cohort studies. However, scientifically founded harmonization initiatives present important advantages. First, it allows study-specific information to be processed to create more similar data. Second, for a given construct (e.g., familial income or level of physical activity), different variables can be generated and modified, providing flexibility during analysis. Third, following harmonization, different approaches are offered to support statistical analysis. Independent analysis-by-study followed by a meta-analysis of study-level estimates can be performed or harmonized data can be analyzed as a collective whole. Fourth, access to harmonized individual participant data provides flexibility for the selection of specific variables and participants or covariates to be included in the statistical analysis. Fifth, it increases the ability to examine heterogeneity and handle missing values. Sixth, it helps to limit the significant and intractable publication bias that is generally fundamental to observational data in the published domain. Finally, it facilitates exploring statistical interactions between risk factors and achieving subgroup analysis.

Addressing DOHaD research questions fundamentally involves exploring the interaction of multiple individual and environmental factors to often explain relatively subtle effects. Large initiatives such as ENRIECO and more recently LifeCycle are good examples of successful harmonization efforts leading to innovative research outputs. If well organized and scientifically founded, small and large retrospective harmonization initiatives can generate valuable harmonized data sets to support research, increase the scientific impact of individual cohort studies, and minimize duplication of research efforts.

Supplementary material. For supplementary material for this article, please visit https://doi.org/10.1017/S2040174422000460

Acknowledgements. We would like to thank the Maelstrom Research and EUCANconnect teams for their contribution to the development of concepts and approaches described in the current article. Additional thanks to the Prenatal Alcohol Exposure research team (including but not limited to: A Bocking, R Wissa, R Schmidt, K McDonald, Nika Zahedi).

Financial support. This work received funding from the European Commission (EUCANconnect, a federated FAIR platform enabling large-scale analysis of high-value cohort data connecting Europe and Canada in personalized health, Grant Agreement No 824989) with its Canadian project partners being funded by the Canadian Institutes of Health Research (CIHR) and the Fonds de la Recherche du Québec (FRQ). The RECAP Preterm project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733280. R Wilson is a UKRI Innovation Fellow with HDR UK [MR/S003959/1].

Conflicts of interest. None.

References

1. Wadhwa PD, Buss C, Entringer S, Swanson JM. Developmental origins of health and disease: brief history of the approach and current focus on epigenetic mechanisms. *Semin Reprod Med.* 2009; 27(5), 358–368.
2. Gillman MW. Developmental origins of health and disease. *N Engl J Med.* 2005; 353(17), 1848–1850.
3. Godfrey KM, Lillycrop KA, Burdge GC, Gluckman PD, Hanson MA. Epigenetic mechanisms and the mismatch concept of the developmental origins of health and disease. *Pediatr Res.* 2007; 61(7), 5–10.
4. Fortier I, Raina P, Van den Heuvel ER, *et al.* Maelstrom research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol.* 2017; 46(1), 103–105.
5. Lesko CR, Jacobson LP, Althoff KN, *et al.* Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. *Int J Epidemiol.* 2018; 47(2), 654–668.
6. Granda P, Blasczyk E. Data harmonization guidelines for best practice in cross-cultural surveys. In *Guidelines for Best Practice in Cross-Cultural Surveys*, 2010, Survey Research Center, Institute for Social Research, University of Michigan.
7. Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016; 3(1), e1002295.
8. Jaddoe VWV, Felix JF, Andersen AMN, *et al.* The LifeCycle project-EU child cohort network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. *Eur J Epidemiol.* 2020; 35(7), 709–724.
9. Gehring U, Casas M, Brunekreef B, *et al.* Environmental exposure assessment in European birth cohorts: results from the ENRIECO project. *Environ Health.* 2013; 12(1), 8.

10. Zeitlin J, Sentenac M, Morgan AS, *et al.* Priorities for collaborative research using very preterm birth cohorts. *Arch Dis Child - Fetal Neonatal Ed.* 2020; 105(5), 538–544.
11. Bergeron J, Massicotte R, Atkinson S, *et al.* Cohort profile: research advancement through cohort cataloguing and harmonization (ReACH). *Int J Epidemiol.* 2021; 50(2), 396–397.
12. Myatt L, Roberts JM, Redman CWG. Availability of COLLECT, a database for pregnancy and placental research studies worldwide. *Placenta.* 2017; 57(3), 223–224.
13. Tollånes MC, Strandberg-Larsen K, Forthun I, *et al.* Cohort profile: cerebral palsy in the Norwegian and Danish birth cohorts (MOBAND-CP). *BMJ Open*. 2016; 6(9), e012777.
14. Voerman E, Santos S, Inskip H, *et al.* Association of gestational weight gain with adverse maternal and infant outcomes. *JAMA*. 2019; 321(17), 1702–1715.
15. Bousquet J, Anto J, Sunyer J, *et al.* Pooling birth cohorts in allergy and asthma: European Union-funded initiatives – a MeDALL, CHICOS, ENRIECO, and GA2LEN joint paper. *Int Arch Allergy Immunol.* 2013; 161(1), 1–10.
16. Butters OW, Wilson RC, Burton PR. Recognizing, reporting and reducing the data curation debt of cohort studies. *Int J Epidemiol.* 2020; 49(4), 1067–1074.
17. Maelstrom Research, Accessed February 17, 2022. *Available at:* http://www.maelstrom-research.org/.
18. EUCAN Connect., Accessed February 17, 2022. *Available at:* https://eucanconnect.com/.
19. DataSHIELD, Accessed February 17, 2022. *Available at:* https://www.datashield.org/.
20. Cummings SR, Browner WS, Hulley SB. Conceiving the research question and developing the study plan. In *Designing Clinical Research*, Fourth, Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB), 2013; pp. 14–22. Lippincott Williams & Wilkins.
21. Larsen PS, Kamper-Jørgensen M, Adamson A, *et al.* Pregnancy and birth cohort resources in Europe: a large opportunity for aetiological child health research. *Paediatr Perinat Epidemiol.* 2013; 27(4), 393–414.
22. Fortier I, Dragieva N, Saliba M, Craig C, Robson PJ, Canadian Partnership for Tomorrow Project's scientific directors and the Harmonization Standing Committee. Harmonization of the health and risk factor questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis. *CMAJ Open*. 2019; 7(2), E272–E282.
23. Wey TW, Doiron D, Wissa R, *et al.* Overview of retrospective data harmonisation in the MINDMAP project: process and results. *J Epidemiol Community Health.* 2021; 75(5), 433–441.
24. de Moira APinot, Haakma S, Strandberg-Larsen K, *et al.* The EU Child Cohort Network's core data: establishing a set of findable, accessible, interoperable and re-usable (FAIR) variables. *Eur J Epidemiol.* April 21, 2021
25. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol.* 1999; 28(1), 1–9.
26. 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(6), 1929–1944.
27. Carter KW, Francis RW, Carter K, *et al.* ViPAR: a software platform for the virtual pooling and analysis of research data. *Int J Epidemiol.* 2016; 45(2), 408–416.
28. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010; 340(feb05 1), c221–c221.
29. Taylor K, Elhakeem A, Nader JLT, *et al.* The effect of maternal pre-/early-pregnancy BMI and pregnancy smoking and alcohol on congenital heart diseases: a parental negative control study. *MedRxiv Prepr Serv Health Sci*. November 4, 2020, 2020.09.29.20203786.
30. Benet M, Albang R, Pinart M, *et al.* Integrating clinical and epidemiologic data on allergic diseases across birth cohorts: a harmonization study in the mechanisms of the development of allergy project. *Am J Epidemiol.* 2019; 188(2), 408–417.
31. Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS ONE*. 2012; 7((10)), e46042.
32. 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.
33. Murtagh MJ, Turner A, Minion JT, Fay M, Burton PR. International data sharing in practice: new technologies meet old governance. *Biopreserv Biobanking.* 2016; 14(3), 231–240.
34. Shabani M, Thorogood A, Murtagh M. Data access governance. In *The Cambridge Handbook of Health Research Regulation*, Cambridge Law Handbooks, Laurie G, Dove E, Ganguli-Mitra A, *et al.*), 2021, Cambridge University Press.
35. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data Sci J*. 2015; 14(0), 2.
36. Dixon BE, Wen C, French T, Williams JL, Duke JD, Grannis SJ. Extending an open-source tool to measure data quality: case report on Observational Health Data Science and Informatics (OHDSI). *BMJ Health Care Inform.* 2020; 27(1), e100054.
37. Schmidt BM, Colvin CJ, Hohlfeld A, Leon N. Defining and conceptualising data harmonisation: a scoping review protocol. *Syst Rev*. 2018; 7(1), 226.
38. Cohen J. The cost of dichotomization. *Appl Psychol Meas.* 1983; 7(3), 249–253.
39. Avraam D, Wilson RC, Burton P. Synthetic ALSPAC longitudinal datasets for the Big Data VR project. *Wellcome Open Res*. 2017; 2.
40. Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. *J Priv Confidentiality.* 2016; 7(3), 67–97.
41. Friedenreich CM. Methods for pooled analyses of epidemiologic studies. *Epidemiology.* 1993; 4(4), 295–302.
42. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol Methods.* 2009; 14(2), 81–100.

https://doi.org/10.1017/S2040174422000460 Published online by Cambridge University Press