**Supplementary Material S1. Harmonization initiatives informing the paper content**

The harmonization initiatives listed include collaborative projects requiring harmonization of epidemiological data from multiple individual studies. The initiatives illustrate a variety of research questions, project sizes (number of participating studies, number of variables harmonized), study populations, and analytical approaches, showcasing the diversity of experience that informed the current work. Harmonization initiative names are informal descriptive labels for the purposes of the table. Core variables refer to the DataSchema variables generated across studies. Known dates and numbers were provided when possible but are sometimes approximate, not available (NA), or not applicable (N/A) at time of publication.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Research networks** | **Harmonization initiative** | **Start** | **End** | **Objectives** | **Regions represented (# countries)** | **Number of studies** | **Core variables** | **Analytical approaches** |
| ATHLOS | ATHLOS core variables | 2015 | 2021 | To identify patterns of healthy ageing trajectories and their determinants | Europe (24), North America (5), Asia (4), Africa (2), South America (2), Australasia (1) | 18 | 183 | Pooled |
| BBMRI-LPC | BBMRI-LPC pilot project | 2013 | 2017 | To pilot harmonization of large prospective study data on human health and disease and to develop recommendations for future research infrastructure | Europe | N/A | N/A | Pooled |
| BioSHaRE | Environmental Determinants of Health Project | 2011 | 2019 | To study the effect of environmental exposures, specifically air pollution and traffic noise, on the health of the population | Europe (3) | 4 | 76 | Federated |
| BioSHaRE | Healthy Obese Project | 2011 | 2014 | To gain insights into the characterization, determinants and consequences of metabolically healthy obese individuals | Europe (7), North America (1) | 11 | 100 | Federated |
| CanPath | CanPath core data | 2008 | ongoing | To support leading-edge research on environmental, lifestyle, and genetic risk factors for development and progression of cancer and chronic disease | Canada | 6 | 2826 | Pooled |
| CAPACIty | CAPACIty core data | 2020 | ongoing | To create a national data infrastructure and collaboration to inform interventions to better predict, prevent and treat serious diabetes-related complications | Canada | 10 | NA | Federated |
| CHICOS | Childhood asthma | 2010 | 2013 | To assess associations between pregnancy, birth, and early life factors and childhood wheezing and asthma | Europe (16) | 31 | 25+ | PooledMeta-analysis |
| CHICOS | Light-to-moderate drinking and SGA | 2010 | 2017 | To investigate light drinking during pregnancy in association with preterm delivery and small size for gestational age (SGA) | Europe (7) | 9 | 140 | Pooled |
| CHICOS,ENRIECO | Maternal factors and birth outcomes | 2010 | 2016 | To assess associations between maternal factors and birth weight, and length of gestation | Europe (11) | 13 | 25+ | PooledMeta-analysis |
| Childbirth Network | Women's Health Epidemiology | 2016 | 2017 | To investigate effects of birth interventions provided in different birth settings on maternal and children's health outcomes | N/A | 2 | 30 | Pooled |
| CHPT | CHPT harmonization | 2015 | 2018 | To evaluate harmonization potential across cohorts for research on determinants and etiology of cancer and other major chronic diseases | Europe (11), North America (2) | 12 | 126 | N/A |
| Chronic LBP IPD Meta-Analysis Group | LBP IPDMA | 2019 | 2020 | To assess treatment effects of exercise therapy for reducing pain and functional limitations in adults with persistent low back pain and modifying effects of individual characteristics | Europe (9), Asia (1), Australasia (1), North America (1), South America (1) | 27 | 25 | Pooled |
| CITF | CITF core data | 2020 | ongoing | To catalyse, support, and, where appropriate and feasible, harmonize the design and rapid implementation of population-based studies | Canada | 44 | 95 | Pooled |
| Cohorts.se | Rare diseases project | 2020 | ongoing | To establish risks factors for and protective factors against rare diseases, including variations in associations by sex and other subgroups | Sweden | 35 | 70 | Pooled |
| ENPADASI | ENPADASI harmonization | 2016 | 2020 | To assess associations between increased dietary fat and decreased carbohydrate intake with circulating HDL and non-HDL cholesterol | Europe (4) | 8 | 16 | Federated |
| ESCAPE | Air pollution and low birthweight | 2008 | 2012 | To assess the effect of maternal exposure to low concentrations of ambient air pollution on birthweight | Europe (12) | 14 | 25+ | PooledMeta-analysis |
| EU Child Cohort Network (LifeCycle) | LifeCycle core data | 2017 | 2022 | To bring together data from pregnancy and child cohort studies in an open and sustainable data resource for research on early exposures and health trajectories | Europe (12), Australasia (1) | 17 | 1396 | Federated |
| EUCANConnect | Gestational age and body size pilot | 2019 | ongoing | To study the association between gestational age at birth and body size in early infancy through adulthood | Europe (9), Australasia (1), North America (1) | 16 | 252 | Federated |
| EUCANConnect | Prenatal stress and preterm birth pilot | 2019 | ongoing | To assess the impact of prenatal stress on gestational age at delivery | Europe (2), North America (1) | 5 | 113 | Federated |
| EUCANConnect | Green spaces and child health pilot | 2020 | ongoing | To investigate the associations between exposure to green spaces and multiple child health outcomes across Europe | Europe (8) | 10 | 200+ | Federated |
| EUCANConnect | Childhood respiratory outcomes pilot | 2020 | ongoing | To perform a meta-analysis using personalized risk and prevention models in order to predict the development of respiratory outcomes in later childhood | Europe (12), Australasia (1) | 15 | 60+ | Federated |
| EUCANConnect | Urban exposome and cognitive function pilot | 2020 | ongoing | To estimate the associations between urban exposome during early-life and cognitive and psychomotor function in children | Europe (7), North America (1) | 10 | 130+ | Federated |
| EUCANConnect | Urban exposome and emotional issues pilot | 2020 | ongoing | To investigate the associations between urban exposome during early-life and emotional and behavioral issues in children | Europe (7), North America (1) | 12 | 115+ | Federated |
| euCanSHare | Public health use case | 2020 | ongoing | To assess data harmonization and sharing between European and Canadian cardiovascular cohorts, and to describe and estimate risk factors and outcomes | Europe, North America | NA | NA | NA |
| euCanSHare | RadiOMICS analyse – multiomics use case | 2020 | ongoing | To derive an integrative model of multi-omics data that outperforms conventional mono-source models and demonstrate feasibility for integrative analysis | Europe, North America | NA | NA | NA |
| euCanSHare | Diabetic cardiomyopathy use case | 2020 | ongoing | To demonstrate differences in mortality, cardiovascular death, and heart failure in patients with and without diabetes, and to assess myocardial differences using MRI | Europe, North America | NA | NA | NA |
| GOING-FWD | GOING-FWD harmonization | 2019 | ongoing | To assess gender-related determinants of clinical cost-sensitive outcomes and patient-related outcome measures in noncommunicable chronic diseases | Europe (4), North America (1) | 30 | 300 | Federated |
| GOING-FWD | USGo harmonization | 2021 | ongoing | To conduct a systematic assessment of gender-related variables and quality indicators of health care management on patient-reported outcomes | United States | 4 | 100+ | Pooled |
| HeLTI | HeLTI Prospective Harmonization | 2017 | ongoing | To identify a core set of variables for research on improving health, child development, and prevention of non-communicable diseases | Asia (2), Africa (1), North America (1) | 4 | 1967 | NA |
| IALSA | Healthy Neuroticism | 2017 | 2020 | To rigorously analyze evidence for “healthy neuroticism,” i.e., the significant moderation of the neuroticism-health relationship by conscientiousness | Europe (2), North America (1), Australasia (1) | 15 | 20+ | Meta-analysis |
| IALSA | Handgrip Strength and Cognitive Function | 2017 | 2021 | To conduct a coordinated multi-study analysis of the longitudinal association between handgrip strength and cognitive function in older adults | Europe (4), North America (2) | 9 | 25+ | Meta-analysis |
| IALSA | Trajectories of Big Five Personality Traits | 2017 | 2020 | To assess change in adulthood in self-reported Big Five personality traits using techniques designed to enhance replicability of results | Europe (4), North America (1) | 16 | 35+ | Meta-analysis |
| IALSA,BRAIN | IALSA/BRAIN exploratory harmonization | 2016 | 2018 | To evaluate harmonization potential across cohorts for research on determinants of within-person aging-related changes in cognitive and physical capabilities and mental health | North America (2), Europe (3), Asia (1) | 9 | 72 | N/A |
| ICAD | ICAD harmonization | 2018 | 2019 | To provide a large resource suitable for addressing a wide range of questions on health impacts and determinants of physical activity in young people | Europe (7), Australasia (1), North America (1), South America (1) | 20 | 106 | Pooled |
| ICM cardio | ICM-UKB harmonization | 2019 | ongoing | To investigate the impact of various risk factors associated with cardiovascular diseases | Europe (1), North America (1) | 2 | 125 | Pooled |
| InterConnect | LTPA and anthropometric outcomes | 2014 | 2018 | To examine the association between leisure time physical activity (LTPA) during early and late pregnancy and newborn anthropometric outcomes | Europe, North America | 8 | 30 | Federated |
| InterConnect | Food group intakes and T2D | 2015 | ongoing | To examine the prospective associations of some food group intakes with the risk of incident type 2 diabetes (T2D) | Europe, North America, South America, Middle East, South Asia, South East Asia, Australasia | 30 | 50 | Federated |
| InterConnect | Dietary patterns and type 2 diabetes | 2016 | ongoing | To replicate associations between exploratory dietary patterns and incident type 2 diabetes in cohorts across the world | Europe, North America, South America, Middle East, South Asia, South East Asia, Australasia | 25 | 70 | Federated |
| INTIMIC | Healthy Diet for a Healthy Life | 2018 | 2022 | To assemble available knowledge of the microbiota and other aspects of microbiome research in a FAIR data fashion and share information with various stakeholders | Europe | NA | NA | Federated |
| IPD-MA | IPD-MA core data | 2020 | ongoing | To investigate the impact of the length of storage of red blood cell units as a risk factor for hospital-acquired infections  | Canada | 7 | 154 | Pooled |
| LifeCycle, EarlyNutrition Project, CHICOS | MOCO collaboration | 2017 | 2019 | To examine Maternal Obesity and Childhood Outcomes | Europe (16), North America (2), Australasia (1) | 39 | 15+ | Pooled |
| LITMUS | LITMUS pilot | NA | ongoing | To develop, validate, and qualify better biomarkers for testing Non-Alcoholic Fatty Liver Disease (NAFLD) | Europe | NA | NA | Federated |
| LONGITOOLS | LONGITOOLS pilot | NA | ongoing | To study the interactions between environment, lifestyle, and health in determining risks of chronic cardiovascular and metabolic diseases | Europe | 25 | NA | Federated |
| MINDMAP | MINDMAP harmonization | 2016 | 2020 | To build a data harmonization platform to support cross-national research on ageing, mental well-being, and the urban environment | Europe (7), North America (1) | 7 | 2890 | Pooled |
| MIRACUM | MIRACUM harmonization | 2018 | ongoing | To establish data integration centers in university hospitals to manage, compute, and share data extracted from electronic health records | Germany | 10 | NA | Federated |
| MOBAND | MOBAND-CP study | 2010 | 2017 | To study cerebral palsy aetiology in a prospective design | Europe (2) | 2 | 1000 | Pooled |
| MOBYDIck | MOBYDIck core data | 2021 | ongoing | To investigate if oral high-dosage DHA supplementation in very preterm infants increases composite outcome of BPD free survival | Australasia (1), North America (1) | 2 | 83 | NA |
| MORGAM | MORGAM core data | 1998 | ongoing | To investigate relationships between development of cardiovascular diseases and their classic and genetic risk factors and biomarkers | Europe, Australasia | 30 | 281 | Pooled |
|  |  |  |  |  |  |  |  |  |
| NEAR | Physical function before death | 2019 | ongoing | To investigate inequalities and temporal trends in physical functioning and longevity among older people in Finland and Sweden | Europe (2) | 9 | 40 | Pooled |
| NEAR | Sleep disturbances and neuroimaging correlates | 2021 | ongoing | To investigate the associations between sleep disturbances,neuroimaging correlates, and their association with cognition. | Sweden | 3 | NA | Pooled |
| NFDI4Health | NFDI4Health pilot | NA | ongoing | To enable FAIR data generated in clinical trials, epidemiological, and public health studies to enhance research collaboration while complying with privacy and ethical requirements | Germany | NA | NA | Federated |
|  |  |  |  |  |  |  |  |  |
| P3G | Generic DataShaper | 2006 | 2011 | To evaluate the value of the DataSchema and Harmonization Platform for Epidemiological Research approach for retrospective harmonization of large population-based studies | Europe (14), Asia (4), North America (2), Australasia (1) | 53 | 148 | N/A |
| ReACH | Prenatal Alcohol Exposure | 2018 | 2022 | To identify correlates of maternal alcohol use before and during pregnancy and assess impacts on low birthweight and preterm birth | Canada | 5 | 56 | Pooled |
| ReACH, EUCANConnect | ReACH exploratory harmonization | 2020 | 2022 | To evaluate harmonization potential across Canadian cohorts for research on the impact of prenatal factors on birth outcomes and children's health | Canada | 24 | 349 | N/A |
| RECAP Preterm Consortium | Adult outcomes of VPT and VLBW | 2019 | ongoing | To compare major adult outcomes between Very Preterm (VPT) and Very Low Birth Weight (VLBW) and term born participants | Europe (6), Australasia (2) | 11 | 567 | PooledMeta-analysis |
| RECAP Preterm Consortium | EPT RECAP | 2020 | ongoing | To perform a comparative analysis of extremely preterm (EPT) birth cohorts in Europe to identify similarities and differences for the determination of baseline populations in perinatal epidemiology | Europe (14) | 17 | 631 | Federated |
| ReCoDID | DENGUE project | 2019 | ongoing | To develop a harmonization system and implement data harmonization for dengue research | South America (5), Central America (1), multi-national | 8 | 100+ | NA |
| ReCoDID | ZIKA project | 2019 | ongoing | To develop a harmonization system and implement data harmonization for Zika research | South America, Central America, Africa, Asia, Europe, North America | 8 | NA | NA |
| RESET | RESET core data | 2020 | ongoing | To explore reversing type 2 diabetes with a low energy diet and exercise intervention in adults under 40 years | Europe (1), North America (1) | 2 | NA | NA |
| SPIRIT | SPIRIT harmonization | 2014 | ongoing | To harmonize pregnancy and birth cohort data and establish an infrastructure for research on the intra-uterine determinants of child health and development, and perinatal health services | Asia (1), North America (1) | 4 | NA | Federated |
| STROKOG | Poststroke cognitive impairment | 2016 | NA | To investigate profiles of and risk factors for post-stroke cognitive impairment in diverse ethno-racial groups | Europe, Asia, Africa, Australasia, North America | 13 | 59 | Pooled |
| WHO | Baseline determinants of aging | 2020 | 2020 | To support research on the impact of cross-national differences in policies and actions on global population health and aging | Europe (31), Asia (10), Africa (3), North America (3), South America (2), Australasia (1) | 12 | 60 | Pooled |
| N/A | Multi-Center Sleep Study | 2017 | 2017 | To investigate sleep disturbances in midlife and late life, and their associations with dementia and cognitive status | Europe (2) | 4 | 11 | Pooled |
| N/A | Breastfeeding and diabetes | NA | 2017 | To study the relation between the duration of full and any breastfeeding and risk of type 1 diabetes | Europe (2) | 2 | 10+ | Meta-analysis |
| N/A | Smoking, education, and preterm birth | NA | 2019 | To study to which extent smoking during pregnancy mediates educational disparities in preterm delivery | Europe (3) | 3 | 10+ | Pooled |

Network acronyms and abbreviations used in Table S1.

|  |  |
| --- | --- |
| Acronym/Abbreviation | Full name |
| ATHLOS | Ageing Trajectories of Health: Longitudinal Opportunities and Synergies |
| BBMRI-LPC | Biobanking and Biomolecular Resources Research Infrastructure - Large Prospective Cohorts |
| BioSHaRE | Biobank Standardisation and Harmonisation for Research Excellence in the European Union |
| BRAIN | Broad and Deep Analyses in Neurodegeneration |
| CanPath | Canadian Partnership for Tomorrow's Health |
| CAPACIty | CAnadian PediAtric diabetes ConsortIum |
| Chronic LBP IPDMA Group | Chronic Lower Back Pain Individual Participant Data Meta-Analysis Group |
| CHICOS | Developing a Child Cohort Research Strategy for Europe |
| CHPT | Cross-cohort Harmonization Project for Tomorrow |
| CITF | COVID-19 Immunity Task Force |
| Cohorts.se | Swedish Cohort Consortium |
| EarlyNutrition Project | Long-term effects of early nutrition on later health |
| ENPADASI | European Nutritional Phenotype Assessment and Data Sharing Initiative |
| ENRIECO | Environmental Health Risks in European Birth Cohorts |
| ESCAPE | European study of cohorts for air pollution effects |
| EU Child Cohort Network | A Europe-wide network of cohort studies started in early life |
| EUCAN-Connect | A federated FAIR platform enabling large-scale analysis of high-value cohort data connecting Europe and Canada in personalized health |
| EUCAN-Share | An EU-Canada Joint Infrastructure for Next-Generation Multi-Study Heart Research |
| GOING-FWD | Gender Outcomes INternational Group: to Further Well-being Development |
| HeLTI | Healthy Life Trajectories Initiative |
| IALSA | Integrative Analysis of Longitudinal Studies of Aging and Dementia |
| ICM | Montreal Heart Institute - Institut de Cardiologie de Montréal |
| ICAD | International Children's Accelerometry Database |
| InterConnect | Global data for diabetes and obesity research |
| INTIMIC | Intestinal Microbiomics Knowledge Platform |
| IPD-MA | Length of Storage of Red Blood Cell Units as a Risk Factor for Hospital-Acquired Infections: An Individual Patient Data Meta-Analysis |
| LifeCycle | Early-life stressors and LifeCycle health |
| LITMUS | Liver Investigation: Testing Marker Utility in Steatohepatitis |
| LONGITOOLS | LONGITOOLS Health and Environment Dynamics |
| MINDMAP | Promoting mental well-being and healthy ageing in cities |
| MIRACUM | Medical Informatics in Research and Care in University Medicine |
| MOBAND | MOthers and BAbies in Norway and Denmark |
| MOBYDIck | Maternal omega-3 supplementation to reduce bronchopulmonary dysplasia in very preterm infants: A randomized controlled trial |
| MOCO | Maternal Obesity and Childhood Outcomes |
| MORGAM | MONICA Risk, Genetics, Archiving and Monograph |
| NEAR | National E-infrastructure for Aging Research |
| NFDI4Health | National Research Data Infrastructure for Personal Health Data |
| P3G | Public Population Project in Genomics and Society |
| ReACH | Research Advancement through Cohort Cataloguing and Harmonization |
| RECAP Preterm Consortium | Research on European Children and Adults Born Preterm |
| ReCoDID | Reconciliation of Cohort Data in Infectious Diseases |
| RESET | REmission of diabetes and improved diastolic function by combining Structured Exercise with meal replacemenT and food reintroduction |
| SPIRIT | Sino-Quebec Perinatal Initiative in Research and Information Technology |
| STROKOG | Stroke and Cognition consortium |
| UKB | UK Biobank |
| USGo | United States Gender outcomes |
| WHO | World Health Organization |

**Supplementary material S2. Life course of the ReACH Prenatal Alcohol Exposure (PAE) project**

This supplementary material presents the life course of a harmonization initiative initiated in 2018 and concluded in 2022. It includes information related to the process undertaken to achieve the project, the timeline, and the approximate budget.

1. **Life course of the Prenatal Alcohol Exposure (PAE) project**
	1. **Initiation**

The PAE project was initiated under the umbrella of the Research Advancement through Cohort Cataloguing and Harmonization (ReACH) initiative [1] by investigators from the Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON (A Bocking) and the Research Institute of the McGill University Health Center, Montreal, QC (I Fortier). The initial objectives of the project were to examine the risk factors for maternal alcohol consumption before, during, and after pregnancy and the effects of prenatal alcohol exposure (PAE) on infant outcomes. Mother-and-child cohorts collecting the information required to support the analyses foreseen were identified using the ReACH study and variables catalogue (https://www.maelstrom-research.org/mica/network/reach). Six of the 15 studies documented in the catalogue when the PAE project was initiated met all selection criteria (i.e., collected information on alcohol intake during pregnancy; collected information on birthweight and gestational age; recruited at least 500 mothers) and were approached for collaboration. In parallel, the project team developed a research protocol, sought, and obtained funding from the ReACH network (funded by the Canadian Institutes of Health Research) and the Canada Fetal Alcohol Spectrum Disorder Research Network. The research protocol was submitted for ethics approval to research ethics committees of the PIs institutions (Lunenfeld-Tanenbaum Research Institute and McGill University Health Centre).

|  |
| --- |
| **Highlights** |
| * The project’s objectives and protocol were defined.
* Funding and ethics approval were obtained.
 |
| **Challenges** |
| * As the ReACH catalogue was under development, the content was incomplete, limiting access to information on the specific data collected by the studies. Studies were thus selected without validating the true potential of the studies to generate the harmonized data required by the PAE project.
* Ethics approvals were complicated by the novelty of this type of collaborative project for the ethical committees. Many questions had to be answered to explain the specificities of harmonization projects using existing data (secondary use of data).
 |

* 1. **Implementation**

The principal investigators (PIs) of each cohort were contacted to confirm their interest in collaborating on the project. Of the six cohorts identified, five PIs agreed (3D Study - Design, Develop, Discover[[1]](#footnote-1); All Our Families[[2]](#footnote-2); Alberta Pregnancy Outcomes and Nutrition[[3]](#footnote-3); Family Atherosclerosis Monitoring in Early Life[[4]](#footnote-4); and Ontario Birth Study[[5]](#footnote-5)). In parallel, a working group including technical (research assistants, data analysts, statisticians, informaticians) and scientific (clinicians, PhD students and epidemiologist working on alcohol exposure) experts was implemented. However, the initial team members changed through time.

A list of core variables (i.e., DataSchema) required to answer the research questions addressed was defined by the working group. With this list in hand, the data analysts further explored study-specific data content and identified, for each cohort, the specific variables required to generate the core variables. This exercise was essential to inform demands for access to data. Cohort-specific demands for access were then submitted following the process in place at each institution. Following submission, some questions or issues were raised by the local ethics and access committees. The project coordinator answered questions and, where required, revised the protocol submitted. Following granted access to data, the project coordinator then worked in collaboration with the cohorts’ host institutions to develop and sign cohort-specific data transfer agreements.

During the access procedures, data analysts further explored the potential for each participating cohort to generate the core variables (i.e., estimated the harmonization potential). With the support of the scientific experts of the working team, the variable definitions were refined to ensure each core variable could be harmonized across at least 3 of the cohorts. This led to revisiting the demands for access submitted to the cohorts.

The project protocol planned to pool data on a central server in Alberta, at PolicyWise for Children & Families [2]. Informaticians from the project working group collaborated with the PolicyWise team to install the software (OPAL [3], R server) selected to support the harmonization process. Following data transfer from each cohort to the central server, secured remote access was given by the PolicyWise team to the analysts participating to the harmonization process and later, to the statisticians performing the analyses.

|  |
| --- |
| **Highlights** |
| * Even if delayed, access to all cohorts’ data was granted.
* The list of core variables to be harmonized was defined, and secured data infrastructure was implemented.
 |
| **Challenges** |
| * The project coordinator and data analysts changed through the lifespan of the project leading to challenges in knowledge transfer and productivity.
* Data access procedures differed greatly across cohorts and were not always documented, leading to difficulties understanding rules in place and unexpected delays.
* Updated definitions of the core variables were generated after initiating the data access processes meaning variable requests had to be adjusted. This led to unexpected delays in obtaining all study-specific data required to support analysis.
 |

* 1. **Production**

Following transfer of study-specific data to the central data repository, each data set was explored for completeness (e.g., all variables requested were received) and quality (e.g., variables in expected format and including no outliers). In some cases, variables were missing, errors were identified, or documentation received was too limited to understand the data content. The project coordinator was thus in regular contact with cohort-specific data managers to ensure proper understanding and validation of the data set. A close collaboration with cohorts was pursued during all the harmonization and analysis processes. Data transfers occurred at different times for each cohort, and for some data sets, updated versions (e.g., following correction or addition of data content) were sent through time.

Once the data analysts were confident with the data received, processing of the study-specific data under the DataSchema format was initiated. A first data analyst wrote processing scripts for each variable to be generated using in combination R and Excel, and a second analyst validated the work. The analysts had regular meetings with the project coordinator and the scientific experts of the working group to ensure adequacy and consistency of the decisions taken. During this phase, problems with data received preventing generation of some DataSchema variables were identified and some variables definitions had to be modified. The ability to generate complex longitudinal alcohol-use variables was also explored with the support of expert statisticians. Unfortunately, the data were not complete enough across timepoints and cohorts to use the intended modeling approach.

Finally, quality control of the harmonized data was performed by looking at the participants distributions within and across each data set. An overview of the PAE project and harmonization potential is documented on the Maelstrom website (<https://www.maelstrom-research.org/study/pae-hp>).

|  |
| --- |
| **Highlights** |
| * Study-specific data was harmonized allowing pooling of data from 11,448 participants across 5 studies.
 |
| **Challenges** |
| * A lot of back and forth between the project coordinator and cohorts’ data managers was necessary to get information and ensure proper usage of the study-specific data.
* Turnover in cohort data managers led to disruptions in communication and project continuity.
* Ability to generate some key variables was limited, resulting in lower harmonization potential then expected and complicating the patterns of missing data across cohorts.
 |

* 1. **Analysis**

Based on the harmonization results (i.e., final list of DataSchema variables and harmonization potential), the objectives of the papers and the analyses plans had to be revised. A new question was also addressed, the objective being to explore the influence of different analytical approaches on the association observed between prenatal alcohol exposure and low birthweight for gestational age. Statisticians from two content expert teams performed exploratory analyses to test proof of concept and develop the analytical models. Multiple adjustments to the analyses plan were made along the way, influenced by various factors, including but not limited to the quality of the harmonized data. Each team identified target journals and drafted a manuscript to disseminate their results, which went through multiple rounds of revision before being circulated to obtain feedback from the participating cohorts’ PIs. Once approval from cohorts was obtained, each article was submitted to targeted journals.

|  |
| --- |
| **Highlights** |
| * Papers providing new insight into (1) patterns of prenatal alcohol exposure and (2) methodological approaches to explore robustness of findings were submitted.
 |
| **Challenges** |
| * Limitations in the harmonization potential across cohorts led to important modifications in the original analyses plan.
* The delay in the project made it necessary to renew the ethics and data transfer agreements, as well as extend server rental from Policywise.
 |

1. **Timeline**



1. **Approximative budget**



**References**

1. Bergeron J, Massicotte R, Atkinson S, Bocking A, Fraser W, Fortier I, et al. Cohort Profile: Research Advancement through Cohort Cataloguing and Harmonization (ReACH). Int J Epidemiol. 2021 May 17;50(2):396–7.

2. PolicyWise [Internet]. PolicyWise for Children & Families. [cited 2022 Feb 15]. Available from: https://policywise.com/

3. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. Int J Epidemiol [Internet]. 2017 Sep 2 [cited 2017 Oct 3]; Available from: http://academic.oup.com/ije/article/doi/10.1093/ije/dyx180/4102813/Software-Application-Profile-Opal-and-Mica

**Supplementary Material S3. Options for joint analysis across multiple studies**

The selection of an operating model to support statistical analysis is obviously informed by the analytical requirements of the project, but also by nature of access to the harmonized Individual Participant Data (IPD). The possible approaches include study-level or two-stage IPD meta-analysis, centralized pooled or one-stage IPD meta-analysis, and federated analysis. Each of these presenting distinct advantages and challenges as outlined below.1–3

**Study-level meta-analysis or two-stage IPD meta-analysis**

Using this approach, study-specific analysis is performed by individual studies on their own IPD, followed by a meta-analysis of the study-level estimates. At present, two-stage IPD meta-analysis is perhaps the method that is most often adopted.4 In fact, in the absence of technology enabling secure federated analysis, it is the *only* option when access to IPD is restricted. Crucially, this approach avoids the need to physically transfer the IPD to a central warehouse for analysis. Only aggregate data - non-disclosive summary statistics – generated in the first stage of analysis are required for the second stage analysis and because meta-analytical methods for such aggregate data are well established, this approach is almost always feasible and is reassuring for data access committees that do not wish their IPD to be physically shared. However, careful harmonization can still be helpful to improve comparability across studies. Furthermore, it is important to recognize that standardizing and coordinating analyses across a group of studies requires substantial effort. In addition, a meta-analysis approach can sometimes lead to a loss of statistical power. However, because many approaches to meta-analysis are based on aggregate results that are close to being sufficient statistics (*i.e.,* they carry *all* the available information needed to underpin the required analysis)5 the loss in power is often considerably less than might be anticipated.

**Centralized pooled analysis or one-stage IPD meta-analysis**

Alternatively, IPD can be transferred from each study to a central server where they are analyzed as a collective whole (a full joint likelihood one-step IPD meta-analysis with appropriate center effects4). This will ensure optimum statistical power and flexibility, and in particular, the potential for enhanced insights into interactive or heterogeneous effects and to the interpretation of joint estimates. However, data must be centrally warehoused, and this can demand robust well managed infrastructural hardware6 to ensure data can be held safely and securely. It may also require high performance computing facilities to facilitate the analysis of large amounts of data within a single centralized repository. Furthermore, in the real world, data governance is often more challenging than the technical requirements for data analysis. Thus, it can often prove very time consuming – or indeed impossible – to obtain permission to transfer the required IPD from a set of original studies to a central data warehouse.

**Federated analysis**

The various trade-offs between the first two approaches described and sensible strategies for choosing between them have been discussed in detail elsewhere4,7. However, there is also a third approach – federated data analysis – that can provide an alternative option.2,8 Under federated data analysis, statistical inference is still based on the underlying harmonized IPD, but those IPD remain on the IT systems behind the firewalls at the studies where they are normally held. Until recently, this approach was rarely adopted. This was because it requires the implementation of a distributed, but interoperable data infrastructure that enables unified co-analysis of the harmonized data across all the studies. This, in turn, demands sophisticated technology and appropriate informatics knowledge or support to implement and use. But there are now a number of approaches to federated analysis that are becoming increasingly straightforward to implement and use. These include DataSHIELD2,9,10, under which the study-level IPD is both invisible and physically unavailable and can neither be copied nor downloaded by the analyst. In addition, active algorithmic traps mitigate the risk of disclosure via targeted analysis. To meet these requirements, each new analytic function implemented in DataSHIELD demands substantive recoding of otherwise standard analytic routines to introduce the disclosure traps. Consequently, at present, DataSHIELD analysis can only be undertaken using the R statistical package11. Other federated approaches include, but are not limited to, those developed under the BIRO12, ViPAR3, Personal Health Train (PHT)13, Vantage614 and OpenSAFELY15 projects. These latter approaches do not generally build in algorithmic disclosure traps into their functional code, and this means that users can typically use whatever software they would like for their analyses. But, in consequence, methods for dealing with more sensitive data are less well developed and sometimes require all analytic output to be scrutinized by human readers before being released to analysts. Because these different approaches are all complementary, on-going attempts are being made to explore their combination with the ultimate aim that users will be able to select the particular approach that optimally meets their needs both for analytic flexibility and disclosure control.

Although federated analysis may appear quite different to the traditional approaches described earlier, in fact it comes in a variety of flavors that may reasonably be viewed as: (i) true one-stage IPD analysis; (ii) effective one-stage IPD analysis; or (iii) two-stage IPD analysis (study-level meta-analysis). Thus, for example, using the ViPAR approach, the IPD are held by the original studies but are transiently pooled to enable co-analysis and then rapidly destroyed. This approach is very flexible from an analytic perspective and represents true one-stage IPD analysis. But as it involves the IPD from each study being transferred to a central analysis site – albeit very transiently – data access committees that do not want their data to be physically transferred outside their own study may be wary of this solution. The other approaches identified send the analysis to the data2 rather than taking data to the analysis, either in the form of scripts calling algorithms on the data servers13,14 or by direct calls to analytic functions2,9. As an example, OpenSAFELY15 is based on a sophisticated IT infrastructure for real-time shared data analysis. Pseudonymized patient records remain at the data centers of the electronic health records software companies that hold them, and those centers also host the first-stage analyses. Like DataSHIELD, the four other approaches identified above all generate first-stage aggregate results at each data source and are then returned to an analysis center where they are combined in the second stage of analysis. This is permitted under governance rules because the aggregate results are fundamentally non-disclosive. If the first-stage analysis generates aggregate results that are sufficient statistics5, this two-stage approach can be made mathematically equivalent to pooling all the data centrally and applying a standard analysis directly to those pooled data16– *i.e.,* it represents “*effective* one-stage IPD analysis”. On the other hand, if the first-stage results are instead combined using random (or fixed) effects meta-analysis in the second stage this corresponds to two-stage meta-analysis. In this latter situation the difference between undertaking a traditional two-stage meta-analysis and carrying out the same analysis using a well-designed federated data-analysis system is that, because the first-stage analysis in the latter case is entirely centrally controlled, new analyses that are recognized as being required because of initial results (*e.g*., the need to add interactions between variables demonstrated to have a substantive impact in early analysis) can be commanded directly by the analyst in real time, rather than having to ask each study to undertake the extra analyses required and wait for them to return the results. This can greatly reduce the time to carry out the additional analyses, as in waiting for all studies to return new results the slowest study always sets the pace.

**References**

1. 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.

2. 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), 1929–44.

3. 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), 408–16.

4. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010; 340, c221.

5. Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*. 4th ed. Oxford, UK, Blackwell Science Ltd., 2002.

6. Ford DV, Jones KH, Verplancke J-P, *et al.* The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res*. 2009; 9(1), 157.

7. 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.

8. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. *Int J Epidemiol*. 2017; 46(5), 1372–8.

9. 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(5), 1372–82.

10. Marcon Y, Bishop T, Avraam D, *et al.* Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. *PLOS Comput Biol*. 2021; 17(3), e1008880.

11. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing, 2019.

12. Di Iorio CT, Carinci F, Brillante M, *et al.* Cross-border flow of health information: is ‘privacy by design’ enough? Privacy performance assessment in EUBIROD. *Eur J Public Health*. 2013; 23(2), 247–53.

13. Beyan O, Choudhury A, van Soest J, *et al.* Distributed analytics on sensitive medical data: the personal health train. *Data Intell*. 2020; 2(1–2), 96–107.

14. Moncada-Torres A, Martin F, Sieswerda M, van Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. In: *AMIA. Annu. Symp. Proc.*, pp. 870–7. 2020.

15. Forbes H, Morton CE, Bacon S, *et al.* Association between living with children and outcomes from covid-19: OpenSAFELY cohort study of 12 million adults in England. *BMJ*. 2021; 372, n628.

16. Jones EM, Sheehan NA, Masca N, Wallace SE, Murtagh MJ, Burton P. DataSHIELD -- shared individual-level analysis without sharing the data: a biostatistical perspective. *Nor Epidemiol*. 2012; 21(2), 231–9.

**Supplementary Material S4. Assessment of input study-specific data sets**

Before generating harmonized data, it is crucial to ensure completeness and quality of the input study-specific data set (the data collected by the study). An example of procedures is provided below.

|  |
| --- |
| **Verify the data set format** |
| *Procedures* | Ensure that the input study-specific data set is in a format compatible for processing. |
| *Example* | The input study-specific data set is sent in an SPSS file, read in R, and examined for encoding issues, such as unreadable character formats. |
| **Verify the list of identifiers (IDs)** |
| *Procedures* | Verify that the correct participants or entities are included in the data set. |
|  | Verify that there are no duplicated IDs. |
|  | If a reference list is available, compare the IDs against the reference IDs. |
|  | Verify that minimal information for inclusion, e.g., age and sex, is adequate and correspond to inclusion criteria. |
|  | Identify and document any inconsistent or incomplete data. |
| *Example* | The study-specific data set was expected to include 2000 mothers but only includes 1800. |
| **Verify the list of variables and format** |
| *Procedures* | Verify that the variables required to generate the DataSchema variables are included, and that metadata (information about the variables) is adequate. |
|  | Identify and document any inconsistent or incomplete data. |
| *Example* | The variable ‘gestational age at birth’ was requested from a study but not received. |
| **Perform univariate checks** |
| *Procedures* | Generate univariate distributions of each variable required to produce the DataSchema variables. |
|  | Identify and document any inconsistent or incomplete data. |
| *Example* | Infant birth weight >5 kg is flagged as an outlier to be checked. |
| **Perform multivariate checks** |
| *Procedures* | Generate cross-references across relevant variables. |
|  | Identify and document any inconsistencies. |
| *Example* | Mother's age at follow-up is checked for consistency with age at baseline. |
| **Resolve issues, where possible** |
| *Procedures* | If relevant, contact the study data managers with questions or requests for more information about input study-specific data sets. |
|  | Where possible, resolve questions and update documentation. |
| *Example* | The study data manager confirms that an unlabeled category “999” for the variable “number of drinks per week” should be treated as missing. |

**Supplementary Material S5. Data processing approaches to generate harmonized variables**

The data processing methods used to generate harmonized variables will depend on the content and format of targeted DataSchema variables, the content and format of input study-specific variables, and the data infrastructure implemented. Some processing approaches and considerations in their application are outlined below.

|  |
| --- |
| **Algorithmic transformation** |
| *Definition* | A finite set of unambiguous processing rules performed in a prescribed sequence, especially a mathematical rule or procedure. |
| *Application* | To harmonize continuous or categorical variables with different but combinable ranges or categories. |
| *Example(s)* | Processing household income level from a continuous variable to a dichotomous variable of above provincial average (yes/no).Processing frequency of alcohol use before pregnancy from total days in the previous 3 months to average per month categories (0: none, 1: <once/month, 2: once/month, 3: 2-3/month, 4: once/week, 5: 2-3 times/week, 6: 4-6 times/week, 7: everyday) |
| *Considerations* | Simple to define, apply, document, and reproduce.Can require specialized domain expertise (e.g., defining appropriate threshold values). Often very case-specific (not generalizable).Often results in loss of information (e.g., processing continuous variables to categorical).Connection to original variable often lost. |
| **Simple calibration model** |
| *Definition* | A mathematical model that predicts the value of one continuous variable from another variable to operate at the same units. |
| *Application* | To harmonize variables measuring the same metric with different scales.The calibration function can be known or derived from the data if bridging variables are present. |
| *Example(s)* | Processing infant birth weight from pounds to kilograms.Processing score on one memory test to another based on a conversion factor predicted from the relationship between scores for participants who had both tests. |
| *Considerations* | Requires known calibration or bridging variables.Bias may be present in estimated calibration.Results are only reliable if prediction error is low. |
| **Standardization model** |
| *Definition* | A model that centralizes, standardizes, or normalizes data to process all variables into a common scale with or without stratification or regression of other variables. |
| *Application* | To harmonize equivalent constructs measured using different scales with no known calibration method or bridging items.Typically used for discrete interval measures. |
| *Example(s)* | Processing cognitive performance measured with different tests to *z*-scores. |
| *Considerations* | Simple to define, apply, document, and reproduce.Result often ‘unitless’, i.e., not corresponding to original or other external units of measure.Eliminates population differences as well as methodological differences.Interpretation of effect sizes and statistical outputs more difficult. |
| **Latent variable model** |
| *Definition* | A statistical model that infers a set of latent variables (constructs that are not directly observable) from a set of observed variables. e.g., factor analysis and latent trait analysis (item response theory). |
| *Application* | To harmonize equivalent constructs measured using different scales with no known calibration method but with bridging items present. |
| *Example(s)* | Processing repeated measures of childhood body mass index to childhood growth trajectories using growth mixture models.Extracting a personality measure (e.g., extraversion) from factor analysis of multiple behavioral dimensions. |
| *Considerations* | More complicated to define, apply, document, and reproduce.Requires access to pooled individual participant data.Requires extensive exploration of study-specific data and assessment of results.Connection to original variables often lost. |
| **Multiple imputation model** |
| *Definition* | A statistical technique for imputing missing values from a set of plausible values that represent the uncertainty about the true values.  |
| *Application* | To harmonize data sets (not variables) with the same set of variables using bridging variables. |
| *Example(s)* | Missing survey responses are imputed from values of other survey responses. |
| *Considerations* | May reduce bias, and loss of power and precision due to missing data.More complicated to define, apply, document, and reproduce.Requires access to pooled individual participant data.Requires extensive exploration of study-specific data and assessment of results.Computationally intensive. |

**Supplementary Material S6. Validation of harmonized data sets**

After harmonized data is generated, the accuracy of harmonization processing and consistency of the harmonized data content must be validated. Example of procedures used to explore harmonized variables content within and across study-specific data sets are presented below.

|  |
| --- |
| **Verify the list of identifiers (IDs) in each study-specific harmonized data set** |
| *Procedures* | Verify that the correct participants or entities are present in the study-specific harmonized data set.Compare the IDs to the list of IDs in the input study-specific data set or reference list available. |
|  | Where relevant, make corrections and document any modifications. |
| *Example* | The list of participants in the study-specific harmonized data set is confirmed to match the participants in the input study-specific data set, excluding participants that were withdrawn. |
| **Verify the list of variables and format in each study-specific harmonized data set** |
| *Procedures* | Verify that the list of study-specific harmonized variables and variable metadata (e.g., name, description, format) correspond to the DataSchema. |
|  | Verify that all harmonization statuses are correctly assigned. Variables considered as ''complete'' for the study-specific data set must include information, and variables considered as ''impossible'' should be empty. |
|  | Where relevant, make corrections and document any modifications. |
| *Example* | The harmonized variable “age started smoking” is missing the category label “-7” indicating non-smokers, so the variable metadata is updated. |
| **Perform univariate checks in each study-specific harmonized data set** |
| *Procedures* | Generate univariate distributions of each study-specific harmonized variable. |
|  | For continuous variables: Identify any improbable values or outliers. |
|  | For categorical variables: Identify any values that do not correspond to the categories defined in the DataSchema. |
|  | For all variables, verify the algorithm and script used to generate each variable. |
|  | Document issues to verify in the input study-specific data. |
|  | Where relevant, make corrections and document any modifications. |
| *Example* | The values for “volume of blood taken” are improbably high, and it is confirmed that the units were in microliters instead of milliliters. |
| **Perform multivariate checks in each study-specific harmonized data set** |
| *Procedures* | Generate cross-references across relevant variables. |
|  | Document issues to verify in the input study-specific data. |
|  | Where relevant, make corrections and document any modifications. |
| *Example* | The variable “ever drank alcohol during pregnancy” is cross-tabulated with “frequency of drinking alcohol during pregnancy” to check consistency of 0 and non-zero values. |
| **Compare data content across study-specific harmonized data sets**  |
| *Procedures* | Generate distributions for harmonized variables across study-specific data sets. |
|  | Compare distributions and identify any improbable results. |
|  | Document issues to verify in the input study-specific data. |
|  | Where relevant, make corrections and document any modifications. |
| *Example* | Comparing sitting height across studies shows that values are systematically 30 cm higher in two cohorts compared to the others.  |
| **Resolve remaining issues, where possible** |
| *Procedures* | If needed, contact the study data managers with remaining questions about input study-specific data sets. |
|  | Where possible, resolve questions and update documentation. |
| *Example* | Measurement methodology for sitting height is verified with study managers, and it is discovered that the methodology used in two cohorts added 30 cm. |

**Definitions:**

* Input study-specific data set: Data set provided by studies prior to harmonization
* Harmonized study-specific data set: Harmonized data set generated from a given study-specific data set
* DataSchema: List of core variables to be generated
1. https://www.maelstrom-research.org/study/3d [↑](#footnote-ref-1)
2. https://www.maelstrom-research.org/study/aob [↑](#footnote-ref-2)
3. https://www.maelstrom-research.org/study/apron [↑](#footnote-ref-3)
4. https://www.maelstrom-research.org/study/family [↑](#footnote-ref-4)
5. https://www.maelstrom-research.org/study/obs [↑](#footnote-ref-5)