Multinational cohort integration

International Clinical Trials Day 2019
Paris, May 20th 2019

Josep Maria Haro
Parc Sanitari Sant Joan de Déu
Barcelona, Spain
Cohorts (and clinical trials) are invaluable resources to obtain detailed description of individual biological variations in connection with a variety of environmental, pathogenic, occupational, societal, and lifestyle determinants that influence the onset and evolution of diseases.

Europe is extremely well served by a rich variety of population, and patient cohorts, including well annotated clinical trial cohorts. Several large cohorts have also been developed in various parts of the world.

Despite recent efforts to network cohorts, the level of integration need to be escalated in order to optimise the exploitation of these resources, essential to underpin and facilitate the development of stratified and personalised medicine.
About the IHCC

The International 100K Cohort Consortium (IHCC) was established in 2018 at the request of the leaders of the Heads of International Research Organizations (HIROs) through a collaboration between the Global Genomic Medicine Collaborative (G2MC) and the Global Alliance for Genomics and Health (GA4GH).

National Medical Genome Projects and Cohorts

IHCC 2nd International Cohorts Summit Reykjavik, Iceland

April 23-24, 2019, Reykjavik, Iceland

IHCC hosted its second International Cohorts Summit in Reykjavik, Iceland. The conference was very successful, with the presentation of 23 collaborative research proposals. 117 representatives from 67 cohorts and 29 countries attended.
We are IALSA

The Integrative Analysis of Longitudinal Studies of Aging (IALSA) network is an international collaboration for reproducible longitudinal research on life course studies.

The study of aging and health-related change demands an integrative developmental framework, involving interdisciplinary collaborations and advanced methodological approaches to understand how and why individuals change with age, in both normative and idiosyncratic ways. Longitudinal studies also provide a basis for the early detection of change related to neurodegenerative disorder and the identification of periods in the lifespan when interventions will potentially have their greatest impact.

IALSA studies include research on change in cognitive and physical capabilities, health, personality, and well-being along the lifespan. IALSA has also provided core funding for the development of research tools central to the Maelstrom research catalogue (https://www.maelstrom-research.org), facilitating metadata discoverability and harmonization projects.
Clouston et al., under review. Famine exposure in adolescence and lifetime cognition
Analyses from a natural experiment of 87,080 older Europeans.

Duggan et al., under review. Meta-analysis: pulmonary function and cognition in aging.
Multi-study coordinated meta-analysis of pulmonary functioning and cognitive ability.

Graham et al., under review. Does personality change?
A replication study of trajectories of big-five personality traits across 15 longitudinal studies on aging.

Zammit et al., under review. Association between grip strength and cognitive function in older adults.
A coordinated multi-study analysis of the longitudinal between cognition and grip strength in aging.

Karr et al., 2018. When does cognitive decline begin?
Older adults who ultimately develop dementia experience accelerated cognitive decline long before diagnosis. A similar acceleration in cognitive decline occurs in the years before death as well. To evaluate preclinical and terminal cognitive decline, past researchers have incorporated change points in their analyses of longitudinal data, identifying point estimates of how many years prior to diagnosis or death that decline begins to accelerate.

Zammit et al., 2018. Associations between aging-related changes in grip strength and cognitive function in older adults: A systematic review.
Objectives. Grip strength and cognitive function reflect upper body muscle strength and mental capacities. Cross-sectional research has suggested that in old age these two processes are moderately to highly associated, and that an underlying common cause drives this association. Our aim was to synthesize and evaluate longitudinal research addressing whether changes in grip strength are associated with changes in cognitive function in healthy older adults.
We facilitate collaborative epidemiological research through rigorous data documentation, harmonization, integration, and co-analysis.
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825884.
“To establish a sustainable European strategy for the development of the next generation of integrated cohorts, thereby contributing to an international strategic agenda for enhanced coordination of cohorts globally, in order to address the practical, ethical and legal, and the methodological challenge to optimising the exploitation of current and future cohort data, towards the development of stratified and personalised medicine as well as facilitating health policy.”
Objective

More specifically:

1. To map the cohort landscape in Europe and large international initiatives.
2. To identify best methods for integrating cohort data in order to enable the harmonisation of past and future data collection.
3. To identify solutions for addressing practical, ethical and legal challenges in integrating data across patient, clinical trial and population cohorts.
4. To take stock of emerging and new data collection technologies and types of data, including new exposures and health risks, and their potential impact.
5. To prepare strategy briefs and conduct stakeholder dialogues to discuss
   i) best practise to overcome practical, ethical, legal and
   ii) methodological challenges in networking and integrating data across patient, clinical trial and population cohorts.
6. To provide, disseminate and make sustainable recommendations on standards to improve future sample and data collection and thereby contribute to define an international strategic agenda for better coordination of cohorts globally.
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825884
Challenges in integrating data from clinical trials

Better coding systems and management of clinical data (electronic data collection).

But:

• Little coordination among clinical trials.
• Challenges in information integration.
• Lack of integration of real world data (for example, patient cohorts) and clinical trial data.
• ....

➢ Role of ECRIN
### Timeline

<table>
<thead>
<tr>
<th>WP</th>
<th>Description</th>
<th>Start</th>
<th>End</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP1</td>
<td>Mapping the cohort landscape in Europe and large international initiatives for data harmonization and integration.</td>
<td>M2</td>
<td>M18</td>
<td>UAM</td>
</tr>
<tr>
<td>T1.1</td>
<td>Definition of the methodology for the mapping.</td>
<td>M2</td>
<td>M16</td>
<td>UAM</td>
</tr>
<tr>
<td>T1.2</td>
<td>Mapping of existing population cohorts.</td>
<td>M6</td>
<td>M15</td>
<td>UAM</td>
</tr>
<tr>
<td>T1.3</td>
<td>Matrix of existing patient and clinical trial cohorts.</td>
<td>M6</td>
<td>M15</td>
<td>SINNAM</td>
</tr>
<tr>
<td>T1.4</td>
<td>Mapping of existing cohort studies assessing the impact of new exposures on health.</td>
<td>M6</td>
<td>M15</td>
<td>NIOM</td>
</tr>
<tr>
<td>WP2</td>
<td>Methodological issues in harmonising and integrating cohort data.</td>
<td>M6</td>
<td>M24</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T2.1</td>
<td>Systematic review and documentation of state-of-the-art methods for harmonisation and integration of cohort data.</td>
<td>M6</td>
<td>M24</td>
<td>UAM</td>
</tr>
<tr>
<td>T2.2</td>
<td>Documenting available methods for analysis of cross-cohort data including specific challenges in analysing data across patient cohorts, including clinical trial cohorts, and population cohorts.</td>
<td>M18</td>
<td>M24</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T2.3</td>
<td>Detailing the existing infrastructures for analysis of cross-cohort data including specific challenges in analysing data across patient cohorts, including clinical trial cohorts, and population cohorts.</td>
<td>M18</td>
<td>M24</td>
<td>ECNIN</td>
</tr>
<tr>
<td>T2.4</td>
<td>Analysing issues related to methodological aspects for the harmonisation and integration of population-based and patients cohorts.</td>
<td>M18</td>
<td>M24</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T2.5</td>
<td>Preparing a strategic brief on the methodological issues regarding harmonisation/integration of cohorts.</td>
<td>M18</td>
<td>M24</td>
<td>SPF</td>
</tr>
<tr>
<td>WP3</td>
<td>Practical, ethical and legal issues in integrating cohort datasets.</td>
<td>M6</td>
<td>M24</td>
<td>SPF</td>
</tr>
<tr>
<td>T3.1</td>
<td>To prepare a systematic mapping of practical obstacles across Europe accessing cohort data, and other preconditions to data integration across cohort studies.</td>
<td>M6</td>
<td>M18</td>
<td>ECNIN</td>
</tr>
<tr>
<td>T3.2</td>
<td>To identify issues related to ethical aspects of the collection, use, and storage of patients and population cohort data. These issues will relate to both current ethical concerns and those raised by new data collection technologies and kinds of cohort data.</td>
<td>M18</td>
<td>M24</td>
<td>SPF</td>
</tr>
<tr>
<td>T3.3</td>
<td>Completion and production of a strategy brief on practical, legal and ethical issues in optimisation of cohort data in Europe.</td>
<td>M18</td>
<td>M24</td>
<td>SPF</td>
</tr>
<tr>
<td>WP4</td>
<td>Taking stock of emerging and new data collection technologies and their potential impact on the development of future cohort studies and the need to optimise the integration of data.</td>
<td>M2</td>
<td>M18</td>
<td>SYNER</td>
</tr>
<tr>
<td>T4.1</td>
<td>To prepare a systematic review of the emerging and new data collection technologies.</td>
<td>M2</td>
<td>M18</td>
<td>SYNER</td>
</tr>
<tr>
<td>T4.2</td>
<td>Mapping of existing population and patients cohorts using social media and new communication technologies.</td>
<td>M6</td>
<td>M12</td>
<td>NIOM</td>
</tr>
<tr>
<td>T4.3</td>
<td>Analyse the strengths, limitations and barriers of using new communication technologies in population-based and patients cohort studies.</td>
<td>M12</td>
<td>M18</td>
<td>SPF</td>
</tr>
<tr>
<td>WP5</td>
<td>Strategy briefs and stakeholder dialogues.</td>
<td>M8</td>
<td>M30</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T5.1</td>
<td>Identification of the participating stakeholders and creation of two steering committees (for practical, legal and ethical issues and for methodological issues).</td>
<td>M6</td>
<td>M28</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T5.2</td>
<td>Stakeholder dialogues.</td>
<td>M25</td>
<td>M30</td>
<td>PSSUID</td>
</tr>
<tr>
<td>WP6</td>
<td>Developing a strategy for optimisation of cohort data in Europe and to define a strategic agenda for cohorts globally.</td>
<td>M24</td>
<td>M38</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T6.1</td>
<td>Produce a report documenting the outcomes from the two stakeholder completations.</td>
<td>M24</td>
<td>M30</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T6.2</td>
<td>Prepare a report documenting the standards to improve future sample and data collection in patient and population cohort studies.</td>
<td>M24</td>
<td>M30</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T6.3</td>
<td>To prepare a report to set the strategic agenda for better coordination of cohorts globally.</td>
<td>M30</td>
<td>M36</td>
<td>PSSUID</td>
</tr>
<tr>
<td>WP7</td>
<td>Dissemination and Communication.</td>
<td>M1</td>
<td>M38</td>
<td>NTNU</td>
</tr>
<tr>
<td>T7.1</td>
<td>Formulation of dissemination plan.</td>
<td>M1</td>
<td>M38</td>
<td>NTNU</td>
</tr>
<tr>
<td>T7.2</td>
<td>Scientific and strategic dissemination of results.</td>
<td>M1</td>
<td>M38</td>
<td>NTNU</td>
</tr>
<tr>
<td>T7.3</td>
<td>Communication activities.</td>
<td>M1</td>
<td>M38</td>
<td>NTNU</td>
</tr>
<tr>
<td>T7.4</td>
<td>Organization of a final meeting with all stakeholders.</td>
<td>M1</td>
<td>M38</td>
<td>NTNU</td>
</tr>
<tr>
<td>WP8</td>
<td>Project management and coordination.</td>
<td>M1</td>
<td>M36</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T8.1</td>
<td>Day to day Coordination and Management.</td>
<td>M1</td>
<td>M30</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T8.2</td>
<td>Reporting, finances and financial management.</td>
<td>M1</td>
<td>M36</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T8.3</td>
<td>Project quality and risk management.</td>
<td>M1</td>
<td>M38</td>
<td>PSSUID</td>
</tr>
<tr>
<td>T8.4</td>
<td>Data management and coordination of ethical issues.</td>
<td>M3</td>
<td>M36</td>
<td>PSSUID</td>
</tr>
</tbody>
</table>

---

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825884
AGEING TRAJECTORIES OF HEALTH: LONGITUDINAL OPPORTUNITIES AND SYNERGIES

World Mental Health Survey Initiative
Annual Collaborators’ Meeting
Boston, July 2016

http://athlosproject.eu
The objective of the ATHLOS project is to achieve a better understanding of ageing by identifying patterns and determinants of healthy ageing:

- **Objective 1:** Produce a harmonized dataset from 20 recent longitudinal studies of ageing from European and other International projects.
- **Objective 2:** Identify patterns of healthy ageing trajectories and their determinants (including cohort effects) in order to predict trajectories of healthy ageing both at the individual and population levels.
- **Objective 3:** Determine and operationalize a more valid definition of ‘old age’ based on characteristics of people that change with age—in particular health status, cognitive function, life expectancy and other measures—rather than chronology alone.
- **Objective 4:** Translate ATHLOS findings, by means of knowledge translation methodologies and microsimulation exercises, into evidence-informed policy recommendations.
- **Objective 5:** Disseminate and make available the resources created by ATHLOS, the pooled datasets, the findings, and policy recommendations arising from the project to facilitate the use of those resources and the scalability of ATHLOS products.
2. **Data Harmonisation**

- **Step 0:** Define the research questions, objectives and protocol
- **Step 1:** Assemble information and select studies
- **Step 2:** Define variables and evaluate harmonization potential
- **Step 3:** Process data
- **Step 4:** Estimate quality
- **Step 5:** Disseminate and preserve

In summary
The aim is to produce harmonized datasets making use the **existing data**, where studies usually have **differences** on:

- Design of studies
- Research questions, procedures, measures, etc
- Sources of information, timelines, etc
Step 0: Define the research questions, objectives and protocol
Step 1: Assemble information and select studies
Step 2: Define variables and evaluate harmonization potential
Step 3: Process data
Step 4: Estimate quality of the harmonized dataset(s) generated
Step 5: Disseminate and preserve final harmonization products
End user

Web data portal

Study catalogue:
- Structure description of studies (metadata)
- Data access management

Mica

Data management:
- Store databases
- Data dictionaries
- Data harmonization

Opal

Agate

User directory management

Datashield

Data analyses

https://athlos.pssjd.org:8446

https://opal.pssjd.org:8443
Draw up a list of research questions.

Examples:

- To assess the role of nutrition and other lifestyles on quality of life, use of health care services, disability and health in older adults.
- By means of a single health metric in terms of functioning and cognition, identify patterns of healthy ageing trajectories.
- To identify the role of social inequalities in health related outcomes and healthy ageing trajectories.
- Etc...
Data Harmonisation

- Step 0: Define the research questions, objectives and protocol
- **Step 1:** Assemble information and select studies
- Step 2: Define variables and evaluate harmonization potential
- Step 3: Process data
- Step 4: Estimate quality
- Step 5: Disseminate and preserve
- In summary
<table>
<thead>
<tr>
<th>Study / Countries</th>
<th>Year of interview</th>
<th>Sample size</th>
<th>Targeted ages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10-66</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cuba</td>
<td></td>
<td>2813</td>
<td>&gt; 65</td>
</tr>
<tr>
<td>India</td>
<td></td>
<td>2004</td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td>2162</td>
<td></td>
</tr>
<tr>
<td>Dominican Rep.</td>
<td></td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td></td>
<td>1965</td>
<td></td>
</tr>
<tr>
<td>Peru</td>
<td></td>
<td>1933</td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>Puerto Rico</td>
<td></td>
<td>2009</td>
<td></td>
</tr>
<tr>
<td>ALSA</td>
<td>W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, W11, W12, W13</td>
<td>2087</td>
<td>&gt; 65</td>
</tr>
<tr>
<td>ATTICA</td>
<td>W1, W2</td>
<td>3037</td>
<td>18-89</td>
</tr>
<tr>
<td>CHARLS</td>
<td>W1, W2, W3, W4</td>
<td>18245</td>
<td>&gt; 45</td>
</tr>
<tr>
<td>COURAGE</td>
<td>W1, W2, W3</td>
<td>4753</td>
<td>&gt; 18</td>
</tr>
<tr>
<td>Poland</td>
<td>W1</td>
<td>4071</td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>W1</td>
<td>1976</td>
<td></td>
</tr>
<tr>
<td>ELSA</td>
<td>W1, W2, W3, W4, W5, W6, W7</td>
<td>12099</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>ENRICA</td>
<td>W1, W2, W3</td>
<td>3289</td>
<td>&gt; 60</td>
</tr>
<tr>
<td>HAPPIE</td>
<td>W1, W2</td>
<td>10728</td>
<td>45-69</td>
</tr>
<tr>
<td>Poland</td>
<td>W1, W2</td>
<td>8857</td>
<td></td>
</tr>
<tr>
<td>Czech Republic</td>
<td>W1, W2</td>
<td>7111</td>
<td></td>
</tr>
<tr>
<td>Lithuania</td>
<td>W1, W2</td>
<td>8028</td>
<td>&gt; 30</td>
</tr>
<tr>
<td>H2000</td>
<td>W1, W2, W3</td>
<td>37317</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>Finland</td>
<td>W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, W11, W12</td>
<td>37317</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>HRS</td>
<td>W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, W11, W12</td>
<td>37317</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>SHARE 20 Eu countries</td>
<td>W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, W11, W12, W13</td>
<td>30816</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>TILDA Ireland</td>
<td>W1, W2, W3</td>
<td>8504</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>UBCOS</td>
<td>W1, W2</td>
<td>61613</td>
<td>whole life</td>
</tr>
<tr>
<td>Acronym</td>
<td>Name</td>
<td>Populations</td>
<td>Waves</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>10-66</td>
<td>The 10/66 Dementia Research Group Population-Based Cohort Study</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>ALSA</td>
<td>Australian Longitudinal Study of Ageing</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>ATTICA</td>
<td>The ATTICA Study</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CHARLS</td>
<td>The China Health and Retirement Longitudinal Study</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>COURAGE</td>
<td>Collaborative Research on Ageing in Europe</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>ELSA</td>
<td>English Longitudinal Study of Ageing</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>ENRICA</td>
<td>Seniors-ENRICA</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>HAPIEE</td>
<td>The Health, Alcohol, and Psychosocial Factors in Eastern Europe Study</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Health2000</td>
<td>The Health 2000 and 2011 Surveys</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>HRS</td>
<td>Health and Retirement Study</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>JSTAR</td>
<td>Japanese Study of Aging and Retirement</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>KLoSA</td>
<td>The Korean Longitudinal Study of Ageing</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>LASA</td>
<td>Longitudinal Aging Study Amsterdam</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>LASI</td>
<td>The Longitudinal Aging Study in India</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MHAS</td>
<td>Mexican Health and Aging Study</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>SAGE</td>
<td>WHO Study on Global Ageing and Adult Health</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>SHARE</td>
<td>Survey of Health, Ageing and Retirement in Europe</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>TILDA</td>
<td>Irish Longitudinal Study on Ageing</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>UBCoS Multigen</td>
<td>The Uppsala Birth Cohort Multigenerational Study</td>
<td>Multigenerational study</td>
<td>&gt;1,000</td>
</tr>
</tbody>
</table>
1. Open to the general public in the **Maelstrom** website:
   https://www.maelstrom-research.org/mica/network/athlos

2. Restricted to ATHLOS researchers
   https://athlos.pssjd.org:8446
Data Harmonisation

- Step 0: Define the research questions, objectives and protocol
- Step 1: Assemble information and select studies
- Step 2: Define variables and evaluate harmonization potential
- Step 3: Process data
- Step 4: Estimate quality
- Step 5: Disseminate and preserve
- In summary
Select and define all those potential variables for the final harmonised dataset = DataSchema

1) Socio-demographic and economic variables (13 variables): sex, birth, education, work, income, etc. 
2) Lifestyle and health behaviors (13 variables): smoking, alcohol, physical activity, etc. 
3) Health status and functional limitations (44 variables): mobility, ADL, IADL, cognitive items, etc. 
4) Diseases (22 variables): diagnostic and onset of asthma, diabetes, etc. 
5) Life expectancy (2 variables): date of death and living status. 
6) Physical measures (12 variables): BMI, blood pressure, grip strength, etc. 
7) Psychology measures (3 variables): cognitive status, depression and anxiety symptoms 
8) Laboratory measures (7 variables): glucose, fibrinogen, cholesterol, etc. 
9) Social environment (33 variables): social network, social support, etc. 
10) Administrative variables (6 variables): id’s, date of interview, interview status, etc.
DataSchema variables were distributed among project-partners to:

1. **map and evaluate** the harmonization potential of study-specific variables across studies, and

2. **propose** a harmonization algorithm in raw text.

Example: [depression](#)
Flexibility in harmonisation decision:

- **Quantity** = number of studies to include
- **Precision** = good content equivalence

<table>
<thead>
<tr>
<th>Variable X</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat. 1</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cat. 2</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Other</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Missing</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat. 1</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cat. 2</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cat. 3</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cat. 4</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cat. 5</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cat. 6</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Other</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Missing</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Process study-specific data under a **common-format** variables to generate the harmonised datasets.

**Number of glasses of red wine currently consumed/week**

1. **Study-specific questions**
   - Study 1: In a typical week during the past 12 months, how many drinks of red wine did you drink on weekdays and on weekends?
   - Study 2: How many glasses of red wine did you drink during the last week?
   - Study 3: In an average week, how many glasses of red wine would you drink per day?
   - Study 4: How often do you drink alcohol in a week?

2. **Harmonization Potential**
   - Complete

3. **Application of algorithms**
   - #glasses/weekday + #glasses/weekend = #glasses/week
   - #glasses last week = #glasses/week
   - (#glasses/day) X 7 = #glasses/week

Could also be based on statistical models...

**Common format**

**Number of glasses of red wine currently consumed/weeks**
Data Harmonisation

- Step 0: Define the research questions, objectives and protocol
- Step 1: Assemble information and select studies
- Step 2: Define variables and evaluate harmonization potential
- Step 3: Process data
- Step 4: Estimate quality
- Step 5: Disseminate and preserve

In summary
Dataschema variable: Current depressive status of the participant (0 no depression, 1 depression)

10-66 (Cuba – wave 1):
EURO-D scale

Processing rationale:
Each item is scored 0 (symptom not present) or 1 (symptom present), generating a simple ordinal scale with a maximum score of 12. The cut-point is ≥4.
CHANGE: 'non-case (score of 4 or less)' from 0.0 to 0, 'probable EURO-D depression (score of 5 or more)' from 1.0 to 1

R algorithm:
Tdep <- tibble(id = dep$id)
Tdep$current_depre <- car::recode(dep[[2]], "0=0; 1=1; NA=999")

R algorithm:
# reversed items
dep5p45<-car::recode(dep5pstone45,"1=3; 3=1; 4=0")
(...)
# direct items
dep5p45<-car::recode(dep5pstone45,"1=0; 2=1; 3=2; 4=3")
(...)
dep$miss <- apply(dep[,2:41], 1, function(x) sum(is.na(x)))
dep$score<-apply(dep[,2:41],1, function(x) sum(x, na.rm=T))
dep$score[which(dep$miss > 4)] <- 999
dep$score[which(dep$miss==3)]<-999
dep$score[which(dep$miss==3)]=999
dep$score[which(dep$miss==2)]=999
dep$score[which(dep$miss==1)]=999
dep$score[which(dep$miss==0)]=999
depp <- tibble(id = dep$id)
depp$current_depre <- car::recode(dep[[2]],"0:15.999 = 0; 16:999 = 1")

JSTAR (5 cities – wave 1):
CES-D 20-item version test

Processing rationale:
1=not at all, 2=1-2 days, 3=3-4 days, 4= >=5 days.
CHANGE: 'not at all' from 1 to 3, '3-4 days' from 3 to 1, '5 days or more' from 4 to 0.

Harmonized variable: Current depressive status of the participant (0 no depression, 1 depression)
Algorithms and statistical processes:

- **Algorithmic transformation**
  Continuous and categorical variables or both with different but **combinable ranges** or categories (e.g. education level)

- **Simple calibration model**
  Continuous metrics with **known calibration model** (e.g. weight in kg or pounds)

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

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

- **Multiple imputation models**
  Continuous or categorical constructs measured using **overlapping scales** permitting to impute missing values (e.g. activities of daily living)
DATA HARMONISATION

- Step 0: Define the research questions, objectives and protocol
- Step 1: Assemble information and select studies
- Step 2: Define variables and evaluate harmonization potential
- Step 3: Process data
- **Step 4: Estimate quality**
- Step 5: Disseminate and preserve
- In summary
Estimate **quality** of the harmonised variables generated:

- **Validation** by comparing descriptive results between harmonised and study-specific variables.
- Check **missing** values, assess and perform imputations if necessary.
- **Quality assessment** of each harmonized variable in order to fulfil the actual definition of the variable in the DataSchema.
Data Harmonisation

- Step 0: Define the research questions, objectives and protocol
- Step 1: Assemble information and select studies
- Step 2: Define variables and evaluate harmonization potential
- Step 3: Process data
- Step 4: Estimate quality
- Step 5: Disseminate and preserve
R-markdown is being used to **process data** and **document** the quality of the harmonized variables **all at once**.

The report will include, for each **DataSchema variable** and for each **study/population/wave**, the following:

1. Description of the DataSchema variable.
2. Description of the study-specific variable and statistics.
3. R algorithm to create the new harmonised variable.
4. Statistics of the new harmonised variable and validation.
5. Document the quality assessment of the new harmonised variable.
Website with the catalogues and harmonisation reports (registration is required):

https://athlos.pssjd.org:8446

Website with the harmonised datasets stored (registration is required):

https://opal.pssjd.org:8443

GitHub account with the Rmarkdown codes (free access):

https://github.com/athlosproject/athlos-project.github.io/
In Summary

- **Understanding input data** about what and how data was collected and quality of study-specific data...

- **Ensure rigour** in the systematic harmonisation process and quality control
  Given the heterogeneity of the studies, **forcing too much** harmonisation can produce variables with too much **bias**. Therefore, many variables cannot be harmonised in all studies.

- **Ensure proper documentation**
  Document how the harmonised variables are created, to support reproducibility and long term usage.
Thank you!

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 635316