www.longitools.org

@LonglTools
LonglTools Project

Harmonising data in collaborative research projects

Justiina Ronkainen[1], Teija Juola[1] & Claire Webster[2]
[1]Research Unit of Population Health, University of Oulu, Finland, [2]Beta Technology Ltd, United Kingdom

Background

In LonglTools, we have access to data on over 11 million EU citizens, from 24 different research studies. These studies were built independently from each other so there are differences in the way the data has been collected. Collaboration between studies increases statistical power and ensures a greater diversity of people, environments and life situations are considered when making conclusions about any phenomenon of interest.

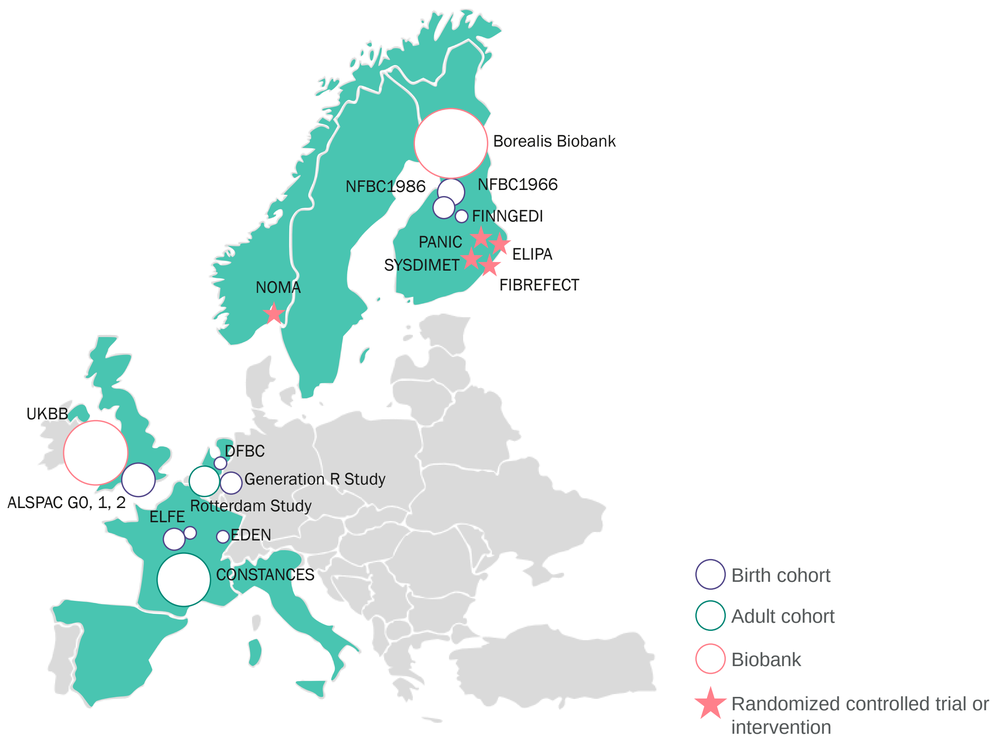

Figure 1. LonglTools studies

Collaboration facilitates cross-study comparisons, validation, and replication making research more reliable and applicable more widely. However, there are complexities when bringing together different studies or data sets to answer specific research questions. One way to overcome some of the complexities is to try to harmonise the data.

Data harmonisation is the process of making different variables consistent and comparable across different sources or study populations. The goal of data harmonisation is to make it easier to combine and analyse data from different sources, and to reduce the risk of errors or inconsistencies that can arise when working with non-standardised data. It is commonly used in research, especially in fields such as epidemiology, genetics, and social sciences, where data from multiple sources may be combined to form a larger dataset.

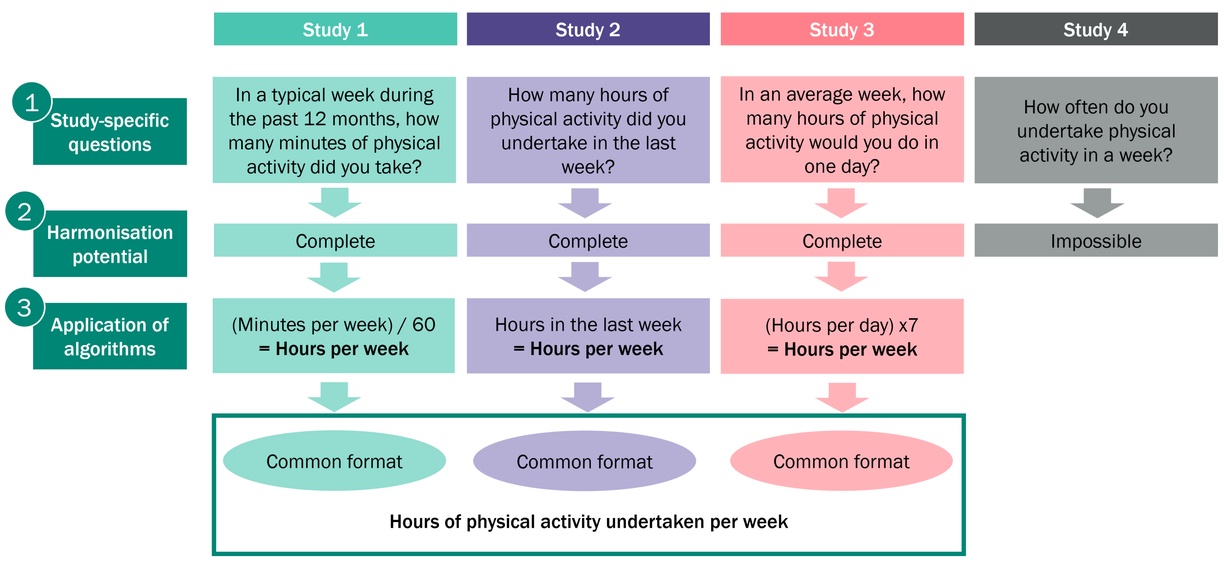

Figure 2. Example of harmonisation of physical activity between four different studies (modified from https://www.measurementtoolkit.org/concepts/harmonisation)

Harmonisation steps

1. Identifying the variables of interest - There are hundreds if not thousands of variables in each study so instead of trying to harmonise everything, together we discussed the research questions we wanted to focus on first and created a list of variables most relevant to those questions.

2. Defining variable names and labels in a consistent manner across datasets - We wanted to have a uniform way of naming the variables so, for example, the name of variables related to participant's physical activity start with letters pa (from physical activity) followed by underscore and the level of activity, e.g., pa_mod refers to moderate physical activity and pa_vig refers to vigorous physical activity.

3. Converting variables to a consistent unit of measurement and recoding categorical variables - It is important to use the same units for continuous variables and same groups for categorical variables. For example, we decided to group a person's education in high, medium, and low based on the International Standard Classification of Education 97/2011. Most of the studies had more than three groups in their education variable so they had to recode their existing groups into three levels.

4. Applying quality control checks to ensure the accuracy and completeness of the data - Harmonisation is a pedantic and sometimes laborious process as we often need to combine several variables into one. It is only human to make errors during the process so checking the harmonised data regularly and comparing it to the original is crucial to obtain reliable and good quality data.

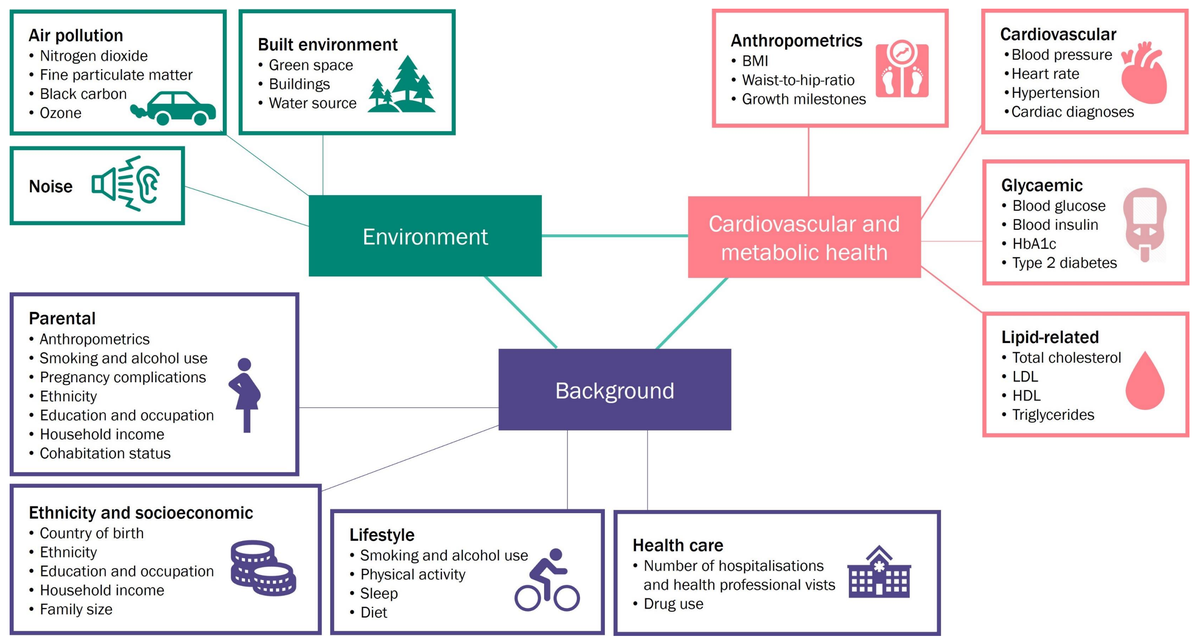

Figure 3. Harmonised variables in LonglTools

Conclusions

Data harmonisation is a vital step in multi-study collaboration. Now that we have this valuable resource of harmonised data, demonstrated in one of LonglTools' key outputs the Metadata Catalogue, we can work to unravel the complexity of the exposome. By harmonising the data and promoting collaboration, we aim to strengthen the research efforts and enhance the validity of our findings.

UNIVERSITY OF OULU

For additional information please contact:
Justiina Ronkainen
Research Unit of Population Health
Faculty of Medicine, University of Oulu
justiina.ronkainen@oulu.fi

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874739.

MEMBER OF THE European Human Exposome NETWORK

longITools HEALTH & ENVIRONMENT DYNAMICS