

[Click here to view linked References](#)

## The EU Child Cohort Network's core data: establishing a set of findable, accessible, interoperable and re-usable (FAIR) variables

Angela Pinot de Moira<sup>1\*</sup>, Sido Haakma<sup>2</sup>, Katrine Strandberg-Larsen<sup>1</sup>, Esther van Enckevort<sup>2</sup>, Marjolein Kooijman<sup>3,4</sup>, Tim Cadman<sup>5,6</sup>, Marloes Cardol<sup>7</sup>, Eva Corpeleijn<sup>7</sup>, Sarah Crozier<sup>8,9</sup>, Liesbeth Duijts<sup>3,4</sup>, Ahmed Elhakeem<sup>5,6</sup>, Johan G Eriksson<sup>10,11,12,13</sup>, Janine F Felix<sup>3,4</sup>, Sílvia Fernández-Barrés<sup>14,15,16</sup>, Rachel E Foong<sup>17,18</sup>, Anne Forhan<sup>19</sup>, Veit Grote<sup>20</sup>, Kathrin Guerlich<sup>20</sup>, Barbara Heude<sup>19</sup>, Rae-Chi Huang<sup>17</sup>, Marjo-Riitta Järvelin<sup>21,22</sup>, Anne Cathrine Jørgensen<sup>1</sup>, Tuija M. Mikkola<sup>11,23</sup>, Johanna L. T. Nader<sup>24</sup>, Marie Pedersen<sup>1</sup>, Maja Popovic<sup>25</sup>, Nina Rautio<sup>21</sup>, Lorenzo Richiardi<sup>25</sup>, Justiina Ronkainen<sup>21</sup>, Theano Roumeliotaki<sup>26</sup>, Theodosia Salika<sup>8</sup>, Sylvain Sebert<sup>21</sup>, Johan L Vinther<sup>1</sup>, Ellis Voerman<sup>3,4</sup>, Martine Vrijheid<sup>14,15,16</sup>, John Wright<sup>27</sup>, Tiffany C Yang<sup>27</sup>, Faryal Zariouh<sup>19</sup>, Marie-Aline Charles<sup>19,28</sup>, Hazel Inskip<sup>8,29</sup>, Vincent W. V. Jaddoe<sup>3,4</sup>, Morris A. Swertz<sup>2,30</sup>, Anne-Marie Nybo Andersen<sup>1</sup> for the LifeCycle Project Group

---

<sup>1</sup> Section for Epidemiology, Department of Public Health, University of Copenhagen, Denmark

<sup>2</sup> University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands

<sup>3</sup> Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA, Rotterdam, The Netherlands

<sup>4</sup> Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA, Rotterdam, The Netherlands

<sup>5</sup> Population Health Science, Bristol Medical School, Bristol BS8 2BN, UK

<sup>6</sup> MRC Integrative Epidemiology Unit at the University of Bristol, Bristol BS8 2PS, UK

<sup>7</sup> Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

<sup>8</sup> MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, UK

<sup>9</sup> NIHR Applied Research Collaboration Wessex, Southampton Science Park, Innovation Centre, 2 Venture Road, Chilworth, Southampton, SO16 7NP

<sup>10</sup> Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

<sup>11</sup> Folkhälsan Research Center, Helsinki, Finland

<sup>12</sup> Obstetrics and Gynecology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore

<sup>13</sup> Singapore Institute for Clinical Sciences (SICS), Agency for Science and Technology (A\*STAR), Singapore, Singapore

<sup>14</sup> ISGlobal, Barcelona, Spain

<sup>15</sup> Universitat Pompeu Fabra (UPF), Barcelona, Spain

<sup>16</sup> CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

<sup>17</sup> Telethon Kids Institute, Perth, WA, Australia

<sup>18</sup> School of Physiotherapy and Exercise Science, Curtin University, Perth, WA, Australia

<sup>19</sup> Université de Paris, Centre for Research in Epidemiology and Statistics (CRESS), INSERM, INRAE, Paris, France

<sup>20</sup> Division of Metabolic and Nutritional Medicine, Department of Pediatrics, Dr. von Hauner Children's Hospital, LMU University Hospital Munich, Germany

<sup>21</sup> Center for Life-Course Health Research, Faculty of Medicine, University of Oulu, P.O. Box 5000 FIN-90014, Oulu, Finland

<sup>22</sup> Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom

<sup>23</sup> Clinicum, Faculty of Medicine, University of Helsinki, Helsinki, Finland

<sup>24</sup> Department of Genetics and Bioinformatics, Division of Health Data and Digitalisation, Norwegian Institute of Public Health, Oslo, Norway

<sup>25</sup> Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Turin, Italy

<sup>26</sup> Department of Social Medicine, Faculty of Medicine, University of Crete, Heraklion, Crete, Greece

<sup>27</sup> Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK

<sup>28</sup> ELFE Joint Unit, French Institute for Demographic Studies (Ined), French Institute for Medical Research and Health (INSERM), French Blood Agency, Aubervilliers, France

<sup>29</sup> NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton, UK

<sup>30</sup> Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

\* Corresponding author: [anpi@sund.ku.dk](mailto:anpi@sund.ku.dk)

**Word Count: Abstract: 250**

**Text body: 4,275**

## **Abstract**

The Horizon2020 LifeCycle Project is a cross-cohort collaboration which brings together data from multiple birth cohorts from across Europe and Australia to facilitate studies on the influence of early-life exposures on later health outcomes. A major product of this collaboration has been the establishment of a FAIR (findable, accessible, interoperable and reusable) data resource known as the EU Child Cohort Network.

Here we focus on the EU Child Cohort Network's core variables. These are a set of basic variables, derivable by the majority of participating cohorts and frequently used as covariates or exposures in lifecourse research. First, we describe the process by which the list of core variables was established. Second, we explain the protocol according to which these variables were harmonised in order to make them interoperable. Third, we describe the catalogue developed to ensure that the network's data are findable and reusable. Finally, we describe the core data, including the proportion of variables harmonised by each cohort and the number of children for whom harmonised core data are available.

EU Child Cohort Network data will be analysed using a federated analysis platform, removing the need to physically transfer data and thus making the data more accessible to researchers. The network will add value to participating cohorts by increasing statistical power and exposure heterogeneity, as well as facilitating cross-cohort comparisons, cross-validation and replication. Our aim is to motivate other cohorts to join the network and encourage the use of the EU Child Cohort Network by the wider research community.

**Keywords:** birth cohort, cross-cohort collaboration, lifecourse epidemiology, data harmonisation, FAIR (findable, accessible, interoperable and reusable) principles.

## **Declarations**

### Funding

The LifeCycle project received funding from the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 733206 LifeCycle). All study specific acknowledgements and funding are presented in the supplementary material. This manuscript reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains

### Conflicts of interest/Competing interests

None to declare

### Ethics approval

Study-specific ethics approval statements are available in the supplementary material

### Consent to participate

Study-specific informed consent statements are available in the supplementary material

### Availability of data and material

Proposals for research based on EU Child Cohort Network data can be put forward by contacting the coordinating centre (lifecycle@erasmusmc.nl)

### Code availability

Available on the LifeCycle GitHub ( <https://github.com/lifecycle-project>)

## Introduction

Non-communicable diseases (NCDs) such as cardiovascular disease, cancer, chronic respiratory disease and diabetes represent a major global health challenge and are the leading cause of death worldwide. Of the 56.9 million deaths that occurred in 2016, 40.5 million (71%) were from NCDs (1); this number is estimated to rise to 52 million by 2030 (2). To address the growing economic and health burden that NCDs represent, the United Nations' Sustainable Development Goal (SDG) target 3.4 aims to reduce premature mortality due to NCDs by one third by 2030 through prevention, treatment and promotion of mental health and wellbeing (1).

Early-life offers an important window of opportunity for achieving this target. Evidence strongly suggests that environmental conditions and exposures during intrauterine and early postnatal life can influence anatomical, physiological and biochemical processes and, in so doing, impact future health (3). Longitudinal pregnancy and child cohort studies provide a means of investigating this phenomenon, including how early-life exposures influence health trajectories, and identifying potential early-life interventions to improve health outcomes (4). However, such studies are expensive to establish and maintain, which often prohibits the large-scale studies required to investigate rare outcomes or exposures, or conduct more advanced statistical analyses to investigate, for example, causality or lifecourse health trajectories.

Cross-cohort collaborations offer a cost-effective approach to increase the statistical power of such analyses. They also provide other benefits such as increased exposure heterogeneity, facilitated cross-cohort comparisons, the ability to cross-validate, replicate and establish the generalisability of findings, and the opportunity to share expertise and knowledge. In recent years, a number of such collaborations have been successfully established, for example the CHICOS ([www.chicosproject.eu](http://www.chicosproject.eu)), BioSHARE (5), HELIX (6) ([www.projecthelix.eu](http://www.projecthelix.eu)), PACE (7), EGG/EAGLE (8), ESCAPE (9) ([www.escapeproject.eu](http://www.escapeproject.eu)) and Enrieco (10) ([www.enrieco.org](http://www.enrieco.org)) projects, which have led to the identification of a number of associations that may have otherwise gone unobserved (11-28). More recently, in 2017, building on expertise gained from these collaborations, the Horizon 2020-funded LifeCycle project was established (29) ([www.lifecycle-project.eu](http://www.lifecycle-project.eu)).

LifeCycle aims to facilitate the utilisation of data from mainly European, but also some non-European, cohort studies for research. It has a particular focus on preconception, fetal and early childhood exposures and their influence on cardio-metabolic, respiratory and mental health trajectories. To achieve its aim, LifeCycle has established the EU Child Cohort Network, a sustainable data resource and infrastructure which is built around making each participating cohort's data findable, accessible, interoperable and reusable (FAIR) (30). The network currently holds data on

approximately 250,000 children and their parents from an initial 16 European and one Australian cohort.

An overview of the EU Child Cohort Network, including the data management and governance structure on which the network is based, plus its primary research themes, was provided by Jaddoe *et al.* in a previous edition of this journal (29). Here we provide a detailed description of the EU Child Cohort Network's core variables, which are a set of basic variables, derivable by the majority of participating cohorts and required for most analyses in lifecourse research. We describe firstly the process by which the list of core variables was established; secondly the protocol developed to harmonise these core data, which defines the harmonisation process adopted generally within LifeCycle; thirdly the catalogue developed to ensure that all EU Child Cohort Network data are both findable and reusable; finally the core data themselves, including the variables harmonised by each cohort and the total number of children with harmonised data. Our aims are to: i) enable an accurate assessment of the quality and validity of the harmonised core data through transparency of our methods; ii) motivate other cohorts to contribute to the network; iii) encourage the use of the EU Child Cohort Network's data by the wider scientific community.

## Methods

### Participating cohorts

An overview of the 17 cohorts that established the EU Child Cohort Network is provided in Table 1. Further details of each cohort can be found in Jaddoe *et al.* 2020 (29), the EU Child Cohort Network Variable Catalogue (<http://catalogue.lifecycle-project.eu>) and each cohort's profile paper (31-49). The network is open for other cohorts to join, provided they meet the following criteria: i) commenced before or during pregnancy or in infancy; ii) plan to follow-up or already have followed-up the cohort throughout childhood; iii) are willing to harmonise data and make them available to researchers using the network. Cohorts can join the network by contacting the coordinating centre ([lifecycle@erasmusmc.nl](mailto:lifecycle@erasmusmc.nl)). Similarly, proposals for research based on EU Child Cohort Network data can be put forward by both LifeCycle partners and external researchers by also contacting the coordinating centre ([lifecycle@erasmusmc.nl](mailto:lifecycle@erasmusmc.nl)). Proposals for research may be based on all EU Child Cohort Network cohorts or a subset of cohorts with available data; they may also include requests for further data harmonisation, which can likewise be restricted to a subset of cohorts with data.

### Harmonisation

The EU Child Cohort Network's core variables are a set of basic, predominantly "lowest common denominator" variables, derivable by the majority of participating cohorts and frequently needed as covariates or exposures in lifecourse research. The process adopted in LifeCycle to establish and harmonise these core variables for the EU Child Cohort Network can be broken down into eight steps; an overview of these steps is displayed in Fig 1. A glossary of the key elements and concepts described in this paper is also provided in Box 1.

#### *Step 1: Establishing a preliminary list of target core variables*

LifeCycle partners with expertise in a wide range of fields including lifecourse epidemiology, public health, environmental epidemiology, biology, statistics, paediatrics, obstetrics, economics, demography, epigenomics and data science, met in a dedicated workshop (June 2017) to identify a preliminary list of core early-life stressors and exposures related to cardio-metabolic, respiratory and mental health outcomes using a consensus approach. This initial list was then further modified by drawing on experiences from other previous collaborative efforts such as MOBAND (50) and CHICOS

([www.chicosproject.eu](http://www.chicosproject.eu)), and through consulting the literature and experts in the field, before being circulated amongst LifeCycle partners for further comment.

*Steps 2, 3 & 4: Collating codebooks, evaluating the harmonisation potential of each variable and finalising a list of target core variables*

All cohorts participating in LifeCycle were requested to provide the coordinating team with cohort metadata (codebooks, questionnaires, instrument documentation, etc.). From these, the potential for each cohort to derive each target variable was established. The core variable list was then adapted in an iterative manner to achieve a balance between precision and inclusivity, ensuring a maximum number of cohorts could contribute data for numerous variables while maintaining data validity. Where possible, international standards and classification schemes were applied. For example, the International Standard Classification of Occupation 1988 1-digit codes (51) were used to categorise parental occupation; the International Standard Classification of Education 97/2011 schemes (52, 53) were used to classify parental education; the WHO fetal growth charts (54) were used to establish size-for-gestational-age; the EUROCAT guide was used for classifying congenital anomalies. For some key exposures such as maternal smoking, breastfeeding, childcare attendance and gestational age, several variables were included, with some variables capturing more information but at the cost of fewer cohorts being able to derive the variables. Repeated measures were also included, to capture the dynamic, longitudinal nature of many variables.

*Step 5: Pilot harmonisation*

Data harmonisation was staggered across cohorts. First, an initial pilot harmonisation was conducted among four cohorts covering the majority of target core variables (the Danish National Birth Cohort, the EDEN mother-child cohort, the Generation R study and the Southampton Women's Survey). This enabled any potential issues in the core variable list to be identified and rectified. During the pilot harmonisation, the core variable list was revised iteratively through electronic communication, a workshop and a final teleconference.

*Step 6: Data harmonisation and local quality control*

Harmonisation for the EU Child Cohort Network was carried out locally by each participating cohort. This avoided any transfer of data but carried the risk of harmonisation protocols being interpreted



differently by different cohorts. To limit this possibility, a detailed harmonisation manual was drawn up by the coordinating team, and supervision and feedback was maintained between the coordinating centre and each of the cohorts. The harmonisation manual is available to download from the LifeCycle website (<https://lifecycle-project.eu>); it includes: i) a final, annotated list of core variables, which, for each variable, includes: a variable name, a precise definition, a label, units, data type, permissible values and guidelines for what constitutes partial vs. complete harmonisation (see Box 1 for definitions of partial vs. complete harmonisation); ii) relevant scale conversions; iii) relevant reference tables (e.g. WHO fetal growth charts, the EUROCAT guide for classifying congenital anomalies etc.). The harmonisation manual was circulated to cohorts in May 2018 and harmonisation of core variables by all cohorts was completed by May 2020. The duration of time that it took a cohort to harmonise all core variables ranged from three to eight months.

Once data were harmonised, each cohort was provided with detailed quality control instructions and scripts to check: i) that variables matched the descriptions provided in the core variable list (name, datatype, values); ii) for outliers or improbable values; iii) for inconsistencies between non-repeated measures (e.g. all mothers coded as not smoking during pregnancy were also coded as smoking zero cigarettes during pregnancy); iv) for inconsistencies between repeated measures (e.g. children reducing height over time). Any inconsistencies identified were investigated on a cases-by-case basis to establish which values were legitimate and which were errors, also in light of the other data available.

#### *Step 7a: Uploading harmonisation descriptions to the EU Child Cohort Network Variable Catalogue*

To facilitate the utilisation of EU Child Cohort Network data for research, and ensure the complete and accurate documentation of harmonisation, an online catalogue of EU Child Cohort Network variables was developed using the Molgenis platform (55) (<http://catalogue.lifecycle-project.eu>). This open source, searchable catalogue includes detailed descriptions of each variable included in the EU Child Cohort Network (variable name, data type, values, unit and description), as well as details of which cohorts have harmonised each variable, whether that harmonisation was complete or partial, an explanation of how the variable was harmonised, plus the syntax and descriptions of the source variables used by each cohort to derive the variable (Fig 2). For the core variables, documentation of harmonisation was conducted by each cohort and uploaded to the catalogue after harmonisation was complete.

The catalogue has been built using a logical tree structure, but variables can also be located using a search function (Fig 3). There are plans to also incorporate descriptive summary statistics for each harmonised variable. Thus, the EU Child Cohort Network Variable Catalogue provides a comprehensive overview of the EU Child Cohort Network's data, ensuring they are both findable and reusable, as well as contributing to the longer-term sustainability of the network.

#### *Step 7b: Uploading data to a data management platform for the federated analysis of data*

To help ensure the sustainability and accessibility of the EU Child Cohort Network, an IT infrastructure has been implemented enabling the federated analysis of data. Full details of this infrastructure are given elsewhere (29, 56, 57). Briefly, this infrastructure consists of secure Opal servers (58) located either at each host institution or on outsourced IT infrastructures. Once harmonisation is complete, each cohort uploads their harmonised data to their Opal server, where they remain stored, behind secure firewalls. Individual-level data are accessed via an RStudio Open Source central analysis server (<https://rstudio.com/products/rstudio/#rstudio-server>) using the R-based platform DataSHIELD (56), which sends blocks of code to each Opal server and then combines the summary statistics that are sent back by each Opal server. There is no transfer of individual participant data to the researcher and a number of disclosure control filters ensure analyses are non-disclosive, thus the many ethical, legal and societal implications of transferring data from one site to another are avoided.

#### *Step 8: Central quality-control*

Quality of harmonised data was assessed at the central level by creating summary statistics for each core variable in R/DataSHIELD. This was to identify outliers and improbable values and inconsistencies in data as outlined above, but also to identify large inconsistencies between cohorts. Where large inconsistencies were found, sampling and recruitment methods and differences in the instruments used to collect data were investigated, as well as the harmonisation process itself, in order to establish to what extent these differences were real vs. an artefact of differing methodology.

## Results

Table 1 provides an overview of the 17 cohorts currently contributing data to the EU Child Cohort Network. As of June 2020, the network holds data on just under 250,000 children and their parents, with contributing cohorts ranging in size from 967 to 76,569 children. This is an initial number and will increase as new cohorts and their parent-child triads join the network.

First and last year of recruitment of cohorts ranged between 1934 (HBCS) and 2016 (NINFEA) respectively. Mean age of children at recruitment ranged from -1084 days before birth (approximately -3 years, in SWS, which recruited mothers before conception) to 17 days postpartum (in CHOP). The majority of mothers enrolled in the cohorts were recruited during pregnancy (13 of the 17 currently participating cohorts).

Tables 2 and 3 summarise some key characteristics of the mother-child dyads from each cohort currently contributing data to the EU Child Cohort Network. Of note is the variation in the proportion of children born small and large for gestational age (ranging from 2.2% in CHOP to 11.2% in BiB and from 2.7% in CHOP to 14.2% in NFBC1986 for SGA and LGA respectively) and the proportion of children ever breastfed (ranging from 73.4% in EDEN to 99.6% in HBCS). Also of note is the variation in the proportion of mothers with a high level of education (ranging from 3.3% in NFBC1966, most likely reflecting the earlier year of recruitment of this cohort, to 67.5% in MoBa) and the proportion of mothers who smoked during their pregnancy (ranging from 7.6% in NINFEA which is based in Italy, where the prevalence of smoking among women and especially pregnant women is known to be lower (59), to 33.1% in Rhea). Multiparity ranged between 27% in NINFEA and 69% in NFBC1966.

Although we focus here on describing the EU Child Cohort Network's core variables, the network also includes variables relating to the early-life exposome, encompassing both the external environment (socio-economic, migration, urban environment and lifestyle factors) and internal environment (determined from biological markers such as DNA methylation, RNA expression and metabolomics), and outcome variables relating to cardio-metabolic, respiratory and mental health. An overview of all the themes of the EU Child Cohort Network is provided in Fig 3, together with estimates of the total number of variables included in each theme. Due to the fact that new variables are continuously being added to the network with the inception of new research projects, these numbers are highly conservative.

The core variables consist of a set of 130 basic, principally lowest common denominator variables, available in the majority of participating cohorts and required for many analyses within the scope of LifeCycle and other lifecourse epidemiology research themes. Of these, seven are so-called "meta

variables”, consisting of mother, child, pregnancy, and cohort identifiers, and variables providing the age of recruitment and country of cohort. The remaining variables consist of 96 non-repeated variables and 17 yearly-repeated variables with up to 18 measures between the ages of 0 and <18 years, together capturing maternal, paternal and child health, lifestyle, socio-demographic characteristics, mother’s obstetric history, birth outcomes and household exposures. There are also two trimester-repeated variables capturing maternal smoking and alcohol consumption during pregnancy, four yearly-repeated variables with up to four measures between the ages of 0 and <4 years capturing childcare and four monthly-repeated variables with up to 216 height or weight measures between the ages of 0 and 215 months. The full list of EU Child Cohort Network core variables is provided in Online Resource 1 and also in the EU Child Cohort Network Variable Catalogue (<http://catalogue.lifecycle-project.eu>). Since the EU Child Cohort Network Variable Catalogue is dynamic and regularly expanded with both new variables and newly participating cohorts, the statistics reported there may differ from what is presented here.

Excluding the seven meta-variables, the percentage of core variables harmonised by cohorts ranged from 21% for HBCS to 92% for ELFE (Fig 4). Missing variables are due to cohorts not having the data required to harmonise the variable. Twelve of the 17 cohorts currently included in the EU Child Cohort Network were able to harmonise at least 50% of core variables completely, and 12 of the 17 cohorts were able to harmonise at least 75% of core variables either completely or partially.

Figs 5-7 give an overview of the number of EU Child Cohort Network children (i.e. from all cohorts combined) with harmonised core data. Of the non-repeated core variables (Fig 5), themes with the most complete data are those relating to maternal characteristics (specifically, age at birth, height, smoking during pregnancy, parity) and child-related characteristics (specifically, sex, gestational age at birth, birth weight, birth length, size for gestational age and death of the child), with more than 217,000 children as of June 2020 having harmonised data relating to these exposures. Notably fewer children have data relating to mother and father’s country of birth and ethnic background, perhaps due to their sensitive nature (60).

An overview of the number of EU Child Cohort Network children with harmonised yearly-repeated core variables, which allow for time-varying exposure statuses, is displayed in Fig 6. Over 80% of children in the network have at least one harmonised measure of cohabitation status, mother’s occupational status, mother’s education level, father’s occupational status, father’s education level, and child’s exposure to pets and cigarette smoke, whilst relatively few children (<10%) have harmonised data on household income. For growth data (Fig 7), the greatest density of measures in the network is between the ages of 0 and <1 year, with a total of 780,993 and 732,202 weight and

height measurements available between these ages respectively, an average of three weight and height measures per child. Large amounts of growth data are also available for ages 1 - <2 years and 7 - <8 years, with over 72% and 47% of children having harmonised weight and height data at these ages respectively, whilst relatively few children currently have weight and height data from 14 years and onwards, partly because many cohorts have not yet reached that age.

## Discussion

The Horizon 2020 LifeCycle Project is a collaboration between scientists from more than 17 pregnancy and birth cohorts from across Europe and Australia. It builds upon the expertise gained from previous collaborations such as the CHICOS, Enrieco and BioSHARE projects in order to establish an open and sustainable data resource known as the EU Child Cohort Network so as to facilitate research on the influence of early-life stressors on later health outcomes.

Here we have described the EU Child Cohort Network, focussing on its core variables, including the protocol developed to harmonise these data and thus make them interoperable. We have also described the EU Child Cohort Network Variable Catalogue, developed to ensure that these and other data in the network are both findable and re-usable. These data will be analysed using a federated analysis platform, meaning there is no need to physically transfer data, and so data are ultimately more accessible to the researcher.

As well as the harmonised core data described here, the EU Child Cohort Network also contains data relating to the early-life exposome, and repeated measures of cardio-metabolic, respiratory and mental health. An additional feature of the network is the varied social, cultural and political environments of the cohorts. Thus, the EU Child Cohort Network constitutes an invaluable data resource, not only in terms of the number of participants included, but also in terms of its breadth, depth and diversity. This will ultimately enable the application of a range of analytical approaches to help infer causality, and identify possible target groups for improved cardio-metabolic, respiratory and mental health across the lifecourse.

However, the creation of such a data resource is not without its limitations. Firstly, the resources required to create a common dataset, i.e. harmonise data, should not be underestimated. Harmonising data is difficult, time consuming and requires considerable investment by all involved. Although central harmonisation, whereby individual participant data are sent to one coordinating centre which harmonises all variables, is often viewed as the more optimal approach, this is not without its drawbacks. Firstly, there are many ethico-legal challenges surrounding the transfer of data; secondly, it takes considerable investment by the data manager to become acquainted with a cohort's data, scaled up 17 times in the case of the EU-Child Cohort Network, potentially leading to errors. It is for these reasons, and the fact that the EU Child Cohort Network is an open network, such that new cohorts are invited to join and are continually joining, that LifeCycle opted for local harmonisation. Here, harmonisation is carried out locally by each cohort, coordinated by a central coordinating centre. This of course has the limitation that harmonisation protocols may be interpreted differently by different cohorts. We have tried to limit this possibility in LifeCycle by

providing detailed instructions and maintaining regular contact with data managers. We have also implemented a number of data quality checks, applied both locally and centrally. These include checks to ensure that harmonised variables match those detailed in the harmonisation manual and to identify outliers or improbable values, or any inconsistencies in measures within or between cohorts. Good documentation of all harmonisation steps is key to diagnosing any inconsistencies, which we have ensured in LifeCycle by establishing the EU Child Cohort Network Variable Catalogue.

Another drawback of data harmonisation is that the end product is often the “lowest common denominator”. For any given variable, some cohorts will inevitably have more detailed variables than other cohorts. In an attempt to create a common variable achievable by all cohorts, more detailed variables are stripped down to simpler versions, inevitably resulting in some loss of information. This may also involve deciding that in some cohorts there is insufficient data to harmonise a variable. Harmonisation is thus a balancing act between retaining as much information as possible while ensuring data are fully comparable (61).

So, if the creation of a common dataset is such a tremendous task and the end product may, in some instances, be less detailed than the original data, why bother? Increased statistical power is one obvious advantage. Combining data from several cohorts to increase power allows rarer, but equally important and often more devastating (62), diseases and rare determinants to be studied. Larger sample sizes also allow for more powerful statistical analyses, such as exploring multiple interactions, complex nonlinear relationships, small effects or dose responses (63). While national registers offer the possibility of creating birth cohorts of an order of magnitude larger than the EU Child Cohort (for e.g. Nordic register-based cohort studies (64, 65)), these typically lack the in-depth lifestyle and behavioural data obtained from questionnaires, or physiological data obtained from detailed clinical examinations. National register data are in addition likely to offer less diversity with respect to social, cultural and political environment. Cross-cohort collaborations also allow fine resolution biological data to be shared, such as medical images or metagenomic data, that may be prohibitively costly to obtain from the entire cohort and therefore only collected from a sub-population of the cohort.

A larger sample size is not the only benefit of cross-cohort collaborations. Combining data also offers the opportunity to study populations typically under-represented in cohort studies, for example individuals from lower socio-economic backgrounds or ethnic minority groups. Heterogeneities between cohorts can be utilised to strengthen causal inference. For example, differing confounding structures allows the untangling of true associations, whilst replication of findings across different populations with differing gene pools, and cultural and socio-economic structures, helps to rule out

chance findings while also establishing the generalisability of results. Geographical, intergenerational and period effects can also be examined to find new associations and generate new hypotheses.

While it could be argued that an easier and potentially less time-consuming approach to combining data from several studies is the more conventional systematic review and meta-analysis of published data, this has a number of disadvantages compared to individual participant data (IPD) meta-analysis. Published data are often subject to selective reporting and publication bias, lack harmonised measures, and offer limited scope and flexibility in terms of statistical analysis, and few opportunities, if any, for data checking (66, 67).

The added value that the collaboration itself brings should also be highlighted: the opportunity to share ideas and methodology, learn from each other, and ultimately strengthen research outputs. Also the increased use of data and exchange opportunities for researchers. Scientific collaboration also facilitates the dissemination of both results and ideas/hypotheses, as well as creating opportunities for interdisciplinary research.

In conclusion, the EU Child Cohort Network offers an invaluable data resource for studying how early-life exposures influence health trajectories throughout the lifecourse. This is both in terms of the number of its participants, and the breadth and depth of its data. Here we share the approach taken within LifeCycle to harmonise the network's core data and describe the EU Child Cohort Network Variable Catalogue established to ensure that the network's data are both findable and reusable. We also highlight some of the great benefits of cross-cohort collaboration. Having hopefully convinced the reader of the benefits of the EU Child Cohort Network and similar cross-cohort collaborations, we end with a plea to other cohorts to join the network and share their data, and to researchers to utilise this incredible resource. Both cohorts and researchers can join the network by contacting [lifecycle@erasmusmc.nl](mailto:lifecycle@erasmusmc.nl).



## References

1. GBD 2017 SDG Collaborators. Measuring progress from 1990 to 2017 and projecting attainment to 2030 of the health-related Sustainable Development Goals for 195 countries and territories: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018;392(10159):2091-138. doi:10.1016/S0140-6736(18)32281-5
2. Mendis S. Global Status Report on Noncommunicable Diseases 2014: World Health Organisation 2014.
3. Hanson MA, Gluckman PD. Early developmental conditioning of later health and disease: physiology or pathophysiology? *Physiol Rev*. 2014;94(4):1027-76. doi:10.1152/physrev.00029.2013
4. Larsen PS, Kamper-Jorgensen M, Adamson A, et al. Pregnancy and birth cohort resources in Europe: a large opportunity for aetiological child health research. *Paediatric and perinatal epidemiology*. 2013;27(4):393-414. doi:10.1111/ppe.12060
5. 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. doi:10.1186/1742-7622-10-12
6. Maitre L, de Bont J, Casas M, et al. Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. *BMJ open*. 2018;8(9):e021311. doi:10.1136/bmjopen-2017-021311
7. Felix JF, Joubert BR, Baccarelli AA, et al. Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. *International journal of epidemiology*. 2018;47(1):22-3u. doi:10.1093/ije/dyx190
8. Middeldorp CM, Felix JF, Mahajan A, consortium EAGLE, Early Growth Genetics c, McCarthy MI. The Early Growth Genetics (EGG) and EARly Genetics and Lifecourse Epidemiology (EAGLE) consortia: design, results and future prospects. *European journal of epidemiology*. 2019;34(3):279-300. doi:10.1007/s10654-019-00502-9
9. Pedersen M, Giorgis-Allemand L, Bernard C, et al. Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *Lancet Respir Med*. 2013;1(9):695-704. doi:10.1016/S2213-2600(13)70192-9
10. Vrijheid M, Casas M, Bergstrom A, et al. European birth cohorts for environmental health research. *Environmental health perspectives*. 2012;120(1):29-37. doi:10.1289/ehp.1103823
11. Birks L, Casas M, Garcia AM, et al. Occupational Exposure to Endocrine-Disrupting Chemicals and Birth Weight and Length of Gestation: A European Meta-Analysis. *Environmental health perspectives*. 2016;124(11):1785-93. doi:10.1289/EHP208
12. Casas M, den Dekker HT, Kruithof CJ, et al. Early childhood growth patterns and school-age respiratory resistance, fractional exhaled nitric oxide and asthma. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology*. 2016;27(8):854-60. doi:10.1111/pai.12645
13. Casas M, den Dekker HT, Kruithof CJ, et al. The effect of early growth patterns and lung function on the development of childhood asthma: a population based study. *Thorax*. 2018;73(12):1137-45. doi:10.1136/thoraxjnl-2017-211216
14. LifeCycle Project-Maternal Obesity Childhood Outcomes Study Group, Voerman E, Santos S, et al. Association of Gestational Weight Gain With Adverse Maternal and Infant Outcomes. *Jama*. 2019;321(17):1702-15. doi:10.1001/jama.2019.3820
15. Gruziova O, Xu CJ, Yousefi P, et al. Prenatal Particulate Air Pollution and DNA Methylation in Newborns: An Epigenome-Wide Meta-Analysis. *Environmental health perspectives*. 2019;127(5):57012. doi:10.1289/EHP4522
16. Haworth S, Shapland CY, Hayward C, et al. Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. *Nat Commun*. 2019;10(1):357. doi:10.1038/s41467-018-07863-x
17. Horikoshi M, Beaumont RN, Day FR, et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature*. 2016;538(7624):248-52. doi:10.1038/nature19806


18. Kupers LK, Monnereau C, Sharp GC, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nat Commun.* 2019;10(1):1893. doi:10.1038/s41467-019-09671-3
19. Leventakou V, Roumeliotaki T, Martinez D, et al. Fish intake during pregnancy, fetal growth, and gestational length in 19 European birth cohort studies. *Am J Clin Nutr.* 2014;99(3):506-16. doi:10.3945/ajcn.113.067421
20. Patro Golab B, Santos S, Voerman E, et al. Influence of maternal obesity on the association between common pregnancy complications and risk of childhood obesity: an individual participant data meta-analysis. *Lancet Child Adolesc Health.* 2018;2(11):812-21. doi:10.1016/S2352-4642(18)30273-6
21. Santos S, Eekhout I, Voerman E, et al. Gestational weight gain charts for different body mass index groups for women in Europe, North America, and Oceania. *BMC Med.* 2018;16(1):201. doi:10.1186/s12916-018-1189-1
22. Santos S, Voerman E, Amiano P, et al. Impact of maternal body mass index and gestational weight gain on pregnancy complications: an individual participant data meta-analysis of European, North American and Australian cohorts. *BJOG.* 2019;126(8):984-95. doi:10.1111/1471-0528.15661
23. Sharp GC, Salas LA, Monnereau C, et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. *Hum Mol Genet.* 2017;26(20):4067-85. doi:10.1093/hmg/ddx290
24. Sonnenschein-van der Voort AM, Arends LR, de Jongste JC, et al. Preterm birth, infant weight gain, and childhood asthma risk: a meta-analysis of 147,000 European children. *The Journal of allergy and clinical immunology.* 2014;133(5):1317-29. doi:10.1016/j.jaci.2013.12.1082
25. Strandberg-Larsen K, Poulsen G, Bech BH, et al. Association of light-to-moderate alcohol drinking in pregnancy with preterm birth and birth weight: elucidating bias by pooling data from nine European cohorts. *European journal of epidemiology.* 2017;32(9):751-64. doi:10.1007/s10654-017-0323-2
26. Stratakis N, Roumeliotaki T, Oken E, et al. Fish Intake in Pregnancy and Child Growth: A Pooled Analysis of 15 European and US Birth Cohorts. *JAMA pediatrics.* 2016;170(4):381-90. doi:10.1001/jamapediatrics.2015.4430
27. Voerman E, Santos S, Patro Golab B, et al. Maternal body mass index, gestational weight gain, and the risk of overweight and obesity across childhood: An individual participant data meta-analysis. *PLoS Med.* 2019;16(2):e1002744. doi:10.1371/journal.pmed.1002744
28. Warrington NM, Beaumont RN, Horikoshi M, et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat Genet.* 2019;51(5):804-14. doi:10.1038/s41588-019-0403-1
29. Jaddoe VVW, Felix JF, Andersen AN, et al. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. *European journal of epidemiology.* 2020;35(7):709-24. doi:10.1007/s10654-020-00662-z
30. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018. doi:10.1038/sdata.2016.18
31. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *International journal of epidemiology.* 2013;42(1):111-27. doi:10.1093/ije/dys064
32. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International journal of epidemiology.* 2013;42(1):97-110. doi:10.1093/ije/dys066
33. Wright J, Small N, Raynor P, et al. Cohort Profile: the Born in Bradford multi-ethnic family cohort study. *International journal of epidemiology.* 2013;42(4):978-91. doi:10.1093/ije/dys112
34. Koletzko B, von Kries R, Closa R, et al. Lower protein in infant formula is associated with lower weight up to age 2 y: a randomized clinical trial. *Am J Clin Nutr.* 2009;89(6):1836-45. doi:10.3945/ajcn.2008.27091

35. Olsen J, Melbye M, Olsen SF, et al. The Danish National Birth Cohort--its background, structure and aim. *Scandinavian journal of public health*. 2001;29(4):300-7. doi:10.1177/14034948010290040201
36. L'Abée C, Sauer PJ, Damen M, Rake JP, Cats H, Stolk RP. Cohort Profile: the GECKO Drenthe study, overweight programming during early childhood. *International journal of epidemiology*. 2008;37(3):486-9. doi:10.1093/ije/dym218
37. Chatzi L, Plana E, Daraki V, et al. Metabolic syndrome in early pregnancy and risk of preterm birth. *American journal of epidemiology*. 2009;170(7):829-36. doi:10.1093/aje/kwp211
38. Eriksson JG, Forsen T, Tuomilehto J, Osmond C, Barker DJ. Early growth and coronary heart disease in later life: longitudinal study. *Bmj*. 2001;322(7292):949-53. doi:10.1136/bmj.322.7292.949
39. Jaddoe VW, van Duijn CM, Franco OH, et al. The Generation R Study: design and cohort update 2012. *European journal of epidemiology*. 2012;27(9):739-56. doi:10.1007/s10654-012-9735-1
40. Guxens M, Ballester F, Espada M, et al. Cohort Profile: the INMA--Infancia y Medio Ambiente--(Environment and Childhood) Project. *International journal of epidemiology*. 2012;41(4):930-40. doi:10.1093/ije/dyr054
41. Magnus P, Irgens LM, Haug K, et al. Cohort profile: the Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology*. 2006;35(5):1146-50. doi:10.1093/ije/dyl170
42. Jarvelin MR, Hartikainen-Sorri AL, Rantakallio P. Labour induction policy in hospitals of different levels of specialisation. *Br J Obstet Gynaecol*. 1993;100(4):310-5. doi:10.1111/j.1471-0528.1993.tb12971.x
43. Jarvelin MR, Sovio U, King V, et al. Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension*. 2004;44(6):838-46. doi:10.1161/01.HYP.0000148304.33869.ee
44. Richiardi L, Baussano I, Vizzini L, et al. Feasibility of recruiting a birth cohort through the Internet: the experience of the NINFEA cohort. *European journal of epidemiology*. 2007;22(12):831-7. doi:10.1007/s10654-007-9194-2
45. Newnham JP, Evans SF, Michael CA, Stanley FJ, Landau LI. Effects of frequent ultrasound during pregnancy: a randomised controlled trial. *Lancet*. 1993;342(8876):887-91. doi:10.1016/0140-6736(93)91944-h
46. Inskip HM, Godfrey KM, Robinson SM, et al. Cohort profile: The Southampton Women's Survey. *International journal of epidemiology*. 2006;35(1):42-8. doi:10.1093/ije/dyi202
47. Magnus P, Birke C, Vejrup K, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology*. 2016;45(2):382-8. doi:10.1093/ije/dyw029
48. Heude B, Forhan A, Slama R, et al. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *International journal of epidemiology*. 2016;45(2):353-63. doi:10.1093/ije/dyv151
49. Charles MA, Thierry X, Lanoe JL, et al. Cohort Profile: The French national cohort of children (ELFE): birth to 5 years. *International journal of epidemiology*. 2020;49(2):368-9j. doi:10.1093/ije/dyz227
50. Tollanes 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. doi:10.1136/bmjopen-2016-012777
51. International Labour Organization. ISCO International Standard Classification of Occupations. 2004. <https://www.ilo.org/public/english/bureau/stat/isco/isco88/index.htm>.
52. Schneider S. The International Standard Classification of Education 2011. *Comparative Social Research*. 2013;30:365-79. doi:10.1108/S0195-6310(2013)0000030017
53. United Nations Educational, Scientific and Cultural Organisation. International Standard Classification of Education ISCED1997.

54. Kiserud T, Piaggio G, Carroli G, et al. The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight. *PLoS Med.* 2017;14(1):e1002220. doi:10.1371/journal.pmed.1002220
55. Swertz MA, Dijkstra M, Adamusiak T, et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics.* 2010;11 Suppl 12:S12. doi:10.1186/1471-2105-11-S12-S12
56. Gaye A, Marcon Y, Isaeva J, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology.* 2014;43(6):1929-44. doi:10.1093/ije/dyu188
57. Wilson R, Butters O, Avraam D, et al. Burton PR. DataSHIELD—New Directions and Dimensions. *Data Science Journal* 2017, 16: 21. *Data Science Journal.* 2017;16(21):1-21.
58. Open Source Software for BioBanks. <http://www.obiba.org/>.
59. Chatenoud L, Chiaffarino F, Parazzini F, Benzi G, La Vecchia C. Prevalence of smoking among pregnant women is lower in Italy than England. *Bmj.* 1999;318(7189):1012. doi:10.1136/bmj.318.7189.1012
60. Hasnain-Wynia R, Baker DW. Obtaining data on patient race, ethnicity, and primary language in health care organizations: current challenges and proposed solutions. *Health Serv Res.* 2006;41(4 Pt 1):1501-18. doi:10.1111/j.1475-6773.2006.00552.x
61. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: consolidating data harmonization--how to obtain quality and applicability? *Am J Epidemiol.* 2011;174(3):261-4; author reply 5-6. doi:10.1093/aje/kwr194
62. The Lancet Diabetes E. Spotlight on rare diseases. *Lancet Diabetes Endocrinol.* 2019;7(2):75. doi:10.1016/S2213-8587(19)30006-3
63. Lin MF, Lucas HC, Shmueli G. Too Big to Fail: Large Samples and the p-Value Problem. *Inform Syst Res.* 2013;24(4):906-17. doi:10.1287/isre.2013.0480
64. Bengtsson J, Dich N, Rieckmann A, Hulvej Rod N. Cohort profile: the DANish LIFE course (DANLIFE) cohort, a prospective register-based cohort of all children born in Denmark since 1980. *BMJ open.* 2019;9(9):e027217. doi:10.1136/bmjopen-2018-027217
65. Mortensen LH, Cnattingius S, Gissler M, et al. Sex of the first-born and obstetric complications in the subsequent birth. A study of 2.3 million second births from Denmark, Finland, Norway, and Sweden. *Acta Obstet Gynecol Scand.* 2020. doi:10.1111/aogs.13872
66. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof.* 2002;25(1):76-97. doi:10.1177/0163278702025001006
67. Stewart LA, Tierney JF, Clarke M, on behalf of the Cochrane Individual Patient Data Meta-analysis Methods Group. Reviews of individual patient data. Higgins JPT, Green S, editors. Chichester: Wiley & Sons Ltd; 2008.





 CATALOGUE USER GUIDE ▾ COHORT DESCRIPTIONS DATASHIELD ▾ CONTACT

[Back to catalogue](#)

Harmonisation

DescriptionVariables usedScript syntax

**(Cohort 1)**  
Participants asked to state whether they have had asthma "recently" or "ever". Coded 1 if they indicate having either recently or ever, 0 if not. Partially harmonised as based on self-report

DescriptionVariables usedScript syntax

**(Cohort 2)**  
asthma\_m = 1 if a053=1.  
Mothers reporting no history of asthma (a052=2), or mothers without doctor diagnosed asthma (a053=2) or who were unsure of a history of asthma (a052=3) coded as having no history of asthma (asthma\_m = 0).  
Mothers who were unsure of whether their asthma was doctor diagnosed (a053=3) coded as missing.

LifeCycle variable: **Maternal history of asthma before pregnancy**

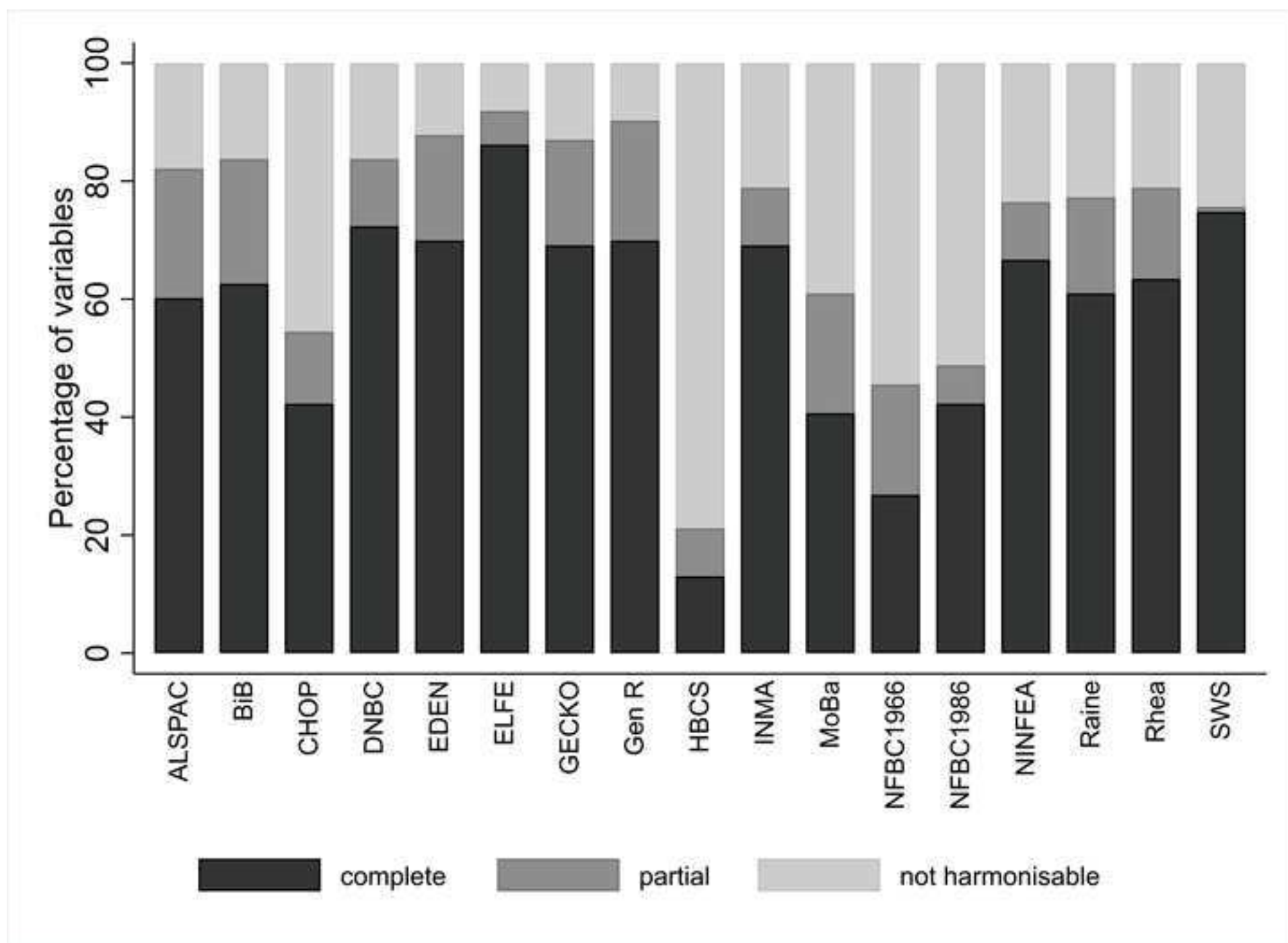
Variable	asthma_m
Label	Maternal history of asthma before pregnancy
Data type	Binary
Values	0 = No 1 = Yes
Comments	Where data are available, asthma should be doctor diagnosed. If no information is available on doctor diagnosis, the variable is partially harmonised. Mothers who were asked whether their asthma was diagnosed by a doctor but who did not know or were unsure, should be coded as missing.

**LifeCycle** CATALOGUE USER GUIDE ▾ COHORT DESCRIPTIONS DATASHIELD ▾ CONTACT

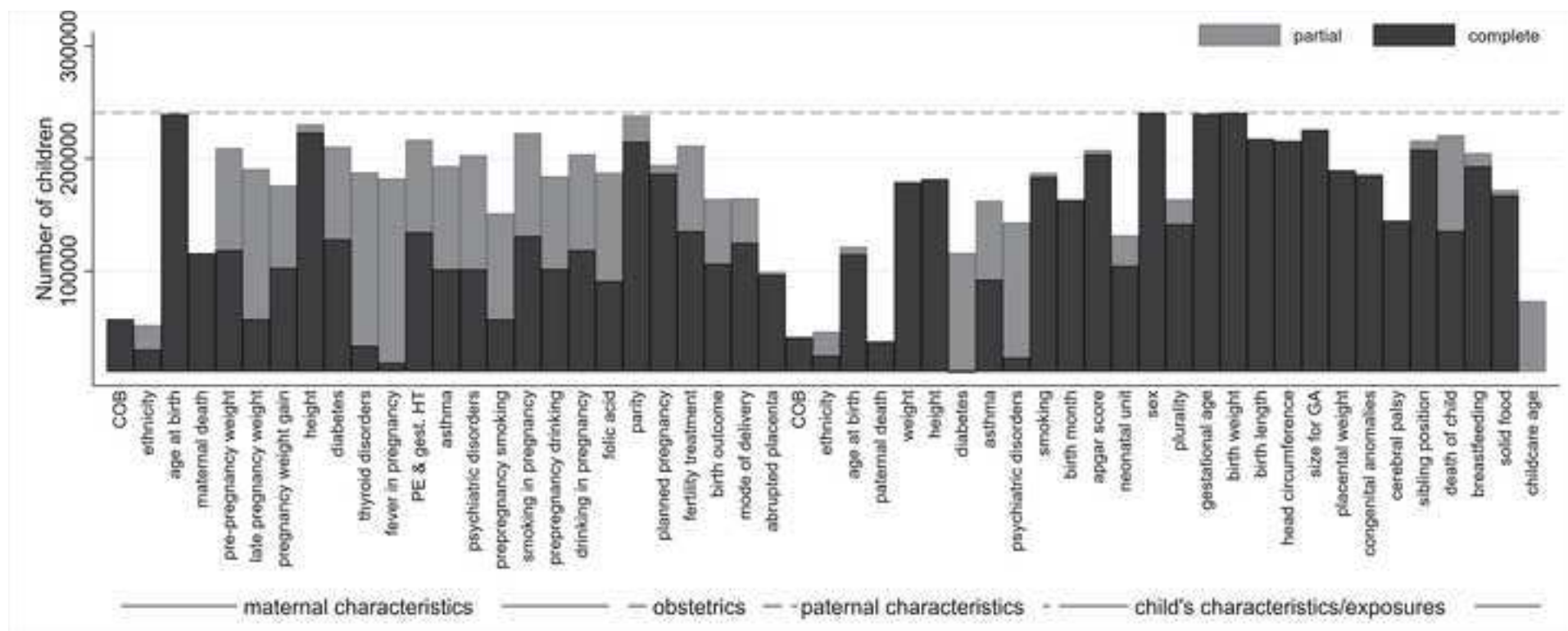
## Catalogue

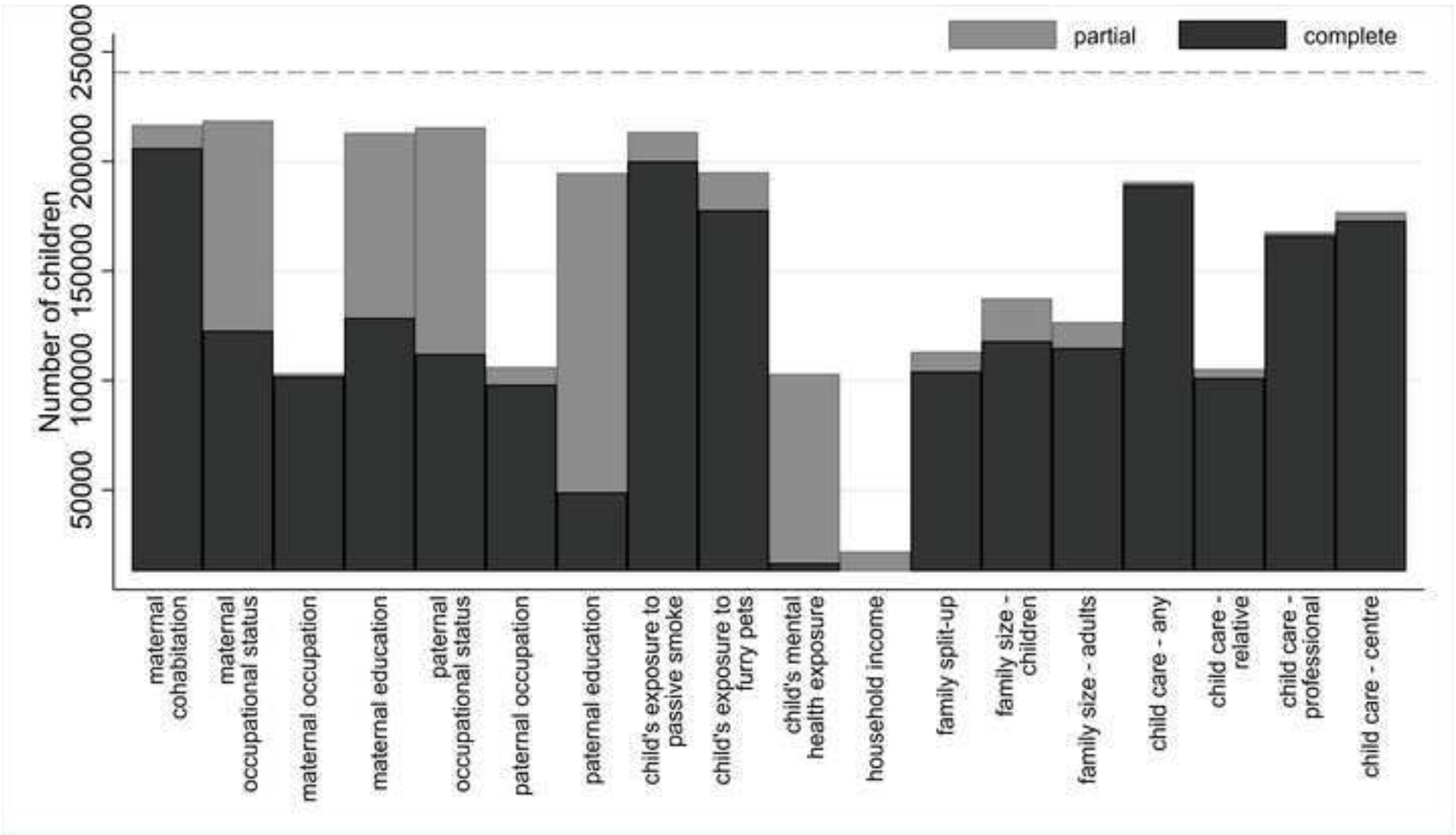
Search catalogue

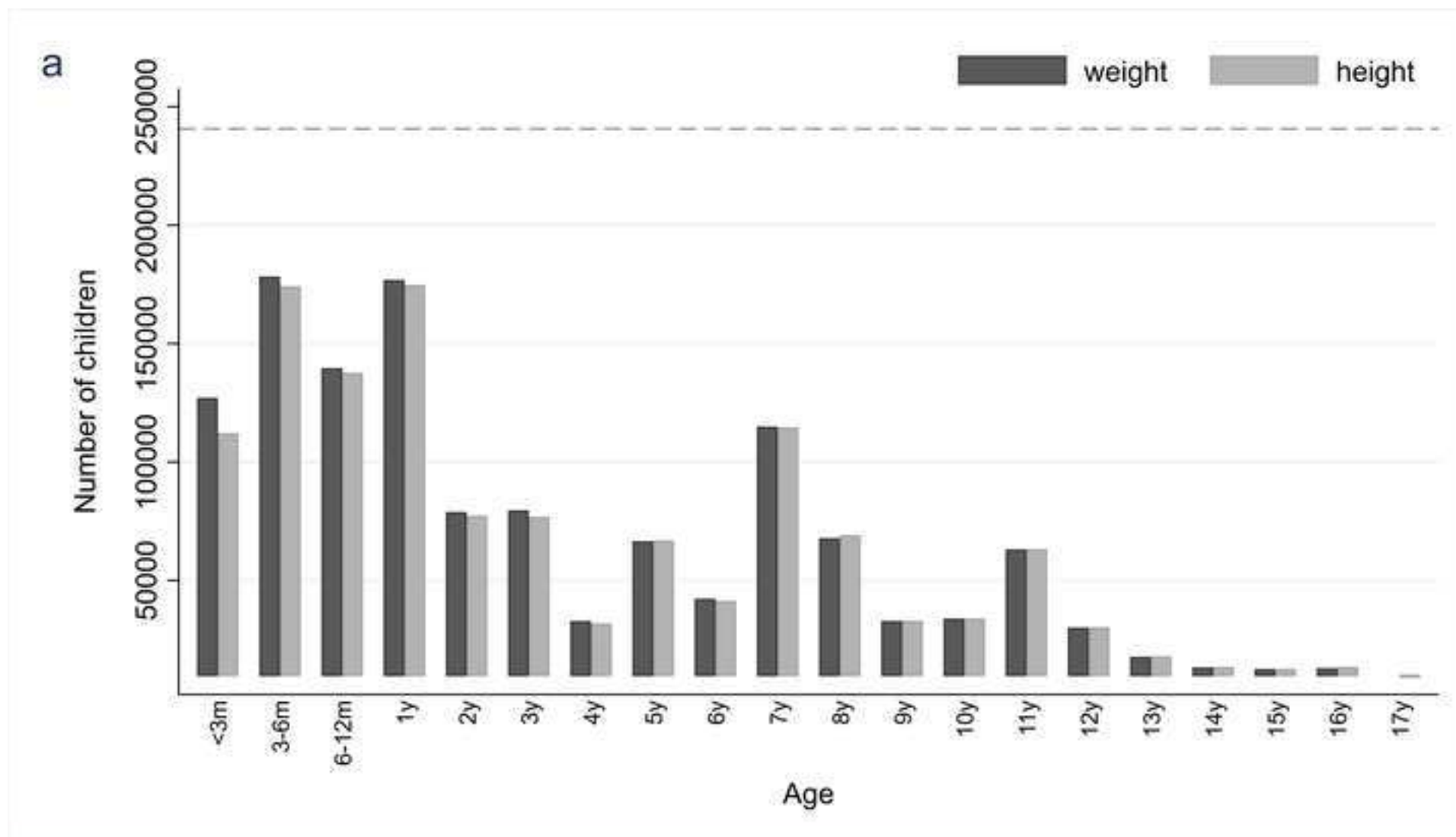
- + ... **Meta variables** (8 variables)
- ... **Maternal Characteristics**
  - + ... **Socio-demographic characteristics** (15 variables covering: maternal cohabitation status<sup>1</sup>, occupational status<sup>1</sup>, occupational code<sup>1</sup>, education<sup>1</sup>, country of birth, ethnicity, age at birth, death)
  - + ... **Health-related characteristics** (47 variables covering: maternal weight, height, diabetes, thyroid disorders, fever during pregnancy, preeclampsia, gestational hypertension, cardiovascular health indicators<sup>2</sup>, asthma, eczema, allergy, psychiatric disorders<sup>1</sup>)
  - + ... **Lifestyle characteristics** (17 variables covering: maternal smoking<sup>3</sup>, alcohol consumption<sup>3</sup>, folic acid supplementation)
  - + ... **Diet** (33 variables covering: method & timing of assessment, food groups, supplements, nutrients, DASH diet index)
  - + ... **Obstetric characteristics** (7 variables covering: maternal parity, family planning, fertility treatment, mode of delivery, birth outcome, placental abruption)
- ... **Paternal Characteristics**
  - + ... **Socio-demographic characteristics** (27 variables covering: paternal occupational status<sup>1,4,5</sup>, occupational code<sup>1,4,5</sup>, education<sup>1,4,5</sup>, country of birth<sup>4</sup>, ethnicity<sup>4</sup>, age at birth<sup>4</sup>, death<sup>4</sup>)
  - + ... **Health-related characteristics** (11 variables covering: paternal weight<sup>4</sup>, height<sup>4</sup>, diabetes, asthma, psychiatric disorders, cardiovascular health indicators<sup>2</sup>)
  - + ... **Lifestyle characteristics** (3 variables covering: paternal smoking)
- ... **Child**
  - + ... **Socio-demographic characteristics** (2 variables covering child's country of birth)
  - + ... **Birth outcomes** (19 variables covering: year & month of birth, Apgar score, sex, neonatal unit transfer, plurality, gestational age, birth weight, birth length, birth head circumference, size for gestational age, placenta weight, congenital anomalies, cerebral palsy, sibling position)
  - + ... **Health-related characteristics** (185 variables covering: child's height<sup>6</sup>, weight<sup>6</sup>, head circumference<sup>6</sup>, cardiovascular health indicators<sup>6</sup>, wheeze<sup>1</sup>, asthma<sup>1</sup>, asthma medication<sup>1</sup>, upper and lower respiratory tract infections<sup>1</sup>, lung function<sup>1</sup>, allergy<sup>1</sup>, allergic sensitisation<sup>1</sup>, eczema<sup>1</sup>, death)
  - + ... **Adult health-related characteristics** (4 variables covering: asthma, COPD)
  - + ... **Exposures/Lifestyle/Environment** (31 variables covering: breastfeeding, solid food introduction, childcare<sup>7</sup>, passive smoking<sup>1</sup>, pets<sup>1</sup>, family mental health<sup>1</sup>, sleep, outdoor play, screen time, multi-behavioural profiles)
  - + ... **Cognitive domains** (40 variables covering: gross motor<sup>1</sup>, fine motor<sup>1</sup>, non-verbal intelligence<sup>1</sup>, working memory<sup>1</sup>, language<sup>1</sup>)
  - + ... **Behavioural domains** (34 variables covering: internalising problems<sup>1</sup>, externalising problems<sup>1</sup>, ADHD<sup>1</sup>, ASD<sup>1</sup>)
- + ... **Household Characteristics** (6 variables covering: household income<sup>1</sup>, family split up<sup>1</sup>, family size<sup>1</sup>)
- + ... **Urban environment** (248 variables covering: air pollution<sup>3,8</sup>, natural spaces<sup>3,8</sup>, built environment<sup>3,8</sup>, social context<sup>3,8</sup>, traffic<sup>3,8</sup>, noise<sup>3,8</sup>, unhealthy food environment<sup>3,8</sup>, meteorology<sup>3,8</sup>)

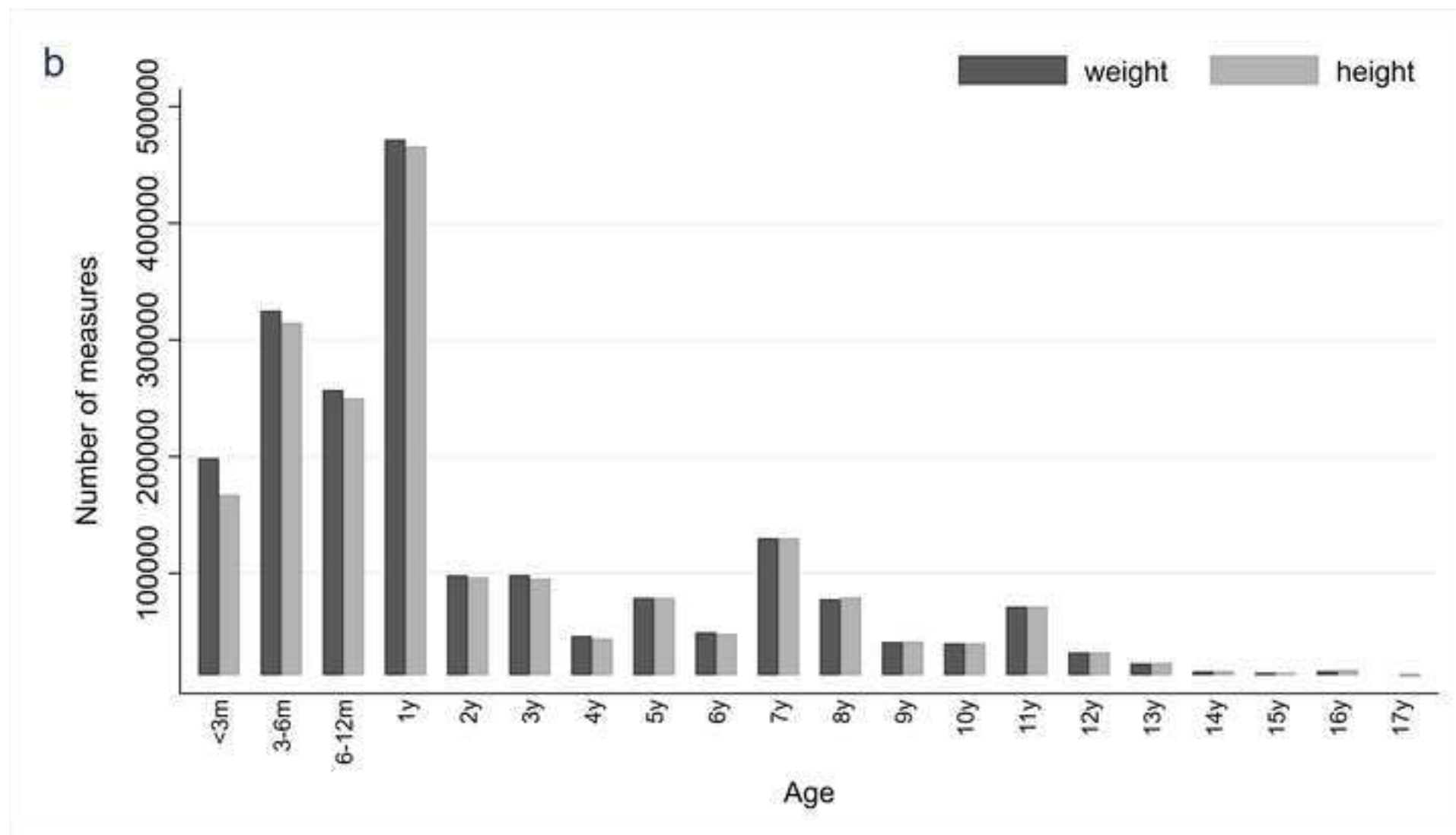












## Figures Titles and Legends

### **Fig 1 The process adopted in LifeCycle to establish and harmonise the core variables for the EU Child Cohort Network**

### **Fig 2 An illustration of the EU Child Cohort Network Variable Catalogue displaying the LifeCycle variable “maternal history of asthma before pregnancy”**

Displayed is a description of the target EU Child Cohort Network variable and how the variable was harmonised in two separate cohorts. Note: descriptions from two separate cohorts are displayed on the same page for illustrative purposes only.

### **Fig 3 An illustration of the EU Child Cohort Network Variable Catalogue’s menu structure giving an overview of the themes included in the EU Child Cohort Network and the number of variables included in each theme**

<sup>1</sup> Including yearly-repeated variables with up to 18 measures between the ages of 0 and <18 years

<sup>2</sup> Including weekly-repeated variables with up to 43 measures taken between gestational weeks 0 and <43

<sup>3</sup> Including trimester-repeated variables with separate measures for the first, second and third trimesters

<sup>4</sup> Including separate variables indicating the type of father the variable relates to (biological, social father, social mother, unknown)

<sup>5</sup> Including separate variables relating to secondary father-figures

<sup>6</sup> Including monthly-repeated variables with up to 216 measures between the ages of 0 and <216 months

<sup>7</sup> Including yearly-repeated variables with up to four measures between the ages of 0 and <4 years

<sup>8</sup> Including yearly-repeated variables with up to 13 measures between the ages of 0 and <13 years

### **Fig 4 Percentage of EU Child Cohort Network core variables harmonised by each cohort**

The figure displays the percentage of the 123 core variables listed in Online Resource 1 (excluding meta-variables) harmonised by each cohort. Shading of bars displays the degree of matching within each cohort: black bars represent percentage of completely harmonised variables; dark grey bars represent percentage of partially harmonised variables; light grey bars represent percentage of variables that were not harmonizable (impossible harmonisation).

**Fig 5 Harmonised non-repeated core variables in the EU Child Cohort Network**

Bars display the number of children with either a partially (grey bars) or completely (black bars) harmonised core variable for each of the main themes/exposures. The dashed line represents the total number of children (240,684), as of June 2020, contributing data to the EU Child Cohort Network with all three of the following variables harmonised: i) birth weight, ii) sex, iii) at least one height or weight measurement taken at  $\geq 1$  year.

COB, country of birth; PE, pre-eclampsia; gest. HT, gestational hypertension; size for GA, size for gestational age.

**Fig 6 Number of children in the EU Child Cohort Network with yearly-repeated measure core variables**

Bars display the number of children with at least one measure between the ages of zero and three (child-care variables) or zero and seventeen (all other variables), either partially (grey bars) or completely (black bars) harmonised. The dashed line represents the total number of children (240,684), as of June 2020, contributing data to the EU Child Cohort Network with all three of the following variables harmonised: i) birth weight, ii) sex, iii) at least one height or weight measurement taken at  $\geq 1$  year.

**Fig 7 Weight and height data in the EU Child Cohort Network**

Graphs display a) number of children in the network with at least one weight (dark grey bars) or height (light grey bars) measure at  $<3$  months, 3-6 months, 6-12 months and yearly intervals from 1 – 17 years; b) total number of weight (dark grey bars) and height (light grey bars) within each age band (i.e. one child may contribute multiple measurements within each age band).

## Tables

**Table 1 Pregnancy and child cohorts contributing data to the EU Child Cohort Network as of June 2020**

<b>Cohort (full name)</b>	<b>Country</b>	<b>Recruitment</b>	<b>Enrolment period</b>	<b>Age at last follow-up (y)</b>	<b>N<sup>a</sup></b>
ALSPAC (Avon Longitudinal Study of Parents & Children)	UK	1991-1992	Pregnancy	25	10,742
BiB (Born in Bradford)	UK	2007-2011	Pregnancy	9	12,397
CHOP (The EU Childhood Obesity Programme)	Germany, Belgium, Italy, Spain and Poland	2002-2004	Birth	11	1,280
DNBC (Danish National Birth Cohort)	Denmark	1996-2002	Pregnancy	18	72,157
EDEN (Study on the pre- & early postnatal determinants of child health & development)	France	2003-2005	Pregnancy	8	1,676
ELFE (Etude Longitudinale Francaise depuis l'Enfance)	France	2011	Birth	7	10,825
GECKO (Groningen Expert Center for Kids with Obesity Drenthe Cohort)	The Netherlands	2006-2007	Pregnancy	10	2,682
Gen R (Generation R)	The Netherlands	2002-2006	Pregnancy	17	8,534
HBCS (Helsinki Birth Cohort Study)	Finland	1934-1944	Birth	76	13,343
INMA (INMA-Infancia y Medio Ambiente (Environment and Childhood Project))	Spain	1997-2008	Pregnancy	18	1,900
MoBa (Norwegian Mother, Father and Child Cohort Study)	Norway	1999-2008	Pregnancy	14	76,569
NFBC1966 (Northern Finland Birth Cohort 1966)	Finland	1966	Pregnancy	46-48	7,810
NFBC1986 (Northern Finland Birth Cohort 1986)	Finland	1985-1986	Pregnancy	33-35	8,372

NINFEA (Nascita e INFanzia: gli Effetti dell'Ambiente)	Italy	2005-2016	Pregnancy	13	6,018
Raine (The Raine Study)	Australia	1989-1992	Pregnancy	26	2,491
Rhea (Mother Child Cohort in Crete)	Greece	2007-2008	Pregnancy	7	967
SWS (Southampton Women's Survey)	UK	1998-2007	Preconception	9	2,921

---

<sup>a</sup> Number of children from the cohort contributing data to the EU Child Cohort Network and with all three of the following variables harmonised: i) birth weight, ii) sex, iii) at least one height or weight measurement taken at  $\geq 1$  year



**Table 2 Child-related characteristics of cohorts contributing data to the EU Child Cohort Network**

Cohort	N <sup>a</sup>	Female, n (%)	GA (weeks), mean (SD)	Birth weight (g), mean (SD)	SGA <sup>b</sup> , n (%)	LGA <sup>c</sup> , n (%)	Ever breastfed, n (%)
ALSPAC	10,742	5,313 (49.5)	40.0 (1.9)	3,408 (555)	644 (6.0)	1,015 (9.5)	7,213 (75.8)
BiB	12,397	5,980 (48.2)	39.5 (1.8)	3,212 (557)	1,385 (11.2)	562 (4.5)	3,228 (78.7)
CHOP	1,280	659 (51.5)	40.4 (1.2)	3,297 (351)	28 (2.2)	34 (2.7)	901 (70.4)
DNBC	72,157	35,464 (49.1)	39.9 (1.8)	3,565 (582)	2,281 (3.2)	10,046 (14.0)	55,214 (98.3)
EDEN	1,676	802 (47.9)	39.7 (1.7)	3,283 (506)	118 (7.0)	60 (3.6)	1,230 (73.4)
ELFE	10,825	5,277 (48.7)	39.6 (1.5)	3,322 (488)	644 (6.0)	535 (5.0)	7,858 (74.8)
GECKO	2,682	1,332 (49.7)	39.8 (1.6)	3,542 (548)	87 (3.3)	357 (13.4)	1938 (79.4)
Gen R	8,534	4,229 (49.6)	40.3 (1.9)	3,400 (576)	615 (7.4)	541 (6.5)	6,013 (91.8)
HBCS	13,343	6,369 (47.7)	39.8 (1.8)	3,407 (479)	NA	NA	11,110 (99.6)
INMA	1,900	923 (48.6)	39.9 (1.6)	3,263 (467)	139 (7.3)	70 (3.7)	1,648 (88.6)
MoBa	76,569	37,390 (48.8)	39.8 (1.9)	3,576 (578)	2,725 (3.6)	7,377 (9.6)	71,768 (93.7)
NFBC1966	7,810	3,628 (46.5)	40.5 (1.9)	3,491 (530)	378 (5.3)	703 (9.9)	4,550 (86.0)
NFBC1986	8,372	4,112 (49.1)	39.8 (1.7)	3,560 (546)	259 (3.1)	1,186 (14.2)	NA
NINFEA	6,018	2,951 (49.0)	39.7 (1.7)	3,238 (493)	471 (7.9)	200 (3.3)	5,502 (92.1)
Raine	2,491	1,218 (48.9)	39.1 (2.3)	3,299 (602)	142 (7.0)	146 (7.2)	2,082 (89.7)
Rhea	967	459 (47.5)	38.7 (1.5)	3,183 (455)	56 (5.9)	51 (5.3)	805 (86.5)
SWS	2,921	1,411 (48.3)	39.7 (1.8)	3,441 (547)	126 (4.3)	259 (8.9)	2,376 (82.5)

Values are mean (standard deviation) or n (valid percent)

<sup>a</sup> Number of children from the cohort contributing data to the EU Child Cohort Network and with all three of the following variables harmonised: i) birth weight, ii) sex, iii) at least one height or weight measurement taken at  $\geq 1$  year

<sup>b</sup> Birth weight  $\leq 5$ th percentile for gestational age (in completed weeks) using the WHO fetal growth charts (52) as the growth standard

<sup>c</sup> Birth weight  $\geq 95$ th percentile for gestational age (in completed weeks) using the WHO fetal growth charts (52) as the growth standard

GA, gestational age at birth; SGA, small for gestational age; LGA, large for gestational age; NA, data not available

**Table 3 Mother-related characteristics of cohorts contributing data to the EU Child Cohort Network**

Cohort	N <sup>a</sup>	Maternal age at birth (y), mean (SD)	Education level, n (%)			Ethnicity, n (%)			Multiparous, n (%)	Smoked in pregnancy, n (%)
			high	medium	low	White	Black, Asian or minority ethnic	mixed		
ALSPAC	10,742	29.2 (4.6)	1,444 (14.2)	6,954 (68.6)	1,741 (17.2)	9,874 (98.3)	169 (1.7)	-	5,629 (54.8)	2,468 (26.0)
BiB	12,397	27.6 (5.6)	2,534 (26.8)	1,502 (15.9)	5,420 (57.3)	4,290 (41.8)	5,783 (56.3)	200 (1.9)	7,259 (60.8)	1,659 (16.2)
CHOP	1,280	30.2 (5.0)	336 (26.3)	640 (50.2)	300 (23.5)	1,232 (96.4)	46 (3.6)	-	652 (51.0)	416 (32.6)
DNBC	72,157	30.1 (4.2)	33,700 (52.3)	14,067 (21.8)	16,655 (25.9)	NA	NA	NA	37,964 (52.6)	17,580 (24.7)
EDEN	1,676	29.7 (4.8)	938 (56.2)	636 (38.1)	94 (5.6)	1437 (99.1)	7 (0.5)	6 (0.4)	911 (54.5)	413 (24.7)
ELFE	10,825	30.8 (4.7)	7,240 (66.9)	3,063 (28.3)	521 (4.8)	8,706 (83.9)	963 (9.3)	705 (6.8)	5,673 (53.0)	1,779 (16.6)
GECKO	2,682	30.7 (4.4)	900 (35.9)	724 (28.9)	885 (35.3)	2,400 (95.5)	70 (2.8)	43 (1.7)	1,591 (59.9)	411 (15.4)
Gen R	8,534	30.7 (5.2)	3,448 (45.3)	3,380 (44.4)	778 (10.2)	4,606 (57.1)	2,665 (33.0)	799 (9.9)	3,691 (44.8)	1,888 (25.9)
HBCS	13,343	28.4 (5.4)	NA	NA	NA	NA	NA	NA	6,861 (51.4)	NA
INMA	1,900	31.8 (4.2)	661 (35.2)	768 (40.9)	449 (23.9)	1,802 (95.7)	80 (4.3)	-	810 (44.5)	588 (31.4)
MoBa	76,569	30.4 (4.4)	48,804 (67.5)	22,166 (30.6)	1,354 (1.9)	NA	NA	NA	39,262 (51.7)	6,194 (8.1)
NFBC1966	7,810	28.1 (6.7)	254 (3.3)	1,033 (13.5)	6,387 (83.2)	NA	NA	NA	5,387 (69.1)	1,569 (20.7)
NFBC1986	8,372	27.8 (5.5)	1,735 (23.7)	2,744 (37.4)	2,856 (38.9)	NA	NA	NA	5,499 (65.9)	1,975 (23.7)
NINFEA	6,018	33.2 (4.2)	3,799 (63.6)	1,923 (32.2)	253 (4.2)	NA	NA	NA	1,548 (27.0)	453 (7.6)
Raine	2,491	27.9 (5.8)	465 (20.1)	633 (27.3)	1,221 (52.7)	2,175 (89.2)	264 (10.8)	-	1,275 (52.3)	666 (27.3)
Rhea	967	29.7 (4.9)	304 (32.1)	481 (50.7)	163 (17.2)	926 (99.8)	2 (0.2)	-	524 (54.9)	290 (33.1)
SWS	2,921	30.2 (3.8)	837 (28.7)	1,730 (59.2)	345 (11.8)	2,799 (95.8)	105 (3.6)	16 (0.5)	1,409 (48.3)	428 (15.4)

Values are mean (standard deviation) or n (valid percent)

<sup>a</sup> Number of children from the cohort contributing data to the EU Child Cohort Network and with all three of the following variables harmonised: i) birth weight, ii) sex, iii) at least one height or weight measurement taken at >= 1 year. Mothers who contributed more than one child to a cohort are counted more than once in the table.

**Box 1 A glossary of the key elements and concepts in LifeCycle**

Term	Definition
Complete harmonisation	The ability to derive the variable as described in the harmonization manual, both in definition and format
Data harmonisation	The process of creating a common dataset from disparate datasets
DataSHIELD	An infrastructure and series of R packages that enables the remote and non-disclosive analysis of individual participant data
EU Child Cohort Network	A network bringing together existing data from more than 250,000 European and Australian children and their parents
Federated data analysis	Centralised analysis of individual participant data where data are stored on local servers and do not leave the host institution
Harmonisation manual	A manual containing a list of target variables together with instructions for their harmonisation
Impossible harmonisation	The complete inability to derive the variable due to no or limited information
Horizon2020 LifeCycle Project	A collaboration between scientists from more than 17 existing pregnancy and child cohort studies
EU Child Cohort Network Variable Catalogue	An online catalogue providing an overview of available data in the EU Child Cohort Network, including details of how data have been created ( <a href="http://catalogue.lifecycle-project.eu">http://catalogue.lifecycle-project.eu</a> )
LifeCycle core variables	A set of basic variables, derivable by the majority of cohorts participating in LifeCycle and frequently required in lifecourse analyses
Opal	A data warehouse that is integrated with R and the DataSHIELD platform, allowing the analysis of data without the physical sharing or disclosing of individual participant data
Partial harmonisation	The ability to derive the variable as described but with some loss of information