



Grant Agreement 101155977

Milestone 2
Detailed protocol for data harmonization

Version 2.5 | October 2026

Milestone Title	Detailed protocol for data harmonization
Brief Description	The aim of this document is to provide guidelines on the harmonization strategy of data on socio-environmental, inter-personal, and intra-personal determinants of health and health behaviour as well as on relevant health behaviours and health outcomes that are available in the four cohorts/studies (ABCD, BiB, BREATHE, WALNUTS)
Work Package No	1
Milestone No	2
Author(s)	V1.0: Paulina Jedynak, Nerea Moreno, Carmen Peuters, Boris Cheval, Andrea Dietrich, Darren Healy, Daniel Phipps, Nikos Ntoumanis, Payam Dadvand, Jordi Julvez V2.5: Nerea Moreno, Darren Healy, Fabienne Fox, Huiyun Ding, Paulina Jedynak, Mai Nguyen, Carmen Peuters, Payam Dadvand, Boris Cheval, Jordi Julvez
Milestone Due Date	March 31, 2025
Actual Delivery Date	V1.0: March 31, 2025 V2.5: October 1, 2026
Nature of the Milestone	Protocol
Dissemination Level	Openly accessible

Document History

Date	Version	Contributors
31/01/2025	1.0	Paulina Jedynak, Nerea Moreno, Carmen Peuters, Boris Cheval, Andrea Dietrich, Darren Healy, Daniel Phipps, Nikos Ntoumanis, Payam Dadvand, Jordi Julvez

10/06/2025	2.0	Content update: Paulina Jedynak, Nerea Moreno, Darren Healy, Fabienne Fox, Huiyun Ding
15/09/2026	2.1	Review and revision: Nerea Moreno
23/09/2026	2.2	Review and revision: Mai Nguyen, Carmen Peuters, Payam Dadvand, Jordi Julvez
29/09/2026	2.3	Revision based on review: Nerea Moreno
30/09/2026	2.4	Review and revision: Boris Cheval, Mai Nguyen, Nerea Moreno
01/10/2026	2.5	Review and revision: Mai Nguyen

Table of Contents

1. Introduction	5
2. Aim	5
3. Method	5
3.1. Selection of the variables to be harmonized	5
3.2. Age groups targeted by the intervention	6
3.3. Comparability across data from different cohorts/studies	6
3.4. Variable transformations	7
3.5. Prioritization	7
3.6. Cross-validation	8
3.7. Naming of the variables	8
3.8. Outlier and extreme values detection	8
3.9. Harmonized dataset structure	9
4. Variables to be harmonized	9
4.1. Health behaviours and health outcomes	9
1. Lifestyle	9
2. Mental health	10
3. Weight status	11

4. Health-related quality of life	11
4.2. Determinants of health behaviours and health	15
1. Living environment and urban exposome	15
2. Intra-individual determinants	15
3. Inter-individual determinants	16
4.3. Potential main confounding factors	19
5. Next steps	20
6. Appendix 1	21
Data preparation	22
Naming	24
Variable Naming Example 1 (categorical variable): Child's ethnicity	29
Variable Naming Example 2 (continuous variable): Child sleep	29
Transformation and scoring	30
Validation of measures	31
Quality control	32
Variable Harmonization Example 1 (categorical variable): child's ethnicity	34
Variable Harmonization Example 2 (continuous variable): Child's sleep	36
Harmonization sharing	37
Script standardization	39
Cross-validation of the scripts in GitHub	39
Output	39
Output Example 1 (categorical variable): Child's ethnicity	39
Output Example 2 (continuous variable): Child's sleep	41
Output Example 2 (continuous variable): Child's sleep	43
Repository of working documents	43

1. Introduction

The aim of the WPI Tasks 1.2 and 1.3 of YEAH! project is to perform hypothesis driven analyses and data driven exposome analyses to identify determinants of health behaviours and health of children and adolescents using existing cohort data. We will use data from the following four European cohorts or studies: Amsterdam Born Children and their Development (ABCD, the Netherlands), Born in Bradford (BiB, UK), BRain dEvelopment and Air polluTion ultrafine particles in scHool childrEn (BREATHE, Spain) and Long-chain Polyunsaturated Fatty Acids and Adolescent Brain Development (WALNUTS, Spain). A prerequisite to conduct the analyses in these Tasks is the harmonization of the data generated within each cohort/study so that different measurements are comparable across cohorts.

2. Aim

The aim of this document is to provide guidelines on the harmonization strategy of data on socio-environmental, inter-personal, and intra-personal determinants of health and health behaviour as well as on relevant health behaviours and health outcomes that are available in the four cohorts/studies (ABCD, BiB, BREATHE, WALNUTS), aligning with the WPI tasks 1.2 and 1.3.

3. Method

3.1. Selection of the variables to be harmonized

YEAH! project members involved in T1.2 and T1.3 have developed a list of health outcomes and their determinants along with related variables that are necessary to accomplish the project objectives and pertaining to the following domains: childhood/adolescence health outcomes, socio-environmental, inter-, and intra-personal determinants of health outcomes. The selection of these has been decided upon by different members of the YEAH! project and are furthermore dependent on the data availability in the participating cohorts. The definition of each of the determinants and health outcomes are based on scientific literature, expert knowledge, and previous European projects pooling child and adolescent

exposome and health data from any of the four involved cohorts/studies. The harmonization strategy will be based on the work previously done in large European Commission-funded projects, namely LifeCycle (<https://lifecycle-project.eu>), Equal-Life (<https://www.equal-life.eu>) and ATHLETE (<https://athleteproject.eu>) projects, that included ABCD, BiB, BREATHE, and/or WALNUTS.

3.2. Age groups targeted by the intervention

The YEAH! intervention will target two different age groups: children (8–10y) and adolescents (14–16y). Due to the nature of the planned analyses in T1.3 (cross-sectional exposome study) and the availability of data in each cohort, the variables selected for harmonization will be derived from age windows that are closest to these predefined focus age groups. The specific age windows used for harmonization will therefore be determined based on the age ranges available within the participating cohorts, with the aim of maximising data availability while remaining as close as possible to the predefined focus age groups. For harmonization purposes, only variables available for at least two cohorts will be considered. However, if an important determinant/outcome is available for one cohort only, it will also be included in the current document for it to also serve as an overview of all available data.

3.3. Comparability across data from different cohorts/studies

To enable or improve comparability among the harmonized variables, new summarising/composite variables will be created: for continuous variables that have been measured by the same tool/indicator across the cohorts, we will make sure that all are using the same units or we will scale them based on their interquartile range (IQR). For those continuous variables measured by different tools/indicators across the cohorts, we will develop z-scores or standardize them based on their mean and standard deviation. For those variables that are not normally distributed (based on assessment of the distribution using graphical methods and/or statistical tests of normality) we will create percentiles. For categorical variables, we will define optimal categories based on the available categories across the cohorts and distribution of participants in each category.

For some outcomes (e.g., selected psychological constructs measured by different questionnaires and response formats), instead of using observed scores we will create latent factors.

Where appropriate, lagged exposures or change scores will be derived (e.g., if a variable changes over time).

It is crucial for the further analyses that the direction of response categories is the same across cohorts, so that 0/lowest number denotes the lowest level of exposure/no outcome, while the highest number denotes the highest exposure/presence of outcome.

3.4. Variable transformations

Highly skewed variables may be transformed (logarithmic, square root, tertiles) to achieve symmetric distributions with a homogenous range of values. The degree of skewness and the distribution of each variable will be assessed using graphical methods (e.g., histograms and Q-Q plots) and, where appropriate, statistical tests of normality. The decision to apply a transformation will be based on the observed distribution and degree of skewness, with the specific transformation selected according to the characteristics of the variable.

3.5. Prioritization

Regarding the selection of the questionnaire/instrument to retrieve an outcome of a study, there could be differences among the availability of sources in the four cohorts for the same age window. Moreover, multiple measures and even similar measures from different informants (e.g., parent-reported, teacher-reported and child-reported) could be available in a single cohort.

In such cases, we will prioritize the following measures:

- (i) A measure with the best psychometric properties; in the next stages of the project, the experts in each domain will evaluate available questionnaires and decide which items would most accurately reflect the outcome and determinant variables.
- (ii) A measure with the prioritized informant (child, parent, or teacher) based on the expert view.

(iii) A measure that is repeated within the cohort at different ages and available for as many cohorts as possible.

3.6. Cross-validation

In the event that we encounter different measures, instruments, or questionnaires across our four cohorts for a single outcome or determinant, we will undertake a validation process to ensure comparability of results. This process will involve carefully examining the psychometric properties of each instrument, conducting cross-validation studies where possible, and using statistical techniques such as standardization or harmonization of scores. For example, one cohort might use the Perceived Stress Scale (PSS) to assess stress levels, while another might use the Adolescent Discrimination Distress Index (ADDI). Because these tools have different focuses and scoring systems, it is important to ensure that the data obtained from these different instruments are indeed comparable, allowing for meaningful analysis and interpretation across cohorts.

These approaches will help maintain the integrity of our research and allow us to draw robust conclusions despite the variability in measurement tools.

3.7. Naming of the variables

Harmonized variable names will aim to be as informative as possible and will code, except for a concise description, such information as e.g., the unit of interest (**ui**) (e.g., **ch** = child, **m** = mother, **f** = father, **p** = parent/caretaker), and the reporter (**r**) (defined according to the prioritization rule described above and coded as follows: **cr** = child, **mr** = mother, **fr** = father, **pr** = parent, **tr** = teacher, **cl** = clinical). The general structure of the variable name will be: **descriptive name_ui_r**.

3.8. Outlier and extreme values detection

Outlier and extreme values detection will require inspection of each variable distribution (e.g., using boxplots, density plots) and checking for typical changes across time, age, differences in sex, differences between cohorts, etc. Application of predefined thresholds (e.g., IQRs) to flag extreme values and checking for biological plausibility, e.g., by applying predefined thresholds (e.g., BMI cut-off values) to flag implausible values may be used. To

check for the robustness of obtained results, sensitivity analyses excluding potential outliers/extreme data points will be performed.

3.9. Harmonized dataset structure

Harmonization will be performed in R and the source code will be available in the online repository. Harmonized dataset will have a wide format stored in the .RDS file. Each variable will be attributed a label specifying the relevant details e.g., variable format (continuous, categorical, character, date), units for continuous variables, coding for categorical variables, comments, and other relevant information. After harmonization and before further analyses, the newly created database will be subjected to a quality control procedure aiming to identify potential harmonization issues such as mismatches in variable units or coding, creation of implausible values, etc.

4. Variables to be harmonized

4.1. Health behaviours and health outcomes

In what follows, we outline the harmonization strategy for health outcomes of interest. We have considered the following health outcomes domains: lifestyle, mental health, weight status, and health-related quality of life.

Table 1 lists the outcomes of interest with the availability in each of the cohorts/studies.

1. Lifestyle

Data on **physical activity** will be derived from questionnaires on doing sports and other physical activities as well as from composite variables derived from accelerometry.

Sedentary behaviours will be estimated from e.g., time spent on mobile phone, playing video games, and watching TV, overall screen time, composite variables derived from accelerometry, and it will be complemented by information derived from physical activity questionnaires.

Active commuting will be derived from questionnaires on school commuting modes and physical activity.

Data on **diet and eating patterns** will be derived from questionnaires on adherence to Mediterranean diet and healthy/unhealthy foods consumption including key food groups and nutrients and dietary behaviours (e.g., Mediterranean Diet Adherence Screener [MEDAS], Food Frequency Questionnaire [FFQ]). For the construction of dietary scores, composite variables will be considered.

Sleep data will be derived from questionnaires on sleep duration and quality as well as from accelerometer data.

2. Mental health

Data on **anxiety, depressive symptoms, conduct problems, oppositional-defiant, aggressive behaviour, bullying, attention and hyperactive-impulsive (ADHD) symptoms, prosocial behaviour, and somatic complaints** will be derived from standardized questionnaires (e.g., SDQ, the Substance Use Risk Profile Scale [SURPS], Child Behavior Checklist [CBCL], Youth Self-Report [YSR], Behavior Rating Inventory of Executive Function Questionnaire [BRIEF], Utrecht Coping List [UCL-A], MCS Sweep 4) as well as from medical diagnosis and, if necessary, complemented by information from other non-specific questionnaires on mental or general health.

Data on **loneliness** will be retrieved from tools like CBCL, YSR, UCLA Loneliness Scale, MCS Sweep 4, Multidimensional Scale of Perceived Social Support [MSPSS], or derived from surveys assessing perceived social support and connectedness and complemented by information on prosocial behaviour, where decreased prosocial tendencies would correlate with increased loneliness.

Relationships with peers will be derived from standardized tests (e.g., SDQ, MSPSS, CBCL, YSR), or derived from social support and social attitudes assessing surveys.

Stress levels will be estimated based on standardized questionnaires (e.g., PSS, Vragenlijst meegemaakte gebeurtenissen [VMG], Child Revised Impact of Events Scale [CRIES], UCL-A)

or questionnaires on stressful life events and Adolescent Discrimination Distress Index [ADDI]).

Impulsive behaviour will be derived from standardized tests (e.g., Delay Discounting Task, The Substance Use Risk Profile Scale [SURPS], Attention Network Test [ANT], BRIEF, CBCL, SDQ, YSR).

Irritable moods will be derived from standardized questionnaires (BRIEF, CBCL, YSR).

Life satisfaction and mental wellbeing will be derived from general life satisfaction and happiness questionnaires complemented by information from the SDQ, VMG, or general/mental health questionnaires.

Self-esteem levels can be directly measured by Rosenberg Self-Esteem Scale (RSES) or inferred from CBCL, YSR, or general mental health or well-being questionnaires.

Self and body image will be extracted from questionnaires examining self-concept, self-esteem, body satisfaction, eating disorders, physical appearance perceptions, or psychological well-being.

Information on **eating disorders** will be derived from eating behaviours questionnaires (e.g., Child Eating Behaviour Questionnaire [CEBQ], Eating Disorder Examination Questionnaire (EDE-Q) or medical diagnosis, when available. If necessary, complementary information will be derived from mental or general health questionnaires.

3. Weight status

Body mass index (BMI, kg/m²) will be calculated based on data derived from anthropometric measurement records or questionnaires.

4. Health-related quality of life

Health-related quality of life is an outcome that was not directly measured in any cohort but could potentially be inferred from questionnaires that assess multiple domains of health and well-being, including physical functioning, mental health, and social relationships (e.g., general or mental health questionnaires or Activity Questionnaire for Adults and Adolescents [AQuAA]).

Table 1: Availability of outcomes of interest for each cohort/study. **X** = data not available.

Outcome	Category	ABCD		BiB		BREATHE	WALNUTS
		10-12y	15-16y	7-11y	12-16y	7-11y	12-16y
Physical activity	Lifestyle	Sports		Sports	Physical activity	Leisure time physical activity	Sports
		Playing, Outdoor activities	Physical activity	Physical activity			High physical intensity activity
Sedentary behaviours		Mobile phone usage, overall screen time, video games, watching TV	Inferred from physical activity questionnaires	Watching TV	Watching TV, playing video games	Watching TV, playing video games	Mobile phone usage, TV, video games
Active commuting		Inferred from physical activity questionnaires	School commuting mode	School commuting mode	School commuting mode	School commuting mode	X
Diet/ Eating patterns		Detailed info on healthy food intake (amount and frequency)		Detailed info on healthy food intake (amount and frequency)	Eating patterns	X	Mediterranean diet/ unhealthy food intake, vitamins
Sleep		Detailed info on sleep (difficulties, habits, duration, quality, daytime sleepiness, medication, screen use)		Detailed info on sleep (difficulties, habits, duration, quality, daytime sleepiness)		Duration	Duration
Anxiety	Mental health and disorders	X	Anxiety medical diagnosis	Reported chronic mental health/ behavioural condition	X	X	Anxiety medical diagnosis
		Anxiety symptoms from questionnaire	Anxiety symptoms from questionnaire	Anxiety symptoms from questionnaire	Anxiety symptoms from questionnaire	Anxiety symptoms from questionnaire	Anxiety symptoms from questionnaire
Depression symptoms		X	Depression medical diagnosis	Reported chronic mental health/ behavioural condition	X	X	Depression medical diagnosis

		Depression symptoms from questionnaire	Depression symptoms from questionnaire	Depression symptoms from questionnaire	Depression symptoms from questionnaire	Depression symptoms from questionnaire	Depression symptoms from questionnaire
Conduct problems, oppositional-defiant, aggressive behaviour, bullying		X	X	Reported chronic mental health/ behavioural condition	X	X	Reported conduct problems
		Conduct problems symptoms from questionnaire	Conduct problems symptoms from questionnaire	Conduct problems symptoms from questionnaire	Conduct problems symptoms from questionnaire	Conduct problems symptoms from questionnaire	Conduct problems symptoms from questionnaire
Hyperactivity		Hyperactivity medical diagnosis	Hyperactivity medical diagnosis	Reported chronic mental health/ behavioural condition	X	X	X
		Hyperactivity symptoms from questionnaire	Hyperactivity symptoms from questionnaire	Hyperactivity symptoms from questionnaire	Hyperactivity symptoms from questionnaire	Hyperactivity symptoms from questionnaire	Hyperactivity symptoms from questionnaire
Attention problems		X	X	Reported chronic mental health/ behavioural condition	X	X	Inattention medical diagnosis
		Inattention symptoms from questionnaire	Inattention symptoms from questionnaire	Inattention symptoms from questionnaire	Inattention symptoms from questionnaire	Inattention symptoms from questionnaire	Inattention symptoms from questionnaire
Somatic complaints		Somatic complaints score from questionnaire	Somatic complaints score from questionnaire	Somatic complaints score from questionnaire	Somatic complaints score from questionnaire	Somatic complaints score from questionnaire	Somatic complaints score from questionnaire
Peer relationship		Peer relationship problems score from questionnaire	Peer relationship problems score from questionnaire	Peer relationship problems score from questionnaire	Peer relationship problems score from questionnaire	Peer relationship problems score from questionnaire	Peer relationship problems score from questionnaire

Prosocial behaviour		Prosocial behaviour score from questionnaire	Prosocial behaviour score from questionnaire	Prosocial behaviour score from questionnaire	Prosocial behaviour score from questionnaire	Prosocial behaviour score from questionnaire	Prosocial behaviour score from questionnaire
Loneliness		Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from different questionnaires	Loneliness measuring questionnaire	X	Loneliness perception
Stress		Reported stress, questionnaires	Reported stress	X	Inferred from discrimination questionnaire	X	X
Impulsive behaviour		Inferred from mental health questionnaires	Reported impulsivity	Inferred from mental health questionnaires	X	Reported impulsivity	Attention tests
Irritable moods		Inferred from mental health questionnaires	Inferred from mental health questionnaires	X	X	X	X
Life satisfaction and mental wellbeing		Inferred from mental health questionnaires	Reported life satisfaction/happiness	Inferred from general health questionnaire	Inferred from general health questionnaire	Inferred from mental health questionnaire	Reported life satisfaction/happiness
Self-esteem		Inferred from mental health questionnaires	Reported self-esteem	X	Inferred from social attitudes	X	X
Self and body image		X	X	X	Inferred from social attitudes	X	Reported self-image
Eating disorders		Eating behaviour questionnaire	X	Reported eating disorder	Detailed questionnaire on eating disorders	Medical diagnosis	Medical diagnosis
BMI	Weight status	Anthropometric scores	Inferred from general health questionnaires	Anthropometric scores	Anthropometric scores	Anthropometric scores	Anthropometric scores
Health-related quality of life		X	Inferred from mental health questionnaires	X	Inferred from mental health questionnaires	X	X

4.2. Determinants of health behaviours and health

The determinants that were considered of interest for child and adolescent health behaviours and health in the YEAH! project are included in the following domains: living environment and urban exposome, intra-individual determinants, inter-individual determinants, and potential confounding factors. As for the intra- and inter-individual determinants, before the final selection of variables to be harmonized, both domains will be carefully analysed by the experts to choose the items and questionnaires most accurately describing each measured construct. Then, one or more summarizing variables will be created for each construct before the harmonization.

Table 2 lists the determinants of interest with the availability in each of the cohorts/studies.

1. Living environment and urban exposome

The variables concerning living environment and urban exposome include both physical and social factors. These are: **air pollution** (NO₂, PM_{2.5}, PM₁₀), **noise pollution**, **climatic factors** (temperature), exposure to **light at night**, **built environment**, **access to natural spaces** (green, blue spaces), access to **physical activity facilities**, **neighbourhood socioeconomic position**, **safety** and **walkability**, and **urbanicity**. This data will be derived from Geographic Information System (GIS) and modelled data relating to the geocode of the residence(s) of the child and complemented with information derived from questionnaires. This data will be created following the already validated protocols of the [LifeCycle](#), [Equal-Life](#), and [ATHLETE](#) projects.

2. Intra-individual determinants

The intra-individual determinants include: **motivation** (intentions, autonomous motivation, controlled motivation, self-determination), **health literacy**, **behavioural attitudes/beliefs** (behavioural beliefs, outcome expectancies, compensatory health beliefs, risk perceptions), **personality/temperament**, **control constructs** (self-efficacy, perceived behavioural control, locus of control), **self-regulatory skills** (action planning, self-monitoring), **affect toward health behaviours** (remembered affect, anticipated affective response, enjoyment of the behaviour, affective attitudes), **psychosocial skills/emotional intelligence/social cognition** (assertiveness, tolerance, social regulation, emotion regulation, emotional

awareness). Several of these determinants were not measured directly (i.e., using validated questionnaires) but may be derived from other questionnaires non-specific for this domain (e.g., SURPS, YSR, SDQ, BRIEF, CRIES, CBCL, UCL-A, adapted Acculturation Rating Scale for Mexican Americans-II for Children and Adolescents [ARSMA], general and mental health questionnaires, parental lifestyle questionnaires, etc.).

3. Inter-individual determinants

The inter-individual determinants include: **social support, social cohesion/belonging, social norms** (injunctive norms/social pressure/obligation, descriptive norms, moral norms, religion, socio-cultural context), **prosocial behaviour, stigma-related processes/stereotypes, parent-for-child health behaviours /parental modelling, parenting styles, teaching styles, bullying experience**. These determinants will be derived from standardized questionnaires and scales (e.g., SDQ, YSR, MCS Sweep 4, Parenting Styles and Dimensions Questionnaire-32 [PSDQ-32], Multidimensional Scale of Perceived Social Support [MSPSS], Vragenlijst Gezinsfunctioneren Ouders [VGFO], Verkorte Schaal voor Ouderlijk Gedrag [VSOG], Opvoedingsbelasting Vragenlijst [OBVL], Short Questionnaire to Assess Health enhancing physical activity [SQUASH], Social Media Disorder Scale [SMDS], Parental Monitoring, Parent and Peer Support Questionnaire, etc.) examining social support networks including parents, teachers, classmates, and close friends, and various aspects of social support and relationships in children and adolescents. These tools contain items focusing on the socio-ecological framework, which emphasizes the importance of young people's relationships within key microsystems such as home and school. Other questionnaires, not specific for the inter-individual determinants but providing relevant information, may also be used.

Table 2: Availability of determinants of interest for each cohort/study. **X** = data not available.

Determinant	Category	ABCD		BiB		BREATHE	WALNUTS
		11-12y	15-16y	7-13y	12-16y	7-11y	11-16y
Air pollution	Living environment or urban exposome	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Noise pollution		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Temperature		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	X	X

Light at night		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	X	X
Built environment		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Green spaces		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Blue spaces		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Opportunities for physical activity (access to parks...)		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	Questionnaires	X
Neighbourhood socioeconomic position		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Neighbourhood safety		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	X	X
Neighbourhood walkability		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Urbanicity		GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data	GIS and modelled data
Motivation	Intra-individual determinants	X	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X	X
Health literacy		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X	X
Behavioural attitudes/ beliefs		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X	X
Personality/ temperament		Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from general questionnaires	Inferred from mental health questionnaires	Inferred from mental health questionnaires
Control constructs		X	Inferred from several non-specific questionnaires	X	Inferred from several non-specific questionnaires	X	X
Self-regulatory skills		Inferred from several non-	Inferred from several non-	X	Inferred from several non-	X	X

		specific questionnaires	specific questionnaires		specific questionnaires		
Affect toward health behaviours		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X	X	X
Psychosocial skills/emotional intelligence/social cognition		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X	Inferred from several non-specific questionnaires	X	X
Social support	Inter-individual determinants	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X
Social cohesion/belonging		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X
Social norms		X	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X
Prosocial behaviour		Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from general questionnaires	Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from mental health questionnaires
Stigma-related processes/stereotypes		X	X	X	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X
Parent-for-child health behaviours/parental modelling		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X
Parenting styles		Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X
Teaching styles		X	X	X	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X

		Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from criminal behaviour and social attitudes	Inferred from mental health questionnaires	Inferred from mental health questionnaires	Inferred from mental health questionnaires
Being bullied		X	Reported bullying	X	X	X	X

4.3. Potential main confounding factors

Our statistical analyses will be adjusted for relevant confounding factors, selected a priori based on the existing literature and expert knowledge, which will be harmonized across cohorts.

Table 3 lists the main potential confounding factors and their availability in each cohort/study.

The following categories of variables potentially influencing the independent and/or dependent variables will be harmonized: **lifestyle** factors (leisure activities, social media use) and **socio-demographic** determinants (parental education level, employment, and immigration status; child's sex, gender, age, ethnicity, puberty; family income, material deprivation and physical neglect). The data for these variables will be derived from available items and questionnaires (e.g., National Statistics Socio-economic Classification [NS-SEC], Puberty Development Scale [PDS], VGFO, VMG, Internet Gaming Disorder Scale [IGDS], Social Media Disorder Scale [SMDS], etc.). Other potential but not yet defined confounders to be included in particular analyses (e.g., breastfeeding, birth, maternal smoking and alcohol use during pregnancy) will be retrieved at further stages of the project when the detailed research hypotheses are defined.

Table 3: Availability of main potential confounders for each cohort/study. **X** = data not available.

Determinant	Category	ABCD		BiB		BREATHE	WALNUTS
		11-12y	15-16y	7-13y	12-16y	7-11y	11-16y
Cohort							
Leisure activities	Lifestyle	Inferred from physical activity	Leisure activities questionnaire	Leisure activities questionnaire	Leisure activities questionnaire	Inferred from watching TV, playing video games	Leisure activities questionnaire
Social media use		Inferred from mobile phone use	Social media questionnaire	Inferred from physical and leisure activities	Inferred from physical and leisure activities	X	Social media questionnaire
Parental education level	Socio-demographic factors	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Parental employment		Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Parental immigration status		Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Child's sex		Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Child's gender		Demographics questionnaires	Demographics questionnaires	X	Demographics questionnaires	X	X
Child's age		Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Child's ethnicity		Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Puberty		Puberty questionnaire	Physical development questionnaire	X	X	X	Demographics questionnaires
Family income		X	X	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires	Demographics questionnaires
Material deprivation/physical neglect		General health questionnaire	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	Inferred from several non-specific questionnaires	X

5. Next steps

The detailed harmonization protocol has been developed and is provided in **Appendix 1**. The appendix contains the specific variables to be extracted from the datasets and the corresponding harmonization procedures to be applied across cohorts.

For variables that have already been harmonized in the context of the LifeCycle, Equal-Life, or ATHLETE projects, the previously harmonized data will be used where applicable. For cohort(s) that were not part of these projects, the relevant variables will be harmonized based on the data types, measurement instruments, definitions, and tools used to collect the data across the cohorts.

The detailed harmonization decisions, including variable definitions, source variables, coding, categorization, and harmonization procedures, are described in Appendix 1. Additional determinants of health outcomes may be considered where necessary to accommodate specific research questions and hypotheses, including potential confounding variables relevant to particular statistical analyses.

6. Appendix 1

Data preparation

1. Obtain raw data from each cohort: Secure access to cohort datasets and ensure you have variable dictionaries/codebooks for each.
2. Standardize data formats: Convert all files to a consistent format (**.rds**) and ensure consistent encoding (UTF-8) and separator (**comma**).
3. Initial variable review per cohort: choosing the variables to harmonize by each, target and document original variable names, formats, and codes.

The task-involved institutions have agreed on how to distribute the variables and have decided which ones are most important for the harmonization process (Table 1 and Table 2).

Table 1. List of the outcomes to be harmonized, along with the assigned institutions.

Outcomes	INSTITUTION
Physical activity	ENS
Sedentary behaviors	ENS
Active commuting	ENS
Diet/ eating	IISPV
Sleep	ISGlobal
Anxiety symptoms	IISPV
Depressive symptoms	IISPV
Internalizing symptoms	ISGlobal
Conduct problems	ISGlobal
Externalizing symptoms	ISGlobal
Hyperactivity	ISGlobal
Attentional problems	ISGlobal

Impulsivity	ACCARE
Prosocial behavior	ISGlobal
Loneliness and social isolation	ISGlobal
Relationships with peers	ISGlobal
Self-esteem	IISPV
Self-image and body image	IISPV
Disordered eating	IISPV
Somatic symptoms	ENS
Weight status	IISPV
Perceived general physical health - General wellbeing	IISPV
Coping & Optimism	ISGlobal

Table 2. List of the determinants to be harmonized, along with the assigned institutions.

Determinants	INSTITUTION
Age	IISPV
Puberty	IISPV
Sex	IISPV
Gender	IISPV
Ethnicity	IISPV
Immigration status	IISPV
Parental education	IISPV
Parental employment status	IISPV
Leisure activities	ENS
Social media use	ENS
Neighborhood socioeconomical position	ISGlobal

Air pollution	ISGlobal
Noise pollution	ISGlobal
Light at night / Light pollution	ISGlobal
Temperature (heat)	ISGlobal
Green spaces	ISGlobal
Blue spaces	ISGlobal
Built environment	ISGlobal
Opportunities for PA (access to facilities like playgrounds, parks)	ISGlobal
Neighbourhood safety	ISGlobal
Neighbourhood walkability	ISGlobal
Urbanicity (building density, population density)	ISGlobal
Social support	ENS
Parent-for-child health behaviors – PA and SB	ENS
Parent-for-child health behaviors – risky behaviors	ISGlobal
Parent-for-child health behaviors – diet	IISPV
Parenting styles	ACCARE
Being bullied	ACCARE
Emotion dysregulation	ACCARE
Prosocial behavior	ISGlobal
Stress resilience / Control constructs	ACCARE
Self-regulatory skills	ACCARE
Other intra-individual determinants (popularity, perfectionism...)	ACCARE

Naming

Assign harmonized variable names (following our naming conventions):

From section “**3.7. Naming of the variables**”:

Harmonized variable names will aim to be as informative as possible and will code, except for a concise description, such information as e.g., the unit of interest (**ui**) (e.g., **ch** = child, **m** = mother, **f** = father, **p** = parent/caretaker), and the reporter (**r**) (defined according to the prioritization rule described above and coded as follows: **cr** = child, **mr** = mother, **fr** = father, **pr** = parent, **tr** = teacher, **cl** = clinical). The general structure of the variable name will be: **descriptive name_ui_r**.

If a study has both father- and mother-separated reports of a child variable, a mother, father and a combined parent-reported variable should be created. For parent-reported variables, it is advisable that a “parent type” variable would be created, where mother or father reports could be tagged in each case. That enables working with the full child info of both parents, but having the possibility of truncating the data if necessary.

Given the needed age period of the assessments (**t**) are either **childhood** or **adolescence**, in addition to the above, a tag **_c** = childhood or **_a** = adolescence must be added to the variable name. The final structure: **descriptive name_ui_r_t**.

Summary table for coding:

Abbreviation	Meaning	Possible values
ui	Unit of interest	ch (child), m (mother), f (father), p (parent/caretaker)
r	reporter of the information	cr (child), mr (mother), fr (father), pr (parent), tr (teacher), cl (clinical)
t	age period of the child at assessment	c (childhood) or a (adolescence)

In addition to the fixed codings that should be added to each variable, the following codings need to be applied if applicable (from Equal-Life and LifeCycle):

For a given descriptive name Y at age t, you should record the following variables:

- i. age of the child (in years) at assessment of another variable: *Y_age_t*
- ii. specific questionnaire or instrument name: [abbr. of name]_Y_t (e.g. SDQ_)
- iii. cohort-specific instrument: *Y_instr_t*
- iv. total score: *Y_t_t*
- v. raw score: *Y_raw_t*
- vi. prorated total scale score: *Y_pro_t*
- vii. average score: *Y_avg_t*
- viii. percentile: *Y_pc_t*
- ix. any score: *Y_score_t*

Similarly, the **descriptive names** for the full list of determinants and outcomes have been organized according to pre-established codes (Equal-Life and LifeCycle):

Outcomes	Descriptive name	Determinants	Descriptive name
Physical activity	pa	Age	age
Sedentary behaviors	sb	Puberty	puber
Active commuting	ac	Sex	sex
Diet/ eating	diet (fruit / veg / snacks / supp / sweet_drinks)	Gender	gender
Sleep	sleep	Ethnicity	ethn
Anxiety symptoms	anx	Immigration status	mig
Depressive symptoms	dep	Parental education	edu
Internalizing symptoms	int	Parental employment status	occup

Conduct problems	cond	Leisure activities	leisure
Externalizing symptoms	ext	Social media use	socialmedia
Hyperactivity	hyp	Neighborhood socioeconomic position	ses
Attentional problems	att	Air pollution	air
Impulsivity	imp	Noise pollution	noise
Prosocial behavior	proso	Light at night / Light pollution	light
Loneliness and social isolation	lone	Temperature (heat)	temp
Relationships with peers	peers	Green spaces	green
Self-esteem	esteem	Blue spaces	blue
Self-image and body image	image	Built environment	built
Disordered eating	dis_eat	Opportunities for PA (access to facilities like playgrounds, parks)	access
Somatic symptoms	somat	Neighbourhood safety	safety

Weight status	weight (BMI)	Neighbourhood walkability	walkability
Perceived general physical health - General wellbeing	generalhealth	Urbanicity (building density, population density)	urb (pop for population)
Coping & Optimism	coping / optimism	Social support	socialsupport
		Parent-for-child health behaviors - PA and SB	pa / sb / ac
		Parent-for-child health behaviors - risky behaviours	smk / alc / drug
		Parent-for-child health behaviors - diet	diet
		Parenting styles	parenting
		Being bullied	bullied
		Emotion dysregulation	emo_dysreg
		Stress resilience / Control constructs	resil
		Self-regulatory skills	selfreg
		Other intra-individual determinants	[create a name that works for you]

(popularity, perfectionism...)	
-----------------------------------	--

Note that apart from the *descriptive names*, the specifics of each variable may require the inclusion of a *descriptive 'surname'*. The example of sleep at the end of the document illustrates this: instead of simply 'sleep', the descriptive name (and 'surname') becomes 'sleepdur' or 'sleepqual' (in adolescence).

In blue, the names proposed that do not appear on Equal-Life and LifeCycle codebook, thus, open for discussion.

Any abbreviations for scores, instrument names, or variable names should be noted in the script and gathered together in an additional table (see the end of these guidelines to get a head start).

Finally, **document all the following** together with the new variable names given:

- "Label/ Description"
- "Source variable(s)" and
- "Desired coding (values/units)".

Variable Naming Example 1 (categoric variable): Child's ethnicity

Age Group	Harmonized Variable Name	Description
Childhood	ethn_ch_pr_c	Child's ethnicity in childhood (parent-reported or derived)
Adolescence	ethn_ch_pr_a	Child's ethnicity in adolescence (parent-reported or derived)

Variable Naming Example 2 (continuous variable): Child sleep

Age Group	Harmonized Variable Name	Description

Childhood	sleepdur_ch_pr_c	Average daily sleep duration in childhood (parent-reported, hours/night)
Adolescence	sleepdur_ch_cr_a	Average daily sleep duration in adolescence (self-reported, hours/night)
Adolescence	sleepqual_ch_cr_a	Self-reported sleep quality in adolescence (categorical rating/yes-no)

Transformation and scoring

From section “**3.3. Comparability across data from different cohorts/studies**”:

“For continuous variables that have been measured by the same tool/indicator across the cohorts, we will make sure that all are using the same units or we will scale them based on their interquartile range (IQR).”

For those continuous variables measured by different tools/indicators across the cohorts, the decisions on the possibility of a combined item will depend on the distributions and the correlations of the original items.

Continuous variables will follow a streamline of quality checks after their proper harmonization and before the analyses:

1. address missingness based on the variables of interest (e.g. exclusion of missing values; where the primary original variable is missing, an alternative original variable measuring the same construct may be used, when appropriate, to derive the most complete harmonized variable)
2. check variable distributions
3. check model assumptions (normality, homoscedasticity, linearity...)
4. variable transformations (e.g. z-standardisation, log transformation)

“For categorical variables, we will define optimal categories based on the available categories across the cohorts and distribution of participants in each category.”

For some outcomes (e.g., selected psychological constructs measured by different questionnaires and response formats), instead of using observed scores we will create latent factors.

Where appropriate, lagged exposures or change scores will be derived (e.g., if a variable changes over time).

It is crucial for the further analyses that the direction of response categories is the same across cohorts, so that 0/lowest number denotes the lowest level of exposure/no outcome, while the highest number denotes the highest exposure/presence of outcome."

For instance, the 4-point Likert scale used in the ABCD cohort for sleep quality (1 = 'very good', 2 = 'good', 3 = 'poor', 4 = 'very poor') would be reversed, whereas the scale in BiB (1 = 'No', 2 = 'Sometimes', 3 = 'Yes, always') follows the correct established order. Therefore, the scale in ABCD should be reversed and standardized to the smallest scale (BiB) before combining the categories.

Categorical variables should be coded using their labelled values rather than their numeric values; for example, "never", "sometimes", "often" and "always", instead of 1, 2, 3 and 4. The list of ordered categories should be included in the codebook for the created variables (1=never, 2=sometimes, 3=often, 4=always).

Validation of measures

From section "**3.6. Cross-validation**":

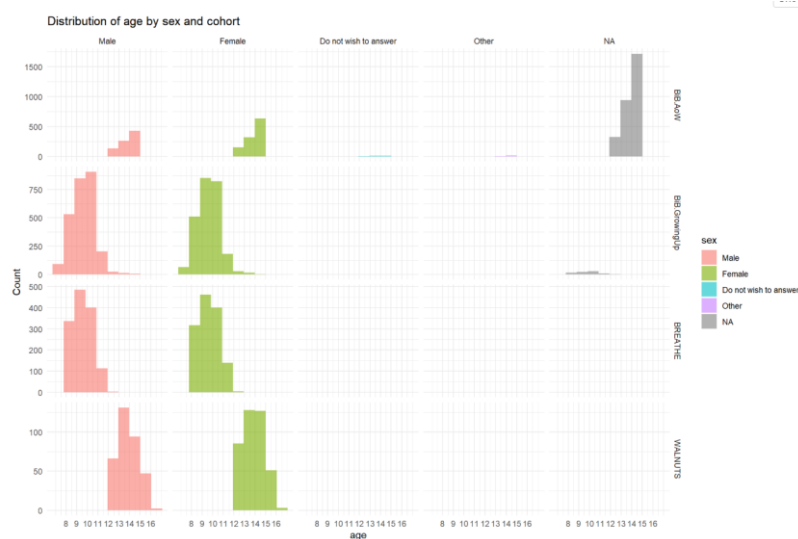
"In the event that we encounter different measures, instruments, or questionnaires for a single outcome or determinant, we will undertake a validation process to ensure comparability of results."

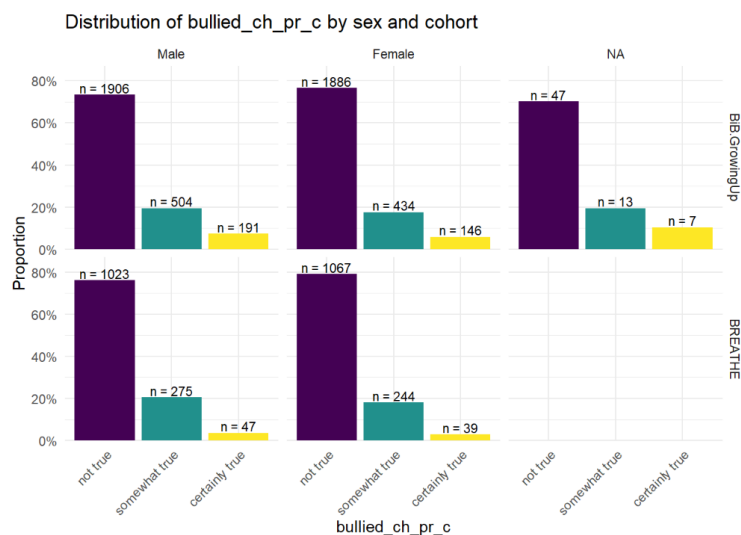
When multiple instruments are available for the same construct, we will ensure comparability through a structured validation process. This will involve reviewing the conceptual alignment and psychometric properties of each measure, assessing concordance through cross-validation where overlapping data exist, and applying statistical harmonization methods such as standardization, rescaling, or latent variable modeling. To test robustness, we will conduct sensitivity analyses by repeating models with each instrument separately and comparing results, ensuring that findings are not driven by the choice of measure.

Quality control

1. Distributions: For each variable (continuous/categorical), plot cohort-wise variable distributions for comparability as well as variable distributions across age and sex. Unless there is a clear plausible explanation of why the distributions would differ between cohorts (e.g. cultural differences), distributions should be comparable between cohorts as this gives strength that the harmonized scaling captures the same construct.

Here is an example:





2. Outliers:

- Continuous: flag values outside plausible range (e.g., $z > 3$, biologically implausible).
- Categorical: check for unexpected codes.

3. Missing values: "Missing values will be coded as NA. Check variable distributions after excluding missings."

4. Unit consistency checks: Ensure all numeric variables are in the same unit and annotate in metadata (e.g., cm vs. m, kg vs. g).
5. Duplicate entries: consider a unique report to avoid having duplicates of the same individual.
6. Variable descriptives:
 - a. Continuous variables: mean, median, sd, min, max, skewness, kurtosis, amount of missing values
 - b. Categorical: percentiles, counts per level

Variable Harmonization Example 1 (categorical variable): child's ethnicity

Childhood:

Cohort	Source Variables	Categories	Notes
ABCD	QZW34A, QZW34B, etn_gen	1="The Netherlands", 2="Surinam", 3="Antilles/Aruba", 4="Turkey", 5="Morocco", 6="Ghana", 7="other country, specify...", 8="N/A"=I don't know	Country of birth for mother and father, combined parental info to infer ethnicity of the child
BiB	guadtPlaceYouBorn, guadtChildFatherBorn, guadtPlaceYouBorn_other, guadtChildFatherBorn_other	1="England", 2="Northern Ireland", 3="Scotland", 4="Wales", 5="Channel Islands", 6="Isle of Man", 7="Republic of Ireland", 8="Czech Republic", 9="Poland", 10="Slovakia", 11="Bangladesh", 12="India", 13="Pakistan", 14="Sri Lanka", 15="Philippines", 16="Don't know", 17="Other"	Country of birth for mother and father, combined parental info to infer ethnicity of the child

BREATH E	pais_mare, pais_mare_espec, pais_pare_espec, ethnicity, adoptat, pais_naix_nen, pais_naix_nen_espec	pais_pare, 1="Espanya"=Spain, 2="Altres"=Others	Use direct ethnicity of the child (parent-reported) in contrast to parental country of origin
-------------	--	--	---

Bold: chosen original variables to create the harmonized variable; Plain: all available original variables possibly harmonizable; ~~crossed-out~~: discarded original variables, not to be included in the harmonized variable.

Adolescence:

Cohort	Source Variables	Categories	Notes
ABCD	QZW34A, QZW34B, ~~eth_gen~~	1="The Netherlands", 2="Surinam", 3="Antilles/Aruba", 4="Turkey", 5="Morocco", 6="Ghana", 7="other country, specify..."	Same as childhood
BiB	awb1_2_ethnicity_r4, ~~ethnicity_1, ethnicity_2~~	1="White", 2="Mixed or Multiple ethnic groups", 3="Asian or Asian British", 4="Black, Black British, Caribbean or African", 5="Arab", 6="Other ethnic groups"	Multiple reported values, choice of self-reported ethnicity rather than school reported
WALNUT S	mother_country, father_country	1="Spain", 2=" Others"	Use combined parental info to infer ethnicity

Bold: chosen original variables to create the harmonized variable; Plain: all available original variables possibly harmonizable; ~~crossed-out~~: discarded original variables, not to be included in the harmonized variable.

The combination of the different categories across cohorts is derived from the inclusion of all ethnicities. These categories are then gathered into two approaches: (i) per sub-region of continents, and (ii) per country of origin same to the cohorts' or not.

This outcome is shown in the Outcome Example 1 provided later in the document.

Variable Harmonization Example 2 (continuous variable): Child's sleep

Childhood:

Cohort	Source Variables	Notes
ABCD	MQ50b_hour_schoolday, MQ50b_min_schoolday, MQ50b_hour_weekend, MQ50b_min_weekend	Duration derived by combining hours and minutes; compute weighted mean (5 weekdays + 2 weekend days)
BiB	guadtWeekdaySleep, guadtWeekendSleep	Average of weekday and weekend hours; weighted mean to obtain nightly duration
BREATH E	hores_son	Direct report of nightly sleep duration on school days

Bold: chosen original variables to create the harmonized variable; Plain: all available original variables possibly harmonizable; crossed-out: discarded original variables, not to be included in the harmonized variable.

Adolescence:

Cohort	Source Variables	Notes
ABCD	F5PQ79 (weekday hours), F5PQ80 (weekend hours)	Weighted mean of weekday/weekend hours for duration
BiB	awb4_3_times2_a5 (school nights fall asleep), awb4_3_times3_a5 (school nights wake up), awb4_3_times4_a5 (non-school nights fall asleep), awb4_3_times5_a5 (non-school nights wake up)	Duration derived from reported times (wake – fall asleep); weighted mean

WALNUT S	week_sleeping_time (weekdays), weekend_sleeping_time (weekend)	Weighted mean of weekday and weekend reported hours
-------------	---	---

Bold: chosen original variables to create the harmonized variable; Plain: all available original variables possibly harmonizable; ~~crossed-out~~: discarded original variables, not to be included in the harmonized variable.

For sleep quality:

Cohort	Source Variables	Categories	Notes
ABCD	F5PQ83	1="very good", 2="good", 3="poor", 4="very poor"	Quality rated on a 4-point Likert scale
BiB	awb4_3_sleep_well_nght_a 5	1="No", 2="Sometimes", 3="Yes, always"	Quality rated on yes/no and sometimes
WALNUT S	-		No quality information available

Bold: chosen original variables to create the harmonized variable; Plain: all available original variables possibly harmonizable; ~~crossed-out~~: discarded original variables, not to be included in the harmonized variable.

Harmonization sharing

From section "**3.9. Harmonized dataset structure**":

"Harmonization will be performed in R and the source code will be available in the online repository."

All YEAH! WP1 data harmonization, transformation, and analysis source code will be made available in an online repository, GitHub. GitHub will be used to create, store, manage, and share scripts, providing version control to track changes and ensure consistency across sites.

The repository is currently private and accessible only to contributing partners, ensuring secure collaboration. All participating institutions will have access to the full repository, enabling consistent application of methods across cohorts. Importantly, no data will be shared directly between institutions; instead, each institution will apply the shared code to its own data.

To further safeguard confidentiality, the repository is set up so that data files cannot be accidentally uploaded (via .gitignore). This approach guarantees transparency, reproducibility, and comparability across the four cohorts, while fully respecting data access restrictions.

Remember to share the **single script** for the harmonization of each variable. This script will proceed a

- (1) merged dataset of the necessary cohort datasets to create the pertinent variable,
- (2) general descriptive analysis/ quality control of that new dataset for possible implausible values,
- (3) harmonization and creation of the new variable (with the new respective name)
- (4) descriptive analysis of that new dataset (a single variable or set of variables) to ensure the results from harmonization

All scripts should take into account that the variables will be unified based on the **IDs**. Therefore, they should all keep the IDs from the different cohorts and the **cohort** to which they pertain (either ABCD, BiB, BREATHE or WALNUTS). This procedure enables the datasets to be unified.

Also, a '**global_id**' variable will be created to unify the datasets of the various cohorts. This variable will contain the IDs from each cohort, preceded by the name of the cohort, coded in lowercase letters and stored under the variable cohort. The **global_id** list will follow the format: [*lowercase cohort name*]_[ID code from the cohort]. For example, **abcd_314** or **bib_B1161199**.

The goal is to obtain **a single dataset for each age-wave** from each harmonization script, one for **children** and one for **adolescents**. The output data frame should contain harmonized IDs, cohorts and both harmonized variables and raw source items.

In addition, to distinguish between sub-studies or phases within the same cohort (e.g., Growing Up vs. Age of Wonder, or ABCD Phase 4 vs. Phase 5), we will retain the same individual ID across age-waves for comparability. This structure ensures both consistency

of IDs over time and clear traceability of participants to their respective cohort and age-wave.

Script standardization

Instructions for the standardization of all scripts for the different variables, and through the institutions.

- “How-to” lines for the usage of the loading dataset(s) script part
- Master file with full harmonization (all variables coming from the institutions’ harmonization process) → output of children, adolescent, and combined datasets
- Quality Control: optionally added to each script / QC master file to run from

Cross-validation of the scripts in GitHub

First of all, each institution will independently test a script from the other three institutions to identify any discrepancies, comments, improvements, or errors. Each partner will select the scripts for review based on their individual preferences.

Output

The output of this Data Harmonization streamline would be a single script for the harmonization of each variable/domain, with the **YEAH! Codebook** providing a summary of the variables and documenting how harmonization was performed (see example table below).

Output Example 1 (categorical variable): Child’s ethnicity

Assessed concept/D omain	Variable name	Label/Desc ription	Values (categoric al)	Un its	Comm ents	Further instruc tions	Mapp ed by
Ethnicity	ethn_region _ch_pr_c	Child’s ethnicity in childhood by region (parent- reported)	Europe, South_Asia, North_Afric a, SubSahara n_Africa, Caribbean, Middle_Eas t,	–	Derive d from parent al birth countri es or child’s birth countr y if		ABCD, BiB and BREAT HE

			Southeast_Asia		present		
	ethn_cohort_ch_pr_c	Child's ethnicity in childhood by cohort country (parent-reported)	yes/no		Derived from parental birth countries or child's birth country if present	Cohort country : yes/no (any parent from the cohort country = yes)	ABCD, BiB and BREATHE
	ethn_region_ch_cr_a	Child's ethnicity in adolescence by region (self-reported)	Europe, South_Asia, North_Africa, SubSaharan_Africa, Caribbean, Middle_East, Southeast_Asia	–	Child-reported ethnicity		ABCD, BiB and WALNUTS
	ethn_cohort_ch_cr_a	Child's ethnicity in adolescence by cohort country (self-reported)	yes/no		Child-reported ethnicity	Cohort country : yes/no	ABCD, BiB and WALNUTS

Output Example 2 (continuous variable): Child's sleep

Assessed concept/D omain	Variable name	Label/Desc ription	Values (categ orical)	Units (nume ric)	Comm ents	Further instructi ons	Mapp ed by
Sleep	sleepdur_c h_pr_c	Child's average nightly sleep duration in childhood (parent- reported)	–	Hours/ night	Derive d as weight ed mean of weekd ay (×5) and weeke nd (×2) report ed sleep	Where only weekday data exist (BREATH E), use weekday directly	ABCD, BiB, BREAT HE
	sleepdur_c h_cr_a	Adolescent' s average nightly sleep duration (self- reported)	–	Hours/ night	Derive d as weight ed mean of weekd ay (×5) and weeke nd (×2) report	For BiB, calculat e from reported sleep/w ake times; for others, direct report/a verage	ABCD, BiB, WALN UTS

					ed sleep		
	sleepqual_ ch_cr_a	Adolescent's sleep quality (self- reported)	1="Poor" , 2="Fair", 3="Good"	–	ABCD uses 4- point Likert scale ("Very good", "Good" , "Poor", "Very poor"); BiB binary yes/no and someti mes	Recode ABCD and BiB into categori es: Poor, Fair, and Good; missing if unavaila ble	ABCD, BiB

The output will also include the descriptive results of each created variable to ensure quality control and reproducibility:

Continuous variable descriptives:

n	mean	sd	Median	min	max	skew	kurt	NA
---	------	----	--------	-----	-----	------	------	----

Categorical variable descriptives:

n	category 1 (n, %)	category 2 (n, %)	category 3 (n, %)	...	NA
---	-------------------	-------------------	-------------------	-----	----

Output Example 2 (continuous variable): Child's sleep

sleepdur_ch_pr_c

n	mean	sd	median	min	max	skew	kurt	NA
10514	8,4	1,64	8,57	1	17	-0,465	3,01	679

Repository of working documents

- Variable_harmonization spreadsheet, containing a) prioritized analyses, b) full list of variables and harmonization assignment, c) original source items for Outcomes, and d) original source items for Determinants
- YEAH! Codebook, containing a) instructions for variable coding, b) codebook and output for Outcomes, c) codebook and output for Determinants, and d) codebook and output for the suggested covariates