Type of manuscript: Original research article

Manuscript title: LongITools: Dynamic longitudinal exposome trajectories in cardiovascular and metabolic non-communicable diseases

Justiina Ronkainen[1], Rozenn Nedelec[1], Angelica Atehortua[2,3], Zhanna Balkhiyarova[4,5,6], Anna Cascarano[2,3], Vien Ngoc Dang[2,3], Ahmed Elhakeem[7, 8], Esther van Enckevort[9], Ana Goncalves Soares[7,8], Sido Haakma[9], Miia Halonen[1], Katharina F. Heil[2,3], Anni Heiskala[1], Eleanor Hyde[9], Bénédicte Jacquemin[10], Elina Keikkala[11,12], Jules Kerckhoffs[13], Anton Klåvus[14], Joanna A. Kopinska[15], Johanna Lepeule[16], Francesca Marazzi[17], Irina Motoc[18], Mari Näätänen[14], Anton Ribbenstedt[19], Amanda Rundblad[20], Otto Savolainen[14,19,21], Valentina Simonetti[22], Nina de Toro Eadie[23], Evangelia Tzala[23], Anna Ulrich[4], Thomas Wright[23], Iman Zarei[14], Enrico d'Amico[22], Federico Belotti[17,24], Carl Brunius[19], Christopher Castleton[25], Marie-Aline Charles[26,27], Romy Gaillard[28,29], Kati Hanhineva[14,29,30], Gerard Hoek[13], Kirsten B. Holven[20,31], Vincent W.V. Jaddoe[28,29], Marika A. Kaakinen[4,5], Eero Kajantie[11,12,32,33], Maryam Kavousi[34], Timo Lakka[35,36,37], Jason Matthews[20], Andrea Piano Mortari[17], Marja Vääräsmäki[11,12], Trudy Voortman[34], Claire Webster[38], Marie Zins[39], Vincenzo Atella[17,24,40], Maria Bulgheroni[22], Marc Chadeau-Hyam[23], Gabriella Conti[41,42], Jayne Evans[38], Janine F. Felix[28,29], Barbara Heude[26], Marjo-Riitta Järvelin[1,23], Marjukka Kolehmainen[14], Rikard Landberg[19], Karim Lekadir[2,3], Stefano Parusso[25], Inga Prokopenko[4,43,44], Susanne R. de Rooij[18], Tessa Roseboom[18,45], Morris Swertz[9], Nicholas Timpson[7,8], Stine M. Ulven[20], Roel Vermeulen[13,46], Teija Juola[1], Sylvain Sebert[1] on behalf of the LongITools Project Group

[1] Center for Life Course Health Research, University of Oulu, Oulu, Finland

[2] Artificial Intelligence in Medicine Lab (BCN-AIM), University of Barcelona, Barcelona, Spain

[3] Department of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain

[4] Section of Statistical Multi-Omics, Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, United Kingdom

[5] Section of Genetics and Genomics, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, United Kingdom

[6] Bashkir State Medical University, Department of Endocrinology, Ufa, Russian Federation

[7] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

[8] MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom

[9] Department of Genetics and Genomics Coordination Center, University of Groningen, Groningen, the Netherlands

[10] University of Rennes, INSERM, School of Advanced Studies in Public Health (EHESP), Research Institute for Environmental and Occupational Health (Isret) UMR_S 1085 , Rennes, France

[11] Finnish Institute for Health and Welfare, Population Health Unit, Helsinki and Oulu, Finland

[12] PEDEGO Research Unit, MRC Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland

[13] Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, the Netherlands

[14] Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland

[15] Department of Social Sciences and Economics, Sapienza University of Rome, Rome, Italy

[16] Grenoble Alpes University, INSERM, CNRS, Institute for Advanced Biosciences, Grenoble, France

[17] CEIS Tor Vergata, Centre for Economic and International Studies, University of Rome Tor Vergata, Rome, Italy

[18] Amsterdam UMC, Epidemiology and Data Science, University of Amsterdam, Amsterdam Public Health, Amsterdam, The Netherlands

[19] Division of Food and Nutrition Science, Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[20] Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Blindern, Oslo, Norway

[21] Department of Biology and Biological Engineering, Chalmers Mass Spectrometry Infrastructure, Chalmers University of Technology, Gothenburg, Sweden

[22] R&D Department, Ab.Acus srl, Milan, Italy

[23] School of Public Health, Dept of Epidemiology and Biostatistics, Imperial College London, St Mary's Hospital, London, United Kingdom

[24] Department of Economics and Finance, University of Rome Tor Vergata, Rome, Italy

[25] CyNexo srl, Trivignano Udinese, Italy

[26] Center for Research in Epidemiology and Statistics (CRESS), INSERM, INRAE, University of Paris, Paris, France

[27] Ined, INSERM, EFS, Elfe Joint Unit, Aubervilliers, France

[28] The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

[29] Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

[30] Department of Biochemistry, University of Turku, Turku, Finland

[31] National Advisory Unit on Familial Hypercholesterolemia, Department of Endocrinology, Morbid Obesity and Preventive Medicine, Oslo University Hospital, Oslo, Norway

[32] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[33] Children's Hospital, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

[34] Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

[35] Institute of Biomedicine/Physiology, University of Eastern Finland, Kuopio, Finland

[36] Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio, Finland

[37] Foundation for Research in Health Exercise and Nutrition, Kuopio Research Institute of Exercise Medicine, Kuopio, Finland

[38] Beta Technology Ltd, Doncaster, United Kingdom

[39] Population-based Epidemiological Cohorts Unit, INSERM UMS 11, Villejuif, France

[40] Stanford University, Stanford, CA, US

[41] Department of Economics, University College London, London, United Kingdom

[42] Social Research Institute, London, United Kingdom

[43] UMR 8199 - EGID, Institut Pasteur de Lille, CNRS, University of Lille, Lille, France

[44] Institute of Biochemistry and Genetics, Ufa Federal Research Centre Russian Academy of Sciences, Ufa, Russian Federation

[45] Gynaecology and Obstetrics, Amsterdam Reproduction and Development Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

[46] Julius Center, University Medical Center Utrecht, Utrecht, the Netherlands

Corresponding author:

Professor Sylvain Sebert

Center for Life Course Health Research

Faculty of Medicine, University of Oulu

P.O. Box 5000

FIN-90014 University of Oulu

Finland

Tel. 00358 503 440842

Email Sylvain.Sebert@oulu.fi

Suggestion for a running head: LongITools Profile Paper

Conflict of interest: The authors declare no conflict of interest

Sources of financial support: This project received funding from the European Commission through the Horizon 2020 research and innovation programme under grant agreement No 874739

How to access the data: LongITools is coordinated by the Centre for Life Course Health Research at the University of Oulu in Finland. Details of the project can be found on the website: www.longitools.org. All cohorts and biobanks involved in LongITools obtained approvals from their institutional ethical committees and participants signed written informed consents. The consortium does not manage a central repository for the data and each cohort has its own data sharing policy. Further information on how to contact each cohort can be found in the Supplementary Table 2.

Acknowledgments

Cohort- or project-specific acknowledgements are available in the Supplementary table 3.

LongITools Project Group

Sylvain Sebert[1], Teija Juola[1], Rozenn Nedelec[1], Justiina Ronkainen[1], Anni Heiskala[1], Miia Halonen[1], Yiyan He[1], Jouko Miettunen[1,2], Ville Karhunen[1,3], Eero Kajantie[4,5,6,7], Marja Vääräsmäki[4,5], Elina Keikkala[4,5], Pia Nyberg[8], Raisa Serpi[8], Janine Felix[9,10], Vincent Jaddoe[9,10], Irene Marques[9,10], Susana Moreira da Silva Santos[9,10], Maryam Kavousi[11], Trudy Voortman[11], Bigina Ginos[11], Marc Chadeau-Hyam[12], Marjo-Riitta Järvelin[1,12], Evangelia Tzala[12], Nina De Toro[12], Thomas Wright[12], Barbara Bodinier[12], Sonia Dagnino[12], Jayne Evans[13], Claire Webster[13], Marjukka Kolehmainen[14], Kati Hanhineva[14,10,15], Timo Lakka[16,17,18], Otto Savolainen[14,19,20], Iman Zarei[14], Mari Näätänen[14], Anton Klåvus[14], Aino-Maija Eloranta[14], Juuso Väistö[16], Anna

Kårlund[14], Santtu Mikkonen[21,22], Merja Atalay[16], Mustafa Atalay[16], Rikard Landberg[19], Carl Brunius[19], Anton Ribbenstedt[19], Morris Swertz[23], Eleanor Hyde[23], Sido Haakma[23], Esther van Enckevort[23], Barbara Heude[24], Marie-Aline Charles[24,25], Marie Zins[26], Johanna Lepeule[27], Bénédicte Jacquemin[28], Lucinda Calas[24], Emeline Lequy-Flahault[29], Wen Lun Yuan[24], Emie Seyve[30], Gabriella Conti[31,32], Roel Vermeulen[33,34], Gerard Hoek[33], Jules Kerckhoffs[33], Inga Prokopenko[35,36,37], Marika Kaakinen[35,38], Zhanna Balkhiyarova[35,38,39], Anna Ulrich[35], Tessa Roseboom[40,41], Susanne De Rooij[40], Irina Motoc[40], Stine Marie Ulven[42], Kirsten B. Holven[42,43], Jason Matthews[42], Amanda Rundblad[42], Siddhartha Das[42], Nicholas Timpson[44,45], Ahmed Elhakeem[44,45], Ana Luiza Goncalves Soares[44,45], Karim Lekadir[46,47], Anna Maria Cascarano[46,47], Angélica Maria Atehortua Labrador[46,47], Vien Ngoc Dang[46,47], Katharina F. Heil[46,47], Catherine Gallin[46,47], Oliver Díaz[46,47], Maria Bulgheroni[48], Valentina Simonetti[48], Enrico D'Amico[48], Laura Giani[48], Fabrizio Manzino[49], Stefano Parusso[49], Christopher Castleton[49], Diego Moimas[49], Vincenzo Atella[50,51,52], Andrea Piano Mortari[50], Joanna Kopinska[53], Francesca Marazzi[50], Matilde Giaccherini[50], Federico Belotti[50,51]

[1] Center for Life Course Health Research, University of Oulu, Oulu, Finland

[2] Medical Research Center Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland

[3] Research Unit of Mathematical Sciences, University of Oulu, Oulu, Finland

[4] Finnish Institute for Health and Welfare, Population Health Unit, Helsinki and Oulu, Finland

[5] PEDEGO Research Unit, MRC Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland

[6] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[7] Children's Hospital, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

[8] Borealis Biobank of Northern Finland, Oulu University Hospital, Oulu, Finland

[9] The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

[10] Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

[11] Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

[12] School of Public Health, Dept of Epidemiology and Biostatistics, Imperial College London, St Mary's Hospital, London, United Kingdom

[13] Beta Technology Ltd, Doncaster, United Kingdom

[14] Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland

[15] Department of Biochemistry, University of Turku, Turku, Finland

[16] Institute of Biomedicine/Physiology, University of Eastern Finland, Kuopio, Finland

[17] Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio, Finland

[18] Foundation for Research in Health Exercise and Nutrition, Kuopio Research Institute of Exercise Medicine, Kuopio, Finland

[19] Division of Food and Nutrition Science, Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[20] Department of Biology and Biological Engineering, Chalmers Mass Spectrometry Infrastructure, Chalmers University of Technology, Gothenburg, Sweden

[21] Department of Applied Physics, University of Eastern Finland, Kuopio, Finland

[22] Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland

[23] Department of Genetics and Genomics Coordination Center, University of Groningen, Groningen, the Netherlands

[24] Center for Research in Epidemiology and Statistics (CRESS), INSERM, INRAE, University of Paris, Paris, France

[25] Ined, INSERM, EFS, Elfe Joint Unit, Aubervilliers, France

[26] Population-based Epidemiological Cohorts Unit, INSERM UMS 11, Villejuif, France

[27] Grenoble Alpes University, INSERM, CNRS, Institute for Advanced Biosciences, Grenoble, France

[28] University of Rennes, INSERM, School of Advanced Studies in Public Health (EHESP), Research Institute for Environmental and Occupational Health (Isret) UMR_S 1085 , Rennes, France

[29] Université Paris-Saclay, Université de Paris, UVSQ, Inserm, UMS 011 "Cohortes en Population", 94800 Villejuif, France

[30] Université Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Development and Respiratory Health, Institute for Advanced Biosciences, La Tronche, France

[31] Department of Economics, University College London, London, United Kingdom

[32] Social Research Institute, London, United Kingdom

[33] Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, the Netherlands

[34] Julius Center, University Medical Center Utrecht, Utrecht, the Netherlands

[35] Section of Statistical Multi-Omics, Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, United Kingdom

[36] UMR 8199 - EGID, Institut Pasteur de Lille, CNRS, University of Lille, Lille, France

[37] Institute of Biochemistry and Genetics, Ufa Federal Research Centre Russian Academy of Sciences, Ufa, Russian Federation

[38] Section of Genetics and Genomics, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, United Kingdom

[39] Bashkir State Medical University, Department of Endocrinology, Ufa, Russian Federation

[40] Amsterdam UMC, Epidemiology and Data Science, University of Amsterdam, Amsterdam Public Health, Amsterdam, The Netherlands

[41] Gynaecology and Obstetrics, Amsterdam Reproduction and Development Institute, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

[42] Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Blindern, Oslo, Norway

[43] National Advisory Unit on Familial Hypercholesterolemia, Department of Endocrinology, Morbid Obesity and Preventive Medicine, Oslo University Hospital, Oslo, Norway

[44] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

[45] MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom

[46] Artificial Intelligence in Medicine Lab (BCN-AIM), University of Barcelona, Barcelona, Spain

[47] Department of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain

[48] R&D Department, Ab.Acus srl, Milan, Italy

[49] CyNexo srl, Trivignano Udinese, Italy

[50] CEIS Tor Vergata, Centre for Economic and International Studies, University of Rome Tor Vergata, Rome, Italy

[51] Department of Economics and Finance, University of Rome Tor Vergata, Rome, Italy

[52] Stanford University, Stanford, CA, US

[53] Department of Social Sciences and Economics, Sapienza University of Rome, Rome, Italy

Abbreviations:

AI: artificial intelligence

ALSPAC: Avon Longitudinal Study of Parents and Children

API: application programming interface

CONSTANCES: Cohorte des Consultants des Centres d'Examens de Santé (in French)

CM-NCDs: cardiovascular and metabolic non-communicable diseases

DFBC: Dutch Famine Birth Cohort

EDEN: *Etude des Déterminants pré et post natals précoces du développement psychomoteur et de la santé de l'Enfant* (In French)

EHEN: European Human Exposome Network

ELFE: *Etude Longitudinale Française depuis l'Enfance* (In French)

ELIPA: Foods for weight maintenance

FAIR: findable, accessible, interoperable and reusable

Fibrefects: Grain fibre modification for gut-mediated health effects

FinnGeDi: Finnish Gestational Diabetes study

GIS: geographical information system

HbA1c: glycosylated haemoglobin

LC-MS: liquid chromatography mass spectrometry

NDVI: normalised difference vegetation index

NFBC: Northern Finland Birth Cohort

NOMA: Fat Quality on Blood Lipids and Immune Response

OULU1935 and 1945: Born in Oulu in 1935 and 1945

PANIC: Physical Activity and Nutrition In Children study

RCGP RSC: Royal College of General Practitioners Research & Surveillance Centre

RCT: randomized controlled trial

RS: Rotterdam Study

SME: small and medium-sized enterprise

SYSDIMET: Health Grain Intervention

UK Biobank: United Kingdom Biobank

WP: work package

Abstract

The current epidemics of cardiovascular and metabolic non-communicable diseases (CM-NCDs) have emerged alongside dramatic modifications in lifestyle and living environments. These correspond to changes in our 'modern' post-war societies globally characterised by rural-to-urban migration, modernisation of agricultural practices and transportation, climate change, and ageing. Evidence suggests that these changes are related to each other; although, the social and biological mechanisms as well as their interactions have yet to be uncovered. LongITools, as one of the nine projects included in the European Human Exposome Network, will tackle this environmental health equation linking multi-dimensional environmental exposures to the occurrence of CM-NCDs.

What this profile paper adds

This consortium profile paper introduces i) LongITools' scientific concepts that are primarily based on longitudinal modelling, ii) the metadata for the project, iii) the expected impact of the project, and finally iv) the strengths and challenges of this endeavour.

Introduction

From one generation to the next, there are vicious circles operating among the rising prevalence of cardiovascular and metabolic non-communicable diseases (CM-NCDs), social inequality, spiralling healthcare costs and varying quality of living environments. If and through which mechanisms these processes relate to each other, is probably one of the greatest epidemiological questions of the 21st century. Undeniably, we are facing complex sociodemographic and medical challenges which can be conceptualised as a network of highly correlated determinants and risk factors. These factors in turn influence longitudinal health trajectories, ultimately contributing to the risk of CM-NCDs and the consequent economic burden. LongITools is one of nine projects included in the European Human Exposome Network (EHEN, www.humanexposome.eu). EHEN is funded by Horizon 2020, the EU Framework Programme for Research and Innovation and it represents the world's largest network of projects created to study the impact of environmental exposure on human health. Within the EHEN, LongITools' task is to study the dynamics of environment and cardiometabolic health and develop tools for exposome research. LongITools brings together European longitudinal data, from prospective cohort studies, randomized controlled trials (RCTs), biobanks and registries, to construct the basis for longitudinal exposome studies. The overarching aim of LongITools is to understand the environmental, biological, and psychosocial dimensions of CM-NCDs, taking life-course factors and a longitudinal approach into consideration. LongITools will improve our understanding on how the exposome, i.e. the combined exposures throughout the life-course of an individual, contributes to the risk of CM-NCDs.

LongITools concepts

The relationships between environmental risk factors such as air pollution, environmental noise, and urbanisation on one hand, and the development of CM-NCDs on the other can be conceptualised in several frameworks as exemplified in Figure 1. LongITools is based on a well-established observational relationship between adiposity and increased risk of adverse glycaemic, lipid-related, and cardiac functions leading to the development of insulin resistance, hypercholesterolemia, hypertriglyceridemia, and high blood pressure which in turn together with environmental risk factors are associated with increased risk of CM-NCDs. Various

stages of development of CM-NCDs share genetic, biological, lifestyle (unhealthy diet and physical inactivity), environmental and sociodemographic causes. However, at all stages of the life-course, which in LongITools are divided into 'early-life', covering the fetal period to childhood, 'adolescence' and 'adulthood and old age', considerable knowledge gaps remain.

Addressing the relationships in Figure 1, LongITools will look to challenge the following assumptions, which are not mutually exclusive, regarding the role of exposome on the life-course development of CM-NCDs:

- **Direct chain of causality:** Variations in environmental risk factors are causally related to changes in lipid and glycaemic trajectories with different relationships to early diseases stages and subsequent development of CM-NCDs. Work within LongITools will attempt to combine data from cohorts as proposed by Hughes et al. and deploy multiple orthogonal analysis designs to challenge the causal chain, such as by implementing cross-cohort comparisons, Mendelian randomisation, and Bayesian path models.[1–3]
- **Joint effects hypothesis:** Associations and potentially causal relationships with a single, or a set of environmental exposures, reflecting underlying commonality, influence the disease trajectories. This hypothesis can be tested by analysing how the environmental factors, in isolation or as latent environmental scores, may modify the core relationship between anthropometry, early disease stages and the onset of CM-NCDs.
- **Bi-directional causality hypothesis:** The relationships between cardiometabolic health trajectories and onset of CM-NCD promote the deterioration of public health and the environment. The assumption of bi-directionality is assumed correct as mutually promoting risk profiles (i.e. of disease status and environmental exposure) are assessed and potentially demonstrated. In this instance, time series/longitudinal data (with possible cross-over events) and bi-directional Mendelian randomisation will be used to explore the possible existence of enforcing feed-forward relationships between disease and environment, for example morbidity, socio-demographic patterns and access and exposure profiles to protective or risky environments.

- **Critical period hypothesis:** There are stages of development within the life-course during which environmental risk factors have apparently greater impact on the development of CM-NCDs. In LongITools we will evaluate the impact of environmental risk factors during the fetal and early childhood period, adolescence, and late adulthood.
- **Biological conversion hypothesis:** Air pollution, climate change, noise and urbanisation can induce biological effects which persist through the modification of regulatory pathways. LongITools will focus on the possible effects of environmental factors on the changes in DNA methylation (epigenomics), gene expression (transcriptomics) and metabolism (metabolomics).
- **Gene-environment hypothesis:** Genetic variation between individuals may modify the induction of biological effects by environmental factors. This relationship may theoretically occur in reverse and effort will be put into the examination of apparent interactions, considering challenges in both statistical power and the true origin of apparent interactions.

The source of life-course data and the opportunity to make these data findable, accessible, interoperable and reusable (FAIR) are core components of this consortium.[4] This consortium profile describes the studies involved in LongITools and the FAIR metadata that the project will build and promote. LongITools is coordinated by the University of Oulu in Finland and includes 15 academic and three small and medium-sized enterprise (SME) partners across Europe (Supplementary table 1). The participants all complement one another, bringing together the full range of technical and specialist expertise in epidemiology, (epi-)genetics, metabolomics, lifestyle, mathematics, economics, policy making and sensor technology that are required to create a critical mass of expertise for the project.

***Who is in the study*?**

LongITools builds upon and leverages prospective birth cohorts, longitudinal studies in adults, register-based follow-ups, randomised controlled trials (RCTs), patient databases as well as maternity and hospital biobanks. Currently, these add up to 25 different studies including 11 million individuals across Europe (Table 1). Birth cohorts within the project will provide substantial longitudinal data from pregnancy to adolescence and early adulthood, complemented by prospective adult cohorts, with multiple follow-ups during adulthood and in

older age. The RCTs involved are focused on the role of nutrition and physical activity in general health and metabolism. These will not only provide comprehensive biological and exposome profiles of study participants but will also allow an in-depth analysis in more controlled settings. Finally, the involved biobanks will be essential for the generalisation of the analyses in large populations. Altogether, the studies involved in LongITools cover the whole life-course, represented by blood samples, metadata, and questionnaires of thousands of cohort participants (Figure 2a, 2b and 3). The data collected in the studies at different time points are summarised in Table 2 and are available in more detail on the LongITools website (www.longitools.org/about).

How do we study?

To optimise findability, all relevant study metadata, i.e., the available variables in the studies as well as how they are harmonised to be made interoperable for pooled and meta-analysis, will be made findable and accessible into a MOLGENIS catalogue,[5] linked to the BBMRI-ERIC Directory of cohorts and biobanks,[6] and integrated with the existing EU Child Cohort Network Variable Catalogue (https://catalogue.lifecycle-project.eu/) created by the Horizon 2020-funded LifeCycle project.[7] LongITools will use recommendations from the LifeCycle project when possible and will establish new harmonisation instructions when needed. By using centrally administered instructions for harmonisation, LongITools aims to ease the collaboration between studies. As all studies historically have their own design and data collection protocols, harmonisation may not always make optimal use of all data available in each study. However, the increased statistical power in the pooled and meta-analyses will be the positive trade-off of possible loss of detail caused by harmonisation. LongITools will use a federated data analysis platform, DataSHIELD, which enables the analysis without need to physically transport the data.

Federated data analysis approach

LongITools will use DataSHIELD, when technically, scientifically and ethically relevant, which was developed as part of the EU-FP7 Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BioSHaRe) Project.[8,9] DataSHIELD enables researchers to analyse data from partner institutions swiftly and securely, respecting current national and European data protection regulations. To briefly summarise its

use, data holders store individual-level data on their own local data warehouse servers and link to the DataSHIELD client portal using MOLGENIS Armadillo server (https://github.com/molgenis/molgenis-service-armadillo). The connection between the data warehouse and the client portal is restricted so that only analysis commands can pass through from the client portal to the data server, and only non-disclosive summary statistics are sent from the data server to the client portal. In this way, analyses using data from multiple studies can be run from a central analyst's computer, thus strongly increasing analysis speed, and decreasing administrative load and local analyst time. Each study controls permissions to identified researchers within LongITools to use their data in any analysis.

What has been and will be measured?

Environmental exposures

LongITools will use existing pan-European models for air pollution, noise, and green space as established within European projects, such as the European Study of Cohorts for Air Pollution Effects (ESCAPE[10,11], www.escapeproject.eu) and the Effects of Low-Level Air Pollution: A Study in Europe (ELAPSE[12,13], www.elapseproject.eu). Following environmental maps will be linked to the individual residential addresses using a geographical information system (GIS):

- **Air pollution** will be assessed using EU-wide air pollution maps at a fine (100*100m) resolution which have been developed within ESCAPE and ELAPSE. These use hybrid land use regression modelling (LUR), incorporating surface air quality monitoring, satellite monitoring, chemical transport modelling and fine scale traffic and land use data;
- **Noise** estimates will be obtained using harmonised pan-European noise exposure models for traffic noise estimates, extending the existing (metropolitan area) maps to the full European population;
- **Green space** will be assessed using satellite-based indices of greenness such as the normalised difference vegetation index (NDVI);
- **Built environment** will be modelled from the GIS and translated into indices of walkability, distances, food and sport outlet density, and accessibility of healthcare services.

These estimates allow LongITools to compose and study the exposomes throughout the life-course. In addition, locally collected exposure data will be applied within RCTs to study the impact of environmental effects and their interaction with intervention target factors in the risk markers of CM-NCD within rather short intervention periods. Although largely available and often collected in a standardised way, the data on environmental exposures have possible intrinsic limitations in terms of i) availability in historical cohorts such as the Dutch Famine Birth Cohort or the Northern Finland Birth Cohort 1966 and ii) heterogeneity of the source (and/or the effects) between countries. This later limitation will be addressed by studying in detail the structure of the data representing the environmental exposures.

Internal exposures

LongITools will analyse the molecular pathways underlying the associations of environmental exposures and cardiometabolic health trajectories by using repeated measures of the internal exposome.

- **Epigenomics** will be studied by using DNA methylation, which has been measured by Illumina Infinium Human Methylation 450K BeadChip and MethylationEPIC BeadChip platforms;
- **Transcriptomics** measures are based on Illumina or Affymetrix arrays and RNA sequencing. For example, we will use transcriptome data from the RCTs Elipa, NOMA and Sysdimet to analyse how air pollution, noise and the build environment may mediate their effect on health via change in specific gene expression;
- **Metabolomics** will be studied using nuclear magnetic resonance (NMR) or liquid chromatography mass spectrometry (LC-MS) based platforms with methods enabling coverage of a wide repertoire of both endo- and exogenous metabolite classes including amino acids, bile acids, steroids, various lipid classes, microbiota-produced metabolites, diet-derived compounds and xenobiotics.[14,15] Non-targeted metabolic profiling will be used to explore the connections between circulating metabolites and the exposure variables, providing metabolic snapshot of the exposome with unique opportunities for molecular epidemiology.[16,17] These analyses will result in semi-quantitative detection of thousands of metabolite features, of which approximately 1000 will be identified *a*

priori. Unidentified metabolites of interest detected from data analysis.[18,19] will be identified using state-of-the-art tools and pipelines.

Information about the availability of the omics data in each LongITools study can be found in Table 2.

Health trajectories

LongITools will use longitudinal, life-course modelling throughout its analyses. LongITools will study how the exposome, linked to geocodes from an individual's birthplace or residential location, is associated with the following cardiovascular and metabolic health trajectories, i.e., the four main outcome phenotypes of LongITools:

- **Anthropometric trajectories,** identified using height and weight measures in infancy, childhood and adolescence, by using longitudinal growth data or latent trajectory modelling supported by adiposity milestones, such as adiposity peak, adiposity rebound, body mass index (BMI) at puberty and life-course BMI trajectories;
- **Glycaemic health trajectories,** identified using repeated measures of glycaemic health, such as fasting glucose, fasting insulin, glycosylated haemoglobin (HBA1c), diabetes diagnosis and diabetes medications;
- **Cardiovascular health trajectories,** identified using repeated measures of blood pressure, heart rate, indices of cardiac structure and function, cardiac diagnoses and cardiovascular medications;
- **Lipid-related health trajectories,** identified using repeated measures of blood lipids, lipoproteins and related medications.

Economic and policy impact

LongITools will build a comprehensive dataset of policy interventions targeting the exposome and healthcare, which were implemented in the time span and locations covered by the birth cohorts (Figure 2). The aim is to investigate if and how such policy interventions have affected the insurgency of CM-NCDs, in terms of both health status and economic implications. Furthermore, LongITools will estimate, within an economic life-course model of health production, the extent to which the economic burden is due to the external exposome

and evaluate policy-relevant 'what-if' scenarios using a dynamic microsimulation model, i.e. the Future Elderly Model.[20–22]

Knowledge exploitation

The theoretical framework will be carried out on existing data from the LongITools consortium to train artificial intelligence (AI) algorithms, such as random forests, support vector machines and deep neural networks, that will enable translation of data and knowledge into simple and available predictive tools for scientists, citizens, policy makers or other end-users. For this later part, co-designing activities are currently on-going with multiple stakeholders, including clinicians, AI technologists, social scientists and exposome experts, to define the functional and user requirements for these AI-powered digital tools. The steps being developed to achieve this are visualised in the Supplementary Figure 1, where interdisciplinary competences converge. Many variables from environmental and personal domains concur to delineate longitudinal trajectories. Some of them are already available, thanks to digital personal healthcare devices, while others will be more specific and will need the inclusion of targeted sensors as part of an embedded system (LongITOolsHub). These tools will be validated in a pilot study.

Strengths and challenges

LongITools comprises a vast amount of prospective data collected in Europe, harnessed to enhance exposome research as well as longitudinal and econometric modelling. When combined, these data offer immense potential to inform future European health policy. Furthermore, the data are organised to enable direct replication under the FAIR principles. While sample size allowing statistical power is deemed essential for robust evidence-based strategies, it is also important to combine study designs to validate findings under different statistical assumptions. Another strength of LongITools is the inclusion of data from RCTs for in-depth sensitivity analyses and to identify novel pathways that could be generalised in the cohort setting. Finally, LongITools includes longitudinal birth cohorts and ageing cohorts from the same geographical location, which enables us to study the changing environment and its association with cardiovascular and metabolic health.

The key challenge faced by LongITools, and more broadly by all epidemiological study, is to translate the findings into meaningful change for global health. To tackle this, LongITools operates in close collaboration with policy makers throughout the project to convert the results into evidence-based policy options. A critical mass of data and expertise brought together in LongITools offers a substantial resource which also leads to another challenge faced by the consortium: how to best combine the characteristics of the cohorts involved. The cohorts were established for their own individual purposes before being brought together under this project, and the methods of data collection have thus not been standardised *a priori* across the consortium. Therefore, consideration is required for the transferability of the statistical models and harmonisation of the data. However, this also gives us the opportunity to examine if similar processes operate in different environments and thus to draw conclusions on generalisability. In addition, owing to the internationality of the project, differences in technology, questionnaire data and biospecimen collection methods, terminology and diagnosis definitions, country-specific measurement techniques, and ethical requirements among the studies exist. This heterogeneity can introduce differences in the results between the studies which can be analysed when necessary; we can generalize where possible and be specific when needed. In addition, the environmental exposures are harmonised by using the same model, which can also mitigate possible inconsistencies between the studies. The consortium has made significant progress in overcoming these challenges by developing and updating harmonisation manual for the key variables and the overarching advantage of LongITools is that all studies provide rich data on similar key exposures and the outcome measures of interest.

Conclusion

LongITools provides a collection of studies across different time periods and encompassing different life stages, which will enable us to use a life-course approach to study the exposome and its role in the trajectories of cardiometabolic health. Valuing the idea of open science, through its innovative data infrastructure, LongITools will spread new knowledge rapidly and efficiently to the other European Human Exposome Network projects and beyond. The generated and combined knowledge can then be used to develop innovative products and services with the potential to create new markets. In this way, LongITools

aims to improve EU citizens' cardiovascular and metabolic health and thereby reduce individual and societal burdens and healthcare costs of CM-NCDs. Through the cooperation between research teams and SMEs and by using our extensive data, we expect to make several breakthrough discoveries. The evidence-based innovation platform developed in collaboration among academic and industrial partners during the project will support the cross-fertilisation of new technologies and stimulate collaborations in developing new products and services within and beyond the European Human Exposome Network. As a proof of concept, LongITools will develop a mobile application for cardio-metabolic risk monitoring, combining computational methods to wearable sensors data, realizing effective cooperation between academic and SME partners.

Funding

This project received funding from the European Commission through the Horizon 2020 research and innovation programme under grant agreement No 874739.

How can I access the data?

LongITools is coordinated by the Centre for Life Course Health Research at the University of Oulu in Finland. Details of the project can be found on the website: www.longitools.org. All cohorts and biobanks involved in LongITools obtained approvals from their institutional ethical committees and participants signed written informed consents. The consortium does not manage a central repository for the data and each cohort has its own data sharing policy. Further information on how to contact each cohort can be found in the Supplementary Table 2.

Acknowledgments

Cohort- or project- specific acknowledgements are available in the Supplementary table 3.

Conflict of interest

The authors declare no conflict of interest.

References

1. Hughes RA, Tilling K, Lawlor DA. Combining Longitudinal Data From Different Cohorts to Examine the Life-Course Trajectory. *Am J Epidemiol*. July 2021. doi:10.1093/aje/kwab190

2. Amaral AFS, Imboden M, Wielscher M, et al. Role of DNA methylation in the association of lung function with body mass index: a two-step epigenetic Mendelian randomisation study. *BMC Pulm Med*. 2020;20(1):171. doi:10.1186/s12890-020-01212-9

3. Detilleux J, Reginster J-Y, Chines A, Bruyère O. A Bayesian path analysis to estimate causal effects of bazedoxifene acetate on incidence of vertebral fractures, either directly or through non-linear changes in bone mass density. *Stat Methods Med Res*. 2016;25(1):400-412. doi:10.1177/0962280212456655

4. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018. doi:10.1038/sdata.2016.18

5. 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

6. Holub P, Swertz M, Reihs R, Van Enckevort D, Müller H, Litton JE. BBMRI-ERIC Directory: 515 Biobanks with over 60 Million Biological Samples. *Biopreserv Biobank*. 2016;14(6):559-562. doi:10.1089/bio.2016.0088

7. Jaddoe VWV, Felix JF, Andersen AMN, et al. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. *Eur J Epidemiol*. 2020;35(2):709-724. doi:10.1007/s10654-020-00662-z

8. 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). doi:10.1186/1742-7622-10-12

9. Gaye A, Marcon Y, Isaeva J, et al. DataSHIELD: Taking the analysis to the data, not the data to the analysis. *Int J Epidemiol*. 2014;43(6):1929-1944. doi:10.1093/ije/dyu188

10. Eeftens M, Beelen R, De Hoogh K, et al. Development of land use regression models for PM2.5, PM 2.5 absorbance, PM10 and PMcoarse in 20 European study areas; Results of the ESCAPE project. *Environ Sci Technol*. 2012;46(20):11195-11205. doi:10.1021/es301948k

11. Beelen R, Hoek G, Vienneau D, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe - The ESCAPE project. *Atmos Environ*. 2013;72:10-23. doi:10.1016/j.atmosenv.2013.02.037

12. de Hoogh K, Gulliver J, Donkelaar A van, et al. Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environ Res*. 2016;151:1-10. doi:10.1016/j.envres.2016.07.005

13. de Hoogh K, Chen J, Gulliver J, et al. Spatial PM 2.5 , NO 2 , O 3 and BC models for Western Europe – Evaluation of spatiotemporal stability. *Environ Int*. 2018;120:81-92. doi:10.1016/j.envint.2018.07.036

14. Klåvus A, Kokla M, Noerman S, et al. "Notame": Workflow for non-targeted LC-MS metabolic profiling. *Metabolites*. 2020;10(4). doi:10.3390/metabo10040135

15. Brunius C, Shi L, Landberg R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics*. 2016;12(11):1-13. doi:10.1007/s11306-016-1124-4

16. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect*. 2014;122(8):769-774. doi:10.1289/ehp.1308015

17. Rappaport SM. Implications of the exposome for exposure science. *J Expo Sci Environ Epidemiol*.

2011;21(1):5-9. doi:10.1038/jes.2010.50

18. Shi L, Westerhuis JA, Rosén J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. *Bioinformatics*. 2019;35(6):972-980. doi:10.1093/bioinformatics/bty710

19. Schillemans T, Shi L, Liu X, Åkesson A, Landberg R, Brunius C. Visualization and Interpretation of Multivariate Associations with Disease Risk Markers and Disease Risk—The Triplot. *Metabolites*. 2019;9(7):133. doi:10.3390/metabo9070133

20. Atella V, Belotti F, Kim D, et al. The future of the elderly population health status: Filling a knowledge gap. *Health Econ*. March 2021. doi:10.1002/hec.4258

21. Atella V, Belotti F, Blaylock B, Goldman D, Piano Mortari A, Tysinger B. *Preventing Ageing Unequally*. OECD Publishing; 2017. doi:https://doi.org/https://doi.org/10.1787/9789264279087-en

22. Goldman DP, Orszag PR. The Growing Gap in Life Expectancy: Using the Future Elderly Model to Estimate Implications for Social Security and Medicare. *Am Econ Rev*. 2014;104(5):230-233. doi:10.1257/aer.104.5.230

23. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol*. 2013;42(1):97-110. doi:10.1093/ije/dys066

24. 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. *Int J Epidemiol*. 2013;42(1):111-127. doi:10.1093/ije/dys064

25. Lawlor DA, Lewcock M, Rena-Jones L, et al. The second generation of The Avon Longitudinal Study of Parents and Children (ALSPAC-G2): a cohort profile. *Wellcome Open Res*. 2019;4:36. doi:10.12688/wellcomeopenres.15087.2

26. Sebert S, Lowry E, Aumüller N, et al. Cohort Profile: The DynaHEALTH consortium - A European consortium for a life-course bio-psychosocial model of healthy ageing of glucose homeostasis. *Int J*

Epidemiol. 2019;48(4):1051-1051K. doi:10.1093/ije/dyz056

27. 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. *Int J Epidemiol*. 2016;45(2):353-363. doi:10.1093/ije/dyv151

28. Vandentorren S, Bois C, Pirus C, Sarter H, Salines G, Leridon H. Rationales, design and recruitment for the Elfe longitudinal study. *BMC Pediatr*. 2009;9(1):58. doi:10.1186/1471-2431-9-58

29. Mustaniemi S, Vääräsmäki M, Eriksson JG, et al. Polycystic ovary syndrome and risk factors for gestational diabetes. *Endocr Connect*. 2018;7(7):859-869. doi:10.1530/EC-18-0076

30. Keikkala E, Mustaniemi S, Koivunen S, et al. Cohort Profile: The Finnish Gestational Diabetes (FinnGeDi) Study. *Int J Epidemiol*. 2020;dyaa039. doi:10.1093/ije/dyaa039

31. Kooijman MN, Kruithof CJ, van Duijn CM, et al. The Generation R Study: design and cohort update 2017. *Eur J Epidemiol*. 2016;31(12):1243-1264. doi:10.1007/s10654-016-0224-9

32. Goldberg M, Carton M, Descatha A, et al. CONSTANCES: a general prospective population-based cohort for occupational and environmental epidemiology: cohort profile. *Occup Env Med*. 2017;74:66-71. doi:10.1136/oemed-2016-103678

33. Ikram MA, Brusselle G, Ghanbari M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. *Eur J Epidemiol*. 2020;35(5):483-517. doi:10.1007/s10654-020-00640-5

34. Karhunen L, Lyly M, Lapveteläinen A, et al. Psychobehavioural factors are more strongly associated with successful weight management than predetermined satiety effect or other characteristics of diet. *J Obes*. 2012;2012. doi:10.1155/2012/274068

35. Lappi J, Mykkänen H, Knudsen KEB, et al. Postprandial glucose metabolism and SCFA after consuming wholegrain rye bread and wheat bread enriched with bioprocessed rye bran in individuals with mild gastrointestinal symptoms. *Nutr J*. 2014;13(1):104. doi:10.1186/1475-2891-13-

104

36. Ulven SM, Leder L, Elind E, et al. Exchanging a few commercial, regularly consumed food items with improved fat quality reduces total cholesterol and LDL-cholesterol: A double-blind, randomised controlled trial. *Br J Nutr*. 2016;116(8):1383-1393. doi:10.1017/S0007114516003445

37. Lakka TA, Lintu N, Väistö J, et al. A 2 year physical activity and dietary intervention attenuates the increase in insulin resistance in a general population of children: the PANIC study. *Diabetologia*. 2020;63(11):2270-2281. doi:10.1007/s00125-020-05250-0

38. De Mello VDF, Schwab U, Kolehmainen M, et al. A diet high in fatty fish, bilberries and wholegrain products improves markers of endothelial function and inflammation in individuals with impaired glucose metabolism in a randomised controlled trial: The Sysdimet study. *Diabetologia*. 2011;54(11):2755-2767. doi:10.1007/s00125-011-2285-3

Figure legends

Figure 1: Directed acyclic graphs representing possible pathways from environmental exposures to adiposity trajectories (A), built environment to glycaemic health trajectories (B) and air pollution to vascular ageing via growth and adiposity trajectories (C). Black arrows indicate potential causal links, grey arrows indicate confounding paths and dashed arrows indicate non-causal, latent class paths. AIR, air pollution; NOISE, environmental noise; BUILT, built environment; E, exposure; O, outcome.

Figure 2: LongITools cohorts (a) and interventions (b).

Figure 3: Map of studies participating in LongITools. The size of the circle indicates the relative size of the study.

Supplementary Figure 1: Design and principles of the LongITools health application.

1 **Tables**

2 Table 1: Summary of studies involved in LongITools. RCT, randomized controlled trial.

Abbreviated study name and reference	Full study name	Study design	Geographic information	Year of establishment*	Sample size at baseline
Prospective birth cohorts					
ALSPAC G0[23]	Avon Longitudinal Study of Parents and Children Generation 0	Prospective adult cohort	Home address geocoded to property and postcode level	1991-1992	14,541
ALSPAC G1[24]	Avon Longitudinal Study of Parents and Children Generation 1	Prospective birth cohort	Home address geocoded to property and postcode level	1991-1992	14,062
ALSPAC G2[25]	Avon Longitudinal Study of Parents and Children Generation 2	Prospective infant/child cohort	Home address geocoded to property and postcode level	2012-2018	850 ongoing
DFBC[26]	Dutch Famine Birth Cohort	Prospective birth cohort	Home address available at each visit	1943	2,414
EDEN[27]	*Etude des Déterminants pré et post natals précoces du développement psychomoteur et de la santé de l'Enfant* (In French)	Prospective birth cohort	Home address geocoded to property and postcode level	2003	2,002
ELFE[28]	*Etude Longitudinale Française depuis l'Enfance* (In French)	Prospective birth cohort	Home address available at each visit	2011	18,329
FinnGeDi[29,30]	Finnish Gestational Diabetes study	Prospective birth cohort	Home address available at baseline	2009-2012	2,212
Generation R[31]	Generation R Study	Prospective birth cohort	Home address available at each visit	2002-2006	9,778
NFBC1966[26]	Northern Finland Birth Cohort 1966	Prospective birth cohort	Home address available from 1966	1966	12,055
NFBC1986[26]	Northern Finland Birth Cohort 1986	Prospective birth cohort	Home address available from 1986	1986	9,432
Prospective adult cohorts					

CONSTANCES[32]	*Cohorte des Consultants des Centres d'Examens de Santé* (In French)	Prospective adult cohort	Home address available	2012-2019	220,000
RS I[33]	Rotterdam Study, first cohort	Prospective adult cohort	Home address available	1989	7,983
RS II[33]	Rotterdam Study, second cohort	Prospective adult cohort	Home address available	2000	3,011
RS III[33]	Rotterdam Study, third cohort	Prospective adult cohort	Home address available	2006	3,932
OULU1935 and 1945	Born in Oulu in 1935 and 1945	Prospective adult cohort	Home address available	1935-1945	2,000
Interventions and trials					
ELIPA[34]	*Elintarvikkeita Painonhallintaan* (In Finnish)	RCT	Home address geocoded to property and postcode level	2008	99
Fibrefects[35]	Grain fibre modification for gut-mediated health effects	RCT	Geocoding in process	2011	25
NOMA[36]	Fat Quality on Blood Lipids and Immune Response	RCT	Home address available at baseline	2012-2014	99
PANIC[37]	Physical Activity and Nutrition In Children study	Controlled intervention	Home address available at baseline	2007-2009	504
SYSDIMET[38]	Health Grain intervention	RCT	Not available	2007	102
Administrative cohorts and Biobanks					
Borealis Biobank	Borealis Biobank of Northern Finland	Biobank	Home address available	2015**	500,000
FMC	Finnish Maternity Cohort (managed by Borealis Biobank)	Biobank	Home address available upon request	1983	950,000
RCGP RSC	Royal College of General Practitioners Research & Surveillance Centre	Primary care sentinel network	Not available	1990-2018	7,000,000
UK Biobank	United Kingdom Biobank	Biobank	Not available	2006	500,000

* Year of establishment in biobanks corresponds to the year from which samples are available

** Diagnostic pathology tissue archives starting from 1978 – 8/2013 transferred to the biobank in addition to ongoing prospective biobank consent and sample collection

Table 2: Data available for general population-based studies and clinical trials in LongITools (x, data available; -, no data collection at this time point; o, data will be collected during LongITools).

Indicator	ALSPAC G0	ALSPAC G1	ALSPAC G2	DFBC	EDEN	ELFE	FinnGeDi	Generation R	NFBC1966	NFBC1986	CONSTANCES	RS	OULU1935/45	ELIPA	Fibrefects	NOMA	PANIC	SYSDIMET
Parental and Pregnancy																		
Anthropometric measures	x	x	x	x	x	x	x	x	x	x	-	-	-	-	-	-	x	-
Blood samples	x	x	x	-	x	x	x	x	x	x	-	-	-	-	-	-	-	-
Lifestyle and health behaviour	x	x	x	-	x	x	x	x	x	x	-	-	-	-	-	-	x	-
Socioeconomic indicators	x	x	x	x	x	x	x	x	x	x	-	-	-	-	-	-	x	-
GIS/living location	x	x	x	-	x	x	x	x	x	x	-	-	-	-	-	-	x	-
Epigenomics	x	x	-	-	x	-	x	-	-	-	-	-	-	-	-	-	-	-
Transcriptomics	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Metabolomics	-	-	-	-	o	-	-	x	-	-	-	-	-	-	-	-	-	-
Childhood																		
Anthropometric measures	-	x	x	-	x	x	x	x	x	x	x	-	-	-	-	-	x	-
Developmental milestones	-	x	x	-	x	x	-	x	x	x	-	-	-	-	-	-	x	-
Growth modelling	-	x	x	-	x	x	-	x	x	x	-	-	-	-	-	-	x	-
Blood samples	-	x	x	-	x	x	-	x	-	-	-	-	-	-	-	-	x	-
Lifestyle and health behaviour	-	x	x	-	x	x	-	x	x	x	-	-	-	-	-	-	x	-
Socioeconomic indicators	-	x	x	-	x	x	-	x	x	x	x	-	-	-	-	-	x	-
GIS/living location	-	x	x	-	x	x	-	x	x	x	x	-	-	-	-	-	x	-
Epigenomics	-	x	-	-	x	-	x	x	-	-	-	-	-	-	-	-	-	-
Transcriptomics	-	x	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-
Metabolomics	-	x	-	-	o	-	-	o/x	-	-	-	-	-	-	-	-	x	-
Adolescence and Early Adulthood																		
Anthropometric measures	-	x	-	-	-	-	-	x	x	x	x	-	-	-	-	-	x	-
Blood samples	-	x	-	-	-	-	-	x	x	x	x	-	-	-	-	-	x	-
Lifestyle and health behaviour	-	x	-	-	-	-	-	x	x	x	x	-	-	-	-	-	x	-

Socioeconomic indicators	-	x	-	-	-	-	-	x	x	x	x	-	-	-	-	-	x	-
GIS/living location	-	x	-	-	-	-	-	x	x	x	x	-	-	-	-	-	x	-
Epigenomics	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
Transcriptomics	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Metabolomics	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	x	-
Adulthood and Old age																		
Anthropometric measures	x	-	-	x	-	-	-	-	x	x	x	x	x	x	x	x	o	x
Blood samples	x	-	-	x	-	-	-	-	x	x	x	x	x	x	x	x	o	x
Lifestyle and health behaviour	x	-	-	x	-	-	-	-	x	x	x	x	x	x	x	x	o	x
Socioeconomic indicators	x	-	-	x	-	-	-	-	x	x	x	x	x	x	x	-	o	x
GIS/living location	x	-	-	x	-	-	-	-	x	x	x	x	x	x	x	x	o	-
Epigenomics	x	-	-	x	-	-	-	-	x	-	-	x	-	-	-	-	-	-
Transcriptomics	x	-	-	-	-	-	-	-	-	-	-	x	-	x	-	x	-	x
Metabolomics	x	-	-	x	-	-	-	-	x	-	-	x	-	x	x	x	o	x

8