More info on LIFECYCLE online:
lifecycle-project.eu

Tutorials on the implementation of causal inference methods and longitudinal modelling in the context of life course trajectory analyses

LifeCycle report D7.4

Authors:

Lorenzo Richiardi (UNITO)

Ghislaine Scelo (UNITO)

Daniela Zugna (UNITO)

Maja Popovic (UNITO)

Deborah Lawlor (UNIVBRIS)

Version 2.0

Delivery date: Month 66

LIFECYCLE Project
Coordination and leadership:

Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands

Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Document Information

Grant Agreement No.	**733206**
Project Title	Early-life stressors and LifeCycle health (The LifeCycle Project)
Project Start Date	01 January 2017
Work package title	WP7 - Methods for causal inference and life course trajectory analyses
Related task(s)	Task 7.4
Lead Organisation	UNITO
Submission date	19 September 2022
Dissemination Level	Public

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Table of contents

List of Figures

List of Tables

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

List of Abbreviations

ALSPAC: Avon Longitudinal Study of Parents and Children
BCG: Barry Caerphilly Growth
BiB: Born in Bradford
CHS: Christ's Hospital School
CI: confidence interval
DAG: directed acyclic graph
IORW: Inverse odds ratio weighting
IPW: Inverse probability weighting
MPES: Multi parameter evidence synthesis
NINFEA: Nascita e Infanzia: l'Effetto dell'Ambiente
PBR: Piedmont Birth Register
PR: prevalence ratio
PROBIT: Promotion of Breastfeeding Intervention Trial
RDD: regression discontinuity design
TMLE: targeted maximum likelihood estimator
WP: work package

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Executive summary

Task 7.4 aimed at developing tutorials and educational material with empirical applications to demonstrate the implementation in specific contexts of causal inference approaches and longitudinal data modelling. During the first part of the project, we identified five topics of specific interest for the birth cohort research community: (i) Mediation analysis extended to multiple mediators; (ii) Development of life course trajectories of health outcomes; (iii) Generalizability and transportability of study results; (iv) Regression discontinuity designs to strengthen causal inference in life course epidemiology; (v) Triangulation of different methods with different sources of bias to improve causal inference. We then prepared five tutorials summarizing the main concepts and assumptions and showing empirical applications. This material was discussed and disseminated at multiple levels, including at the LifeCycle meetings, through the LifeCycle workshops, at conferences, through manuscripts, and using the LifeCycle method database. We started working also on an additional tutorial on mediation analysis in DataSHIELD, sharing the computer code at a workshop organized in conjunction with the last LifeCycle meeting. The overarching aim of Task 7.4 was to facilitate the adoption and use of state-of-the-art methods within and outside LifeCycle and to promote discussion on their advantages and limitations. Ultimately, this will help in identifying causal relationships in birth cohort studies to better inform policies.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

1. Introduction

WP7 of LifeCycle focuses on specific methodological aspects of importance for the EU Child Cohort Network and life course research in general. It aims at developing an integrated analysis strategy to apply causal inference methods, model longitudinal data and health trajectories; assessing approaches to analyse multiple exposure data in the context of longitudinal modelling; and, at the same time, enhancing, inside and outside LifeCycle, also through tutorials and workshops, the knowledge and use of causal inference approaches and methods to model longitudinal data.

Task 7.4 aims at developing tutorials and/or educational material with empirical applications to demonstrate the implementation in specific contexts of causal inference approaches and longitudinal data modelling. This material should be disseminated within LifeCycle and to the broader research community outside LIfeCycle. Task 7.4 is also informed by the work done under Tasks 7.1 and 7.2 that developed the analysis strategy (see deliverables 7.1 and 7.2).

2. Identification of the specific topics and data availability for the tutorials

The specific topics for the tutorials were identified following a process described in the Milestone 12 (submitted in December 2019). Potential topics were discussed during the WP7 sessions at the LifeCycle meetings, the LifeCycle workshops, and the WP7 telephone conferences. We considered: (i) participants' inputs, (ii) the specific aims of WP4 (Early-life stressors and cardio-metabolic health life course trajectory), WP5 (Early-life stressors and respiratory health life course trajectories), WP6 (Early-life stressors and mental health life course trajectory), and WP8 (DNA methylation and gene expression in life-course health trajectory), (iii) emerging topics in the epidemiological literature of interest for life course epidemiology, and (iv) materials available in the searchable database of method materials of importance for life course epidemiology, developed and maintained within the framework of tasks 7.1 and 7.2.

The following topics were identified: 1. Mediation analysis extended to multiple mediators; 2. Developing life course trajectories of health outcomes; 3. Generalizability and transportability of study results; 4. Use of non-genetic instrumental variable approaches (policy-related instrumental variables and regression discontinuity) to strengthen causal inference in life course epidemiology, then further focused on regression discontinuity; 5. Triangulation of different methods with different sources of bias to improve causal inference.

Following the preparation of the *dsMediationClient* DataSHIELD package, in 2021 mediation analysis in DataSHIELD was identified as a topic of interest for an additional tutorial.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

3. Tutorials and applied examples

3.1 Mediation analysis extended to multiple mediators (1)

Mediation analysis aims at estimating to what extent the effect of an exposure on an outcome is explained by a set of mediators on the causal pathway between the exposure and the outcome. In this context, the total effect of the exposure on the outcome can be decomposed into the natural indirect effect, i.e. the effect explained by the mediators jointly, and the natural direct effect, i.e. the effect unexplained by the mediators. However finer decompositions are also possible in presence of independent or sequential mediators. As sequential mediation analysis is increasingly common in epidemiology, applied researchers have to interface with difficulties related to the application, implementation, and interpretation of the methods pro- posed in literature. We review four statistical methods to analyse multiple sequential mediators, all based on the counterfactual framework: the inverse odds ratio weighting approach (2), the inverse probability weighting approach (3), the imputation approach (4), and the extended imputation approach (5). These approaches are described, compared and implemented using a case-study with the aim to investigate the role of adverse reproductive outcomes and infant respiratory infections on infant wheezing in the NINFEA birth cohort. A DAG depicting the case-study used in the tutorial is reported in **Figure 1**.

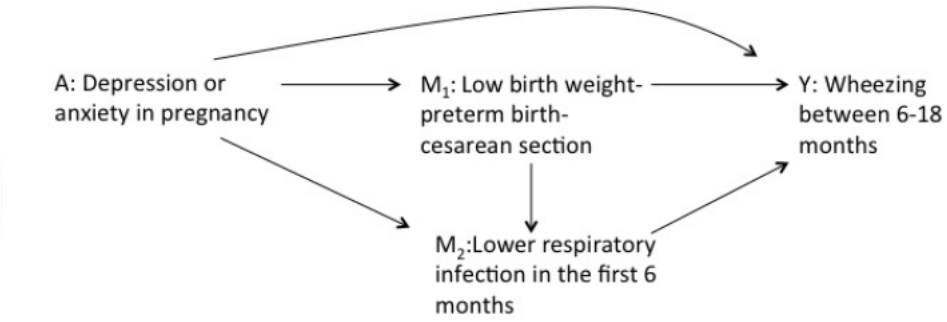

Figure 1. DAG representing the hypothesized structure of the case study. For the sake of simplicity the confounders C are not shown.

Table 1 reports the main results for the different methods: IORW (Inverse odds ratio weighting), IPW (Inverse probability weighting), imputation approach, marginal and conditional effects. In **Table 2**, the imputation approach is used to further decompose between M1 and M2.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Table 1. Estimates of total, direct and indirect effects of maternal depression or anxiety in pregnancy on the risk of infant wheezing between 6 and 18 months of age using inverse odds ratio weighting, inverse probability weighting and the imputation approach.

M1: adverse reproductive outcomes. M2: infant lower respiratory infections. Data from the NINFEA birth cohort.

	Through M_1		Through M_1 and M_2	
	PR*	95% CI*	PR	95% CI*
Conditional effect			**IORW* approach**	
Pure direct effect	1.59	1.27-1.94	1.57	1.25-1.92
Total indirect effect	1.03	0.94-1.12	1.05	0.95-1.15
Total effect	1.64	1.33-2.00	1.64	1.33-1.97
Marginal effect			**IPW** approach**	
Pure direct effect	1.60	1.30-1.94	1.57	1.27-187
Total indirect effect	1.02	0.99-1.04	1.04	0.99-1.09
Total effect	1.63	1.33-1.98	1.63	1.31-195
Conditional effect			**Imputation approach**	
Pure direct effect	1.60	1.31-1.94	1.57	1.26-1.90
Total indirect effect	1.02	1.01-1.05	1.05	1.01-1.09
Total effect	1.64	1.33-1.99	1.64	1.33-1.99
Marginal effect			**Imputation approach**	
Pure direct effect	1.60	1.30-1.91	1.57	1.24-1.88
Total indirect effect	1.02	1.00-1.04	1.04	0.99-1.09
Total effect	1.63	1.33-1.95	1.62	1.29-1.95

PR: prevalence ratio; CI: confidence interval calculated by bootstrap

Table 2. Estimates of conditional total, direct and indirect effects of maternal depression or anxiety in pregnancy on the risk of infant wheezing between 6 and 18 months of age using the extended imputation approach.

M1: adverse reproductive outcomes. M2: infant lower respiratory infections. Data from the NINFEA birth cohort.

	Extended imputation approach	
Conditional effect	**PR***	**95% CI***
Pure direct effect	1.57	1.28-1.86
Total indirect effect through M_1 and M_2 jointly	1.05	1.00-1.09
Total indirect effect through M_1	1.00	0.99-1.00
Partial total indirect effect through M_2	1.05	1.00-1.09
Total effect	1.64	1.34-1.96

PR: prevalence ratio; CI: confidence interval calculated by bootstrap

3.2 Developing life course trajectories of health outcomes (6)

Longitudinal data are necessary to reveal changes within an individual as he or she ages. However, rarely will a single cohort study capture data throughout a person's entire life span. Here we describe in detail the steps needed to develop life-course trajectories from cohort studies that cover different and overlapping periods of life. Such independent studies are probably from heterogenous populations, which raises several challenges, including: 1) data harmonization (deriving new harmonized variables from differently measured variables by identifying common elements across all studies); 2) systematically missing data (variables not measured are missing for all participants in a cohort); and 3) model selection with differing age ranges and measurement schedules. We illustrate how to overcome these challenges using an example which examines the associations of parental education, sex, and race/ethnicity with children's weight trajectories. Data were obtained from 5 prospective cohort studies (carried out in Belarus and 4 regions of the United Kingdom) spanning data collected from birth to early adulthood during differing calendar periods (1936–1964, 1972–1979, 1990–2012, 1996–2016, and 2007–2015). Key strengths of our

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

approach include modelling of trajectories over wide age ranges, sharing of information across studies, and direct comparison of the same parts of the life course in different geographical regions and time periods. We also introduce a novel approach of imputing individual-level covariates of a multilevel model with a nonlinear growth trajectory and interactions. **Figure 2** demonstrates how the information shared by the 5 cohorts is combined to identify the predicted weight trajectories according to selected covariates.

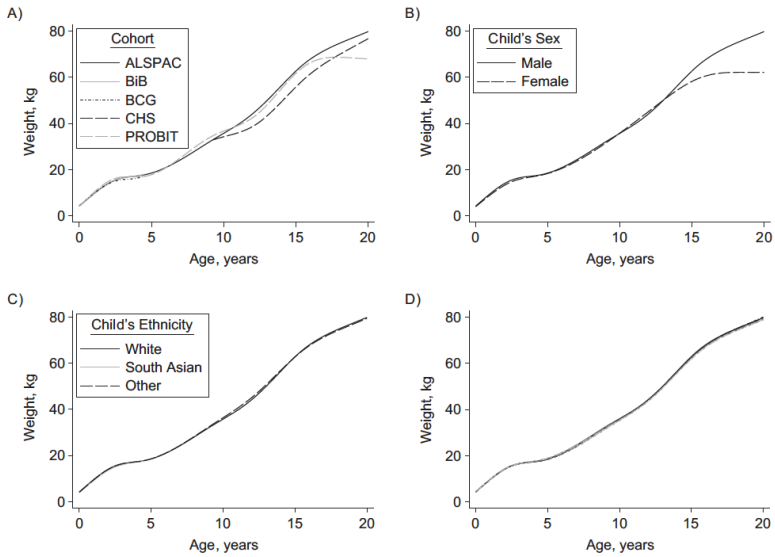

Figure 2. Predicted mean weight trajectories of children in 5 cohort studies according to cohort (A), child's sex (B), child's race/ethnicity (C), and parental education and employment status (D).

3.3 Generalizability and transportability of study results (7)

Interest in external validity of study findings has increased in recent years, leading to a multiplication of theoretical publications in the field, as well as raising concerns on how well target populations of health interventions are represented in study samples. Generalizability addresses the feasibility of applying a study finding to the broader population from which the study sample is a subset, while transportability focuses on the setting where the study sample is at least partly external to the population of interest. The design and timeline of clinical trials often allows to generalize the findings to the source population as long as it is clearly defined. In large population-based, observational epidemiological studies, transportability becomes more relevant as the intent is often to quantify associations between an exposure and an outcome in contemporary populations of interest to inform decision-makers. This is particularly true for life-course epidemiology, where long-term longitudinal data collection is in play. Transportability offers the ability to assess an association in different populations, under certain assumptions often not more stringent than those required for generalizability. In this paper we review the conceptual framework and assumptions behind transportability of results from a population-based study sample to a population of interest in an observational setting. We detail one method through an applied example, where internally validity was constructed for illustrative purposes.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

There are several mechanisms that may affect transportability: the context, the dynamic nature of the populations, the intentional definition of the target and study populations, and the unintentional selection from the source to the study population. Some of these mechanisms, especially the latter but potentially all of them, are also potential sources of selection bias that may affect the internal validity of the study estimates. In these instances, the selection is affected by the variables included in the data generating model, rather than being an external cause of the distribution of the variables.

External validity depends, in essence, on the difference between the target population and the study population, and on how this difference can modify the effect of the exposure on the outcome. Thus, reasoning on external validity goes beyond the structure of the population of interest, namely the target population. It requires the choice of at least one study population from which transportability of the estimates to the target population is possible. It follows that external validity issues involve considering the structure of at least two populations, the target and the study population, and their relevant differences.

Classical DAGs are insufficient to depict the whole process and cannot be used to judge on the validity of the inference across populations. For this reason, Judea Pearl and his colleagues have introduced the notion of selection diagrams, which are causal diagrams augmented with a set of variables that depict the mechanism underlying the relevant differences between the target and the study population (8). The graphical rules that are used in DAGs to reason on internal validity can be adapted to understand in selection diagrams if and under what conditions an estimate can be transported from a study to a target population. The focus is on the presence of open paths between the S-variables and Y. Specifically, an estimate is directly transportable using recalibration techniques if there are no open paths from S to Y after removing all variables pointing towards the exposure A and conditioning on Z. According to this rule the causal effect is transportable in **Figure 3**.

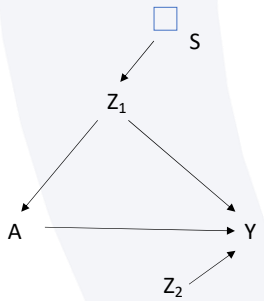

Figure 3. Selection diagram of a transportable causal A-Y estimate. S represents the difference between the target and the study population. Z_1 and Z_2 are measured.

To show the applicability of one selected method for transportability, we used data from the Piedmont Birth Register (PBR) (the target population) and the NINFEA study (the study population). Compulsory computerized birth registration was established in the whole of

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Italy in 2001 and in the Piedmont area it is of particularly high quality and completeness. We used PBR data of 2019 including 27852 pregnancy records. The PBR holds information on maternal and child/delivery characteristics. Specifically, data on parents' age, educational level and occupation were available, as well as maternal smoking, alcohol consumption, weight gain, and intake of folic acid during pregnancy. Information on the reproductive history of the mother (i.e. parity, previous miscarriages and use of infertility treatment) and information on reproductive outcomes (i.e. type, gestational age, birth size) is also recorded in the PBR. NINFEA is a web-based birth cohort with the aim of investigating the effects of early-life exposures on the health of newborns, children, adolescents, and adults. Cohort members are children of mothers recruited between 2005 and 2016 in Italy who completed a first online questionnaire at any time during their pregnancy on general health and exposures before and during pregnancy. Further follow-up information is obtained with repeated questionnaires when their child turns 6 months, 18 months, 4, 7, 10 and 13 years of age. The NINFEA cohort included 4052 (singleton) pregnancies after exclusion of births occurring outside Piedmont.

The aim of the applied example is to transport the risk difference in the prevalence of the outcome of interest between exposed and unexposed subjects among the PBR using data available from the NINFEA cohort. In order to guarantee the internal validity of the causal estimate in the NINFEA cohort, we simulated the exposure and the outcome, both binary, under different scenarios, and we kept the covariates' distribution as observed in the two populations. We selected three common confounders/effect modifiers of the exposure-outcome association, the maternal age, parity and education at the beginning of pregnancy. As opposed to the PBR, members of the NINFEA cohort study originate from a selected sample of the source population, with participation strongly associated with socioeconomic factors, such as a high educational level. As reported in **Table 3**, the participants of NINFEA were more highly educated compared to PBR (64% vs 30%), more likely to be primiparous (74% vs 50%), and slightly older.

Table 3. Observed distribution of covariates in the NINFEA cohort and the Piedmont Birth Register (PBR)

	NINFEA (N=4052)	PBR (N=26909)
Maternal age *(median, IQR)*	33(30-36)	32(28-36)
Parity *(N, %)*		
0	2994 (73.9)	13332 (49.5)
1	868 (21.4)	9869 (36.7)
2+	198 (4.9)	3708 (13.8)
Maternal education *(N, %)*		
High	2614 (64.5)	8124 (30.2)
Medium	1267(31.3)	11989 (44.5)
Low	171(4.2)	6796 (25.2)

We used the targeted maximum likelihood estimator (TMLE) to transport the average causal effect of the exposure on the outcome from the NINFEA cohort to the PBR population (9). It is a semiparametric double/multiple robust method that improves the chances of correct model specification by allowing for flexible estimation using machine-learning algorithms. The risk difference in the NINFEA cohort was 14.0% (95% CI: 11.5%-17.1%). When

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

transported to the PBR population, the risk difference was 16.5% (95% CI: 11.9%; 21.0%). In our setting, we could calculate the estimate directly in the PBR population. The risk difference was 16.6% (95% CI: 15.5%-17.7%), i.e. very close to the estimate using the transporting method.

In summary, reasoning on transportability allows to clearly formulate hypotheses on the generalizability of the results obtained from a study sample. Available methods further offer the potential to transport the estimates to contemporary populations. Applying these methods while transparently listing the assumptions made will be invaluable for health policy decision-makers using results from life-course epidemiology research.

3.4 Regression discontinuity (10)

Regression discontinuity design (RDD) is a quasi-experimental approach, applied in circumstances where an exogenous source of variation arises from a continuously measured assignment variable with a clearly defined cut-off point above or below which the population is at least partially assigned to a treatment or exposure. The assignment variable creates a discontinuity in the probability of exposure at the threshold, where the direction and magnitude of the jump is a direct measure of the causal effect of the exposure on the outcome for subjects near the cut-point. When the assignment rule perfectly determines the exposure (from 0 to 1 at the cut-off) regression discontinuity takes a sharp design. This means that the exposure assignment and the actual exposure status coincide (e.g., imposed policy measures). If the assignment rule affects the probability of exposure creating a discontinuous change at the threshold, without an extreme 0 to 1 jump, regression discontinuity takes a fuzzy design (e.g., optional participation in programs). Given the peculiarities of the design it has been widely applied in the context of natural experiments and it has become extremely popular in the econometrics, educational and social research exploiting the threshold rules often used by educational institutions, public and private insurance schemes, governmental welfare programs and social policies. Although in many of these settings the main outcomes of interest were health outcomes, RDD is still rarely used in health and epidemiology research.
We provide an overview of the RDD, highlighting its underlying assumptions, approaches for testing the assumptions and the validity of the design, its advantages and limitations. Given a growing number of birth cohorts and multiple birth cohort consortia being established to study the effects of early life exposures on later health outcomes, we use and update a recent systematic review on the application of RDD in healthcare research, focusing on the assignment variables and the interventions/exposures that have been investigated and that could serve as design models in the context of perinatal epidemiology and birth cohort research (11). The RDD is depicted using a DAG in **Figure 4**.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

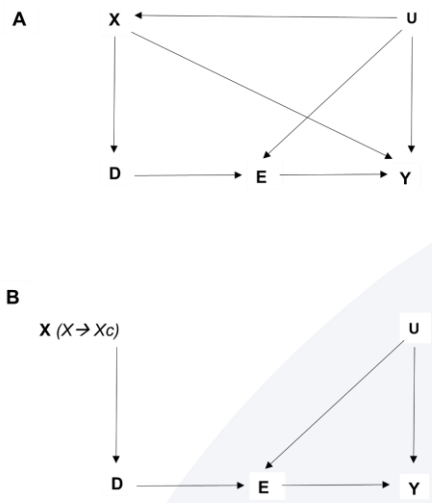

Figure 4. Causal directed acyclic graph of the regression discontinuity design. X represents the assignment variable which cut-off value c determines the eligibility criteria (D) for an exposure / treatment (E). Y denotes an outcome, while U represents measured and unmeasured confounding factors. *Panel A:* Directed acyclic graph representing a causal model underlying a regression discontinuity design. *Panel B:* Causal graph for the regression discontinuity design in the close vicinity around the assignment variable cut-off. Note that if the assignment rule is deterministic (sharp regression discontinuity design) D is equal to E.

The main conditions and assumptions of RDD and information on whether these are testable empirically in the data are summarized in **Table 4**. Estimating the causal effect under the fuzzy design requires additional assumptions, including (i) the continuity of the potential outcomes at the cut-off, (ii) the monotonicity of the treatment selection response to the assignment variable at the cut-off, and (iii) the local exclusion restriction assumption.

Table 4. The main conditions and assumptions of the Regression Discontinuity Design (RDD)

	Description	Empirical testing
Assignment rule condition	A continuous pre-exposure variable with a clearly defined cut-off value for the exposure assignment	Yes
	The same cut-off value is not used to assign the individuals to other exposures	No (theoretical only)
	Lack of discontinuities other than the one at the cut-off	Yes
Lack of manipulation in the assignment variable	The cut-off value is exogenous – unrelated to the individuals' value of the assignment variable, and the individuals' assignment variable values are not determined by the cut-off of the assignment variable	Yes (indirectly)
Exchangeability around the assignment variable cut-off	Similarity of the individuals close to the cut-off in the observed characteristics	Yes
	Similarity of the individuals close to the cut-off in the unobserved characteristics	No
	The outcome probability is continuous at the cut-off in the absence of exposure	Yes

The key step in RDD is the identification of relevant assignment variables that meet the criteria and assumptions reported in **Table 4**. **Table 5** lists examples from the literature that are of relevance in the context of birth cohort research. Advantages and limitation of the RDD approach are summarized in **Figure 5**.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Table 5. Some of the assignment variables and exposures as possible Regression Discontinuity Design (RDD) models in perinatal epidemiology and birth cohort research

Assignment variable	Determined cut-off values	Possible exposures / interventions
Birth weight	Low-birth weight (<2500 grams) Very low birth weight (<1500 grams) Extremely low birth weight (<1000 grams) High birth weight (>4000 grams)	Extra neonatal care Neonatal intensive care unit Rooming-in and mother-child bonding Breastfeeding Caesarean section Treatments (e.g., probiotic supplementation, surfactant therapy) Health insurance and supplemental benefits
Gestational age	Preterm (<37 gestational weeks) Very preterm (<32 gestational weeks) Extremely preterm (<28 gestational weeks)	
Maternal age at conception	<18 years >35 years >40 years	Minimum cigarette/alcohol purchase age Screening and procedures for high-risk pregnancies
Socioeconomic measures (e.g., family income)	Setting-specific	Social, welfare, and cash transfer programs Health insurance policies
Parity	Setting-specific	
Age or date/year of birth	Setting-specific	Introduction of: a. Vaccination campaigns b. Pregnancy-specific guidelines c. Maternity/paternity leave policies d. Child-support grants e. Social and welfare programs
Clinical measures (e.g., blood pressure, blood glucose levels)	Setting-specific	Treatment initiation
Environmental measures (e.g., air-pollution levels)	Setting-specific	Local interventions (e.g., to reduce air-pollution)

Advantages

- Relatively weak and testable assumptions
- Strong internal validity
- Intuitive interpretation
- Transparent and simple graphical representation

Limitations

- Low statistical power
- Limited external validity (geographic- and time-specific settings)
- The estimated causal effect is local
- Contamination by other exposures/interventions
- Rarity of settings outside of policy and program evaluations

Figure 5. Summary of Regression Discontinuity Design (RDD) advantages and limitations.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

3.5 Triangulation of different methods with different sources of bias to improve causal inference (12)

Over the last 5 years, we and others have encouraged the use of triangulation to improve causal understanding in epidemiology, in order to identify better targets for intervention development to prevent and treat disease. The idea behind triangulation is to integrate results from several different approaches (i.e. different study designs, analytical methods and data sources) where each approach has different key sources of potential bias that are unrelated to each other and ideally would bias findings in different directions (13). If the results of different approaches all point to the same conclusion, this strengthens confidence in that conclusion. When results from different approaches are inconsistent, a thorough and specific assessment of risk of bias of each approach should help to determine what subsequent research is needed to obtain a more robust causal answer. Triangulation is increasingly invoked in epidemiological research but even in our own work, the term seems to refer to a wide-range of very different materials and methods, from the comparison of results from two or three methods applied to the same data at one end of a spectrum (we refer to this as 'Internal triangulation) to a much more systematic approach that aims to integrate all currently available evidence ('External validation'). The former is more common, and indeed has been done for decades, often with no declaration that this is triangulation. Given this diverse use of the term 'triangulation', the aim of this paper is to provide a framework for triangulation that could be used across the spectrum. In section 1 we define different types of triangulations and their strengths and limitations. We emphasise the importance of defining the type that is planned, acknowledging that there is a spectrum from the focused internal to focused external triangulation. For example, in a previous study of the effect of maternal circulating 25-Hydroxy Vitamin D and calcium on offspring birth weight, we combined Mendelian randomization analyses in data that we had access to, with a search for randomised controlled trial evidence of maternal pregnancy supplementation for these exposures. Ideally one might combine internal and external triangulation. Section 2 summarises steps in the triangulation process, emphasising the importance of publishing protocols/analysis plans prior to starting data analyses, as we have increasingly done. In this section we emphasise the importance of blind specific risk of bias assessment; blind meaning that the assessment is done before analyses and sharing of results and specific, for example by agreeing and specifying sources of selection bias and specific confounders, rather than defining these based on available data or declaring that residual bias has to be present in any observational study. We provide references/links to existing risk of bias tools for different methods, including Mendelian randomization and modifications of tools for some methods (e.g. within sibship and negative control analyses). In Section 3, the final section, we describe potential future plans, including methods that could be developed for obtaining a quantified causal effect estimate from very diverse design/analytical approaches and for including the potential effects of bias in this quantification and the use of Multi parameter evidence synthesis (MPES) in external triangulation. We use published and some novel examples to illustrate the use of the framework. **Table 6** summarises the strengths and limitations of internal and external triangulation.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Table 6. Relative strengths and limitations of internal and external triangulation

Internal triangulation (applying different approaches to data the researchers have access to)	External triangulation (systematically searching for and integrating all available studies that are relevant to the research question)
Strengths • Efficient (relatively quick and easy) • More robust than relying on just one approach	**Strengths** • Most robust approach • Little cherry picking of which methods/approaches to include
Limitations / challenges • May miss key information that would change conclusions • Given data is known to researchers the approaches that are applied might be influenced by knowledge of the data	**Limitations / challenges** • Likely to be very time consuming • As the search is planned to be very extensive the time needed to review, extract data and undertake analyses may mean that several key new studies have been published by the time this is completed. This cycling could continue for ever. *Open triangulation* could handle this. • The more studies and approaches included the more likely there are to be inconsistencies (even when taking account of risk of bias)

Table 7 summarises key steps in the process for triangulation of an internal and external triangulation. These can be adapted for types that fall between the definite internal and definite (full) external triangulation types. Whilst the steps are linear in time, as will all research, the process is often iterative. For example, applying specific risk of bias assessments to approaches that will be used with existing data (internal) or to papers from systematic searches (external) may reveal additional sources of bias that were not initially thought of when preparing the initial template for reporting this information. Comments from co-authors and reviewers may lead to changes to analysis plans. These and other reasons will mean going back and changing earlier steps. This highlights the importance of publishing the protocol/analysis plan and adding to that any subsequent changes with their justification, thus avoiding unrevealed data driven methods.

Table 7. Steps in triangulation

Internal Triangulation	External Triangulation
Define specific causal question	Define specific causal question
Agree different methods that could be triangulated in available data	Develop systematic search strategy and complete initial search to identify the types of approaches available to address
Protocol for blind risk of bias assessment*: 1. Agreeing **specific** sources of bias for each method 2. Preparing specific risk of bias reporting template. This should include recording the causal question of each method and extent to which that addresses your causal question	Protocol for blind risk of bias assessment*: 3. Agreeing **specific** sources of bias for each method identified in the search 4. Preparing specific risk of bias reporting template for these methods. This should include recording the causal question of each method and extent to which that addresses your causal question.
Agreeing and writing analysis plan, including sensitivity analyses that might be used to explore/mitigate bias.	Agreeing and writing data extraction tool and who will extract data. Agreeing and writing plans for pooling and integrating data from different approaches

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Combining all of the above steps in a protocol and publishing that protocol.	Combining all of the above steps in a protocol and publishing that protocol.
Analysis of data as planned.	Update search, extract data, pool and integrate results as planned.
Interpretation of results and write up of paper	Interpretation of results and write up of paper

* This may include modification of some existing risk of bias tools and/or the need to develop new tools, where there are methods that are not (fully) covered by existing tools.

3.6 Additional tutorial on mediation analysis in DataShield

The recently developed DataSHIELD package dsMediation allows federated mediation analysis across multiple studies in a privacy-preserving way. The *dsMediation* package includes functionalities for mediation analysis involving a single mediator or multiple mediators. It alows four different approaches: the regression-based, the simulation-based, the weighting-based, and the imputation-based approach. Hence it includes nine functions mainly derived from three R packages: *regmedint*, *mediation*, and *medflex*. These functions provide an output for each individual study participating in the analysis and the analysts can meta-analyse the results using the native R packages, such as metaphor or meta. In the tutorial, the functions are applied to real data. Several scenarios are considered with different types of exposure, mediators and outcome (binary, categorical or continuous), and including or not including the interaction between the exposure and the mediators. A manuscript on this topic is in preparation.

4. Conclusions

We have provided in this deliverable a detailed summary of the five tutorials / educational papers that have been prepared within the context of LifeCycle to demonstrate the implementation in specific contexts of causal inference approaches and longitudinal data modelling. The ultimate aim of this work is to facilitate the adoption and use of these methods within and outside LifeCycle and to promote discussion on their advantages and limitations.

5. Contribution of partners

- **UNITO**: led the Task, led tutorials 2.2.1, 2.2.3, 2.2.4, and 2.2.6
- **UNIVBRIS**: led tutorials 2.2.2 and 2.2.5
- **All partners**: contributed to the selection of the topics for the tutorials

6. Deviations from the original plan

This deliverable has been fulfilled fully in line with the original plan as stated in the grant agreement.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

7. Dissemination activities

To reach the research community outside LifeCylce, four tutorials have been published as a journal article (combining longitudinal data from different cohorts) or archived preprints (mediation analysis, regression discontinuity, transportability). The main concepts of the fifth tutorial (triangulation) are summarised in a publicly available slide presentation. The tutorial on mediation analysis is under review in an international journal; the other tutorials will be soon submitted for publication.

The conceptual frameworks and preliminary results of the tutorials were discussed at the General Assembly LifeCycle meetings: Barcelona 2018, life course trajectories; Copenhagen 2019, mediation analysis; Bristol 2019, brief update on mediation analysis, regression discontinuity, transportability; Remote October 2020, life course trajectories, transportability; Remote May 2021, regression discontinuity; Remote October 2021, transportability; Paris 2022, triangulation.

The main reference articles and educational material identified during the ongoing work on the tutorials were added to the LifeCycle method database (https://lifecycle.cpo.it) publicly available online through the Eu Child Cohort catalogue (see Deliverables 7.1 and 7.2). In the database, the LifeCycle tutorials and method review articles are searchable using a specific filter labelled “LifeCycle tutorials”, which currently lists 7 publicly available products:

- Santos S, Zugna D, Pizzi C, Richiardi L. Sources of confounding in life course epidemiology. J Dev Orig Health Dis 2019;10:299-305.
- Santos S, Maitre L, Warembourg C, Agier L, Richiardi L, Basagaña X, Vrijheid M. Applying the exposome concept in birth cohort research: a review of statistical approaches. Eur J Epidemiol. 2020;35:193-204.
- Hughes RA, Tilling K, Lawlor DA. Combining longitudinal data from different cohorts to examine the life-course trajectory. Am J Epidemiol 2021;190:2680-2689.
- Zugna D, Popovic M, Fasanelli F, Heude B, Scelo G, Richiardi L. Applied causal inference methods for sequential mediators, 18 May 2022, Preprint (Version 3) available at Research Square [https://doi.org/10.21203/rs.3.rs-965331/v3]
- Popovic M, Zugna D, Richiardi L. Regression discontinuity design in perinatal epidemiology and birth cohort research. Preprint available at arXiv [https://arxiv.org/abs/2208.110472208.11047]
- Scelo G, Zugna D, Richiardi L. Transporting results in an observational epidemiology setting: purposes, methods, and applied example, 24 August 2022, Preprint available at Research Square [https://doi.org/10.21203/rs.3.rs-1987165/v1]
- Lawlor D. Triangulation Tutorial. Slide presentation. 2022

The concepts, methods and/or examples described in the tutorials were also used to assist in the preparation of the following workshops (see Deliverable 7.5): (i) “Longitudinal modelling in the context of life-course studies”, Barcelona, 2018; (ii) “Mediation analysis”, Online, 2020 (videos available through the LifeCycle Youtube channel); (iii) “Transporting

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

estimates across populations: why, when, and how?", Online 2021 (videos available through the LifeCycle Youtube channel); (iv) "Mediation analysis with DataShield", Paris, May 2022. Analogously the discussions during the workshops contributed to the preparation of the tutorials.

The results of the tutorials were presented at the following congress presentations and seminars:

- Scelo G. Transportability of results in an observational study setting. XI National Congress of the Italian Society of Medical Statistics and Clinical Epidemiology (SISMEC). Remote, September 15th, 2021.
- Popovic M , Zugna D, Gagliardi L, Richiardi L. Regression discontinuity design in lifecourse epidemiology. XI National Congress of the Italian Society of Medical Statistics and Clinical Epidemiology (SISMEC). Remote, September 16th, 2021.
- Lawlor D. Triangulation in evidence synthesis. XLVI Congress of the Italian Epidemiology Association. Padova, June 30th, 2022.
- Lawlor D. Triangulation – from selective qualitative comparisons to systematic quantitative integration. Seminar series of the Piedmont Referring Center for Cancer Epidemiology and University of Turin, Turin, May 6th, 2022.
- Lawlor D. Triangulating epidemiological evidence: From focused qualitative comparison to systematic qualitative integration. Workshop on Triangulation of Evidence in Environmental Epidemiology organised by the US National Academy of Sciences Engineering and Medicine, May 9th, 2022.

8. References

1. Zugna D, Popovic M, Fasanelli F, Heude B, Scelo G, Richiardi L. Applied causal inference methods for sequential mediators. Research Square. 2021.
2. Tchetgen Tchetgen EJ. Inverse odds ratio-weighted estimation for causal mediation analysis. Stat Med. 2013 Nov 20;32(26):4567-80.
3. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. Epidemiol Methods. 2014 Jan;2(1):95-115.
4. Vansteelandt S, Bekaert M, Lange T. Imputation Strategies for the Estimation of Natural Direct and Indirect Effects. Epidemiol Methods. 2012; 1(1): 131-158.
5. Steen J, Loeys T, Moerkerke B, Vansteelandt S. Flexible Mediation Analysis With Multiple Mediators. Am J Epidemiol. 2017 Jul 15;186(2):184-193.
6. Hughes RA, Tilling K, Lawlor DA. Combining longitudinal data from different cohorts to examine the life-course trajectory. Am J Epidemiol. 2021; 190: 2680-9.
7. Scelo G, Zugna D, Richiardi L. Transporting results in an observational epidemiology setting: purposes, methods, and applied example. Research Square. 2022.
8. Pearl J, Bareinboim E. Transportability across studies: A formal approach. Technical Report R-372, Cognitive Systems Laboratory, Dept. Computer Science, UCLA Original, 2018. https://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf. Accessed Sept 2022.
9. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol. 2017 Jan 1;185(1):65-73.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

10. Popovic M, Zugna D, Scelo G, Tilling K, Richiardi L. Use of regression discontinuity to strengthen causal inference in life course epidemiology. arXiv. 2022.
11. Hilton Boon M, Craig P, Thomson H, Campbell M, Moore L. Regression Discontinuity Designs in Health: A Systematic Review. Epidemiology. 2021;32(1):87-93.
12. Lawlor DA, Davey Smith G, Munafo M, Higgins J, Tilling K. A framework for triangulation of evidence to improve causal inference. Advanced draft has been prepared; Slide presentation available through the LifeCycle method database at: https://lifecycle.cpo.it/life_cycle_methods/100.
13. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. Int J Epidemiol. 2016;45(6):1866-1886.

LIFECYCLE Project
Coordination and leadership:
Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands
Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).