Population DNA methylation studies in the DOHaD framework

Janine F. Felix\textsuperscript{1,2,3}, Charlotte A. M. Cecil\textsuperscript{1,2,4,5}

SELF-ARCHIVING VERSION

1. The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
2. Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
3. Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
4. Department of Child and Adolescent Psychiatry, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
5. Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, United Kingdom

Corresponding author: Janine F. Felix MD PhD, The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, the Netherlands, j.felix@erasusmc.nl, Ph.+31 10 70 43997.

ABSTRACT

Epigenetic changes represent a potential mechanism underlying associations of early-life exposures and later-life health outcomes. Population-based cohort studies starting in early life are an attractive framework to study the role of such changes. DNA methylation is the most studied epigenetic mechanism in population research. We discuss the application of DNA methylation in early-life population studies, some recent findings, key challenges and recommendations for future research. Studies into DNA methylation within the DOHaD framework generally either explore associations between prenatal exposures and offspring DNA methylation or associations between offspring DNA methylation in early life and later health outcomes. Only few studies to date have integrated prospective exposure, epigenetic and phenotypic data in order to explicitly test the role of DNA methylation as a potential biological mediator of environmental effects on health outcomes. Population epigenetics is an emerging field which has challenges in terms of methodology and interpretation of the data. Key challenges include tissue specificity, cell type adjustment, issues of power and comparability of findings, genetic influences, and exploring causality and functional consequences. Ongoing studies are working on addressing these issues. Large collaborative efforts of prospective cohorts are emerging, with clear benefits in terms of optimizing power and use of resources, and in advancing methodology. In the future, multidisciplinary approaches, within and beyond longitudinal birth and preconception cohorts will advance this complex, but highly promising, field of research.

Keywords: Epigenetics; Cohort Studies; DNA-methylation; Epidemiology, Molecular
**Introduction**

In the Developmental Origins of Health and Disease framework, many associations between the early-life environment (e.g. prenatal exposure to maternal obesity, smoking and suboptimal diet) and later health outcomes have been described. However, the underlying molecular mechanisms are often not known. Epigenetic changes have recently emerged as a potential mechanism that may at least partially explain these associations.

**What is epigenetics**

The ‘epigenome’ refers to a collection of molecular mechanisms, which do not affect the DNA sequence, but can affect gene regulation by changing the structure and accessibility of the DNA. Three main types of epigenetic mechanisms can be distinguished: DNA methylation, histone modifications and non-coding RNAs. These mechanisms are not independent, but rather function together as part of the larger, complex epigenetic machinery. Here, we focus specifically on DNA methylation, as this is currently the most studied epigenetic mark in large population studies. DNA methylation refers to the binding of a methyl group to the DNA, mainly at positions where a cytosine is located next to a guanine, a cytosine-phosphate-guanine (CpG) site. Typically (but not always), higher levels of DNA methylation at CpG sites in promoter or enhancer regions of a gene impede transcription factors from binding to the DNA, thereby inhibiting gene expression. On the contrary, methylation of CpG sites in a gene body generally leads to increased expression. Because DNA methylation patterns can be passed on mitotically during cell division, they can lead to long-term alterations in gene activity and related phenotypes. At the same time, DNA methylation patterns can also show a considerable degree of flexibility over time, enabling cells to respond to changing internal and external inputs.

**DNA methylation within the DOHaD framework**

Interest in DNA methylation within the DOHaD framework stems from the fact that (a) it has been shown to respond to both genetic and environmental influences, beginning in utero, and (b) it plays a fundamental role in development. Consequently, DNA methylation shows promise as a mechanism for understanding how environmental and genetic influences shape trajectories of health and disease across the lifespan.

Early life, in particular, may be a critical period for epigenetic effects on health outcomes. Extensive reprogramming of DNA methylation patterns takes place in utero. Thus, aberrant modifications that occur at this early stage, before tissue differentiation, could result in alterations in gene expression across many tissue types. In addition, environmental and genetic effects on organ maturation during critical periods in pregnancy and postnatally may be mediated through the epigenome. Evidence for this has come primarily from animal experimental studies, which enable careful manipulation of environmental exposures. In humans, an early example is the work performed in the Dutch Famine Study, which has shown that exposure to famine in the first trimester of pregnancy, but not the last, is associated with DNA methylation changes at the imprinted IGF2 gene, a key modulator of fetal development. Timing-specific effects were later found for further loci, some of which also associated with cardiometabolic phenotypes, such as birth weight and LDL cholesterol levels. Together, this work provided preliminary evidence that (very) early life may be a critical period for epigenetically-mediated effects of environmental exposures on health outcomes. Also, twin studies have been
instrumental in disentangling the relative contributions of genetic, environmental and stochastic factors to DNA methylation and phenotypic variability\textsuperscript{11,17}.

Population-based cohort studies starting in early life offer an attractive platform to build on this early work and characterize associations between early-life exposures, epigenetic mechanisms and later-life health. Below, we discuss the application of DNA methylation in early-life population research, including some recent findings, key challenges and recommendations for future research.

**DNA methylation in early-life population-based cohort studies**

In population-based studies, there is increasing interest in the role of DNA methylation changes in observational associations of adverse exposures in the preconception period, during pregnancy and in early infancy with later-life common diseases. Through the development of high-throughput cost-effective arrays which analyze hundreds of thousands of DNA methylation markers across the genome, it has become possible to measure DNA methylation in (relatively) large samples. Levels of DNA methylation at these markers are then associated with determinants and outcomes of health and disease in “epigenome-wide association studies” or EWASs. In addition, many long-running population-based studies have longitudinal collection of data and biological samples in extensive biobanks. This makes it possible to study DNA methylation over time in relation to repeated measurements of exposures and outcomes and to test key mediational hypotheses.

So far, population-based research examining epigenetics within the DOHaD framework has fallen mainly into two categories – (i) studies investigating associations between prenatal exposures and offspring DNA methylation patterns, e.g. dietary, environmental, and psychosocial exposures (Fig 1a: a path); and (ii) studies investigating associations between offspring DNA methylation patterns in early life and later health outcomes, e.g. neurodevelopment or cardiometabolic health (Fig 1a: b path). In both cases, DNA methylation has typically been examined at birth, either in cord blood or in placental tissue, given its role as a barrier and interface between maternal and fetal environments.

**Early-life exposures and DNA methylation**

To date, the prenatal exposure that has been most robustly linked to DNA methylation changes in the offspring is maternal smoking\textsuperscript{18-23}. In a recent large study based on data from 13 cohorts (n = 6,685), Joubert \textit{et al} found that sustained smoking during pregnancy associated with altered DNA methylation patterns across thousands of CpG sites, an effect that was still observable – albeit to a lesser extent – in older children, although the sample size was around half that of the newborn analysis\textsuperscript{18}. This study’s top finding was a CpG in \textit{AHRR}, which has been associated with both personal and maternal smoking in multiple other studies\textsuperscript{19,24,25}. Pathway analyses found neurological and developmental pathways, which is in line with the maternal smoking-related child outcomes. Alterations in DNA methylation at birth have also been documented in relation to other chemicals, such as prenatal exposure to arsenic\textsuperscript{26}, mercury\textsuperscript{27} and air pollution\textsuperscript{28,29}.

Maternal pre-pregnancy body mass index (BMI) has also been associated with offspring DNA methylation, with findings showing associations with both lower and higher maternal BMI in a recent meta-analysis of 9340 newborns\textsuperscript{30}. Effects were generally small, with a median increase of 0.03% in methylation level per unit increase in maternal BMI, and strong evidence for a direct causal intra-uterine
effect was lacking for most loci. Furthermore, dietary factors have been examined, with a particular focus on folate intake, due to its established role in both epigenetic regulation and fetal brain development. The largest of these studies, based on data from two birth-cohorts (n = 1,988), reported associations between maternal plasma folate levels during pregnancy and differential DNA methylation in cord blood across multiple genes involved in folate biology and neurodevelopmental processes. Interestingly, this study found reduced methylation with higher levels of folate, a methyl donor, at the vast majority of CpG sites, a finding which underlines the complex associations of the one-carbon metabolism pathway with methylation. In contrast, other known risk factors for poorer offspring health have not been found to be strongly associated with neonatal DNA methylation. For example, a recent meta-analysis on maternal alcohol use found no evidence for an association with cord blood DNA methylation. Similarly, a meta-analysis of prenatal maternal stress did not identify associations with DNA methylation in cord blood, although an enrichment for genes related to methyltransferase activity was identified. Besides a true null effect, the lack of a strong signal in these studies may reflect low levels of more severe forms of exposure in the general population, relatively small sample sizes, resulting in decreased power to detect effects, or heterogeneity between the included studies in terms of population, exposure definition or confounding structure.

**DNA methylation and childhood health outcomes**

Compared to prenatal exposures, fewer population-based studies have examined the relationship between DNA methylation patterns and childhood health outcomes. We highlight here two important areas of progress in the field of DNA methylation in population-based studies: neurodevelopment and adiposity.

For neurodevelopmental outcomes, most of the evidence to date on this topic has been drawn from cross-sectional data in smaller, selected samples, e.g. high-risk/clinical samples, although efforts to characterize these associations in the wider population are fast increasing. One such study, based on data from 537 newborns, found that epigenetic patterning of glucocorticoid genes in the placenta prospectively associated with infant neurodevelopmental profiles. Specifically, infant membership to a reactive, poorly regulated profile associated with increased placental methylation of the NR3C1 glucocorticoid receptor gene, and decreased methylation of the HSD11B2 gene involved in limiting fetal exposure to maternal circulating cortisol. Other population studies have examined DNA methylation in relation to risk for neurodevelopmental disorders – particularly ADHD, but also other psychiatric problems with early developmental origins, such as oppositional defiant and conduct problem behavior. For example, a longitudinal epigenome-wide study based on over 800 children found that DNA methylation patterns at birth (but not at age 7) prospectively associated with trajectories of ADHD symptoms from childhood to adolescence. Of the 13 genome-wide significant loci identified at birth, several were annotated to genes involved in fatty acid metabolism and neural tube development. Importantly, one of the loci was annotated to a gene (ST4GAL3) that has been recently identified as the top genetic hit for ADHD in the largest GWAS to date. Studies are also beginning to integrate brain imaging data to investigate the links between DNA methylation, brain development and neuropsychiatric outcomes. For example, a recent EWAS study found that SP6 methylation at birth prospectively associated with lower amygdala-hippocampal volume and higher psychotic symptoms in young adulthood – an association that was replicated in an independent sample of patients with schizophrenia.

Most of the evidence regarding DNA methylation and adiposity stems from adult studies. The largest of these identified and replicated cross-sectional associations between DNA methylation at 187 loci and BMI in over 10,000 samples, showing that these are mostly a consequence rather than a cause of population epigenetics and DOHaD
adiposity\textsuperscript{46}. Less is known about DNA methylation and childhood adiposity. Several studies have described associations of DNA methylation with birth weight, with a limited number of CpG sites identified in more than one study\textsuperscript{47,51}. No strong evidence was found for persistence of associations into later life in the two studies that explored longitudinal relationships\textsuperscript{47,51}. In targeted analyses, a replicable association of RXRA methylation in umbilical cord tissue with childhood fat mass was shown\textsuperscript{52}. Several differentially methylated regions in newborn blood, such as the promoter of the long non-coding RNA ANRIL, have been associated with childhood adiposity measures based on epigenome-wide scans\textsuperscript{53,54}. In cross-sectional analyses in children and adolescents, DNA methylation in several regions has been associated with adiposity measures, including the SOCS3 locus\textsuperscript{55,57}. It remains unclear whether DNA methylation patterns are a risk factor for or consequence of adiposity. Detailed associations in the large adult EWAS show that changes in DNA methylation are mostly a consequence of obesity\textsuperscript{46}. This is supported by smaller studies in children and individuals before and after bariatric surgery\textsuperscript{58-61}.

**DNA methylation as a mediator**

Despite promising findings, very few population-based studies to date have integrated prospective exposure, epigenetic and phenotypic data in order to explicitly test the role of DNA methylation as a potential biological mediator of environmental effects on health outcomes. Notable exceptions include an EWAS study based on 321 cord blood samples, which showed that, in males, DNA methylation of the PON1 gene at birth associated with prenatal mercury levels and predicted cognitive performance scores during childhood\textsuperscript{27}. Using repeated measures of DNA methylation, the authors found that this effect persisted in early childhood (3-5 years) and was attenuated in mid-childhood (6-10 years). Furthermore, DNA methylation of the PON1 locus at birth was shown to associate with gene expression levels in an independent set of cord blood samples. Another exception is a longitudinal EWAS study examining epigenetic correlates of adolescent substance use (tobacco, alcohol and cannabis use; n = 244; age 14–18). This study found that neonatal DNA methylation levels in a tightly interconnected network of genes enriched for neurodevelopmental processes (n = 65 epigenome-wide corrected loci) partially mediated the effect of prenatal tobacco smoking on later substance use risk in adolescence\textsuperscript{62}. Relatedly, maternal prenatal smoking has been found by other studies to associate with lower birth weight partially through alterations in cord blood DNA methylation using both targeted and epigenome-wide approaches\textsuperscript{24,63}.

**Current challenges and future directions**

The field of population epigenetics is still in its infancy. Although great strides have already been taken, there is still much to discover and learn. Here we present some of the key challenges that current research in this area is facing, and how these may be addressed.

**Tissue specificity**

DNA methylation patterns are largely tissue-specific. In large cohort studies, blood is typically the main source of DNA collected. In some cases there may be DNA from buccal swabs, umbilical cord or placental tissue, but other tissues are generally inaccessible for population researchers. This poses challenges, as the extent to which findings may be generalized to the target tissue of interest is often unclear\textsuperscript{64,65}. One way to address this limitation is to refer to publicly available datasets that show cross-tissue concordance in DNA methylation patterns, based on selected samples (e.g. comparison of blood
and brain in postmortem samples or blood and adipose tissue in patients undergoing surgery), for example: [http://epigenetics.iop.kcl.ac.uk/bloodbrain](http://epigenetics.iop.kcl.ac.uk/bloodbrain) and [https://redgar598.shinyapps.io/BECOn](https://redgar598.shinyapps.io/BECOn). However, it is not just the concordance in DNA methylation levels between tissues, but also, and maybe more so, the variability of those DNA methylation levels in association with the exposure or outcome of interest that matters. Concordance between tissues in associations with specific phenotypes has been shown, for example for BMI between blood, adipose and liver tissue, but overall evidence is limited 46,64,69. The NIH Roadmap Epigenomics Project can be used as a reference for epigenetic marks in a large number of tissues ([www.roadmapepigenomics.org](http://www.roadmapepigenomics.org))70.

**Cell type adjustment**

DNA methylation patterns vary not only between, but also within tissues. In blood samples, DNA is collected from leukocytes, which represent a heterogeneous population characterized by cell-type specific DNA methylation. As such, the relative proportions of cell types from which the DNA is extracted may influence epigenetic results. Ideally, detailed measurements of leukocyte subtypes would be available for correction in all study participants, but this is often not feasible. As an alternative, the use of reference-based approaches for estimating cell type proportions has become popular, such as methods developed for whole blood71,72 and, more recently, cord blood73-75. Reference-free methods are also available76-78. These methods, however, are still under development and population-based studies with specific cell type measurements can play a pivotal role in developing and validating new analytic approaches.

**Methodological considerations**

DNA methylation data are multifactorial, high dimensional and inter-correlated, raising questions about how best they should be analyzed79. So far, studies have varied widely in study characteristics, methodology and analytic routines, limiting comparability of findings. Furthermore, many studies have been underpowered to identify effects, especially when examining exposures or outcomes with low prevalence in the general population. To address these challenges, large collaborative efforts of prospective cohorts, such as the Pregnancy And Childhood Epigenetics (PACE) Consortium, have emerged, similar to the field of genome-wide association studies although numbers are still much smaller80. Such collaborations have shown enormous potential in increasing statistical power and optimizing the use of resources81-85, as well as allowing replication and comparison of effects between studies with potentially heterogeneous confounding structures, thus strengthening the evidence for reported associations. In addition, they are a valuable platform for standardization of study design and analysis protocols to maximize comparability and help to delineate best practices in the field. Multidisciplinary teams, bringing together basic, population and clinical scientists, will enable easier exchange of knowledge and ideas, translation between bench, population and bedside, and strengthening of scientific conclusions by bringing together multiple lines of evidence.

**Role of genetics**

It has been reported that a substantial proportion of the variation in DNA methylation between individuals is due to genetic variants, acting in cis or trans (mQTLs)86-88. As such, measured DNA methylation levels may reflect genotype and mediate associations of genetic variants and health outcomes, or, if genotype drives both DNA methylation and the outcome under study, changes in DNA methylation may be an epiphenomenon of the genetic risk rather than part of the causal pathway. Furthermore, genetic factors...
may moderate associations between environmental exposures, DNA methylation and outcomes (Fig 1a). Thus, genetic variability should always be considered when examining DNA methylation – ideally directly, or if unavailable or underpowered, by referring to online mQTL databases (e.g. www.mqtlldb.org).

**Causal inference**

DNA methylation is a dynamic process that may change under the influence of genetic and environmental factors, but also with disease states. Thus, DNA methylation may be a cause, a consequence, or an epiphenomenon of the phenotype under study (Fig 1b). This issue of reverse causality is particularly problematic when using cross-sectional designs. While the use of prospective, longitudinal data partially addresses directionality - especially when DNA methylation is measured before the onset of disease - observational studies in general are subject to confounding. The application of advanced causal inference methods, such as two-step Mendelian randomization offers a valuable tool for probing causal pathways. For example, a recent study applied this method to support a small, causal effect of prenatal maternal B<sub>12</sub> levels on childhood IQ levels, via DNA methylation at birth. The tissue specificity of DNA methylation is an additional challenge to causal inference analyses.

Determining whether DNA methylation plays a causal role in disease is important for understanding the etiology of the disease as well as potentially developing new therapies in future. Yet, even if DNA methylation changes are not causal to a disease, but rather a consequence or an epiphenomenon, they may still serve as biomarkers of the disease or its progression (Fig 1b). In addition, if an association of DNA methylation with a phenotype is non-causal, the DNA methylation level may be used as a proxy for a specific exposure. This may be especially helpful for exposures that are difficult to measure (reliably). Previous studies have shown the potential of using DNA methylation as a biomarker in a number of these instances, including for alcohol consumption, maternal smoking during pregnancy and gestational age.

**Functional characterization**

The buck does not stop at DNA methylation. To understand pathways leading to disease, we need to study the functional effects in terms of gene expression as well as the roles of other epigenetic marks. Effect sizes in EWAS are often small, for example in an EWAS of maternal BMI, the largest effect estimates for the significantly associated CpG sites were an 0.1% decrease or increase in methylation level per unit increase in maternal BMI (kg/m<sup>2</sup>). For a strong exposure such as sustained maternal smoking, the maximum effect size was a 10% decrease in methylation levels for sustained smokers compared to non-smokers. It is currently unclear to what extent statistical significance parallels biological meaningfulness. In future, studies that incorporate both DNA methylation and expression data, preferably within the same population and at the same time point, can shed more light on this question of functionality. Such data are not widely available in longitudinal birth cohorts yet, with some noticeable exceptions (INMA Study, Gambia Study). In addition, the ENCODE database can be used to explore whether identified DNA methylation sites are located in potential regulatory regions. Intensifying collaborations and expanding them to researchers in neighboring fields, such as clinical and basic researchers, will make it easier to build chains of evidence across species, across research fields and across the life course, leading to strong and innovative research.
**Starting at the very beginning**

Most early-life cohorts begin in pregnancy and thus generally do not have prospective and hands-on data about the preconception and early pregnancy periods. Some preconception cohorts do exist (Southampton Women’s Survey: [www.mrc.soton.ac.uk/sws](http://www.mrc.soton.ac.uk/sws), Generation R Next Study: [www.generationr.nl/next](http://www.generationr.nl/next)) and they, together with longitudinal birth cohorts with follow-up long enough to include the next generation, will provide valuable information on exposures in these very early life stages and their epigenetic associations. Furthermore, the vast majority of epigenetic research in birth cohorts is currently focused on maternal exposures during pregnancy. In contrast, the contribution of paternal epigenetics and exposures is currently understudied, but has the potential to lend important new insights in the development of the offspring epigenome and the pathways leading from early life exposures to child and adolescent health.

**Conclusion**

Population epigenetics is an emerging field with strong potential to shed light on mechanisms underlying associations of early-life adverse exposures and later life health. Yet, the field is still under development and poses challenges in terms of methodology and interpretation of the data. The use of multidisciplinary approaches, combining various “omics” studies linking epigenetic changes to functional readouts and phenotypes, within and beyond longitudinal birth and preconception cohorts are needed to advance the field and shed further light on this complex, but highly promising, field of research.

**Acknowledgements:** None

**Financial support**

JFF received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements No 633595 (DynaHEALTH) and No 733206 (LifeCycle). CC is supported by the Economic and Social Research Council (grant ref: ES/N001273/1).

**Conflicts of interest:** None
References


Felix and Cecil

Population epigenetics and DOHaD


Figure 1

**a. DNA methylation as a causal mechanism**

- Potential moderators:
  - Offspring factors (e.g. genetics, sex)
  - Maternal factors (e.g. genetics, health, lifestyle)
  - Exposure factors (e.g. timing, duration, severity)

- DNA methylation (mediator variable)

- Health outcomes (e.g. physical and mental health/disease)

- **Utility & translational applications**
  - Advance understanding of etiological pathways linking prenatal exposures to health outcomes
  - Inform DNA methylation targets for personalized intervention and treatment (e.g. epigenetic therapy)

**b. DNA methylation as a non-causal biomarker**

- Exposure biomarker

- Prenatal exposures

- Risk prediction biomarker
- Early detection biomarker
- Diagnostic biomarker
- Prognostic biomarker
- Response biomarker

- Health outcomes

- No disease
- Prodromal/subclinical
- Disease onset
- Disease progression
- Disease treatment

- **Utility & translational applications**
  - Utility as research and clinical biomarker for exposure detection, disease risk and management as well as response to intervention
  - Non-causal biomarker unlikely to have utility as intervention/treatment target

---

*a. The causal mediation model, whereby prenatal exposures (independent variable) partly influence health outcomes in the offspring (dependent variable) via changes in DNA methylation (mediator variable). Of note, both the a path and the b path are hypothesised to be moderated by genetic effects, as well as additional factors. Furthermore, DNA methylation may also mediate genetic (as well as environmental) effects. b. The alternative non-causal model, whereby DNA methylation can serve as a biomarker of, but not causal mechanism in, exposure-outcome associations. Note that we present here the two models that are most relevant to the DOHaD framework; however, it is important to note that other models have also been proposed. For example, DNA methylation may function as a moderator of genetic and environmental influences on outcomes or as a mediator of genetic influences on outcomes. Also, stochastic changes may influence DNA methylation. More complex models are also possible [see for a more detailed discussion].*