TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

# Statistical Approaches for Quantifying the Mediating Role of Omics Markers Between the Exposome and Health

Noora Kartiosuo

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 755 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2026

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

# STATISTICAL APPROACHES FOR QUANTIFYING THE MEDIATING ROLE OF OMICS MARKERS BETWEEN THE EXPOSOME AND HEALTH

Noora Kartiosuo

TURUN YLIOPISTON JULKAISUJA ANNALES UNIVERSITATIS TURKUENSIS
SARJA SER. AI OSA TOM. 755 | ASTRONOMICA CHEMICA PHYSICA MATHEMATICA | TURKU 2026

## University of Turku

Faculty of Science
Department of Mathematics and Statistics
Statistics
Doctoral Programme in Exact Sciences

## Supervised by

Professor Kari Auranen
Department of Mathematics and Statistics &
Department of Clinical Medicine
University of Turku

Professor Jaakko Nevalainen
Unit of Health Sciences
Faculty of Social Sciences
Tampere University

Professor Olli Raitakari
Centre for Population Health Research &
Research Centre of Applied and Preventive Cardiovascular Medicine
University of Turku

## Reviewed by

Professor Juha Karvanen
Department of Mathematics and Statistics
University of Jyväskylä

Professor Anthony Hannan
Epigenetics and Neural Plasticity Group
The Florey Institute of Neuroscience and Mental Health, University of Melbourne

## Opponent

Professor Manuela Zucknick
Oslo Centre for Biostatistics & Epidemiology
University of Oslo




ISBN 978-952-02-0553-9 (PRINT)
ISBN 978-952-02-0554-6 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)
Painosalama, Turku, Finland, 2026

*To my loved ones*

UNIVERSITY OF TURKU
Faculty of Science
Department of Mathematics and Statistics
Statistics
KARTIOSUO, NOORA: Statistical Approaches for Quantifying the Mediating Role of Omics Markers Between the Exposome and Health
Doctoral dissertation, 169 pp.
Doctoral Programme in Exact Sciences
June 2026

## ABSTRACT

With increasing availability of sequencing methods, a wide range of omics markers have redeemed a central role in health sciences. The exposome, *i.e.* the range of life-course influences from external environment, life style and internal environment, are known to affect health and also influence various omics markers. These markers in turn can play a role in health and disease, and thus they hold promise as mediating links between the exposome and health outcomes.

Causal mediation analysis provides a statistical framework for quantifying indirect effects of exposures carried through mediators, *i.e.* variables that serve as mechanistic pathways transmitting the effects of exposures on health outcomes. Some properties of omics data pose challenges in the conduct of mediation analysis. For example, sequencing count data may not be suitable in their raw form but require appropriate methods to account for their compositional nature. Omics data also often exhibit sparsity (excess zero counts). Furthermore, in high-dimensional omics datasets, the meaningful signals need to be identified from a large number of variables to quantify the mediating role of multiple simultaneous mediators.

This thesis is motivated by empirical research questions concerning the role of the epigenome and gut microbiome in the association between exposome and health, both within individuals and across generations. The objective is to develop statistical approaches in the frameworks of mediation analysis and compositional data analysis for investigating the mediating role of various omics markers in the relationship between the exposome and health. To this end, a method for compositional mediation analysis is presented and its performance is evaluated using extensive simulation studies and empirical data. An asymptotic normal approximation for compositional log-ratio coordinates is derived and its applicability under varying levels of sparsity is investigated in a simulation study. Compositional methods are found to perform well when sparsity is not extreme. In the empirical analyses, the methods presented in this thesis are applied to examine the mediating role of the gut microbiome and DNA methylation between exposures and cardio-metabolic health of an individual. Finally, the potential application of these methods in investigating questions concerning paternal influences via sperm epigenome across generations is discussed.

KEYWORDS: mediation analysis, compositional data analysis, omics, causal inference

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Matematiikan ja tilastotieteen laitos
Tilastotiede
KARTIOSUO, NOORA: Statistical Approaches for Quantifying the Mediating Role of Omics Markers Between the Exposome and Health
Väitöskirja, 169 s.
Eksaktien tieteiden tohtoriohjelma
Kesäkuu 2026

## TIIVISTELMÄ

Sekvensointimenetelmien kehittyminen on kasvattanut kiinnostusta yhä monipuolisempia biologisia muuttujia, kuten omiikoita, kohtaan terveyden tutkimuksessa. Omiikka-aineistot sisältävät suuren määrän tietoa kiinnostuksen kohteena olevista biologisista molekyyleistä, ja niitä tutkimalla pyritään ymmärtämään kokonaisvaltaisesti terveyteen liittyviä tekijöitä. Eksposomi, joka sisältää ulkoisen ja sisäisen ympäristön sekä elintavat koko eliniän ajalta, vaikuttaa sekä yksilön terveyteen että erilaisiin omiikoihin. Omiikoita onkin kiinnostavaa tutkia mahdollisina välittäjinä eli mediaattoreina eksposomin ja terveyden välillä.

Mediaatioanalyysi on kausaalipäättelyn menetelmä, jonka tarkoituksena on arvioida paljonko altisteen ja vasteen yhteydestä johtuu mediaattorista eli muuttujasta, joka välittää altisteen vaikutuksia vasteeseen. Sekvensointiaineistot ovat usein lukumääräarvoisia, ja havainnoissa voi olla runsaasti nollia. Tällaiset aineistot eivät välttämättä sovi mediaaatioanalyysiin sellaisenaan, vaan niitä voidaan tutkia esimerkiksi rakenneosa-aineistoille kehitetyillä menetelmillä. Aineistot ovat usein myös korkeaulotteisia, sisältäen jopa satoja tuhansia muuttujia joiden joukosta mediaattorit on löydettävä.

Tässä väitöskirjassa tutkitaan erilaisten omiikoiden välittävää eli medioivaa roolia eksposomin ja terveyden välillä sekä kehitetään tilastollisia menetelmiä, joilla omiikoiden välittämiä epäsuoria vaikutuksia voidaan arvioida. Työn motivaationa ovat empiiriset tutkimuskysymykset epigenomin ja mikrobiomin roolista. Väitöskirjassa esitetään menetelmä kompositionaalisten muuttujien välittävän roolin tutkimiseen, ja sen sopivuutta sekvensointiaineistoon tutkitaan sekä simulaatioilla että empiirisellä tutkimusaineistolla. Rakenneosa-aineistoa kuvaaville logaritmisuhdekoordinaateille johdetaan asymptoottinen normaaliapproksimaatio, jonka soveltuvuutta nollia sisältävään lukumääräaineistoon tutkitaan simulaatioiden avulla. Nollien määrän pysyessä kohtalaisena nämä menetelmät soveltuvat sekvensointiaineistojen tutkimiseen. Empiirisissä tutkimuksissa esiteltyjä menetelmiä sovelletaan suolistomikrobiomin ja epigenomin roolien tutkimiseen yksilöiden eksposomin ja kardiometabolisen terveyden välillä. Lopuksi pohditaan mediaatioanalyysin kehikon soveltuvuutta epigeneettisen periytymisen tutkimisessa, kun kiinnostuksen kohteena on isän altisteiden vaikutukset jälkeläisten terveyteen siittiöiden epigeneettisten mekanismien välityksellä.

ASIASANAT: mediaatioanalyysi, rakenneosa-aineiston analyysi, omiikat, kausaalipäättely

# Acknowledgements

This thesis was conducted at University of Turku, at the Department of Mathematics and Statistics, in close collaboration with Research Centre of Applied and Preventive Cardiovascular Medicine (CAPC). It has truly been a privilege to work with these two teams and to learn from both method-orientated as well as empirically grounded health-focused research.

I owe my deepest gratitude to my three amazing supervisors, Professors Kari Auranen, Olli Raitakari and Jaakko Nevalainen, whose complementary perspectives have shaped not only my thesis but also myself as a scientist in invaluable ways. Without your support, this thesis would not have been finished. I really admire your scientific expertise, from which I have learned so much. I am deeply grateful to Kari for his patience, dedicated hands-on supervision and constant support. I appreciate that there was always room for questions of any calibre, as well as for other enlightening conversations in our regular meetings. Olli, I am thankful for the opportunity that you gave me over a decade ago to work with the CAPC datasets for the summer, which ultimately set me on this path. Your encouragement and belief in me sparked me to further explore the academic career, and I deeply appreciate your ongoing guidance and sharing your wisdom along the way. Jaakko, your support has been extremely valuable throughout my journey. I sincerely thank you for your insightful perspectives and wise advice throughout this process.

I owe heartfelt thanks to Professor Katja Pahkala, who, by setting a great example, inspired me to pursue this path. You show how kindness can go hand in hand with excellence and ambition. I am grateful for your emotional support, mentorship and friendship, as well as our long talks about science, academia, and life.

I am extremely honoured and grateful that Professor Manuela Zucknick has agreed to be my opponent. I would like to thank my thesis pre-examiners, Professor Anthony Hannan and Professor Juha Karvanen. I am grateful that you took the time to read and review my thesis, and it was an honour to also meet both of you in person to exchange insights on the topics where we share interests.

I am thankful for the opportunity to do my thesis at the Department of Mathematics and Statistics. I thank the Head of Statistics, Professor Henri Nyberg, for leading the team and creating a space to share both successes and setbacks. Thank you for taking the responsibility to serve as a custos at my thesis defence. I am grateful to Assistant Professor Joni Virta for his elegant solutions and insightful contributions

to our collaborative paper, as well as for his support when I was anxiously hoping for my work to get published. I also wish to jointly thank the current Statistics staff – Associate Professor Janne Kujala, PhD Pekka Nieminen, MSSc Jouko Katajisto - as well as the former members Professor Mervi Eerola and MSc Eila Seppänen, for the working environment and encouragement.

Our team was also enriched by many brilliant PhD students. I wish to thank my fellow statistics doctoral researchers: Vida Zamani, Lauri Heinonen, Visa Kuntze, Roope Rihtamo, Samuel Rauhala, Mikko Valtanen and Emil Lehti. It has been fascinating to witness the range of cool science that can be done around statistics, and to have the peer support for both the times of success as well as misfortunes. I wish to thank MSc Juho Pelto for the enlightening discussions around our shared interests. I am very grateful to have had MSc Katariina Perkonoja both as a colleague and a friend. Your compassion and commiseration as well as our honest conversations have made work and life feel lighter.

I thank the Department of Mathematics and Statistics, the Head of the department, Professor Iiro Honkala, and the doctoral programme of exact sciences (EXACTUS) for fostering the environment where this thesis was conducted.

I have been very lucky to work on my own research alongside having a role as a statistician at CAPC. Being part of this team has allowed me to grow into the scientist I had hoped to become. I am grateful for the enormous amount of support that I have received from many scientists that I look up to greatly. I am thankful to Associate Professor Suvi Rovio for setting such a wonderful example. Your guidance and compassion throughout my journey, regardless of whether I faced scientific or personal challenges, have made a great difference to me. I am very grateful to Professor Markus Juonala for his ongoing support and solution-oriented insights. I wish to thank Docent Juha Mykkänen for his encouragement, genuine interest in my work as well as his calming presence during hectic times.

Being a part of the data and statistics team at CAPC has been a pleasure. I am grateful to my present colleagues MSSc Irina Lisinen and MSc Sini Stenbacka for the team spirit, fun times as well as collaborative problem solving. I also wish to thank Arja Kylliäinen, MSc Johanna Ikonen and MSc Johanna Haapala for their previous contributions to the team. Nina Ruotsalainen deserves my special gratitude for always being available for any questions about practical matters, of which there have been many over the years. The entire group of friends and colleagues at CAPC, with whom I have shared both the highlights and the everyday moments of research, deserves my sincere thanks.

I have been fortunate to have the opportunity to work with two valuable cohort studies. I wish to thank Professor Emeritus Jorma Viikari for his groundbreaking work with the Young Finns Study. Thank you Nina Aalto and Tiina Peromaa for your tremendous work over the years to ensure that our cohort studies yield high-quality data for research. I gratefully acknowledge the participants of the Young

Finns Study and the STRIP study, whose data has made this research possible.

One of the highlights of my journey towards PhD has been the research visits to the Murdoch Children's Research Institute in Melbourne. I am grateful for the wonderful collaboration with Professor Richard Saffery, who over the years has repeatedly welcomed me into his brilliant group and taught me a great deal about biological mechanisms. Your unwavering enthusiasm about my research interests has been a source of motivation and strengthened my confidence as a researcher. I wish to thank Associate Professor Boris Novakovic for generously sharing his expertise and his patience in guiding me through the analysis pipelines. I extend my thanks to PhD Toby Mansell for the insightful conversations driven by our shared interests. My heartfelt thanks to the whole Molecular Immunity group for their hospitality, insights and friendship over the years. The friends and colleagues there, as well as in the Inflammatory Origins and Heart Research groups, are the reason why Melbourne nowadays feels like a home away from home for me. I would like to thank Kavindi Gamage, Anna Czajko and Sherly Li for the many adventures and memorable moments during my time spent in Melbourne.

I have had the opportunity to work with many wonderful coauthors and collaborators, from whom I have learnt so much. I want to thank Docent Panu Rantakokko for readily sharing his expertise and insights about environmental pollutants with me. I am thankful for Docent Emma Raitoharju and her team for our collaboration and their fascinating work with epigenetic data. I owe a great deal to Professor Noora Kotaja and her team. In particular, I am grateful for Dr. Matthieu Bourgery and Docent Juho-Antti Mäkelä, whose contribution to our male germline sample collection and data processing and analysis has been invaluable. I greatly value Professor Noora Kotaja's insights as well as her enthusiasm. Finally, I am grateful for my other coauthors – Anne-Louise Ponsonby, Sam Tanner, David Burgner, Marko Elovainio, Mikael Fogelholm, Mirja Hirvensalo, Nina Hutri, Eero Jokinen, Antti Jula, Jari Kaikkonen, Hannu Kiviranta, Juhani S. Koskinen, Mika Kähönen, Tomi P. Laitinen, Terho Lehtimäki, Britt-Marie Loo, Leo-Pekka Lyytikäinen, Costan G. Magnussen, Pashupati P. Mishra, Satu Männistö, Laura Pulkki-Råback, Tapani Rönnemaa, Leena Taittonen, Tuija H. Tammelin, Jorma Toppari and Päivi Tossavainen – for their contributions to the publications of this thesis.

One pivotal collaboration alongside my thesis work over the past decade has been with the International Childhood Cardiovascular Cohort (i3C) consortium. I wish to thank the i3C team for the fruitful collaboration and for generously sharing their expertise in cardiovascular research. Being a part of an international team of this calibre has played an important role in my development as a scientist. In particular, I am profoundly grateful to Professor Terry Dwyer for his mentorship and friendship over the years. Our discussions have shaped my scientific thinking and taught me so much about both research and life.

My thesis work was financially supported by The Emil Aaltonen Foundation,

Väisälä Foundation, The Finnish cultural Foundation, The Alfred Kordelin Foundation, The Diabetes Research Foundation, Oscar Öflunds Stiftelse, Paulo Foundation, Turku University Foundation, TREMENDO, EXACTUS and MATTI doctoral programmes and University of Turku research grants.

I am thankful to all my friends and loved ones – far too many to name here individually – for their friendship and support throughout this journey. I am especially grateful for my cottage board game crew, as well as my little wine tasting group, for light-hearted times spent together, providing much-needed perspective, and reminding me of life beyond the thesis. My warm thanks go to my partner's family for their kindness and continued interest in my research. I would also like to thank one of my primary school teachers, Maija Parmanen, whose dedication laid an early foundation for my academic path and ambition for learning.

Besides the joy and occasional desperation, statistics has introduced many great people into my life. In this regard, I would like to thank my "Oranki" group from my pre-PhD years, with whom I shared countless hours of studying and social activities, and whose presence have continued beyond those years. Specifically, I want to thank my friend Roosa Sandell for the many shared moments and belief in me and my ambitions; Lotta Saros, for the many walks and lunches during which we shared reflections of (doctoral) life and its challenges; and Markus Matilainen, for the regular exchange of thoughts regarding the quirks of academic life. I would also like to acknowledge those with whom I have rolled up our sleeves together in the various statistical student and professional associations.

Someone (likely not a statistician) once said that we are the average of the five people we spend the most time with. I owe my deepest gratitude to the five dearest people in my life for their influence, their love, and the person they have helped me become. Each of you has left an incredible mark on my life. I cannot express how thankful I am to my parents Minna and Pentti for providing me the greatest environment to grow up in, always believing in me, and encouraging me to explore my own paths in life. I wish to express my sincere appreciation to my two best friends – my chosen sisters – Eveliina and Yasmin for adding so much life to my years and keeping me human. Your unwavering acceptance and support has been one of the greatest gifts of my life. I am so grateful for my partner Tuomas for his grounding presence during setbacks and moments of self-doubt. Your patience, everyday acts of kindness and constant companionship mean the world to me.

Turku, May 2026
*Noora Kartiosuo*

# Table of Contents

# Abbreviations

| | |
|---|---|
| alr | Additive logratio |
| ASV | Amplicon sequence variant |
| CDE | Controlled direct effect |
| clr | Centered logratio |
| DACT | Divide-aggregate composite null test |
| DAG | Directed acyclic graph |
| DE | Direct effect |
| DNA | Deoxyribonucleic acid |
| DOHaD | Developmental Origins of Health and Disease |
| EWAS | Epigenome-wide association study |
| IE | Indirect effect |
| ilr | Isometric logratio |
| IQR | Interquartile range |
| miRNA | Micro RNA |
| NDE | Natural direct effect |
| NIE | Natural indirect effect |
| OR | Odds ratio |
| piRNA | Piwi-interacting RNA |
| PC | Principal component |
| PCB | Polychlorinated biphenyl |
| PE | Proportion eliminated |
| PM | Proportion mediated |
| RNA | Ribonucleic acid |
| SAP | Statistical analysis plan |
| SBP | Sequential binary partition |
| SCFA | Short-chain fatty acid |
| sncRNA | Small non-coding RNA |
| STRIP | The Special Turku Coronary Risk Factor Intervention Project |
| T2D | Type 2 diabetes |
| TE | Total effect |
| tsRNA | transfer-RNAs-derived fragments |
| YFS | The Cardiovascular Risk in Young Finns Study |

# Use of AI tools

Generative AI tools were used in accordance with the guideline on the responsible use of generative AI in research (European Commission, 2024). AI tools (ChatGPT) were used for language editing and grammar rules, studying concepts, R code correction and interpretation, and reference suggestions. The doctoral researcher as well as other researchers who participated in the articles of the thesis are accountable for the integrity of all content presented in the dissertation and articles.

# List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Kartiosuo N, Nevalainen J, Raitakari O, Pahkala K, and Auranen, K. Hypothesis-driven mediation analysis for compositional data: an application to gut microbiome. Biostatistics & Epidemiology, 8(1):e2360375, 2024.

II Kartiosuo N, Virta J, Nevalainen J, Raitakari O, and Auranen, K. On the distribution of isometric log-ratio coordinates under extra-multinomial count data. Statistical Papers 66(5):113, 2025.

III Kartiosuo N, Auranen K, Mansell T, Novakovic B, Nevalainen J, Pahkala K, Rovio S, Mykkäanen J, Viikari J, Juonala M, Rantakokko P, Kiviranta H, Kaikkonen J, Lehtimäki T, Raitoharju E, Mishra PP, Kähönen M, Ponsonby AL, Tanner S, Burgner D, Raitakari O and Saffery R. The effect of polychlorinated biphenyls on type 2 diabetes risk is mediated via DNA methylation. Environment International, 203:109779, 2025.

IV Pahkala K, Rovio S, Kartiosuo N, Auranen K, Bourgery M, Elovainio M, Fogelholm M, Haapala J, Hirvensalo M, Hutri N, Jokinen E, Jula A, Juonala M, Kaikkonen J, Kiviranta H, Koskinen JS, Kotaja N, Kähönen M, Laitinen TP, Lehtimäki T, Lisinen I, Loo BM, Lyytikäinen LP, Magnussen CG, Mishra PP, Mykkänen J, Mäkelä JA, Männistö S, Nevalainen J, Pulkki-Råback L, Raitoharju E, Rantakokko P, Rönnemaa T, Stenbacka S, Taittonen L, Tammelin TH, Toppari J, Tossavainen P, Viikari J, Raitakari O. Cohort Profile Update: Expanding the Cardiovascular Risk in Young Finns Study into a multigenerational cohort. International Journal of Epidemiology, 55(1):dyaf206, 2026.

# 1 Background and Motivation

An essential part of population health research is to understand how the exposome, *i.e.* the totality of environmental and other exposures from conception through the life course, affects human health and disease. [1] The exposome consists of interrelated collection of components, such as chemical exposures, physical and built environment, socio-economic context, and lifestyle factors. [2] To support the design of preventive measures or interventions through, for example, individual level support or macro-level health and social policies, it is critical to identify modifiable risk factors that have a causal role in the development of disease. [3; 4]

When aiming to understand the causal role of exposome in health, the mechanistic pathways through which its effects occur are of particular interest. It is unlikely that a single factor in the exposome or a single mechanistic pathway is the sole cause for a disease or condition. More likely several pathways play a causal role, and among them, the rapid development of sequencing methods have made a range of omics markers an important focus. [5; 6] Omics, in brief, are a collection of molecular measurements from biological samples, such as tissues or cells, that provide information of the underlying biological system. [7] They may play a role in phenotypes and in complex pathogenesis of disease and furthermore, they are susceptible to various environmental stressors, thus providing a potential bridge between the exposome and health.

From the perspective of statistical analysis, various omics, serving as potential mechanistic pathways, can be considered as mediators, *i.e.* intermediate variables between exposures and health outcomes. Their role may thus be investigated using methods designed for causal mediation analysis, which aims to quantify the effects transmitted through mediating pathways. [8] While various omics markers are potentially relevant intermediate variables that could transmit the effects of exposures on health, investigating their mediating role presents challenges. One such challenge stems from the high-dimensional nature of the omics data. Methods suitable for such high dimensional settings tend to focus on prediction at the cost of causal interpretation, and as such, have been described as "hypothesis-generating". [9; 10]

Furthermore, sequencing count data are often non-normally distributed and characterised by excess zero-counts (sparsity) and excess variability (overdispersion). As such, they may not be directly suitable for standard regression models for which mediation analysis approaches, often relying on specific functional forms of associa-

tions between the variables of interest, have been developed. As sequencing count data are often constrained by the total reading depth, it has been suggested that they need to be treated as compositional. [11] The mediating role of omics markers, measured as sequencing counts, could thus be investigated using compositional data analysis approaches, such as log-ratio transformations. Usually, these transformations are assumed to produce normally distributed variables suitable for standard regression models. However, under extremely sparse counts with excess zero-count observations, the normality assumption may not hold.

In this thesis, statistical approaches for investigating the mediating role of omics markers are reviewed, developed, examined and applied. These approaches are based on the frameworks of causal mediation analysis and compositional data analysis. The objective is to introduce a statistical framework for mediation analysis with compositional, sparse, and high-dimensional omics mediators. Statistical properties, such as limiting distributions of these omics mediators, are derived and investigated. An empirical implementation of compositional mediation analysis is presented and the performance of the approach is evaluated with respect to some special features of omics data sets, such as sparsity, using extensive simulation studies. Finally, the methods are applied on empirical data to demonstrate their ability in real-life mediation problems pertaining to the role of microbiome and epigenome. The methodological contributions of this thesis are motivated by empirical research questions where omics markers are considered as mediators between the exposome and health outcomes. Below, the three motivating examples are briefly introduced.

First, variation in gut microbiome composition has been suggested as a potential mechanism carrying the effects of the exposome on health. Diet, including various dietary components, can affect the gut microbiome, which in turn has been associated with cardio-metabolic health. [12] In publication I, the gut microbiome is investigated as a mediator in the effect of fibre intake on insulin levels. Microbial sequencing counts are treated as compositional data, and mediation analysis approach is built within the compositional data analysis framework. Taxonomic knowledge is utilised to adopt a hypothesis-driven approach rather than a data-driven one. Impact of sparsity (excess of zero-count observations) on the performance of the mediation analysis is assessed. Publication II further investigates the impact of varying amounts of sparsity on the suitability of compositional methods for sequencing count data.

Second, DNA methylation, an epigenetic mechanism that can up- or downregulate gene expression, plays a key role in development, health, and disease. [13; 14] Furthermore, many aspects of the exposome, such as exposure to environmental toxicants, have been linked to variation in DNA methylation, suggesting that it may act as a mechanistic link between the exposome and health. [15; 16] DNA methylation data are typically very high-dimensional, often encompassing hundreds of thousands of variables, necessitating selection of potential mediators before their joint mediating role can be quantified. In publication III, the mediating role of DNA methylation

in the association between exposure to polychlorinated biphenyls and type 2 diabetes was investigated.

Finally, mounting evidence, especially from animal studies, suggests that fathers contribute to the health of their offspring pre-conceptionally via epigenetic markers in their germ cells. [17] Although more challenging to investigate in humans, preliminary evidence indicates that paternal exposures, such as smoking, environmental chemicals and traumatic events, can impact offspring health. The exposome has also been found to affect the sperm epigenetic profile, but human data linking paternal sperm epigenome to offspring phenotypes remain limited. [18] Investigating inter- or transgenerational epigenetic effects in humans is complicated by potential confounding, various sources of bias and the long follow-up times needed to observe the health and phenotypes of subsequent generations. Publication IV of this thesis reports the field study and dataset collected by the MULTIEPIGEN project, set out to investigate paternal contributions to offspring health via sperm epigenetic markers. A statistical analysis plan, also presented in publication IV, outlines how the role of the sperm epigenome in transmitting the effects between generations can be framed as a mediation question. Compositional and high-dimensional mediation analysis approaches can then be applied to identify mediators and quantify their effects.

The remainder of this thesis is organised as follows. Chapter 2 reviews the main aspects of causal mediation analysis. In Chapter 3, approaches for multiple and high-dimensional mediators are reviewed. In Chapter 4, some principles of compositional data analysis are introduced. Furthermore, a new approach for mediation analysis with compositional data is presented alongside theoretical and simulation-based findings on the applicability of compositional data analysis for omics datasets characterised by sparsity. In Chapter 5, the empirical research questions motivating this thesis and the datasets are presented, and the use of mediation analysis and compositional data analysis in these questions is outlined. Finally, some concluding remarks are provided in Chapter 6. The thesis introduction is followed by four original publications.

# 2 Causal Mediation Analysis

In this thesis, mediation analysis plays a crucial role as the main statistical approach for the empirical research questions (publications I, III and IV). Hence, in this chapter, those key ideas of mediation analysis applied in the publications of this thesis are presented, providing context to the remainder of the work.

## 2.1 Background

In observational studies, establishing causal effects is challenging because treatments or exposures are not randomly assigned, unlike in randomised controlled trials, regarded as the gold standard for causal inference. To uncover causal effects based on observational studies and to address biases, such as those due to confounding, a wide range of approaches are available. These include, but are not limited to, instrumental variable methods, which exploit variables that influence the exposure and provide an unconfounded source of variation, difference-in-differences method, which mimic an experimental design by comparing outcomes over time between exposed and unexposed groups, inverse probability weighting, which balances the study population and makes the exposure independent of underlying confounders, and causal graph theory, which is used to encode causal assumptions. [9] Triangulation, *i.e.* combining multiple lines of evidence based on a number of different approaches that are based on different assumptions and have different, preferrably unrelated, sources of bias, can strengthen causal inference in observational studies. [19; 20]

When the interest is in quantifying the extent to which a causal effect between an exposure and outcome is transmitted through an intermediate variable, mediation analysis techniques are required and thus, of the vast and rich literature on causal inference, the statistical approaches of this thesis focus on causal mediation analysis. The aim is to uncover causal pathways by which the effects of treatment or exposure on the outcomes are transmitted. The interest in mediation analysis thus lies not only in the effect of exposure on an outcome but also in understanding and quantifying the causal (mechanistic) pathway between the two. The intermediate variable that transmits the effects from exposures to outcomes is called the mediator. [21] In the publications of this thesis, various omics markers (gut microbiome, DNA methylation, and small non-coding RNA molecules) are the mediators of interest.

Mediation questions can be presented as directed acyclic graphs (DAG), such

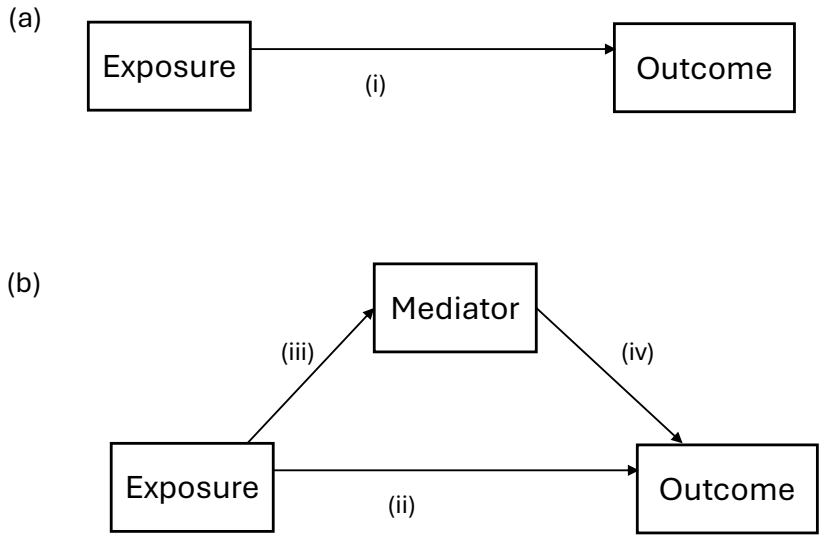


**Figure 1.** Directed acyclic graphs (DAGs) illustrating the pathways of interest in a mediation analysis with a single mediator. Panel (a): pathway (i) corresponds to the total effect of exposure on the outcome. Panel (b): pathway (ii) corresponds to the direct effect, and pathways (iii) and (iv) together constitute the indirect effect via the mediator.

as the one in Figure 1, where a very basic idea of mediation setting with a single mediator is shown. [22; 23] Figure 1 (a) describes the total effect, *i.e.* the effect of exposure on the outcome of interest without accounting for the mediator. In Figure 1 (b), the total effect is decomposed into two pathways: the indirect effect (consisting of paths (iii) and (iv)) through the mediator, and the direct effect (ii), *i.e.* the part of the total effect not transmitted through the mediator. [21]

In this chapter, mediation effects are first defined based on counterfactuals, which do not require specifying a parametric model. Then, the identifiability assumptions under which the effects can be estimated from observational data are introduced. Lastly, parametric forms for the mediation effects are introduced in the case of continuous and binary outcomes and some principles of statistical inference in mediation analysis are presented. The counterfactual definitions and identifiability assumptions allow interpreting the mediation effects through a causal lens. [24]

## 2.2 Counterfactual Effects

In the potential outcomes framework, the causal effect of a treatment or an exposure for an individual is the difference in the outcome in the presence versus in the absence of the treatment of exposure. However, the fundamental problem of causal inference is that these two alternatives cannot be observed simultaneously for one individual, as it is impossible to go back in time to set the individual under an alternative treatment. [25; 26] Hence, the counterfacual framework often referred to as Neyman-Rubin causal model, defines the causal effect as the difference in the outcomes that would have been observed under alternative circumstances. [27; 26] The individual-level causal effect is thus the difference of the individual's outcomes under two different exposure levels or, in causal literature, treatments: the outcome under the treatment that they actually received, and the outcome that would have been observed if they had been the under the treatment they did not receive. [22; 9] The latter of these outcomes is counterfactual, *i.e.* non-observable, as in reality, only one outcome is observed for any one individual. [9] The individual-level causal effects thus cannot be directly identified from the observed data. [9] However, population-level effects, such as average causal effects, can be identified under certain assumptions, described in more detail in Section 2.3.

While the Neyman-Rubin framework serves as a basis for causal inference using potential outcomes, the potential outcomes framework has also been extended to address mediation questions. In the counterfactual framework for mediation, the counterfactual outcomes are defined for both the exposure and the mediator. [24] To formulate mediation effects in the counterfactual framework, denote a binary exposure variable as $X$ with levels $x$ (often considered as "exposed") and $x^*$ ("unexposed"), a continuous outcome variable as $Y$, and a continuous mediator as $M$. Let $x', x'' \in \{x, x^*\}$, and let $Y(x', m)$ be the level of the outcome variable if the exposure was, possibly contrary to fact, set to level $x'$ and the mediator to level $m$. Let $M(x')$ be the level of the mediator that would be observed if the exposure was, possibly contrary to fact, set to level $x'$. Finally, let the nested counterfactual $Y(x', M(x''))$ denote the level of the outcome $Y$ that would occur, possibly contrary to fact, if $X$ was set to level $x'$ and $M$ would obtain the value if would naturally have when exposure was set to $x''$. [21] Note that this counterfactual is a "cross-world" counterfactual, as it involves two different levels of the exposure.

In the standard approach to counterfactual mediation analysis, following the seminal ideas of Pearl (2001, 2014) [22; 8], the total effect, the controlled direct effect, the natural direct effect, and the natural indirect effect are derived as the average changes in the outcome if certain interventions took place over the whole popula-

tion. These effects are defined based on the counterfactual notation as follows: [21]

$$\begin{aligned} \mathrm{TE} &= \mathbb{E}\left[Y(x, M(x))\right] - \mathbb{E}\left[Y(x^*, M(x^*))\right], \\ \mathrm{CDE} &= \mathbb{E}\left[Y(x, m)\right] - \mathbb{E}\left[Y(x^*, m)\right], \\ \mathrm{NDE} &= \mathbb{E}\left[Y(x, M(x^*))\right] - \mathbb{E}\left[Y(x^*, M(x^*))\right], \\ \mathrm{NIE} &= \mathbb{E}\left[Y(x, M(x))\right] - \mathbb{E}\left[Y(x, M(x^*))\right]. \end{aligned} \tag{1}$$

The total effect (TE) is the difference in the average counterfactual outcome $Y$ when all individuals are set to have exposure $x$ versus when all individuals are set to have exposure $x^*$, with the mediator taking the values that would naturally occur under $x$ and $x^*$, respectively. The controlled direct effect (CDE) is the difference in the outcome when the exposure is changed from $x^*$ to $x$, while the level of the mediator is held fixed at a pre-specified value $m$ for each individual.

The natural direct effect (NDE) is the difference in the average counterfactual outcome $Y$ when the exposure changes from $x^*$ to $x$ while the mediator obtains the value that would naturally occur under exposure level $x^*$, corresponding to the effect of exposure on the outcome if the pathway through the mediator was disabled. [8; 24] Lastly, the natural indirect (NIE) effect is the change in the average counterfactual outcome $Y$ when the exposure is held constant at level $x$ and the level of mediator changes from what it would have obtained under exposure level $x$ to what it would have obtained under exposure level $x^*$.

The key distinction between controlled and natural effects is that the controlled direct effect involves an intervention on the exposure while the mediator is fixed at the same level in the whole population, whereas the natural effects allow the mediator to obtain the levels it would naturally have under given (counterfactual) exposure level. Natural effects are applicable specifically in observational studies. [22]

Based on the counterfactual definitions, it is straightforward to see that the total effect decomposes into the NIE and NDE: [21]

$$\begin{aligned} \mathrm{TE} &= \mathbb{E}\left[Y(x, M(x))\right] - \mathbb{E}\left[Y(x^*, M(x^*))\right] \\ &= \mathbb{E}\left[Y(x, M(x))\right] - \mathbb{E}\left[Y(x, M(x^*))\right] + \mathbb{E}\left[Y(x, M(x^*))\right] - \mathbb{E}\left[Y(x^*, M(x^*))\right] \\ &= \mathrm{NIE} + \mathrm{NDE}. \end{aligned} \tag{2}$$

A similar decomposition is not possible for controlled effects as there are no controlled indirect effects. Nevertheless, TE$-$CDE can be interpreted as the portion eliminated, the part of the effect of exposure that could be eliminated if, for every individual, the mediator $M$ was set to the level $m$. [21]

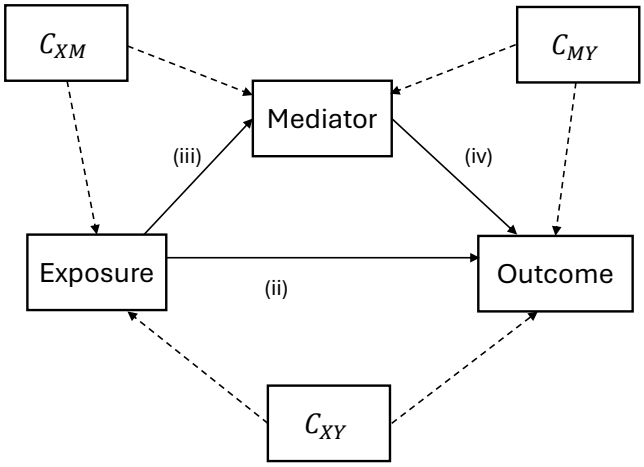


**Figure 2.** Directed acyclic graphs of a mediation setting with confounders for exposure-outcome, exposure-mediator and mediator-outcome relationships.

## 2.3 Identifiability of Causal Effects

As opposed to randomised trials, the estimation of causal effects in observational settings poses more challenges, as the exposure levels are not assigned randomly. In order to estimate the average causal effect of a treatment, the treated and nontreated (or exposed and nonexposed) groups need to be comparable and to this end certain exchangeability assumptions need to be stated. [26] In addition, for the general case of average causal effects, the stable unit treatment value assumption (SUTVA) requires that the assigned treatment has the same effect regardless of what treatment other units received and regardless of how the treatment was assigned, i.e., whether the treatment was randomised or received in an observational setting. [26]

The literature on identification of counterfactual mediation effects is widely focused on assumptions about confounding. Confounding variables (confounders) are common causes of two variables of interest which, if not properly taken into account, can influence the relationship between these two and cause bias in the estimates of causal effects. Confounding is particularly problematic in observational studies, where various underlying factors could affect the exposure, mediator and outcome.

Directed acyclic graphs (DAGs) are non-parametric causal models, widely used in the graph-based causal inference ([22]), describing the (assumed) causal relationships between the different variables without the need to specify their functional forms. [23; 21] The arrows capture causal dependencies between the parent and child nodes (variables) while the unexplained variation (errors) associated with the

nodes are assumed to be mutually independent. [21; 28]

Figure 2 extends the DAG of Figure 1 (b) to include three sets of potential confounders: the exposure-outcome confounders $C_{XY}$, the exposure-mediator confounders $C_{XM}$ and the mediator-outcome confounders $C_{MY}$. The collection of these three sets of confounders is here denoted as $C$. For identification of causal effects, a complete causal DAG, which describes the causal structure and includes all known or assumed causes of the exposures, mediators, and outcomes, is built. [23] In a standard setting for causal analysis, the so-called *backdoor paths* indicating confounding can be identified based on the DAG. Controlling for the sufficient confounders "blocks" these paths, *i.e.* "deconfounds" the relationships. [23]

The causal structure of the DAG implies certain identifiability (also called exchangeability or conditional independence) conditions, based on which the causal effects can be identified from observational data. The conditions that are required to estimate natural direct and indirect effects in observational studies can be delineated as follows: [8; 22; 21]

1. No unmeasured confounding of the exposure-outcome relationship (*i.e.*, $C_{XY}$ deconfounds the relationship)
2. No unmeasured confounding of the mediator-outcome relationship (*i.e.*, $C_{MY}$ deconfounds the relationship)
3. No unmeasured confounding of the exposure-mediator relationship (*i.e.*, $C_{XM}$ deconfounds the relationship)
4. The mediator-outcome confounders $C_{MY}$ must not be affected by the exposure.

In brief, exchangeability here means that conditionally on the confounders, the relationships between exposure, mediator, and outcome can be interpreted as if they were based on a randomised experiment and thus, individuals who have different levels of exposure are comparable in terms of their counterfactual outcomes. The assumptions rule out the role of confounding as the explanation of observed relationships between the exposure, mediator and outcome.

While condition 1 is sufficient for identifying the total effect, and conditions 1 and 2 for identifying controlled direct effects, all four conditions need to be fulfilled for the identification of natural direct and indirect effects. [24] The last assumption is referred to as a *cross-world independence assumption*. Under the absence of mediator-outcome confounders affected by the exposure, the expectation of the cross-world counterfactual $\mathbb{E}\left[Y(x', M(x''))\right]$, $x' \neq x''$ can, based on the exhangeability conditions, be identified. [29]

In addition to the four identifiability conditions, consistency, which links counterfactuals to the observations, and positivity, which ensures that the effects of interest are estimable from the data, are assumed. [21; 29] Consistency means that if $X = x$

then $Y = Y(x)$, meaning that if the individuals's exposure level is in reality $x$, their observed outcome $Y$ corresponds to the potential outcome $Y(x)$. [29] Positivity is defined as $P(X = x|C) > 0$, which means that, conditionally on the confounders, each individual has a non-zero probability of receiving each possible level of exposure. Unlike the other conditions, the positivity assumption only relates to observed data and as such is also testable. [29]

Under the identifiability conditions presented above, the natural direct and indirect effects can be derived from the observed data with the following non-parametric expressions: [30; 8; 22]

$$\text{TE} = \int \{\mathbb{E}[Y|X = x, C = \boldsymbol{c}] - \mathbb{E}[Y|X = x^*, C = \boldsymbol{c}]\} dF_C(\boldsymbol{c}), \tag{3}$$

$$\text{CDE} = \int \{\mathbb{E}[Y|X = x, M = m, C = \boldsymbol{c}] - \mathbb{E}[Y|X = x^*, M = m, C = \boldsymbol{c}]\} dF_C(\boldsymbol{c}),$$

$$\text{NDE} = \int \int \{\mathbb{E}[Y|X = x, M = m, C = \boldsymbol{c}] - \mathbb{E}[Y|X = x^*, M = m, C = \boldsymbol{c}]\}$$
$$dF_{M|X=x^*,C=\boldsymbol{c}}(m) dF_C(\boldsymbol{c}),$$

$$\text{NIE} = \int \int \mathbb{E}[Y|X = x, M = m, C = \boldsymbol{c}]$$
$$\{dF_{M|X=x,C=\boldsymbol{c}}(m) - dF_{M|X=x^*,C=\boldsymbol{c}}(m)\} dF_C(\boldsymbol{c}).$$

In practice, building a causal DAG and recognising all confounders that need to be accounted for needs not only statistical knowledge but also substantive knowledge of the research question at hand. Sensitivity analysis for violations of the no unmeasured confounding assumption is briefly introduced in Section 2.5. In addition, approaches for estimating mediation effects even when condition 4 does not hold have been introduced. [31]

## 2.4 Parametric Formulation

If the identifiability assumptions hold, the non-parametric expressions in Equation (3) can be used to derive parametric expressions for the total, controlled direct, natural direct and natural indirect effects. For this, parametric models need to be specified for the outcome (conditional on the exposure, mediator and confounders) and for the mediator (conditional on the exposure and confounders). [24]

Historically, mediation effects were first considered for continuous outcomes and mediators under the linear model with no exposure-mediator interaction. The advantage of the nonparametric counterfactual approach is that direct and indirect effects can be defined and estimated under various statistical models, including those with non-linear outcomes and mediators and in presence of interactions. [24] In what follows, the parametric formulations are presented for continuous and binary outcomes in the absence of exposure-mediator interaction.

### 2.4.1 Continuous Outcomes

Denote a binary exposure variable as $X$ with levels $x$ and $x^*$ and a continuous outcome variable as $Y$ and the mediator as $M$, and define a set of variables $C$ that contain the exposure-outcome, exposure-mediator and mediator-outcome confounders, and no unmeasured confounders. In addition, assume that there is no exposure-mediator interaction. Now, based on the non-parametric expressions of Equation (3) that follow from the DAG of Figure 2, the following parametric linear models can be defined for the relationship between $M$ and $X$ & $C$ and $Y$, $X$, $M$ and $C$.

$$\mathbb{E}\left[M|X=x, C=\boldsymbol{c}\right] = \beta_0 + \beta_1 x + \beta_2' \boldsymbol{c}, \tag{4a}$$

$$\mathbb{E}\left[Y|X=x, M=m, C=\boldsymbol{c}\right] = \gamma_0 + \gamma_1 x + \gamma_2 m + \gamma_3' \boldsymbol{c}. \tag{4b}$$

Of note, although for continuous outcomes the standard parametric approach in mediation analysis literature is linear regression model (as in Equation 4), which assumes linearity and i.i.d. error terms, more flexible models, such as non-linear models, generalized models or mixed-effects models for clustered data, can be specified to study more complex relationships between the exposure, mediator and outcome.

If the identifiability conditions hold and the regression models are correctly specified, the parametric expressions for the direct and indirect effects can now be derived by plugging models (4a) and (4b) in their corresponding non-parametric expressions. [32] The effects thus become the following:

$$\begin{aligned}
\text{CDE} &= \int \left[(\gamma_0 + \gamma_1 x + \gamma_2 m + \gamma_3' \boldsymbol{c}) dF_C(\boldsymbol{c}) - (\gamma_0 + \gamma_1 x^* + \gamma_2 m + \gamma_3' \boldsymbol{c}) dF_C(\boldsymbol{c})\right] \\
&= \gamma_1 (x - x^*), \\
\text{NDE} &= \int \int \left[(\gamma_0 + \gamma_1 x + \gamma_2 m + \gamma_3' \boldsymbol{c})\right. \\
&\quad \left. - (\gamma_0 + \gamma_1 x^* + \gamma_2 m + \gamma_3' \boldsymbol{c})\right] dF_{M|X=x^*, C=\boldsymbol{c}}(m) dF_C(\boldsymbol{c}) = \gamma_1 (x - x^*), \\
\text{NIE} &= \int \int \left[(\gamma_0 + \gamma_1 x + \gamma_2 m + \gamma_3' \boldsymbol{c}) dF_{M|X=x, C=\boldsymbol{c}} dF_C(\boldsymbol{c})\right. \\
&\quad \left. - (\gamma_0 + \gamma_1 x + \gamma_2 m + \gamma_3' \boldsymbol{c}) dF_{M|X=x^*, C=\boldsymbol{c}} dF_C(\boldsymbol{c})\right] \\
&= (\gamma_0 + \gamma_1 x + \gamma_2 \mathbb{E}\left[M|x, c\right] + \gamma_3' \boldsymbol{c}) - (\gamma_0 + \gamma_1 x + \gamma_2 \mathbb{E}\left[M|x^*, c\right] + \gamma_3' \boldsymbol{c})) \\
&= \gamma_2 \beta_1 (x - x^*).
\end{aligned}$$

Of note, while in this simplified scenario the CDE is equal to the NDE. In the presence of exposure-mediator interaction, their expressions would differ. A more detailed explanation on how the parametric expressions for the natural effects follow from the regression models, independence assumptions and definitions can be found *e.g.* from the Appendix of VanderWeele and Vansteedlandt (2009) [32].

Parameter $\gamma_1$ (*i.e.*, the direct effect) can be seen to correspond to pathway (ii) of Figure 1 (b) while parameters $\beta_1$ and $\gamma_2$ correspond to pathways (iii) and (iv) in

Figure 1 (b), respectively. The expression for the indirect effect, $\beta_1\gamma_2$, corresponds to the product-of-coefficients approach or *the product method*, presented *e.g.* by Baron and Kenny (1986) [33]. Based on the decomposition of Equation (2), the total effect can be written as $\gamma_1 + \beta_1\gamma_2$.

An alternative method for obtaining the total effect, used frequently especially before the development of the counterfactual framework for mediation analysis is to fit a separate regression model: [33]

$$\mathbb{E}\left[Y|X = x, C = \boldsymbol{c}\right] = \alpha_0 + \alpha_1 x + \alpha_2'\boldsymbol{c}, \tag{5}$$

where parameter $\alpha_1$ is the total effect, corresponding to pathway (i) in Figure 1 (a). This approach provides an alternative method for obtaining the indirect effect, referred to as *the difference method*, where indirect effect is given by $\alpha_1 - \gamma_1$, *i.e.* as the difference of the total and direct effects. [33; 22] The indirect effect thus corresponds to the reduction of the total effect when mediator is accounted for. Under the assumption of linear models for the outcome and mediator and no exposure-mediator interaction, the product and difference methods yield the same estimates for indirect effects. However, in the presence of interaction or under non-linear models, the difference method may produce biased results. In the empirical applications of this thesis, the product approach has been used to derive indirect effects.

In alignment with the decomposition of the total effect, the proportion mediated (PM) and proportion eliminated (PE) can both be simply calculated as

$$\text{PM} = \frac{\text{NIE}}{\text{TE}}; \qquad\qquad \text{PE} = \frac{\text{CDE}}{\text{TE}}.$$

The proportion mediated helps summarise the results of a mediation analysis by describing the extent to which the mediator transmits the effect of exposure in an unit-independent manner. However, it only has a meaningful interpretation if the direct and indirect effects have the same sign.

### 2.4.2 Binary Outcomes

For a binary outcome, modelled using logistic regression, and continuous mediator the parametric regression models corresponding to those in Equation (4) are:

$$\mathbb{E}\left[M|X = x, C = \boldsymbol{c}\right] = \beta_0 + \beta_1 x + \beta_2'\boldsymbol{c}, \tag{6a}$$

$$\text{logit}[P(Y = 1|X = x, M = m, C = \boldsymbol{c})] = \gamma_0 + \gamma_1 x + \gamma_2 m + \gamma_3'\boldsymbol{c}. \tag{6b}$$

Similarly to the case with the continuous outcome, making the exchangeability assumptions and assuming no exposure-mediator interaction, the parametric versions

of the direct, indirect and total effects become:

$$\log[OR_{\mathrm{NDE}}] \approx \gamma_1(x - x^*),$$
$$\log[OR_{\mathrm{NIE}}] \approx \beta_1\gamma_2(x - x^*),$$
$$\log[OR_{\mathrm{TE}}] \approx (\gamma_1 + \beta_1\gamma_2)(x - x^*).$$

The approximations hold when the outcome is rare. [34] On the odds ratio level, the total effect is a product, instead of a sum, of the $OR_{\mathrm{NDE}}$ and $OR_{\mathrm{NIE}}$. [24]

Corresponding to model (5) in the continuous outcome scenario, the total-effects model may be formulated as:

$$\mathrm{logit}[P(Y = 1|X = x, C = \boldsymbol{c})] = \alpha_0 + \alpha_1 x + \alpha_2'\boldsymbol{c}.$$

However, some complications may emerge from the non-collapsibility of the odds ratio. In many cases, indirect effects obtained by the difference method do not coincide with those obtained by the product-of-coefficients method. [35; 34] Namely, the difference method could yield too conservative estimates, as the regression coefficients of exposures in logistic regression are not comparable when the set of covariates in the model is changed. [24] When the outcome is rare, the product and difference methods lead to approximately similar results. However, when the outcome is not rare, the results from the two methods differ, and it has been argued that neither the product nor the difference method produces an effect that has an interpretation of a causal indirect effect. [34] For outcomes that are not rare, log-binomial models will yield direct and indirect effects at the risk ratio scale that will coincide if there is no exposure-mediator interaction.

### 2.4.3 Statistical Inference

After specifying the counterfactual effects of interest, setting their identifiability conditions, and constructing the parametric models, the models can be fitted to the observations and the mediation effects estimated.

The standard errors of the total and direct effects are typically readily available from the regression software. By contrast, the indirect effects are based on estimates from two different statistical models and thus, their standard errors need to be calculated. The two most commonly used methods for obtaining standard errors are the delta method and standard bootstrapping. While the more traditional delta method is computationally less demanding, bootstrapping has been suggested to yield more accurate inferences, especially with small sample sizes. [36]

For the indirect effect obtained by the product-of-coefficients approach, an approximate formula for its standard error is derived based on the multivariate delta

method (Sobel 1982) [37]:

$$\text{s.e.}(\hat{\beta}_1\hat{\gamma}_2) = \sqrt{\gamma_2^2\sigma_{\beta_1}^2 + \beta_1^2\sigma_{\gamma_2}^2},$$

where $\sigma_{\beta_1}^2$ and $\sigma_{\beta_2}^2$ are the variances of the estimators of $\beta_1$ and $\gamma_2$, respectively. This can be conveniently rewritten as:

$$\text{s.e.}(\hat{\beta}_1\hat{\gamma}_2) = \sqrt{\sigma_{\beta_1}^2\sigma_{\gamma_2}^2\left\{\left(\frac{\beta_1}{\sigma_{\beta_1}}\right)^2 + \left(\frac{\gamma_2}{\sigma_{\gamma_2}}\right)^2\right\}}. \tag{7}$$

The confidence intervals can then be obtained using the standard error.

In standard bootstrapping, data are resampled repeatedly and, for each sample, mediation models are fitted to estimate the total, direct and indirect effects. The distribution of these effects, based on multiple samples, is then used to derive confidence limits. [38] Bootstrapping does not rely on distributional assumptions, allowing the sampling distribution of the indirect effect to be non-normal, although bias-corrected options for bootstrapping have been suggested to handle the asymmetric sampling distributions of the indirect effects that may occur under small sample sizes. [39]

In practice, whether the mediated effect is null can be inferred based on the confidence intervals obtained via the delta method, which entails an assumption of a normal distribution of the indirect effect, which may be incorrect. Equivalently, if testing the null hypothesis of no indirect effect is of interest, the test can be based on standard z-scores. [40] This test, often referred to as Sobel's test, has been noted to be underpowered. [40; 41] On the other hand, due to the dependence of the standard error of the estimate of the indirect effect on the magnitudes of the $\beta_1$ and $\gamma_2$ effects, it tends to be smaller for indirect effects compared to the corresponding direct or total effects. [42] Thus the statistical power to detect mediation may become larger than that regarding the direct or total effects. [38] Due to this reason, statistical inference and testing based on the standard errors in case of indirect effects has been criticised [42], even though other arguments highlight the importance of characterising the uncertainty around the indirect effect [43].

An alternative method for testing mediation is to consider path-specific p-values in a joint significance (MaxP) test, where the hypothesis of no mediation is rejected if both pathways included in the indirect effect are non-null. [40; 38]. However, it should be noted that the mediation test actually has a composite form, as the null

hypothesis consists of three different elemenents ($H_{01}$, $H_{02}$, $H_{03}$): [40; 41]

$$\begin{aligned} H_{01} &: \beta_1 = 0, \gamma_2 \neq 0, \\ H_{02} &: \beta_1 \neq 0, \gamma_2 = 0, \\ H_{03} &: \beta_1 = 0, \gamma_2 = 0, \\ H_A &: \beta_1 \neq 0, \gamma_2 \neq 0. \end{aligned} \tag{8}$$

The fourth case, where both pathways are non-null, is the alternative hypothesis. [41; 38] If any of the three null hypotheses is not rejected, absence of mediation is concluded. An application of the composite null hypothesis framework in a high-dimensional setting is presented in publication III of this thesis. In previous literature, this approach has been suggested to have a better statistical power than the Sobel's test or MaxP test.

## 2.5 Unmeasured Confounding

Unmeasured confounding may lead to biased estimates of causal effects. A framework of sensitivity analysis for unmeasured confounding, presented in VanderWeele (2010 & 2011) [44; 45] but also applied much earlier in epidemiology [46], provides an approach to assess how much unmeasured confounding would explain away the inferred effects. Biases can come about in any observational setting, not just the ones pertaining to mediation. However, the strong identification assumptions make unmeasured confounding a crucial issue especially in the analysis of mediation.

The bias in total, direct and indirect effects is defined as the difference between the mediation effects of Equation (3) and the true mediation effects that are conditioned on the unmeasured confounders. [44; 45] Obviously, the effect of any unmeasured confounder on the exposure, mediator and the outcome cannot be known, and hence, in practice the true, unbiased effects are not observable.

For simplicity, the basic idea of sensitivity analysis is here presented for the total effects under the linear model. Assume a binary exposure $X$, a continuous outcome $Y$, a set of measured confounders $C_{XY}$ and a binary unmeasured confounder $U_{XY}$, which is not causally associated with the measured confounders and does not have interaction with exposure $X$. The effect of the unmeasured confounder on the outcome, conditional on the exposure and the observed confounders, is denoted as $\theta = \mathbb{E}[Y|x, c, U_{XY} = 1] - \mathbb{E}[Y|x, c, U_{XY} = 0]$, which is further assumed to be constant across the strata defined by $C_{XY}$. In addition, the dependence between the unmeasured confounder $U_{XY}$ and exposure, conditional on the confounders, is defined as $\delta = P(U = 1|x, c) - P(U = 1|x^*, c)$, which is further assumed to be constant across the confounder strata.

In this simple case, the bias of the total effect can be shown to be the product of the two parameters, $\theta\delta$. [24] Subtracting this bias from the estimate of the total effect yields the bias-corrected estimate. [24] In practice, both $\theta$ and $\delta$ are unknown. However, their plausible values can be hypothesised based on expert knowledge or assumptions. A common practice is to specify for both parameters a range of plausible values and calculate the extent of bias under each combination of varying $\theta$ and $\delta$. It is then inspected whether under any combination of the realistic values for the two parameters the unmeasured confounder could explain away the effects that were obtained without bias correction. The standard errors of the bias-adjusted estimates are the same as those of the original ones, and thus bias-corrected confidence intervals are easily obtained. [45] This approach can easily be extended to non-linear outcomes and continuous $U_{XY}$.

Analogously, the sensitivity analysis framework can be used for the controlled direct effect, assuming that $U_{MY}$ confounds the mediator-outcome relationship, by deriving the two aboove-mentioned bias parameters conditionally on the mediator as $\theta_m$ and $\delta_m$. The bias becomes: $\mathrm{Bias(CDE)} = \theta_m\delta_m$. In the absence of exposure-mediator interaction, the sensitivity analysis for the natural direct effect is essentially the same as it is for the controlled direct effect. For an unmeasured mediator-outcome confounder, both natural direct and natural indirect effects can be biased, altough they still sum up to an unbiased total effect, as the unmeasured confounder is not assumed to affect the exposure. It thus naturally follows that the bias of the natural indirect effect is $-\theta_m\delta_m$.

The presence of unmeasured confounding may be assumed *e.g.* based on *a priori* knowledge of potential common causes of the exposure, mediator and/or the outcome. Furthermore, presence of residual confounding can be assessed based on the correlation between the error terms in the models of the mediator and the outcome. [47] In pubication I of this thesis, sensitivity analysis for unmeasured confounding was conducted.

# 3 Multiple and High-Dimensional Mediators

In the empirical research questions of this thesis, the exposure is assumed to affect the outcome through multiple distinct mediators. The omics datasets are often high-dimensional, with $p$ (the number of variables) $\gg$ $n$ (the number of observations), and thus it is necessary to identify the true mediators from a large set of candidates and to estimate their individual and joint mediation effects. In mediation analysis with multiple mediators, both the pathways through individual mediators (consisting of pathways $(iii.1), \ldots, (iii.J)$ and $(iv.1), \ldots, (iv.J)$ in Figure 3) as well as the overall or "total" mediated effect through the collection of all identified mediators (pathways **(iii)** and **(iv)** in Figure 3) may be of interest.

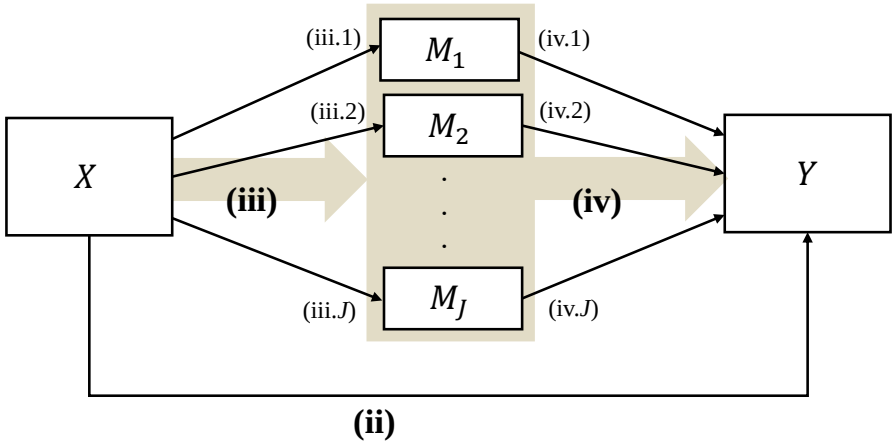


**Figure 3.** Directed acyclic graph visualising a mediation problem with multiple mediators. The individual arrors $iii.j$ and $iv.j$, $j = 1, \ldots, J$, correspond to mediator-specific effects, while the bolded arrows $\boldsymbol{iii}$ and $\boldsymbol{iv}$ represent the overall effects through the entire set of mediators. Pathway (ii) describes the direct effect.

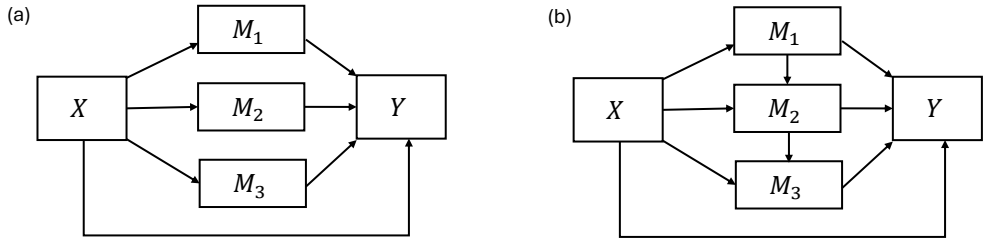


**Figure 4.** Directed acyclic graphs visualising two approaches to mediation analysis with multiple mediators. Panel (a): Contemporaneous mediators: the mediators are not assumed to affect each other causally. Panel (b): Causally ordered mediators.

## 3.1 Multiple Mediators

Mediation analysis can be extended to accommodate multiple mediators, and various approaches for multiple mediators have been proposed in the literature. In the case of multiple mediators, the set of mediators is a $J$-dimensional vector denoted as $\boldsymbol{M} = (M_1, M_2, \ldots, M_J)$. The approaches can be classified as methods for mediators that are contemporaneous, *i.e.* without a causal ordering (visually described for $J = 3$ in Figure 4 (a); *e.g.* [48; 49; 50]) or mediators that are causally ordered (Figure 4 (b); *e.g.* [51; 52; 47]).

In the empirical analyses of this work, omics mediators are considered as contemporaneous rather than causally ordered, as the omics data are obtained from a single time point without *a priori* knowledge of any potential causal interrelationships between the mediators. However, in other settings the omics markers may also be causally ordered or the set of mediators may include both contemporaneous as well as causally ordered mediators.

### 3.1.1 Contemporaneous Mediators

To be able to derive counterfactual mediation effects for multiple contemporaneous mediators (Figure 4 (a)), a nested counterfactual for $J$ mediators is defined as

$$Y(x_0, M_1(x_1), M_2(x_2), \ldots, M_J(x_J)); \; x_0, x_1, x_2, \ldots, x_J \in \{x, x^*\}. \qquad (9)$$

A special case is the counterfactual $Y(x', \boldsymbol{M}(x'')) = Y(x', M_1(x''), \ldots M_J(x''))$, which describes the counterfactual outcome that would occur if exposure was set at $x'$ and the entire vector of mediators would obtain the values that would naturally

occur if exposure was set at $x''$, with $x', x'' \in \{x, x^*\}$. From this definition follows the counterfactual definition of the overall indirect effect (OIE), *i.e.* the indirect effect through the entire vector of mediators:

$$\begin{aligned} \text{OIE} &= \mathbb{E}\left[Y(x, \boldsymbol{M}(x))\right] - \mathbb{E}\left[Y(x, \boldsymbol{M}(x^*))\right] \\ &= \mathbb{E}\left[Y(x, M_1(x), \ldots, M_J(x))\right] - \mathbb{E}\left[Y(x, M_1(x^*), \ldots, M_J(x^*))\right]. \end{aligned} \tag{10}$$

In addition, mediator-specific counterfactual indirect effects can be defined. To this end, special attention should be paid to how the levels of the other mediators, apart from the $j$th mediator of interest, are defined in the counterfactual expression (9), as there are are $2^J$ different ways to choose the levels for $x_1, \ldots, x_J$. One choice, suggested by Wang et al. (2013) [50], results in a particularly convenient way for defining counterfactual mediator-specific indirect effects:

$$\begin{aligned} \text{IE}_j = &\mathbb{E}\left[Y(x, \boldsymbol{M}_{j-}(x), M_j(x), \boldsymbol{M}_{j+}(x^*))\right] \\ &- \mathbb{E}\left[Y(x, \boldsymbol{M}_{j-}(x), M_j(x^*), \boldsymbol{M}_{j+}(x^*))\right], j = 1, \ldots, J. \end{aligned} \tag{11}$$

Here, the counterfactual outcome is the level of outcome when exposure is set to $x$, the first $j-1$ mediators $\boldsymbol{M}_{j-} = (M_1, \ldots, M_{j-1})$ obtain the values they would obtain when exposure is $x$, the $j$th mediator is changed from the value it would take under $x$ to the value it would take under $x^*$, and the rest of the mediators $\boldsymbol{M}_{j+} = (M_{j+1}, \ldots, M_J)$ obtain the values they would obtain under $x = x^*$. Although the mediators here do not have a causal ordering, in the interpretation of these mediator-specific indirect effects the order of coordinates is of importance. The use of this definition is demonstrated in publication I of this thesis, where the mediators were contemporaneous but had a natural ordering due to how they were built. Under this definition, regardless of the order of the mediators, the $J$ mediator-specific indirect effects sum up o the overall indirect effect (OIE), which was also demonstrated in the supplemental material of publication I of this thesis.

An alternative, rather intuitive way is to set all mediators except the $j$th to the level $X = x$ so that the indirect effects are: [49]

$$\mathbb{E}\left[Y(x, \boldsymbol{M}_{j-}(x), M_j(x), \boldsymbol{M}_{j+}(x))\right] - \mathbb{E}\left[Y(x, \boldsymbol{M}_{j-}(x), M_j(x^*), \boldsymbol{M}_{j+}(x))\right],$$

$j = 1, \ldots, J$. While this definition appears straightforward, the mediator-specific indirect effects now do not sum up to the overall indirect effect. However, the interpretation of these counterfactual effects is more intuitive and does not depend on the ordering of the mediators.

The exchangeability conditions for multiple mediators are analogous to those in the case of a single mediator, presented for each mediator separately. [47] Under

these conditions, the counterfactual outcomes in (11) can be expressed as: [50]

$$\int\int \mathbb{E}\left[Y|X=x_0, \boldsymbol{M}_{j-}(x_1)=\boldsymbol{m}_{j-}, M_j(x_2)=m_j, \boldsymbol{M}_{j+}(x_3)=\boldsymbol{m}_{j+}, C=\boldsymbol{c}\right] \tag{12}$$

$$\times\, dF_{\boldsymbol{M}_{j-}(x_1),M_j(x_2),\boldsymbol{M}_{j+}(x_3)}(\boldsymbol{m}_{j-}, m_j, \boldsymbol{m}_{j+}) dF_C(\boldsymbol{c}),$$

$x_0, x_1, x_2, x_3 \in \{x, x^*\}$. As shown in Wang et al. (2013) [50], provided that the joint distribution of the mediators can be estimated, the mediation effects are identifiable. The total effect, direct effect, and overall indirect effect require estimating only the joint distributions of $(M_1(x), \ldots, M_J(x))$ and $(M_1(x^*), \ldots, M_J(x^*))$ and thus, these effects are identifiable without further assumptions. [50]

For the mediator-specific indirect effects, however, the joint distributions of the mediators involve both $x$ and $x^*$ simultaneously, leading to a cross-world problem: for an individual, a vector of simultaneous mediators under both $x$ and $x^*$ cannot be observed. To be able to estimate the joint distributions of the mediators, Wang et al. (2013) [50] suggested pre-specifying (a common) cross-world correlation coefficient $\rho_{kl}$ between each $M_k(x_k)$ and $M_l(x_l)$, $k \neq l$, which allows identification of the mediator-specific indirect effects. In the simple case of linear models for the outcome and mediator and assuming no exposure-mediator interaction, the expression (12) depends only on the marginal distributions of the mediators, as explained in the supplemental material of publication I of this thesis. In this case, parametric mediator-specific indirect effects become identifiable even without pre-specifying the between-mediator correlations. However, if the outcome needs to be modelled using *e.g.* a logistic regression model, the cross-world correlation coefficients need to be specified to estimate the joint distribution of the mediators. [50]

In the case of continuous mediators and outcome, no exposure-mediator interaction, and assumption of linear relationships between the outcome, mediator and exposure, the following parametric linear models for the set of mediators and the outcome can be defined as:

$$\mathbb{E}\left[M_j|X=x, C=\boldsymbol{c}\right] = \beta_{0j} + \beta_{1j}x + \beta_2'\mathbf{c}, \tag{13a}$$

$$\mathbb{E}\left[Y|X=x_0, \boldsymbol{M}=\boldsymbol{m}, C=\boldsymbol{c}\right] = \gamma_0 + \gamma_1 x + \sum_{j=1}^{J}\gamma_{2j}m_j + \gamma_3'\mathbf{c}. \tag{13b}$$

Under the sequential exhangeability conditions, similarly to the case with one mediator, the parametric expressions of the causal effects then follow based on the coun-

terfactual definitions and Equation (12):

$$\text{NDE} = \gamma_1(x - x^*),$$

$$\text{OIE} = \sum_{j=1}^{J} \gamma_{2j}\beta_{1j},$$

$$\text{IE}_j = \gamma_{2j}\beta_{1j},\ j = 1, \ldots, J.$$

Regardless of how the counterfactual definitions of mediator-specific indirect effects ($\text{IE}_j$'s) are formulated, in the case of continuous outcome and mediators and no exposure-mediator interaction, the additivity of the $\text{IE}_j$'s to the OIE holds with the parametric models used here.

Of note, in the empirical modeling, the mediators should be modelled jointly using a multivariate regression, allowing correlated residuals. The covariances of the $\beta_{1l}$ and $\beta_{1k}$ as well as those of $\gamma_{1l}$ and $\gamma_{1k}$ are in fact involved in the standard errors of the OIE when based on the delta method, as demonstrated in the supplementary material A.2 of publication I of this thesis.

### 3.1.2 Causally Ordered Mediators

When the mediators are causally ordered, as in Figure 4 (b), the dependence of $M_2$ on $M_1$, the dependence of $M_3$ on $M_2$ and $M_1$, and so on, leads to more complex nested counterfactuals. For example, in the case of two causally dependent mediators, the nested counterfactual for the outcome is written as $Y(x_0, M_1(x_1), M_2(x_2, M_1(x_3)))$, $x_0, x_1, x_2, x_3 \in \{x, x^*\}$. [52] In this case, the direct effect, indirect effects via $M_1$ only and via $M_2$ only, and the indirect effects via both mediators, can each be defined in 8 different ways depending on how the levels of $x_0 x_1$, $x_2$ and $x_4$ are assigned in the nested counterfactuals. Furthermore, Daniel et al. (2015) [52] demonstrated that for two mediators, there are 4098 ($8^4$) ways to build a sum of these four effects. However, only 24 of these ways allow decomposing the total effect coherently into the four distinct effects. With increasing numbers of mediators, nested counterfactuals become increasingly complex, with $J$ mediators the number of possible decompositions of the total effect being $(2^J)!$. [52]

Also the confounding assumptions are challenged when dealing with causally ordered mediators. In Figure 4 (b), the first mediator $M_1$ is in fact a mediator-outcome confounder for $M_2$. However, as $M_1$ is also affected by exposure $X$, its presence violates the assumption that no mediator-outcome confounders are themselves affected by the exposure. [51] Of note, when investigating the vector of mediators $\boldsymbol{M}$ jointly, the assumption may still hold. In the case of causally dependent mediators, the exchangeability conditions are sequential. [47]

The estimation of natural direct and overall indirect effects proceeds in the same

way as with contemporaneous mediators. However, no straightforward solution for mediator-specific indirect effects exists. Daniel et al. (2015) [52] decomposed the total effect into a direct effect (*i.e.*, through none of the two mediators), indirect effects through only each of the mediators, and indirect effects through all mediators. VanderWeele and Vanseelandt (2014) [51] suggested utilising a known causal order of the mediators by defining indirect effects sequentially. Specifically, subsets of mediators are investigated $J$ times, each model defining the indirect effect as:

$$\begin{aligned}\mathrm{IE}_{j+} = &\left(\mathbb{E}\left[Y(x, \boldsymbol{M}_{(j+1)-}(x))\right] - \mathbb{E}\left[Y(x, \boldsymbol{M}_{(j+1)-}(x^*))\right]\right) \\ &- \left(\mathbb{E}\left[Y(x, \boldsymbol{M}_{j-}(x))\right] - \mathbb{E}\left[Y(x, \boldsymbol{M}_{j-}(x^*))\right]\right), j = 1, \ldots, J,\end{aligned}$$

where $\boldsymbol{M}_{(j+1)-}(x)$ is the vector of mediators $1, \ldots, j$. Thus, $\mathrm{IE}_{j+}$ is the increment in the mediated effect when the $j$th mediator is added. In practice, the first stage corresponds to a mediation analysis for the first mediator $M_1$, providing the indirect effect and proportion mediated through $M_1$. In the next stage, the second mediator in the causal pathway, $M_2$, is added to the models. The difference between the overall indirect effect in the second versus the first model characterises the additional contribution of mediator $M_2$ compared to mediator $M_1$ only. This process can then be carried forward by adding one mediator at a time, following their assumed causal order, to learn the incremental effect of each of the sequential mediators. [51]

### 3.1.3 One Versus Multiple Models for the Outcome

The overall indirect effect through all mediators jointly, based on a single model for $Y$, may differ from the sum of the indirect effects of the mediators considered separately, as highlighted in Vanderweele 2014 [51]. Only if the mediators are independent of each other (conditional on the exposure and the covariates) and do not have interactive effects on the outcome, the single-mediator effects summed up will yield same results as the multiple-mediator analysis. In practice, regardless of whether the mediators are contemporaneous or causally ordered, this means that the model for $Y$ should involve all the mediators simultaneously. In the case of causally ordered mediators (Figure 4 (b)), the path through the first mediator ($X \rightarrow M_1 \rightarrow Y$) would be also included in the pathway through the second mediator ($X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$) and hence, separate mediator-specific models would lead to counting these paths twice when summing the indirect effects into the overall indirect effect. Hence, the proportion mediated could be inflated. [51] Even if the mediators do not causally affect one another but have an additive interactive effect on the outcome or an underlying correlation structure, the overall indirect effect will differ from the sum of mediator-specific indirect effects. [48]

## 3.2 High-Dimensional Mediators

The approaches outlined above for multiple mediators apply to settings where the number of mediators is relatively small, and in particular, smaller than the number of observations. Omics datasets are often high-dimensional, *i.e.* the number of variables $p$ is much larger than the number of observations $n$ ($p \gg n$). [53] If, for example, the relationship between DNA methylation and an exposure or phenotype is of interest, an epigenome-wide association study (EWAS) can be used to assess this relationship across cytosine-guanine dinucletoide (CpG) sites measured throughout the genome. [54] CpG sites are DNA regions in which a cytosine nucleotide is followed by a guanine nucleotide and are susceptible to methylation, a process that can influence gene regulation. [55] In EWAS settings, the number of variables describing DNA methylation levels across CpG sites is often hundreds of thousands.

From the perspective of causal mediation analysis, high-dimensional mediators pose challenges. The causal structure of the effects of the numerous mediators as well as their interrelationships is not likely known *a priori*. In addition, knowledge of mediator-specific confounders is also often lacking. Hence, more data-driven, hypothesis-generating approaches are often needed to begin uncovering potential causal mechanisms.

In the case of high-dimensional mediators, prior to estimating the causal effects of interest, the key mediators need to be sieved from the large number of candidates. Various approaches, summarised by Blum et al [56], have been proposed to assess the mediating role of high-dimensional omics markers. Some approaches, utilised especially in the field of epigenetics, rely on identifying candidate mediators based on a high-dimensional set of exposure-mediator models and then conducting separate mediator-outcome analyses only for the set of identified mediators, appropriately accounting for multiple testing in both stages. This approach is to some extent analogous to the joint significance (MaxP) test, where it is required for each mediator that both the exposure-mediator and mediator-outcome associations separately show strong enough evidence against the null, ignoring the composite form of the mediation null hypothesis. However, considering the effects of each mediator individually does not take into consideration correlation or interactions between mediators. [56]

The mediators may also be considered jointly, treating them as whole (*i.e.*, considering only the overall indirect effect transmitted through all mediators). Such approach however does not allow identifying which specific mediators could play a role in the exposure-outcome associations and thus, the specific mechanistic pathways that could be targeted by, for example, therapeutic interventions remain uncovered. [56] The high-dimensional set of mediators in the outcome model also requires regularisation approaches, which often prioritise prediction yet lack causal interpretation. [57; 9; 56] Thus, this approach is not able provide insights on what interventions would be needed to modify the outcomes.

A high dimension of mediators can also be an advantage, as demonstrated by Liu et al. (2022). [41] They suggest the divide-aggregate composite null test (DACT) for high-dimensional mediators, based on the idea of the composite nature of the null hypothesis as presented in Equation (8). The test leverages the information of the distribution of p-values provided by the large number ($J$) of tests in order to provide a composite p-value for the hypothesis of mediation. This approach was applied in publication III of this thesis.

In practice, two large omics-wide analyses, usually EWAS, encompassing each of the $J$ omics markers are conducted. Both EWAS provide a set of p-values related to each of the $J$ $\beta_1$ and $\gamma_2$ parameters, denoted as $p_{\beta_j}$ and $p_{\gamma_j}$, $j = 1, \ldots, J$, respectively. The distributions of the two high-dimensional sets of p-values testing $\beta = 0$ and $\gamma = 0$ can be investigated and the large number of tests is leveraged to estimate the "proportions of true nulls". [58] These proportions are denoted as $\hat{\pi}_0^{\beta}$ and $\hat{\pi}_0^{\gamma}$ and they describe the proportions of the CpG sites that are not associated with the exposure and outcome, respectively. Typically, both proportions are very close to 1.

Based on $\hat{\pi}_0^{\beta}$ and $\hat{\pi}_0^{\gamma}$, Liu et al. (2022) suggest that the probabilities of the three hypotheses pertaining to the composite null hypothesis (see Definition (8)) and the alternative hypothesis can be calculated as follows: [41]

$$\begin{aligned} P(H_{01}) &= \hat{\pi}_0^{\beta}(1 - \hat{\pi}_0^{\gamma}) \\ P(H_{02}) &= (1 - \hat{\pi}_0^{\beta})\hat{\pi}_0^{\gamma} \\ P(H_{03}) &= \hat{\pi}_0^{\beta}\hat{\pi}_0^{\gamma} \\ P(H_A) &= (1 - \hat{\pi}_0^{\beta})(1 - \hat{\pi}_0^{\gamma}). \end{aligned} \tag{14}$$

The relative proportions of the three null cases are further normalised to sum to 1, denoted by $\hat{w_1}, \hat{w_2}$ and $\hat{w_3}$, and then used as weights to derive the so-called divide-aggregate composite-null test p-value of no mediation for each of the $J$ mediators as:

$$\mathrm{DACT}_j = \hat{w_1} p_{\beta_j} + \hat{w_2} p_{\gamma_j} + \hat{w_3}(\mathrm{Max}(p_{\beta_j}, p_{\gamma_j})^2),\ j = 1, \ldots, J. \tag{15}$$

Here, $\mathrm{Max}(p_{\beta_j}, p_{\gamma_j})^2$ refers to the squared maximum of the two path-specific p-values, which was noted to be the p-value of MaxP test by Liu et al. (2022).

While the test statistic of Equation (15) often follows the uniform distribution due to $w_3$ being close to 1, in some cases $w_1$ or $w_2$ may become larger and hence the distribution of the DACT test statistic may deviate from uniformity. In this case, Liu et al. (2022) [41] suggest calibrating the original DACT p-values of Equation (15) using the empirical null framework developed by Efron (2004) [59]. Briefly, the calibration is done by transforming the DACT p-values into z-scores in order to estimate the empirical null distribution. [58] This approach can aid in correcting

the inflation and bias in the p-values stemming from, for example, correlatedness of multiple tests due to dependencies in the underlying epigenetic data.

In Liu et al. (2022) [41], the more traditional approaches to infer the statistical significance of the mediator, Sobel's test and joint significance test, were concluded to have too low statistical power and too high type I error rates when applied for high-dimensional mediators. These tests were suggested to be overly conservative due to the small numbers of signals in high dimension, multiple testing approaches and for not accounting for the composite nature of the null hypothesis. In contrast, the type I error rate of the DACT test was found to be close to the nominal levels, and its statistical power was superior to the two traditional approaches.

# 4 Compositional Mediators

Over the recent years, the compositional nature of sequencing count data has been acknowledged. [11; 60] Such data consist of counts of reads that are assigned to various groups, for example, taxonomic units in microbial data. As the read counts in sequencing studies are restricted by the capacities of the sequencing instruments and as such, are arbitrary, such data can be interpreted to carry relative information rather than absolute information, as opposed to *e.g.* more traditional ecological settings, where the abundance of species is of key interest. [11]

The patterns inferred from sequencing data depend on whether the data are considered as abundances (counts) or composition (relative information). This is demonstrated in Figure 5, following a similar example presented in Gloor et al. (2017) [11]. The example consists of three samples, each having two features. When considering these data as read counts (left-hand panel), samples 1 and 2 have the same amount of the first feature, whereas the amount of that feature is twofold in sample 3. When treating the data as compositions (relative information; right-hand panel), the relative proportion of the first feature is the same between samples 1 and 3. Thus, even though the total abundances between samples 1 and 3 were clearly different, the relative information carried is the same, *i.e.* the two samples are compositionally equivalent. Thus, approaches treating the features as two independent count variables (*e.g.* those relying on negative binomial models) would lead to different conclusions compared to the methods that treat the data as relative abundances.

## 4.1 Introduction to Compositional Data Analysis

Compositional data comprise of elements that together constitute a whole entity, summing to unity or, more generally, to a constant. Compositional data often arise as non-negative observations, such as counts that are meaningful when scaled by their total. [61] For example sequencing-based omics data are constrained by their sequencing depth, daily time use is constrained to 24 hours, and election polls are constrained to sum to 100%. An important property of compositional data follows from this: compositions only carry relative, not absolute, information, and as the interest lies in the parts of composition that are scaled into proportions of the whole, compositional data are scale invariant. [61; 62] The unit-sum constraint also leads to negative correlation between the parts of composition, sometimes referred to as

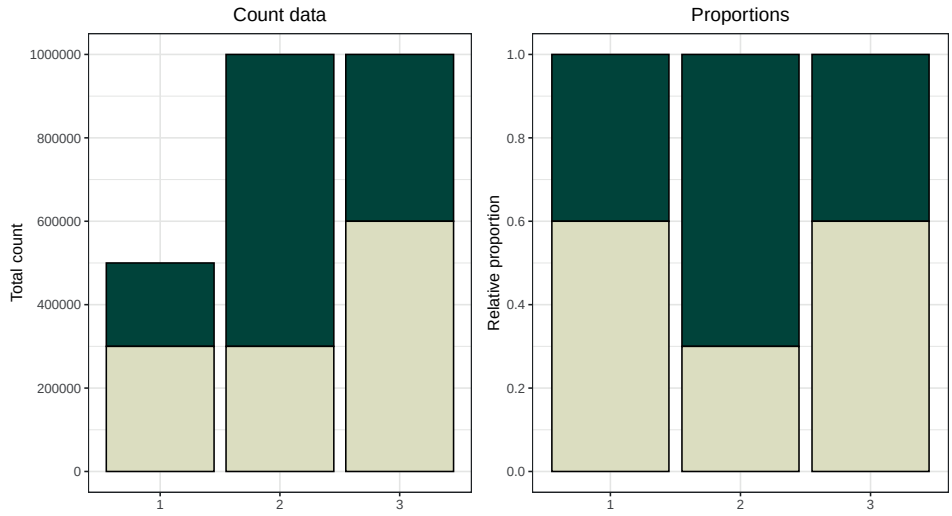


**Figure 5.** Example of sequencing count data treated as counts (left-hand panel) or proportions (right-hand panel). The graph follows a similar example presented in Gloor et al. (2017) [11].

*spurious correlation* [61]. If the relative amount of one part (for example in daily time use, sedentary time) increases, the relative amount of at least one another part (for example, time spent for physical activity or sleep) inevitably decreases.

A $J$-part composition is defined as a vector $\boldsymbol{z} = (z_1, \ldots, z_J)$, where all components are positive and represent only relative information, while the absolute numerical values of the components are not of interest. The closure operator is used to scale the composition $\boldsymbol{z}$ to a constant sum $\kappa$ [62]:

$$\mathcal{C}(\boldsymbol{z}) = \left( \frac{\kappa z_1}{\Sigma_{i=1}^{J} z_i}, \ldots, \frac{\kappa z_J}{\Sigma_{i=1}^{J} z_i} \right).$$

Often $\kappa$ is set to 1 so that the composition represents proportions. However, in sequencing count data, $10^6$ is a common choice, allowing the parts of composition to represent reads per million.

Due to the constant-sum constraint, compositional data reside in the simplex space, where this constraint is naturally built in. A $J$-part simplex is denoted by $\mathcal{S}^J$ and defined as

$$\mathcal{S}^J = \left\{ \boldsymbol{z} = (z_1, \ldots, z_J);\ z_j > 0,\ j = 1, \ldots, J;\ \sum_{j=1}^{J} z_j = \kappa \right\}.$$

This simplex is the sample space of compositional data. A three-part composition that resides in $\mathcal{S}^3$ can be visually presented using a ternary diagram, as demonstrated in Figure 6. Figure 6 also presents the locations of three distinct compositions, $z_1 =$

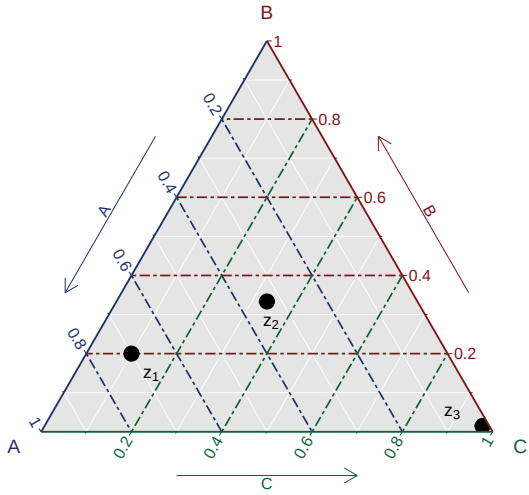


**Figure 6.** Ternary plot in $\mathcal{S}^3$ representing three distinct compositions: $z_1 = (0.70, 0.20, 0.10)$, $z_2 = (1/3, 1/3, 1/3)$, $z_3 = (0, 0, 1)$.

$(0.70, 0.20, 0.10)$, $z_2 = (1/3, 1/3, 1/3)$ and $z_3 = (0, 0, 1)$. Of note, in some older treatises (*e.g.* Aitchison (1982) [63]), the simplex $\mathcal{S}^J$ is defined with $J$ proportions for which $\sum_{j=1}^{J} z_j < 1$ with an additional fill-up value $z_{J+1} = 1 - \sum_{j=1}^{J} z_j$. This convention emphasises the fact that in compositional data, due to the constant-sum constraint, knowing all parts but one is enough to know the full composition.

Due to the constraints and properties of compositional data, neither the vector operators defined in Euclidean space nor standard multivariate methods, which often assume a real-valued sample space, are directly applicable to compositional vectors and their analysis. Instead, the framework of compositional data analysis, suggested initially by Aitchison, provides a collection of statistical methods suited for handling compositional data. [63; 64] The Aitchison geometry on the simplex refers to the algebraic-geometric structure of the simplex, and the set of operators defined on the simplex, Aitchison inner product, norm, and distance, are based on logratios of the parts of the composition, taking into account the relative nature of the data. [65]

## 4.2 Log-Ratio Transformations

Because the algebraic-geometric structure of compositional data residing on the simplex differs from that of observations residing on the real space, using standard statistical methods for compositional data may lead to incoherent inferences. [65] The

principle of working on coordinates refers to representing the data in log-ratio coordinates, and treating these coordinates as random variables when applying standard statistical methods in the Euclidean geometry. [65] Log-ratio transformations are tools to map the data from the simplex to Euclidean space. They allow handling the relative nature of the data as well as the constant-sum constraint so that standard statistical models are applicable.

There are three widely-used log-ratio transformations for compositional data: additive and centered log-ratio transformations, both first suggested by Aitchison (1986) [64], and isometric log-ratio transformation, presented by Egozcue et al. (2003) [66]. The additive and centered log-ratio transformations are here covered briefly before focusing on the isometric log-ratio transformation, which was the focus in Publications I and II of this thesis, in more detail.

### 4.2.1 Additive and Centered Log-Ratio Transformations

Additive log-ratio (alr) transformation is used to transform a $J$-part composition $\boldsymbol{z} = (z_1, \ldots, z_J)$ in $\mathcal{S}^J$ into $\mathbb{R}^{J-1}$. It is defined as the logarithm of the ratios of the parts of a composition to a chosen reference part, as follows:

$$\mathrm{alr}(\boldsymbol{z}) = \left( \ln \frac{z_1}{z_J}, \ldots, \ln \frac{z_{J-1}}{z_J} \right).$$

The alr transformation is not symmetric as it depends on the reference part. Choosing another part of the composition as reference would mean a different transformation and hence, a different interpretation. Especially when the reference part has a meaningful role in the composition in terms of the empirical research question, alr coordinates carry a rather straightforward interpretation. [67] The alr transformation is an isomorphism, *i.e.* it preserves the relative relationships between the parts. However, it is not isometric, *i.e.* it does not preserve the distances between the data points when moving from the simplex to real space. [67; 66].

The centered log-ratio (clr) transformation is defined as a logarithm of the ratios of each part of the composition to the geometric mean of all parts:

$$\mathrm{clr}(\boldsymbol{z}) = \left( \ln \frac{z_1}{\mathrm{gm}(\boldsymbol{z})}, \ldots, \ln \frac{z_J}{\mathrm{gm}(\boldsymbol{z})} \right),$$

where $\mathrm{gm}(\boldsymbol{z})$ denotes the geometric mean of the components: $\left( \prod_{j=1}^{J} z_j \right)^{\frac{1}{J}}$. [67; 66] While the clr transformation is symmetric in the components, it keeps the same dimension, *i.e.* transforms the $J$-part composition from $\mathcal{S}^J$ into $\mathbb{R}^J$. The transformation is both isomorphic and isometric. However, its components are constrained to sum to zero, leading to a singular covariance matrix, and thus, to a degenerate distribution. [67]

Initially, the alr transformation was used by Aitchison for statistical modeling of compositional data and the clr transformation for methods relying on distances, such as principal components analysis. However, their issues regarding preservation of distances and singular covariance matrices, respectively, motivated the development of the isometric log-ratio transformation (ilr). [67; 66] The ilr transformation was of interest in publications I and II of this thesis due to its straightforward applicability in mediation setting and flexibility in choosing the parts to contrast against each other.

### 4.2.2 Isometric Log-Ratio Transformation

The isometric log-ratio (ilr) transformation was first introduced by Egozcue et al. (2003) [66] to overcome the shortcomings of the additive and centered log-ratio transformations. The transformation, built on an orthonormal basis, is isometric due to its ability to preserve the metric properties (*i.e.* angles and distances) when mapping the compositions from the simplex to the real space. The ilr transformation preserves the Aitchison operators in the simplex as ordinary Euclidean operators in the real space, and the results of statistical analyses using the ilr coordinates are the same as could be obtained using compositions and Aitchison geometry. [67; 66].

The $J$-dimensional composition $\boldsymbol{z}$ is transformed from the simplex into a $J-1$ dimensional vector of isometric log-ratio coordinates, denoted here as $\boldsymbol{m}$, as follows:

$$\boldsymbol{m} = \mathrm{ilr}(\boldsymbol{z}) = \boldsymbol{V}'\ln(\boldsymbol{z}).$$

Here, the $J \times (J-1)$ -dimensional orthonormal basis matrix $\boldsymbol{V}$ is called a contrast matrix, and its elements are based on a given sign matrix $\boldsymbol{\Psi}$, which consists of values +1, -1, and 0, as follows:

$$(\boldsymbol{V})_{jk} = \psi_{jk}\sqrt{\frac{n_k^+ n_k^-}{n_k^+ + n_k^-}}\left[\frac{1}{n_k^+}\right]^{\mathrm{I}[\mathrm{sign}(\psi_{jk})=+]}\left[\frac{1}{n_k^-}\right]^{\mathrm{I}[\mathrm{sign}(\psi_{jk})=-]}, \qquad (16)$$

$j = 1, \ldots, J$; $k = 1, \ldots, J-1$. Here, $n_k^+$ and $n_k^-$ are the numbers of cells with values $+1$ and $-1$ in the $k$th column of $\boldsymbol{\Psi}$. The choice of the sign matrix $\boldsymbol{\Psi}$ is further discussed below. Thus, the ilr coordinates are obtained as

$$m_k = \sqrt{\frac{n_k^+ n_k^-}{n_k^+ + n_k^-}}\ln\frac{\mathrm{gm}(z_k^+)}{\mathrm{gm}(z_k^-)}, k = 1, \ldots, J-1,$$

where $\mathrm{gm}(z_k^+)$ and $\mathrm{gm}(z_k^-)$ denote the geometric means of the proportions in the classes denoted by $+1$ and $-1$ in the $k$th column of the sign matrix, respectively.

#### Choosing the contrasts

There is no a simple canonical basis to construct ilr coordinates. Instead, various alternatives have been suggested, and the best choice often depends on the research

question at hand. [61] One way to build the coordinates is to use a sequential binary partition (SBP), where the composition is recursively split into two groups that are contrasted against each other. This choice leads to so called *balance* coordinates. [68] Natural examples for building balance coordinates based on *a priori* knowledge stem from applications that contain natural hierarchies. For example, in election polls the parties can be roughly categorised to left-wing and right-wing parties, and further categorised within these categories; whereas the analysis of microbial data can rely on the known taxonomic/phylogenetic hierarchies within the data. [68; 69] If the data represent natural groups or hierarchies, the SBP can be built based on these hierarchies so that both the relationships between distinct groups and the relationships between members within groups can be investigated. These two relationships are referred to as *inter-group* and *intra-group* analysis by Egozcue and Pawlowsky-Glahn (2005). [68]

For illustration, an example regarding a simplified question on taxonomic data from publication I of this thesis is used. In a five-part taxonomic composition, assume that there are three groups and within these groups, five distinct taxa. Taxon *A* belongs to the first group; taxa *B* and *C* to the second group; and taxa *D* and *E* to the last group, as visualised in Figure 7(a). One example of a set of balance coordinates with intuitive interpretation is obtained by building the $5 \times 4$ sign matrix for the sequential binary partition (SBP) as follows:

$$\boldsymbol{\Psi} = (\psi_{jk}) = \begin{bmatrix} +1 & 0 & 0 & 0 \\ -1 & +1 & +1 & 0 \\ -1 & +1 & -1 & 0 \\ -1 & -1 & 0 & +1 \\ -1 & -1 & 0 & -1 \end{bmatrix}. \tag{17}$$

Each column of matrix (17) defines how the components are contrasted against each other when building the four isometric log-ratio coordinates. The first coordinate is the logratio of taxon *A* against all other taxa, describing the role of this specific taxon in the composition; it can also be considered as an inter-group comparison of group 1 against groups 2 and 3 combined. The second coordinate contrasts the taxa $\{B, C\}$ against taxa $\{D, E\}$ into an inter-group balance between the second and the third group. Lastly, the third and fourth coordinates correspond to intra-group balances of the second and the third groups, respectively. The hierarchy as encoded by the matrix (17) is visualised in Figure 7(b). In addition to this example approach, presented in Publication I of this thesis, also more data-driven approaches that utilise the phylogenetic information have been suggested. [69]

A special case of balance coordinates is the *pivot* coordinates, where the parts of composition are organised in a particular order and each part of the composition

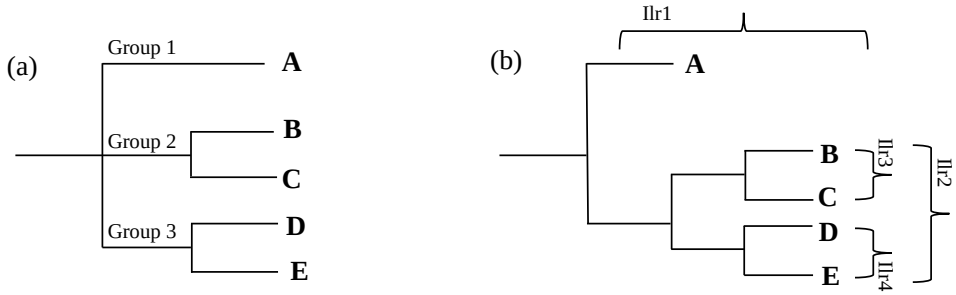


**Figure 7.** An illustrative example of a taxonomic hierarchy among five distinct taxa. (a) Taxonomy describing a known or an assumed relational structure between the five taxa. (b) Hierarchy encoded by SBP matrix (17). The example is from publication I of this thesis

is sequentially contrasted against the remaining parts of the composition. [70] For a $J$-dimensional composition, the $J \times (J-1)$-dimensional pivotal SBP matrix is

$$\boldsymbol{\Psi} = (\psi_{jk}) = \begin{bmatrix} +1 & 0 & 0 & \dots & 0 \\ -1 & +1 & 0 & \dots & 0 \\ -1 & -1 & +1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \vdots & +1 \\ -1 & -1 & -1 & \vdots & -1 \end{bmatrix}. \tag{18}$$

Thus, the pivotal choice of contrasts leads to the following coordinates: [70]

$$m_k = \sqrt{\frac{(J+1)-k}{(J+1)-k+1}} \ln \left( \frac{z_k}{\sqrt[((J+1)-k)]{\prod_{h=k+1}^{(J+1)} z_h}} \right), k = 1, \dots, J.$$

The first pivot coordinate has a particularly straightforward interpretation as the first part of composition contrasted against all other parts, and it is useful in the analysis of compositional data especially if one part of the composition is of interest.

In addition to the more hypothesis-driven approaches presented here and utilised in publication I of this thesis, also data-driven approaches have been suggested. For example, in the context of sequencing data, often characterised by high dimension, the logratios with best predictive abilities can be identified. [71]

## 4.3 Mediation Analysis with Compositional Data

Due to the increasing interest in sequencing count markers as mediators in the effects of exposome on health and disease, and the observation that sequencing count data may be compositional, various approaches have been suggested for mediation analysis with compositional mediators, especially in the context of microbial data. [72; 73; 74; 75] One of these approaches, as well as comparison with some of the other approaches, is presented in publication I of this thesis.

In general, these approaches are based on transforming microbial data from the simplex to ilr or alr coordinates and using the ensuing coordinates as mediators in a multiple-mediator analysis. In addition to taking into account the compositional nature of the data, log-ratio coordinates are better suited for simple regression models than raw sequencing counts characterised by large variability and sparsity.

In publication I of this thesis, both hypothesis-driven taxonomic and pivotal approaches for building mediating ilr coordinates were introduced, corresponding to the matrices (17) and (18). In the former, each coordinate was treated as the mediator of interest whereas in the latter, the first pivotal coordinate was the only causal mediator of interest. It was also noted that while each different contrast matrix used to build ilr coordinates entails a different causal hypothesis and interpretation, the overall indirect effect (*i.e.* the inner product of the two vectors of coefficients) remains the same, as alternative contrast matrices correspond to choosing different orthonormal basis for the simplex. In publication I of this thesis, the mediator-specific indirect effects were defined following Equation (11). Conveniently, this definition complies with the sequential binary partition method used to build the coordinates.

While the approach suggested in publication I of this thesis is based on low-dimensional data and an *a priori* hypothesis, other, high-dimensional approaches have been presented. In these approaches, the aim is often to sieve potential mediating taxa from a large number of candidate mediators. Sieving approaches and regularisation methods, such as Lasso regression, have been suggested to identify the signals of mediation from a high-dimensional composition. [73; 72] In addition to ilr coordinates, also the use of alr coordinates as well as modelling microbial taxa in the simplex space using Aitchison geometry have been suggested. [73; 74] A more comprehensive comparison of the different approaches is presented in Publication I of this thesis.

In previous literature, the use of pivotal contrast matrices in a data-driven manner, rather than hypothesis-driven, has also been suggested. If each part of composition is of interest separately, the pivot coordinates are built $J$ times, using each part of the composition at a time as the pivot element. [72] Mediation analysis is then used to assess the mediating role of each part of the composition. However, as noted in Publication I of this thesis, under these so-called *alternating pivot coordinates*, the sum of individual indirect pivot effects would not sum up to the overall indirect

effect. Each of the $J$ analyses relies on a separate contrast matrix and corresponds to a different hypothesis, aiming to identify or sieve signals of mediation from a number of potentially misspecified matrices.

In Chapter 3, methods for multiple mediators were classified into methods for contemporaneous and causally ordered mediators. In case of compositional mediators, the set of mediators is inherently contemporaneous rather than causally ordered. [76] However, these mediators are correlated (cf. Section 4.4 for their covariance structure), also conditionally on the exposure and confounders. This correlation, although not causal, arises from how the coordinates are built and encompasses also the underlying correlations between the parts of the composition.

One important point noted by Arnold et al. (2020) [76] is that causal inference with compositional data is framed differently depending on whether absolute or relative values are used. The absolute values, such as microbial counts or energy from dietary components, can affect the outcome of interest either directly or indirectly through the increase in the total quantity (*e.g.* total energy intake). However, when scaled to relative abundances, the total becomes a collider variable and builds a dependence between the parts of the composition. Thus, the effect of the relative abundance of one component reflects both the change in that component as well as in all other components.

## 4.4 Asymptotic Normality of Ilr Coordinates

Sometimes in the context of compositional mediation analysis, the normality of ilr or alr coordinates is directly assumed when modelling the relationships between exposure, mediators and outcome based on a standard linear model. [72] However, if the underlying sequencing data exhibit excess variability, this assumption can be wrong. In compositional mediation analysis, the coordinates have a twofold role as both dependent variables (cf. Equation (13a)) as well as explanatory variables for the response (cf. Equation (13b)). Highly non-normal coordinates can thus violate the assumptions of standard linear regression models and hence, in case of extremely non-normally distributed residuals, affect the efficiency of estimation and produce misleading confidence intervals and distort the relationships. The asymptotic normality of ilr coordinates under simple multinomial sampling has been shown before and is reviewed below. [77; 78] In publication II of this thesis, the normality of ilr coordinates was investigated under compositional data exhibiting sparsity and excess variability.

### 4.4.1 Multinomial Counts

The asymptotic normality of the ilr coordinates when the composition is based on multinomial sampling of counts was noted by Graffelman in 2011 [77] and formally proven by Graffelman et al. in 2015 [78]. Assuming a multinomial experiment with $J$ possible categories and total size $K$, denote the vector of counts in the categories as $\boldsymbol{K} = (K_1, \ldots, K_J)$ and their relative frequencies, *i.e.* proportions as $\boldsymbol{f} = (f_1, \ldots, f_J)$; $f_j = K_j/K$, $j = 1, \ldots, J$; $\Sigma_{j=1}^{J} f_j = 1$. The vector $\boldsymbol{f}$ is hence a composition in $\mathcal{S}^J$. The proportions are maximum likelihood estimates of the underlying multinomial probabilities $\boldsymbol{p} = (p_1, \ldots, p_J)$. Asymptotically, the distribution of $\boldsymbol{f}$ approaches a multivariate normal distribution with expected value $\mathbb{E}[\boldsymbol{f}] = \boldsymbol{p}$ and variance-covariance matrix $\mathrm{Cov}(\boldsymbol{f}) = \boldsymbol{\Sigma_f} = \boldsymbol{D}_p - \boldsymbol{p}\boldsymbol{p}'$, where $\boldsymbol{D}_p$ is a diagonal matrix with elements $p_1, \ldots, p_J$.

Following the presentation of Casella and Berger (2002) [79] and Graffelman et al. (2015) [78], the centered asymptotic distribution of maximum likelihood estimator of $\boldsymbol{g}(\boldsymbol{f})$ for a function $g$ that has continuous first partial derivatives is:

$$\sqrt{K}(\boldsymbol{g}(\boldsymbol{f}) - \boldsymbol{g}(\boldsymbol{p})) \approx \mathcal{N}\left(0, \left(\frac{\partial \boldsymbol{g}(\boldsymbol{p})}{\partial \boldsymbol{p}}\right) \boldsymbol{\Sigma_f} \left(\frac{\partial \boldsymbol{g}(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)'\right). \tag{19}$$

When $\boldsymbol{g}$ is the ilr transformation, *i.e.* $\boldsymbol{g}(\boldsymbol{p}) = \mathrm{ilr}(\boldsymbol{p}) = \boldsymbol{V}'\ln(\boldsymbol{p})$, the partial derivatives become:

$$\frac{\partial \boldsymbol{V}'\ln(\boldsymbol{p})}{\partial \boldsymbol{p}} = \boldsymbol{V}'\boldsymbol{D}_p^{-1},$$

where $\boldsymbol{D}_p^{-1}$ is a diagonal matrix with elements $1/p_1, \ldots, 1/p_J$. In this case, the covariance matrix in (19) is $\boldsymbol{V}'\boldsymbol{D}_p^{-1}\boldsymbol{\Sigma_f}\boldsymbol{D}_p^{-1}\boldsymbol{V} = \boldsymbol{V}'\boldsymbol{D}_p^{-1}\boldsymbol{V}$. The covariance matrix of the ilr coordinates thus only depends on the multinomial probabilities $\boldsymbol{p}$ and the contrast matrix $\boldsymbol{V}$. [78]

Based on (19), under large enough total count ($K \to \infty$) it holds for the ilr coordinates that

$$\mathrm{ilr}(\boldsymbol{f}) \to \mathcal{N}(\mathrm{ilr}(\boldsymbol{p}), \frac{1}{K}\boldsymbol{V}'\boldsymbol{D}_p^{-1}\boldsymbol{V}'). \tag{20}$$

### 4.4.2 Counts with Sparsity

Sequencing count data often involve excess heterogeneity compared to purely multinomial counts. Especially taxon-specific probabilities are characterised by excess of zero-count observations (*sparsity*) and, at the extreme levels of such heterogeneity, counts are concentrated into specific classes and hence, the distributions of class-specific proportions may become multimodal. Figure 8 demonstrates how the compositions as whole (left-hand panels) and distributions of the proportions of each

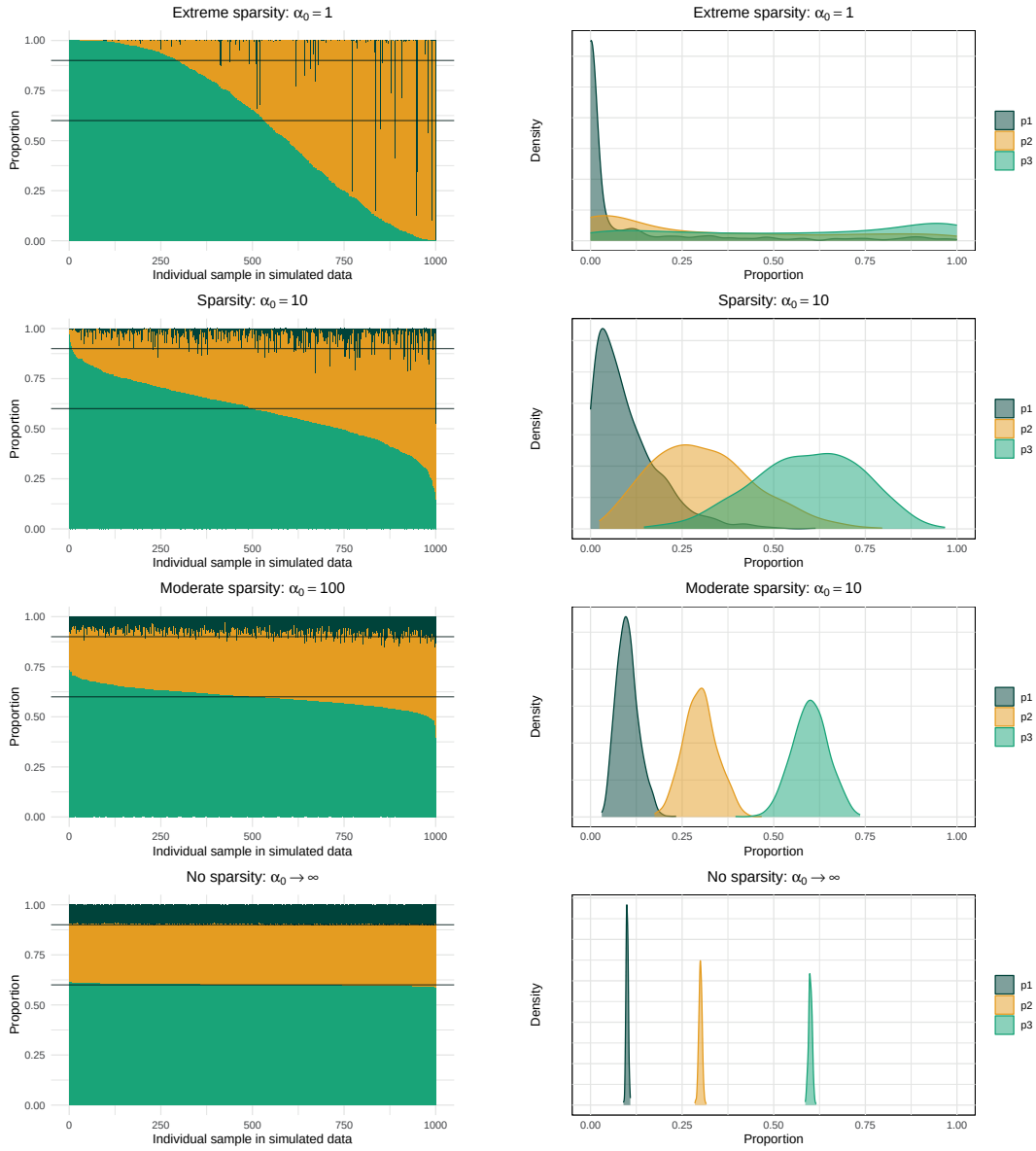


**Figure 8.** Three-part compositions under varying extents of sparsity. For each scenario, $n = 1000$ observations were simulated from the underlying Dirichlet-multinomial distribution with varying extents of Dirichlet-multinomial sparsity (cf. Equation (21)) and class-specific probabilities of 0.10, 0.30, and 0.60. In the left-hand panels, each vertical line corresponds to one observation scaled into proportion. For each observation, the proportion of counts belonging to each of the classes are presented in different colors. The observations are presented in descending order of the third class. The horizontal lines represent the expected proportions. The right-hand figures present the smoothed marginal distributions of the simulated class-specific proportions.

class (right-hand panels) behave under varying extents of sparsity. In each case, the three-part compositions are based on counts in classes that have probabilities of 0.10, 0.30 and 0.60, respectively. When data are extremely sparse, the first of the classes is absent from majority of the samples, and some of the samples consist of counts from only one of the classes. The distributions or the class-specific proportions thus range from 0 to 1 and do not have a clear mode. When sparsity becomes more moderate, the samples become more similar with each other, and especially the two most abundant classes lack zero-count observations. The distributions of the class-specific proportions are well-defined and have modes. With no sparsity, for each sample, the class-specific proportions are very close to the underlying probabilities and the distributions of the class probabilities are extremely concentrated.

In publiction II of this thesis, the asymptotic normality of ilr coordinates was investigated under a general compound multinomial distribution exhibiting excess variability. The derivation of the asymptotic normality of the ilr coordinates under multinomial counts, presented by Graffelman et al. [78] and summarised above, is rather straightforward due to the known asymptotic normality of the proportions. However, under a compound multinomial distribution with extra-multinomial variation in the probabilities, the distribution of the proportions may deviate from the normal distribution.

In order to investigate the normality of the coordinates under overdispersed compound multinomial counts, the normality of the proportions based on the general compound multinomial model was investigated first. Briefly, it was found that when the total of the counts increases and the variation induced by the mixing distribution becomes negligible, the compound multinomial proportions reach asymptotic normality.

In addition to the general compound multinomial distribution, the special case of Dirichlet-multinomial distribution was investigated. Briefly, the Dirichlet distribution describes the variation in (compositional) proportions. [70] The Dirichlet probability density function for $J$ categories is

$$\frac{\Gamma(\alpha_S)}{\prod_{j=1}^{J}\Gamma(\alpha_j)}\prod_{j=1}^{J}p_j^{\alpha_j-1},\ \ j=1,\ldots,J,$$

where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_J)$, $\alpha_j > 0$, $\Sigma_{j=1}^{J}\alpha_j = \alpha_S$ is the vector of the so-called concentration parameters. [80] Their total, $\alpha_S$, defines the sparsity of the distribution: the smaller its value, the more the data are concentrated to a specific class. The expectation of $j$th class is $\alpha_j/\alpha_S$ and the variance-covariance matrix is $(\mathrm{diag}(\alpha - \alpha\alpha'))/(\alpha_S+1)$. The class-specific mode, $(\alpha_j-1)/(\alpha_0-J)$, is only defined when $\alpha_j > 1$.

Dirichlet-multinomial distribution is a compound multinomial distribution where class-specific probabilities are first drawn from the Dirichlet distribution, and the

sample of counts $\boldsymbol{x}$ follows a multinomial distribution with the Dirichlet probabilities and number of trials $K$:

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_J) \\ \boldsymbol{x}|K, \boldsymbol{\pi} &\sim \text{Multinomial}(K, \boldsymbol{\pi}). \end{aligned} \tag{21}$$

For the Dirichlet-multinomial distribution, $\mathbb{E}[x_i] = K\alpha_j/\alpha_S$ and $\text{Cov}(\boldsymbol{x}) = K\frac{K+\alpha_S}{1+\alpha_S}(\text{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}')$. Again, the parameter $\alpha_S$ controls the sparsity, and when $\alpha_S \to \infty$, the Dirichlet-multinomial distribution approaches the multinomial distribution. [81]

Based on the results for general compound multinomial distribution, in Publication II the normal approximation was formulated for Dirichlet-multinomial proportions when both $K \to \infty$ and $\alpha_S \to \infty$ under two scenarios: $K/\alpha_S = o(1)$ and $\alpha_S/K = o(1)$, *i.e.* when $\alpha_S \to \infty$ faster than $K$ or when $K \to \infty$ faster than $\alpha_S$. The normal approximation for the proportions $\boldsymbol{p}$ based on Dirichlet-multinomial counts is:

$$\boldsymbol{p} \approx \mathcal{N}\left(\alpha, \frac{1}{K}\left(\frac{\alpha_S + K}{\alpha_S + 1}\right)(\text{diag}(\alpha) - \alpha\alpha')\right) \tag{22}$$

when $\alpha_S \to \infty$ and $K \to \infty$, irrespective of which converges faster.

Based on the result (22), following the approach by Graffelman et al. (2015) [78], in Publication II of this thesis, the asymptotic normal approximation of ilr coordinates when the counts follow a Dirichlet-multinomial distribution was derived. This asymptotic normal approximation for the coordinates is

$$\text{ilr}(\boldsymbol{p}) \approx \mathcal{N}(\text{ilr}(\alpha), \boldsymbol{V}'\boldsymbol{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{p}}\boldsymbol{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{V}), \tag{23}$$

where the variance-covariance matrix can be also expressed as $\frac{1}{K}\frac{\alpha_S+K}{\alpha_S+1}\boldsymbol{V}'\boldsymbol{D}_p^{-1}\boldsymbol{V}$. Due to the excess variation in $\boldsymbol{p}$ induced by sparsity, it was noted that in practice, $\text{ilr}(\alpha)$ may considerably deviate from $\mathbb{E}[\text{ilr}(\boldsymbol{p}))]$. To overcome this, it was furthermore suggested using a first-order Taylor approximation for the log-transformation.

The performance of the normal approximation (23) was investigated using a simulation study under varying extents of sparsity. When the total count $K$ was large enough and the extra-multinomial variability was not extreme, ilr coordinates were indeed found to follow well a normal distribution. Furthermore, the approximation of Equation (23) perfomed well and under any amount of extra-multinomial variability, better than the normal approximation presented in Equation (20) by Graffelman et al. [78]. The structure of the composition and choice of the contrast matrix affected the performance, and especially if the coordinates involved rare classes that were more likely to involve (an excess amount of) zero-count observations, the moments of the normal approximation did not correspond to simulated data so well.

In addition to extra-multinomial variability, the variability in the total count $K$ was also investigated. In sequencing studies, the total count is often highly variable,

although in the analysis of compositional data it is discarded when the counts are scaled into relative abundances. While the total count is non-informative, it can affect the precision of the estimated proportions. However, inducing additional variation to $K$ did not largely affect the performance of the approximation.

## 4.5 Zero-Count Observations

The previous sections have relied on the assumption that a composition consists of strictly positive values, *i.e.* the data do not contain zeroes. In practice, zero-count observations are common in real-world data, such as sequencing counts. As the logarithm of zero is undefined, the log-ratio methods are not directly applicable to compositional data with zero-count observations. Instead, zeroes need to be treated in an appropriate manner. Zero values can occur in compositional data for multiple reasons, and hence multiple approaches on how to handle zeroes exist.

### 4.5.1 Types of Zeroes

In the literature of compositional data analysis, at least three types of zero observations have been introduced, even though the terminology varies. The following terminology follows the presentation of Pawlowsky-Glahn et al. (2015) [62] and Pawlowsky-Glahn and Buccianti (2011) [82].

**Rounded zeroes.** Rounded zeroes refer to values that are reported as zero even though the underlying data contain a non-zero, such as values below the detection limit. These zeroes are thus considered as indicating left censoring.

**Structural zeroes.** Structural zeroes, sometimes also called as essential or absolute zeroes, refer to components that are truly zero. A classical example of structural zeroes in compositional data analysis literature is the household budget patterns, where a subset of households report a zero-budget for *e.g.* tobacco products. Similarly, in microbial data, various taxa can be expected to be entirely absent from a subset of samples. [83]

**Count zeroes.** Count zeroes are zero-count observations that can occur due to a limited sample size. For example, in microbiome studies, certain taxa remain unobserved in some samples due to a limited sequencing depth, but could have been observed if the total count was larger. Being largely dependent on the sample size, the interpretation of count zeroes can be either rounded or structural: with a large sample size, zero-count is more likely structural whereas under smaller sample size the possibility of rounded zeroes cannot always be ruled out.

### 4.5.2 Treatment of Zeroes

**Replacement.** One widely-used solution for zero-count observations in sequencing count data is to replace them with a fixed value, so-called pseudo-count. [84] Quite frequently, a fixed value of 0.5 is used. [73; 72] However, as this approach may distort the structure of the data, other imputation techniques have been suggested, such as multiplicative zero replacement, which aims to preserve the compositional structure of the data. The zeroes can, for example, be imputed in a Bayesian manner by replacing them by their posterior expectations under a Dirichlet prior. [85] This imputation choice is claimed to preserve better ratios between the parts of the composition. [86] The zero-replacement and imputation techniques suit best for such zeroes that can be considered as rounded (count) zeroes. As such zeroes are not considered as "true" zeroes, replacing them with a well-justified small value can be a sensible solution. [84; 87].

In publication II of this thesis, the distributions of the ilr coordinates were investigated under both fixed pseudo-counts and using the Bayesian multiplicative zero replacement. Unsurprisingly, the normal approximation of the ilr coordinates corresponded better to the empirical moments under Bayesian zero treatment than to the empirical moments under fixed pseudo-counts. The fixed pseudo-counts may distort the structure of the data and ratios, and thus using more sophisticated approaches could be more adequate. [86]

**Filtering.** In some cases, removing very sparse, zero-inflated classes (and then re-scaling the remaining classes into a new composition) may be justifiable. For example, in the context of microbiome data, in publication II of this thesis it was suggested that the taxa that are non-unimodal due to their sparsity should be discarded prior to compositional data analysis using log-ratio coordinates, especially if modelling will be based on assuming normality. However, in some cases if sparse taxa are informative although being sparse, it may be meaningful to retain them in the data and rely on models that accommodate highly sparse overdispersed data. In general, taxa can be filtered out based on some heuristics, for example due to a low abundance or prevalence. Keeping only classes that can be considered as informative can improve the reliability of the results of statistical analyses. [88]

**Amalgamation.** Amalgamation, *i.e.* summing parts of the composition together, has been suggested as a method to reduce the dimension of compositional data and a tool to investigate relationships between subcompositions. [68] However, amalgamation is non-linear in the simplex: for example, it does not preserve Aitchison distances. Thus, when analysing amalgamations as well as the original non-amalgamated parts of the composition, the results of the two analyses may lead to incoherent interpretations. [68] It is important to note that amalgamations should be applied over all observations in the data. If classes are amalgamated over only a subset of observations (*e.g.* in elections, parties form a coalition in only some

municipalities), data analysis becomes problematic. [62]

**Presence vs. absence.** Especially essential zeroes can also be considered as grouping factors so that the rest of the composition is stratified based on the presence or absence of one part and then continuing the analysis of the remaining parts of the composition in a standard manner. In the context of microbiome data, it has been suggested that investigating the presence/absence of taxa using logistic regression yields more replicable results than some benchmark methods for analysis of microbial abundances. [83] While the compositional mediation analysis methods presented here rely on compositional log-ratio transformations of the microbial counts, mediation analysis can be also extended for dichotomous mediators and thus, it is possible to treat the presence/absence of a taxa as a mediator. [35]

# 5 Applications on Empirical Data

## 5.1 Mediating Role of the Gut Microbiome Between Fibre Intake and Insulin Levels

### 5.1.1 Background

Publication I of this thesis focuses on developing a compositional mediation analysis approach to quantify the mediating role of the gut microbiome in the effects of exposures on health, utilising *a priori* knowledge of the potential mediators and their taxonomic interrelationships. The motivating question concerns the role of the microbiome as a mediator of the effects of sufficient fibre intake on insulin levels. [12; 89] The gut microbiome is sensitive to changes in dietary patterns, and various dietary components as well as the overall diet have been found to influence the gut microbiome composition and abundance of numerous microbial species. These, in turn, can affect for example inflammation, insulin sensitivity, adiposity and, subsequently, cardio-metabolic health. Overall, the gut microbiome has been linked with various phenotypes and diseases, such as obesity and type 2 diabetes and, through gut-brain axis, even neurodevelomental outcomes. It thus serves as a plausible mechanistic link, *i.e.*, a mediator, between the exposome and health. [90]

### 5.1.2 Study Setting

The hypothesis was investigated using data from The Special Turku Coronary Risk Factor Intervention Project (STRIP) study, which is a prospective randomised intervention study promoting heart-healthy diet with a specific focus on low intake of saturated fat and cholesterol. [91; 92] The subjects ($n = 1116$), born in 1989–1991, were recruited from well-baby clinics in Turku at the age of 5 months. The intervention group was followed up and received intervention at 1–3 month intervals until age 2, and thereafter twice a year, while the control group was followed up twice a year until age 7 and yearly thereafter. The intervention period and regular study visits lasted until the participants were 20 years of age. [91] The first follow-up visit took place 6 years after the intervention period had ended when the participants were 26 years of age. [92]

In publication I of this thesis, data from the age-26 follow-up were used. Prior to the study visit, the participants filled a four-day food diary, including at least one

weekend day. Based on the diary, fibre intake was quantified and, in the analyses, treated as binary indicator of sufficient fibre intake ($\geq 25$ grams per day). [93; 94] Insulin levels were measured from the serum samples taken at the follow-up visit and treated as $\log_e$ transformed in the statistical analyses. After excluding participants with missing data on diet, faecal sample or insulin levels and those who had type 1 diabetes, did not fast before the blood sample or had used antibiotics in the past three months, or were obese, the analysis sample size was $n = 264$. Sex and intervention group were considered as confounders in the analyses. Furthermore, a sensitivity analysis for physical activity as an unmeasured confounder was conducted.

The raw gut microbiome sequence data were processed into an amplicon sequence variant (ASV) table that describes the abundances of distinct microbial sequences. [95; 96] The ASVs were further taxonomically classified into phyla, classes, orders, families, and genera. The individual-specific total read counts varied between 11800 and 839000, the median being 160000. [95] In total, 6591 unique ASVs were identified, assigned to 20 phyla and 291 genera. The genus-level dataset contains information on how many times different microbial genera were found in each participants' faecal sample. In this thesis, genera from the *Actinobacteria* phylum were investigated as mediators.

### 5.1.3 Statistical Analysis and Results

In the empirical analyses of publication I, the mediating role of the gut microbiome between fibre intake and insulin levels was investigated. The *Actinobacteria* phylum was considered as the *a priori* mediator of interest. The use of both taxonomic and pivotal SBP matrices was demonstrated by investigating two hypotheses related to the role of the *Actinobacteria* phylum and and the genera within it. First, the genera within the *Actinobacteria* phylum were considered in an intra-group analysis, where the genera were contrasted against each other using a SBP matrix based on the taxonomy within the phylum. Second, an inter-group analysis was conducted by building a pivotal SBP matrix, using *Actinobacteria* phylum as the pivot element, contrasted against the other phyla.

The inter- and intra-group mediation analyses highlight differences between coordinates based on taxonomic (cf. Matrix (17)) and pivotal (cf. Matrix (18)) SBP matrices. In the intra-group analysis, where the coordinates were built based on taxonomic knowledge, each ilr coordinate was given a relevant biological interpretation. Each of the coordinates was also treated as a potential causal mediator in a multiple-mediator analysis, and both the mediator-specific and overall indirect effects were of interest (cf. Equations (11) and (10) of Chapter 3). On the contrary, in the second analysis with the pivotal SBP matrix, only the first pivot coordinate was of interest, while the remaining coordinates were treated as nuisance parameters.

In the simulation studies conducted in publication I, the sparsity of counts was

found to affect the performance of the mediation analysis in a complex manner. Briefly, the presence of sparsity increases the variation of the coordinates. Subsequently, the standard error of the effect of exposure on mediator ($\sigma_{\beta_1}$) increases whereas the standard error of the effect of mediator on outcome ($\sigma_{\gamma_2}$) decreases. Hence, both extreme cases of sparsity (*i.e.*, extrene sparsity and high variability of coordinates, and multinomial counts and very little variation in the coordinates) led to inflation in one of the coefficient-specific standard errors and deflation in the other and thus affected the statistical power to identify the indirect effects in an unexpected manner. To avoid excess sparsity in the empirical analyses, taxa that appeared in less than 10 % of the samples were filtered out. To be able to calculate the isometric log-ratio coordinates, the remaining zero-count observations were replaced with a pseudo-count of 0.5.

In the empirical mediation analysis, sufficient fibre intake was found to be associated with insulin levels, log-transformed insulin levels being $0.106$ units lower in individuals with sufficient fibre intake ($90\%$ CI $[-0.202; -0.011]$).

Of the mediators in the intra-group analysis, the most prominent was a logratio that contrasted the genera *Enterohabdus* to the other genera within the order *Coriobacteriales*. The ratio was larger within participants who had higher fibre intake ($\beta = 0.409$, $[0.060; 0.759]$). In addition, this log-ratio coordinate was associated with lower log-transformed insulin levels ($\gamma = -0.041$, $[-0.074; -0.008]$. The indirect effect of this log-ratio coordinate was $-0.017$, $[-0.036; 0.003]$. Of note, while the MaxP test here would suggest presence of mediation, the Sobel-type confidence intervals do not support this conclusion. Some of the mediator-specific effects counteracted, having different directions of mediation effects. Thus, in the intra-group analysis, the overall indirect effect of the genera within the *Actinobacteria* phylum was $-0.011$, $[-0.073; 0.051]$.

In the inter-group analysis, sufficient fibre intake was associated with the log-ratio coordinate of the entire *Actinobacteria* phylum contrasted against the other phyla ($-0.451$, $[-0.762; -0.140]$). The effect of this pivot coordinate on the log-transformed insulin levels was $0.042$ $[-0.004; 0.088]$ and thus the indirect effect for this mediator was $-0.019$ $[-0.043; 0.006]$. No mediation via the *Actinobacteria* phylum contrasted agains the other phyla was concluded.

## 5.2 Mediating Role of DNA Methylation Between Environmental Toxicant Exposure and Type 2 Diabetes

### 5.2.1 Background

Epigenetic mechanisms regulate gene expression without alterations in the underlying DNA sequence. While the genetic sequence remains unchanged through the life, epigenetic mechanisms are dynamic, playing a pivotal role during development in cell differentiation. [97] The dynamic changes in the epigenome continue throughout the life-course in response to aging as well as external stressors. [15] Epigenetic mechanisms include DNA methylation, histone modifications, and small RNA molecules, each of them affecting gene expression in concert by up- or downregulating gene activity. [97] Of these, DNA methylation, *i.e.* the addition of a methyl group to a DNA molecule, is perhaps the most widely studied. [15] In addition to playing a role in development, epigenetic mechanisms are also involved in health and disease, especially non-communicable diseases such as cardio-metabolic and autoimmune diseases, as well as cancer. [13; 14] Due to their responsiveness to the exposome as well as their role in disease pathogenesis and health, epigenetic mechanisms serve as a plausible link between the two. [16]

In publication III of this thesis, the mediating role of DNA methylation in the association between exposure to environmental toxicants and type 2 diabetes (T2D) is investigated. Environmental toxicants, such as polychlorinated biphenyls (PCBs), can function as endocrine-disruptors, and mounting evidence suggests their role as both obesogens and diabetogens. [98; 99] Variation in DNA methylation has been linked with PCBs and other environmental toxicants [100], and the (causal) role of DNA methylation in the pathogenesis of T2D has also been established. [101; 102] Thus, DNA methylation is a plausible mediator between exposure to environmental toxicants and cardio-metabolic health.

### 5.2.2 Study Setting

The Cardiovascular Risk in Young Finns Study (YFS) is a prospective study initiated in 1980 to follow risk factors of cardiovascular disease in Finnish children and adolescents in the five Finnish cities with medical schools (Helsinki, Kuopio, Oulu, Tampere, Turku) and the rural areas nearby. [103] The baseline study included $n = 3596$ participants, males and females, who were 3, 6, 9, 12, 15 and 18 years old in year 1980. The YFS cohort has been subsequently followed up in the years 1983, 1986, 1989, 1992, 2001, 2007, 2011 and 2018. Figure 9 presents an overview of the study follow-up visits and participant ages in each of the six birth-year cohorts.

In publication III, data from the follow-up visits from year 2001 onward were

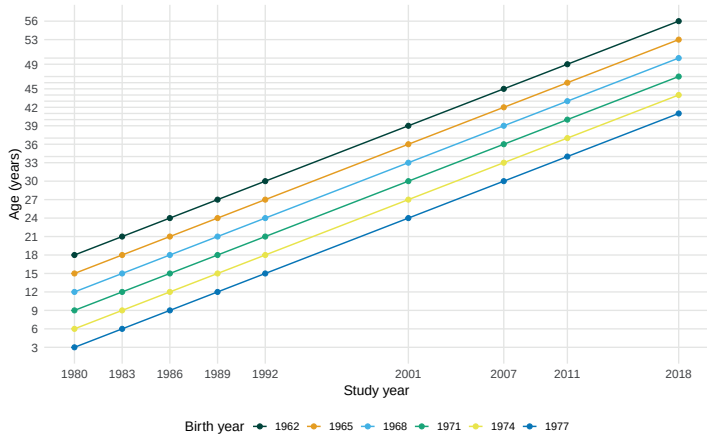


**Figure 9.** Overview of the follow-up of the Cardiovascular Risk in Young Finns Study, showing the ages of the participants in the six birth-year cohorts at each of the 9 study visits.

used. Serum PCB concentrations were measured in 2001 as proxies for participants' exposure to these chemicals. Altogether ten PCB congeners were assessed. Because the congeners were highly correlated, principal component analysis was applied to reduce the dimensionality of the exposure variables. The first three principal components (PC), which together explained 96.1 % of the total variance, were used in the subsequent mediation analyses. All congeners had positive loadings of approximately 0.30 on the first PC, which describes the overall exposure level. The second PC was characterised by high positive loadings of low-chlorinated PCBs and negative loadings of certain higher-chlorinated congeners, thus capturing differences between the low-chlorinated (lower half-lives and less correlated with age) and high-chlorinated (more resistant to degradation and more correlated with age) congeners.

Type 2 diabetes follow-up data were used from year 2011 onward. The diagnostic criteria included fasting glucose $\geq$ 7 mmol/L, haemoglobin A1c level $\geq 6.5\%$, self-reported T2D, self-reported use of glucose-lowering mediation, diagnoses in registry data, and registered drug reimbursement. T2D diagnosis was treated as binary. The analysis sample consisted of $n = 1216$ participants, of whom 71 were T2D cases.

Genome-wide DNA methylation levels were assessed from whole blood samples taken at the 2011 follow-up visit. After pre-processing, altogether 770881 probes remained in the data set, each pertaining to a specific CpG site. The level of DNA methylation, called $\beta$ value, represents the proportion of methylation detected at the specific CpG site within a sample, calculated by dividing the methylated signals from the sample by the sum of methylated and unmethylated signals. The $\beta$ value is constrained between 0 and 1 and is often concentrated near the extreme values, showing multimodality. Thus, for statistical analyses, the $\beta$-values were transformed into M-values by $M = \log_2(\beta/(1-\beta))$. [104]

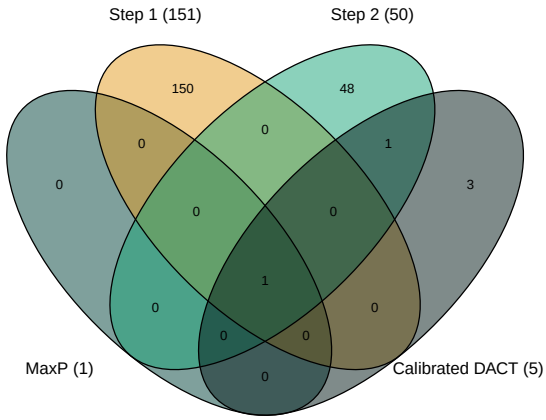


**Figure 10.** Venn diagram describing the overlap in the number of CpG sites with p-value $< 0.00005$ in Step 1 EWAS, Step 2 EWAS, MaxP test and DACT test.

### 5.2.3 Statistical Analysis and Results

Briefly, the analysis of total effects showed that especially the second principal component (PC2) as well as the three low-chlorinated PCB congeners that had high loadings to PC2, PCBs 74, 99 and 118, were associated with an increased risk of T2D. Subsequently, the mediation analyses focused on the second principal component as an exposure and DNA methylation as the mediator. Two large ($p = 770881$ variables) epigenome-wide association studies were conducted for the effects of the principal components on DNA methylation (Step 1 EWAS) and for the effects of DNA methylation on risk of T2D (Step 2 EWAS). The divide-aggregate composite null test (DACT) was used to identify potential mediators. [41] Altogether five mediating CpG sites were identified, including CpG sites from the *CPT1A* and *ABCG1* genes, which have been linked with T2D in previous studies. A multiple-mediator analysis was conducted to quantify their combined indirect effect. Together, DNA methylation in these five CpG sites corresponded to a proportion mediated of 40 %. While these results were reassuringly consistent with some previous studies, it is noteworthy that the number of T2D cases in our sample was relatively small for a large epigenome-wide association study compared with prior literature. [101]

The Venn digram of Figure 10 presents the overlaps in the number of CpG sites with p-values smaller than the nominal significance level of 0.00005 for the two distinct EWASs (Step 1 and Step 2), MaxP test, and DACT. One of the identified CpG sites (cg00574958) was identified in each of the four sets. In addition, DACT was able to identify four additional CpG sites as potential mediators, whose p-values

were not below the nominal threshold in the path-specific EWASs.

To gain better insight into the behaviour of DACT-derived p-values obtained in publication III, Figure 11 presents the empirical distributions of the p-values from the two EWASs (first row), from the MaxP test (second row), and both raw and corrected DACT p-values (third row). The EWAS p-values (first row) appear slightly enriched towards 0, indicating some signals of association. Based on these two sets of p-values, the proportions of true nulls were estimated as 0.84 and 0.95 for the PC2 and T2D EWASs, respectively. Obviously, the distribution of the p-values based on the MaxP test (second row, left panel) is extremely right-skewed, demonstrating its conservativeness in concluding mediation. The distribution of the squared MaxP p-values ($\text{MaxP}^2$) (second row, right panel), on the other hand, resembles more those of Step 1 and Step 2.

Based on the null proportions for step 1 and step 2 EWAS, the probabilities of the three composite null cases (cf. Equation ((8)) presented in Equation (14) were respectively 0.154, 0.0404 and 0.798, the non-null probability being estimated as 0.0078. The DACT p-values, built as linear combinations of path-specific p-values and squared MaxP p-values, using weights based on the probabilities, exhibits a nearly-uniform behaviour. However, the DACT p-values that are corrected using the empirical null framework to account for the correlatedness between the methylation sites, are more strongly enriched towards 0.

## 5.3 Paternal Sperm Non-Coding RNAs as Mediators Between the Paternal Exposome and Offspring Health

### 5.3.1 Background

Beyond the life-course exposome of an individual, also embryogenesis and *in utero* conditions have been shown to influence health in adulthood. [105] To most extent, the role of mothers has been a focus in studies of developmental origins of health and disease, with plenty of evidence pointing to the adverse effects of various maternal exposures during pregnancy on offspring health. [105] For example, the offspring of mothers who suffered from the famine during pregnancy had an increased risk of for example cardio-metabolic disease, potentially due to developmental programming. [106; 107] Various other *in utero* exposures, such as smoking, environmental toxicants, and traumatic experiences have also been associated with a wide range of offspring health phenotypes. [108; 109; 110; 111; 112] Even transgenerational effects, persisting to subsequent generations, have been observed. [113; 114]

In addition to intrauterine conditions, paternal contributions to offspring health have been acknowledged and are gaining increasing interest. [17] In brief, the hypothesis of paternal epigenetic inheritance refers to non-genetic inheritance of phe-

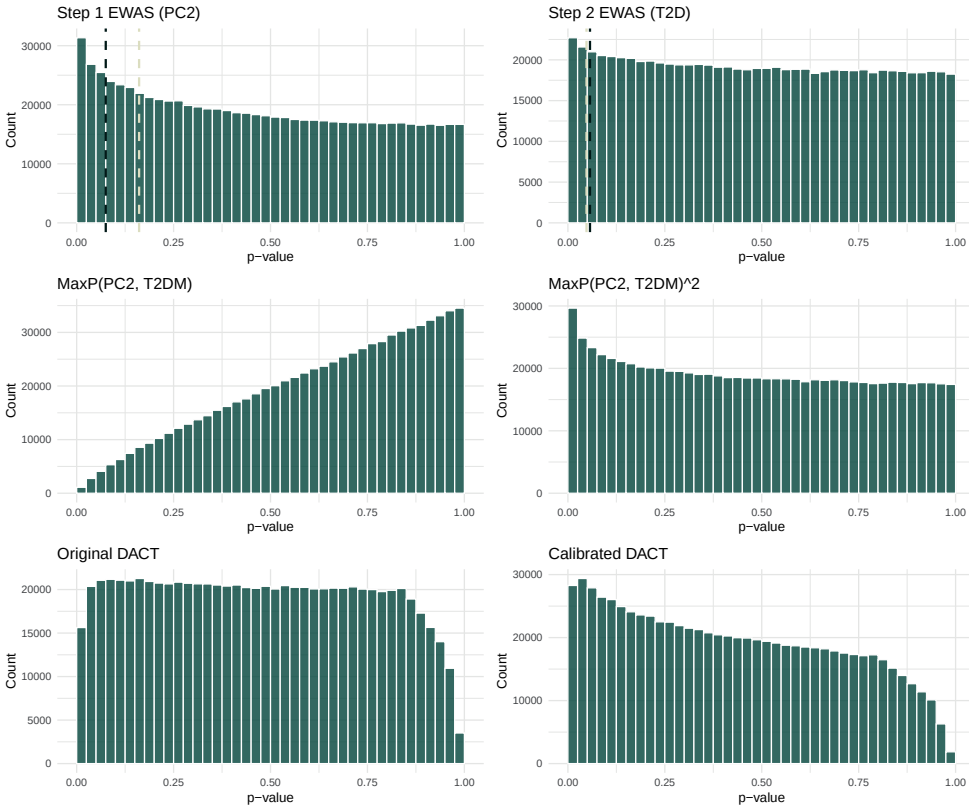


**Figure 11.** Distributions of the p-values for Step 1 EWAS (PC2), Step 2 EWAS (T2D), MaxP test and its square (as used in DACT), and DACT p-values both without and with the empirical null distribution correction. For the Step 1 and Step 2 EWAS p-value distributions, the dark line represents p-values that were $< 0.05$ and the light line p-values that were below the empirical non-null proportions.

notypes and diseases via germ cells, *i.e.* paternal sperm. This type of inheritance is not based on changes in the DNA sequence but instead relies on epigenetic marks in the sperm that contribute to gene expression. [115] The epigenetic information carriers of interest include DNA methylation, histone modifications and sperm non-coding RNA molecules that can contribute to early emryonic development. [116] While research often focuses on adverse effects of paternal exposures, they can also confer adaptive benefits to subsequent generations. [17]

In animal studies, epigenetic inheritance has been demonstrated for induced traits, such as pre-diabetic phenotype, as well as for the effects of exposure to various stressors, such as environmental toxicants. [117; 118; 119; 120] Some evidence on human studies also suggests that paternal exposures, such as abundance or surfeit of food, smoking especially at puberty, as well as environmental chemicals, can influence both paternal sperm epigenome and health in the subsequent generations. [121; 122; 123; 124; 18] However, to date, no single study in humans has been able to combine data of paternal exposures, offspring health, and the paternal sperm epigenome, and thus paternal epigenetic inheritance in humans remains to be established. [18]

From the perspective of statistical analysis, the role of paternal sperm epigenome in epigenetic inheritance, *i.e.* in the association between paternal exposures and offspring health, can be thought as a mediation question and thus investigated using mediation analysis approaches applied to omics markers.

### 5.3.2 Study Setting

The MULTIEPIGEN project is a three-generational extension of the YFS, where, in addition to the original study participants, denoted as G1 ($n = 2127$), also their parents (G0, $n = 2452$) and offspring (G1, $n = 2762$) took part in the follow-up study of year 2018. The aim of the project is to investigate familial transmission of phenotypes and inter- and transgenerational effects of paternal exposome on offspring health via sperm epigenetic mechanisms. The dataset includes 2385 separate three-generational triads (G0-G1-G2) from 817 families. There were in total 1879 G0-G1, 2499 G1-G2 and 2520 G0-G2 dyads.

Previously, life-course exposome data, including smoking, environmental toxicants, socio-economic status, life style habits, and cardiometabolic health, have been collected from the G1 participants. Life-style and socio-economic factors from the G0 participants have been measured at the early follow-up visits during the childhood/youth of G1 participants. Current exposure to various stressors and life style habits was assessed from all participants in the 2018 follow-up. In addition, history of tobacco smoking and stressful or traumatic life events were assessed from the G0 and G1 participants using a retrospective questionnaire. In these questionnaires, participants were asked to report stressors with respect to the development and life-

course stages of their participating offspring (around the time of conception, during pregnancy, early childhood, and later childhood/adolescence of the offspring). The pre-specified offspring phenotypes on cardio-metabolic health and subclinical atherosclerosis, cognitive function, and psychosocial well-being were also measured from all participants at the 2018 follow-up.

In publication IV of this thesis (cohort profile), the multigenerational field study is reported in detail. The supplementary statistical analysis plan characterises how the inter- and transgenerational questions can be addressed as mediation questions. [125] The intergenerational questions here refer to transmission of the paternal exposures to the subsequent generations via the sperm epigenome. In this scenario, the father and his sperm cells are exposed to the stressors of interest, and the subsequent generation is therefore considered "directly" or intergenerationally exposed. The transgenerational questions extend this framework to multiple successive generations, and the effects of grandfathers (G0) exposure levels on G2 phenotypes are of interest. In this case, the sperm cells of the G0 participants are exposed, while those of the G1 generation are not considered directly exposed and thus, the G2 generation is not considered directly exposed either.

**Sperm non-coding small RNA molecules** Altogether $n = 1410$ semen samples were received. From these samples, altogether 823 were sequenced to obtain data on sperm small non-coding RNAs (sncRNAs) (G0 $n = 111$, G1 $n = 418$, G2 $n = 294$). The remaining samples were not sequenced mostly due to biological reasons, such as insufficient amount of RNA, low sperm count, or too much somatic cells.

The sperm small non-coding RNA molecules are short RNA molecules that do not participate in protein synthesis but can play a role *e.g.* in gene regulation and expression by RNA interference. [126] For each participant, four distinct sets of non-coding RNA molecules were obtained: transfer RNA-derived fragments (tsRNAs), micro RNAs (miRNAs), piwi-interacting RNA (piRNA) clusters, and piRNA sequenes. While each of these sncRNA types have been suggested to play a role in offspring development and outcomes, especially tsRNAs are enriched in sperm and have been proposed as key contributors in inter- and transgenerational inheritance. [127; 115]

Table 1 presents for each sncRNA type the dimension of the data set (number of variables), characteristics of the distribution of the total count, and summaries of proportions of variable-specific zero-count observations. The tsRNAs and piRNA clusters are clearly less sparse, the median proportion of zero-counts for tsRNAs being 10 % (IQR 1 % – 62 %), whereas the miRNAs and piRNA sequences are characterised by higher sparsity, *i.e.* larger number of zero-count observations. In fact, the vast majority of the nearly 300000 piRNA sequences assessed are zeroes for nearly all samples. Of note, the counts of the sncRNAs are often scaled into reads per million, giving the data compositional interpretation. [128; 129]

**Table 1.** Characteristics of the four distinct sets of non-coding RNA molecules in the three generations of males in the MULTIEPIGEN project ($n = 823$).

| RNA class | Dimension | Median | IQR | Min | Max |
|---|---|---|---|---|---|
| tsRNA | 276 | 457602 | [457602; 704531] | 4851460 | 4851460 |
| prop zeroes | 276 | 10 % | [1 %; 62 %] | 0 % | 99.4 % |
| miRNA | 1218 | 137692 | [137692; 221780 ] | 4117 | 1634365 |
| prop zeroes | 1218 | 89 % | [41 %; 97 %] | 0 % | 99.9 % |
| piRNA clusters | 180 | 56573 | [30187; 77280] | 214 | 654512 |
| prop zeroes | 180 | 4 % | [0.6 %; 19 %] | 0 % | 58 % |
| piRNA sequences | 262320 | 67168 | [35488; 119369] | 225 | 816549 |
| prop zeroes | 262320 | 99 % | [96 %; 99%] | 0 | 99.9% |

### 5.3.3 How to Investigate Multigenerational Questions?

The statistical analysis plan (SAP), presented as a supplementary material of publication IV of this thesis, details how the multigenerational epigenetic hypotheses motivating the MULTIEPIGEN project can be framed as mediation questions and analysed using mediation analysis approaches. The analysis plan is mostly based on the methodological results from publications I–III and therein we suggest using the compositional and multiple mediator frameworks in analysing the sperm non-coding RNA molecules as intermediates between the paternal exposome and offspring phenotypes. Briefly, each of the four separate sets of the ncRNAs will be investigated separately using mediation analysis within the compositional data analysis framework. As outlined in Table 1, the dimension of each separate dataset varies from some hundreds to hundreds of thousands potential molecules, and thus, approaches for high-dimensional mediators and multiple-mediator mediation analysis are required. The extent of sparsity varies between the datasets, tsRNAs and piRNA clusters having, on average, less sparse variables, whereas in the miRNAs and piRNA sequences most variables have large numbers of zero-count observations.

However, to investigate epigenetic hypotheses in an observational multigenerational setting, aspects beyond the mediating role of omics markers, relating to *e.g.* the structure of the dataset, need to be taken into consideration. The analysis plan included in publication IV also considers these aspects from the perspective of the three-generational MULTIEPIGEN dataset. In addition, a few more general thoughts are provided below.

In multigenerational studies, the sampling schemes and data structure require special methodological considerations. The correlatedness of phenotypes within families needs to be accounted for. The cluster (family) sizes can be informative if an underlying factor or the exposure of interest affects both the offspring phenotypes as well as the number of offspring, for instance due to subfertility. [130; 131] This could manifest as selection bias require specific statistical considerations to avoid bi-

ased estimates and misinterpretation. [131; 132; 133] The family data often remain incomplete in multigenerational settings. For example, in the MULTIEPIGEN study, only the G1 index participants were recruited while data from the second parent of the G2 participants were not obtained and hence, contributions from G1 mothers and fathers simultaneously cannot be investigated. In addition, the siblings of G1 participants were not studied.

In retrospective assessment of the exposome, recall bias is possible. Also confounding bias can occur if, for example, environmental and social factors, shared across the generations, are unmeasured. [134] Both the shared environment as well as genes need to be taken accounted for. Not only inherited alleles but also alleles not transmitted from the parent to the offspring can influence the offspring phenotypes due to genetic nurturing, *i.e.* the parents' genetic influence to the offspring environment. [135]

The timing of measurements also plays a key role. To rule out reverse causality, there should be a clear temporal order of paternal exposure, mediating epigenetic markers, and offspring phenotype. Due to the dynamic nature of the epigenome, in the optimal case epigenetic markers would be measured at the time of conception. However, obtaining human data from pre-conceptional exposome to sperm epigenome at the time of conception and long-term health of subsequent generations requires decades of effort and hence, such data are currently unavailable.

Mendelian randomisation has been suggested as a tool to strengthen causal inference in epigenetic mediation, and genetic instruments for especially DNA methylation are widely known. [136] While Mendelian randomisation has been recently applied to circulating miRNAs, to date, large population cohort studies with both genetic information and data on sperm sncRNAs are sparse and thus, to our knowledge, instrumental variables for the sperm ncRNAs are not available. [137] Specifically, when investigating transgenerational effects (*e.g.* the effect of G0 exposure on G2 phenotype), the G1 exposome, epigenome and phenotype represent an alternative pathway that must be considered in order to assess the causal effect of interest.

# 6 Concluding Remarks

This thesis was motivated by the question of epigenetic inheritance, namely how the paternal exposome may contribute to offspring health via sperm epigenetic modifications. Due to the overdispersed and compositional nature of sequencing count data, frameworks of mediation analysis, particularly those designed for multiple and high-dimensional mediators, were applied within the framework of compositional data analysis to enable treating non-normally distributed sequencing counts, constrained by their reading depth, as mediators in a formal mediation analysis framework. Publications I, II and III of this thesis focus on those aspects of mediation analysis and compositional data analysis that are relevant in investigating the mediating role of omics markers. They thus provide methodological grounds for investigating the motivating research questions about paternal contributions on offspring health via the sperm epigenetic markers, especially the non-coding RNAs, as outlined in publication IV.

The contributions of this thesis are briefly as follows. An approach for causal hypothesis-driven mediation analysis with compositional mediators was developed and its performance under different underlying data-generating assumptions was investigated using a simulation study (publication I). The applicability of the compositional data analysis framework for sparse sequencing count data was investigated and conditions under which the compositional isometric log-ratio coordinates are asymptotically normal were derived (publication II). The appropriateness of the normal approximation for a special case of sparse counts was investigated using a simulation study (publication II). The methods for compositional mediation analysis, mediation analysis with multiple mediators, sensitivity analysis for unmeasured confounding, and high-dimensional mediator identification were applied on empirical data sets (publications I and III), and novel CpG sites mediating the effects of environmental toxicants to type 2 diabetes were identified (publication III). Finally, the lessons learned from both the theoretical and empirical aspects of studies I–III were synthesised in a statistical analysis plan outlining how inter- and transgenerational questions of can be investigated within the three-generational MULTIEPIGEN project (publication IV).

The methods presented in this thesis are applicable for a wide range of omics markers. Furthermore, the collection of approaches are broadly applicable across multiple stages of the research continuum, from the hypothesis-generating exploratory

sieving studies, which aim to uncover any potential mediators, to more targeted, hypothesis-driven stages where understanding specific mechanistic pathways are of interest. When it comes to investigating epigenetic inheritance, where very little research on humans has been conducted to date, this whole continuum is naturally of importance. Although the ultimate aim is to understand the molecular mechanisms through which the paternal exposome contributes to the health of subsequent generations, the causal roles of the large number of omics markers as well as their complex interactions and underlying confounders remain widely unknown. [56] Hence, the current research is still inevitably hypothesis-generating.

In this thesis, the focus has been in causal, hypothesis-driven approaches, while data-driven high-dimensional methods were contrasted as alternatives for sieving studies that lack causal interpretation. However, the gap between the two is decreasing as causal thinking is more often being taken into consideration in high-dimensional techniques and machine learning approaches. [138] The use of causal machine learning has been advocated, for example, in assessing individualised treatment effects. [139] In observational life-course epidemiology, machine learning methods have been suggested to contribute to identifying causal pathways. [140]

It is also worth noting that the statistical methods presented in this thesis stem from a set of very specific empirical research questions. As such, the analytical approaches may not be directly applicable in different settings. In publications I and II, the simulation studies relied on an underlying Dirichlet-multinomial distribution. However, due to the spurious correlations between groups, other distributions with more flexible variance-covariance matrix could have been more realistic. Furthermore, one of the focal points here concerned the normality of the ilr coordinates in mediation analysis. However, instead of forcing sparse overdispersed count data into models assuming normality, future studies should investigate whether these features of count data could be accommodated also within the framework of mediation analysis.

Although the limitations of the MULTIEPIGEN setting have been to some extent covered in the statistical analysis plan and in Section 5.3.3, it is likely that some unanticipated challenges may arise when proceeding from the planning of the multigenerational analyses to their execution, stemming from the complexities of the research question and the data at hand. In fact, while the MULTIEPIGEN project may be able to provide valuable insights into the role of the paternal life-course exposome on offspring health and epigenetic mechanisms, to be able to fully understand the phenomenon of epigenetic inheritance in humans, a prospective periconceptional setting with a focus on paternal sperm epigenome at the time of conception, long offspring follow-up times, a carefully characterised (pre-conceptional) exposome, and control for confounding is necessary.

With increasing availability of omics data in epidemiological research and mounting evidence of their role both in pathogenesis of health and disease as well as the

burden of exposome on them, the mediating role of various omics can be expected to gain more interest. When replication in independent samples and triangulation yield deeper understanding of potential mechanisms, the shift from hypothesis-generating exploratory studies to more formally formulated causal investigations is incumbent upon the field.

# Summaries of Original Publications

I The role of the gut microbiome as a transmitter of effects of exposures or lifestyle on health is of increasing interest. In Publication I of this thesis (Kartiosuo et al. (2024)), we suggest a causal mediation analysis approach for investigating the mediating role of the gut microbiome composition when *a priori* hypothesis or knowledge is available. In this approach, the hierarchies between the taxa can be utilised to build hypothesis-driven isometric log-ratio coordinates that are subsequently used as mediators in a framework for multiple contemporaneous mediators. A simulation study is used to demonstrate use of the method and investigate its performance under various scenarios. Sparsity, often present in microbial abundance data, affects the precision and efficiency of estimation of mediation effects. Finally, the use of the approach is demonstrated in practice to investigate and quantify the mediating effects of the taxa within the *Actinobacteria* phylum in the association between sufficient fibre intake and insulin levels.

II Isometric log-ratio transformation is a common tool in compositional data analysis, used to map compositions from the simplex to the Euclidean space. If the underlying data follow a multinomial distribution, the distribution of isometric log-ratio coordinates is known to asymptotically follow a multivariate normal distribution. Publication II (Kartiosuo et al. (2025a)) investigates the asymptotic distribution of ilr coordinates when the underlying observations follow a compound compound multinomial distribution, characterised by overdispersion. Conditions for asymptotic normality are derived. Under the special case of Dirichlet-multinomial distribution, a normal approximation for the coordinates is derived and investigated using a simulation study. The simulation study demonstrates that the approximation works well asymptotically unless overdispersion is high.

III DNA methylation can be affected by a range of exposures and, through its regulatory effects on gene expression, can in turn affect health outcomes, making it a mechanistically plausible mediating pathway for effects of exposures on health. Publication III of the thesis (Kartiosuo et al. (2025b)) investigates the mediating role of DNA methylation in the association between polychlorinated biphenyls and cardio-metabolic health. Based on

two high-dimensional epigenome-wide association studies, the composite nature of the null hypothesis of no mediation is utilised to identify candidate mediating CpG cites. Altogether five potential mediators are identified and their combined indirect effect is quantified using mediation analysis for multiple contemporaneous mediators. The proportion mediated through these mediators is 40 % of the total effect of exposure to polychlorinated biphenyls on type 2 diabetes. The mediating CpG sites are, for example, from the genes *CPT1A* and *ABCG1*, which have previously been linked to type 2 diabetes risk.

IV The MULTIEPIGEN project is an extension of the Cardiovascular Risk in Young Finns Study (YFS), set out to investigate the multigenerational effects of paternal exposome on health of subsequent generations through epigenetic changes in sperm. The original YFS participants as well as their parents and offspring were invited to the clinics and semen samples were collected for all volunteering males over 18 years of age to assess their sperm epigenome. Publication IV of this thesis (Pahkala et al., 2026) is the cohort profile of this project. The contributions of this thesis relate especially to the statistical analysis plan (SAP) reported in the supplement of the publication. The SAP conceptualises how the questions on trans- and intergenerational epigenetic inheritance can be formulated as mediation problems, where the sperm epigenetic marks are considered as mediators. Two types of data on sperm epigenome are available: the sequencing counts of the sperm non-coding RNAs of four distinct types (transfer RNAs-derived fragments, microRNAs, piwi-interacting RNA clusters ans sequences), and DNA methylation. The analysis pipeline for investigating the multigenerational hypotheses is outlined and statistical methods pertaining to each stage of analysis are explained.

# List of References


[1] C. P. Wild. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers and Prevention*, 14(8):1847–1850, 2005. doi: https://doi.org/10.1158/1055-9965.EPI-05-0456.

[2] J. W. J. Beulens, M. G. M. Pinho, T. C. Abreu, N. R. den Braver, T. M. Lam, A. Huss, J. Vlaanderen, T. Sonnenschein, N. Z. Siddiqui, Z. Yuan, J. Kerckhoffs, A. Zhernakova, M. F. Brandao Gois, and R. C. H. Vermeulen. Environmental risk factors of type 2 diabetes—an exposome approach. *Diabetologia*, 65:263–274, 2022. doi: https://doi.org/10.1007/s00125-021-05618-w.

[3] M. Silberberg, V. Martinez-Bianchi, and M. J. Lyn. What is population health? *Primary Care: Clinics in Office Practice*, 46(4):475–484, 2019. ISSN 0095-4543. doi: https://doi.org/10.1016/j.pop.2019.07.001. Population Health.

[4] J. E. Manson and S. S. Bassuk. Population health: The power of prevention. *NAM Perspect.*, 10, 2024. doi: https://doi.org/10.31478/202411b.

[5] Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, 2017. doi: https://doi.org/10.1186/s13059-017-1215-1.

[6] L. Maitre, , M. Bustamante, C. Hernández-Ferrer, D. Thiel, C.-H. E. Lau, A. P. Siskos, M. Vives-Usano, C. Ruiz-Arenas, D. Pelegrí-Sisó, O. Robinson, D. Mason, J. Wright, S. Cadiou, R. Slama, B. Heude, M. Casas, J. Sunyer, E. Z. Papadopoulou, K. B. Gutzkow, S. Andrusaityte, R. Grazuleviciene, M. Vafeiadi, L. Chatzi, A. K. Sakhi, C. Thomsen, I. Tamayo, M. Nieuwenhuijsen, J. Urquiza, E. Borràs, E. Sabidó, I. Quintela, Á. Carracedo, X. Estivill, M. Coen, J. R. González, H. C. Keun, and M. Vrijheid. Multi-omics signatures of the human early life exposome. *Nature Communications*, 13(1):7024, 2022. doi: https://doi.org/10.1038/s41467-022-34422-2.

[7] Institute of Medicine. *Evolution of translational omics: Lessons learned and the path forward*. The National Academies Press, Washington, DC, 2012. ISBN 978-0-309-22418-5. doi: https://doi.org/10.17226/13297.

[8] J. Pearl. Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–20, 2001.

[9] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL `https://miguelhernan.org/whatifbook`.

[10] M. Vailati-Riboni, V. Palombo, and J. J. Loor. *What are omics sciences?*, pages 1–7. Springer International Publishing, Cham, 2017. ISBN 978-3-319-43033-1. doi: https://doi.org/10.1007/978-3-319-43033-1_1.

[11] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8(NOV):1–6, 2017. ISSN 1664302X. doi: https://doi.org/10.3389/fmicb.2017.02224.

[12] R. K. Singh, H.-W. Chang, D. Yan, K. M. Lee, D. Ucmak, K. Wong, M. Abrouk, B. Farahnik, M. Nakamura, T. H. Zhu, T. Bhutani, and W. Liao. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, 15:73, 2017. doi: https://doi.org/10.1186/s12967-017-1175-y.

[13] A. P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447:433–440, 2007. doi: https://doi.org/10.1038/nature05919.

[14] M. G. Danieli, M. Casciaro, A. Paladini, M. Bartolucci, M. Sordoni, Y. Shoenfeld, and S. Gangemi. Exposome: Epigenetics and autoimmune diseases. *Autoimmunity Reviews*, 23 (6):103584, 2024. ISSN 1568-9972. doi: https://doi.org/10.1016/j.autrev.2024.103584.

[15] L. Hou, X. Zhang, D. Wang, and A. Baccarelli. Environmental chemical exposures and human epigenetics. *International Journal of Epidemiology*, 41:79–105, 2012. doi: https://doi.org/10.1093/ije/dyr154.

[16] H. Wu, C. M. Eckhardt, and A. A. Baccarelli. Molecular mechanisms of environmental exposures and human disease. *Nature Reviews Genetics*, 24(5):332–344, 2023. doi: https://doi.org/10.1038/s41576-022-00569-3.

[17] M. Pembrey, R. Saffery, L. O. Bygren, and Network in Epigenetic Epidemiology. Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research. *Journal of Medical Genetics*, 51(9):563–572, 2014. doi: https://doi.org/10.1136/jmedgenet-2014-102577.

[18] L. Senaldi and M. Smith Raska. Evidence for germline non genetic inheritance of human phenotypes and diseases. *Clinical Epigenetics*, 12:136, 2020. doi: https://doi.org/10.1186/s13148-020-00929-y.

[19] M.R. Munafò and G. Davey Smith. Robust research needs many lines of evidence. *Nature*, 553: 399–401, 2018. doi: https://doi.org/10.1038/d41586-018-01023-3.

[20] D. A. Lawlor, K. Tilling, and G. Davey Smith. Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, 45(6):1866–1886, 2016. doi: https://doi.org/10.1093/ije/dyw314.

[21] C. H. Miles. On the causal interpretation of randomised interventional indirect effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1154–1172, 2023. doi: https://doi.org/10.1093/jrsssb/qkad066.

[22] J. Pearl. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4): 459–481, 2014. ISSN 1082989X. doi: https://doi.org/10.1037/a0036434.

[23] J. M. Rohrer. Thinking clearly about correlations and causation: graphical causal models for observational data. *Advances in methods and practices in psychological science*, 1(1):27–42, 2018. doi: https://doi.org/10.1177/2515245917745629.

[24] T. J. Vanderweele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York, NY, 2015.

[25] P. W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[26] D. B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: https://doi.org/10.1198/016214504000001880.

[27] J. Pearl. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

[28] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. doi: https://doi.org/10.2307/2337329.

[29] T. H. Nguyen, I. Schmid, E. L. Ogburn, and E.A. Stuart. Clarifying causal mediation analysis: Effect identification via three assumptions and five potential outcomes. *Journal of Causal Inference*, 10(1):246–279, 2022. doi: https://doi.org/10.1037/met0000299.

[30] K. Imai, L. Keele, and T. Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010. doi: https://doi.org/10.1214/10-STS321.

[31] E.J. Tchetgen Tchetgen and T. J. Vanderweele. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, 25(2):282–291, 2014. doi: https://doi.org/10.1097/EDE.0000000000000054.

[32] T. J. VanderWeele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4):457–468, 2009. ISSN 19387989. doi: https://doi.org/10.4310/SII.2009.v2.n4.a7.

[33] R. M. Baron and D. A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986. doi: https://doi.org/10.1037//0022-3514.51.6.1173.

[34] T. J. VanderWeele and S. Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172(12):1339–1348, 2010. doi: https://doi.org/10.1093/aje/kwq332.

[35] T. J. VanderWeele. Mediation analysis: a practitioner's guide. *Annual Review of Public Health*, 37(1):17–32, 2016. ISSN 0163-7525. doi: https://doi.org/10.1146/annurev-publhealth-032315-021402.

[36] L. Valeri and T. J. VanderWeele. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods.*, 18(2):137–150, 2013. ISSN 03010422. doi: https://doi.org/10.1037/a0031034.

[37] M. E. Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *American Sociological Association*, 13:290–312, 1982. doi: https://doi.org/10.2307/270723.

[38] H. P. O'Rourke and D. P. MacKinnon. When the test of mediation is more powerful than the test of the total effect. *Behavior Research Methods*, 47:424–442, 2015. doi: https://doi.org/10.3758/s13428-014-0481-z.

[39] D. Chen and M. S. Fritz. Comparing alternative corrections for bias in the bias-corrected bootstrap test of mediation. *Eval Health Prof.*, 44(4):416–427, 2021. doi: https://doi.org/10.1177/01632787211024356.

[40] D. P. MacKinnon, C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104, 2002. doi: https://doi.org/10.1037/1082-989x.7.1.83.

[41] Z. Liu, J. Shen, R. Barfield, J. Schwartz, A. A. Baccarelli, and X. Lin. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, 117(537):67–81, 2022. doi: https://doi.org/10.1080/01621459.2021.1914634.

[42] J. W. R. Twisk. Indirect effects in mediation analyses should not be tested for statistical significance. *Journal of Clinical Epidemiology*, 171(111393), 2024. doi: https://doi.org/10.1016/j.jclinepi.2024.111393.

[43] A. G. Cashin and T.-T. Vo. Indirect effects in mediation analyses should still include measures of unertainty and, when appropriate, test for statistical significance. *Journal of Clinical Epidemiology*, 172(111395), 2024. doi: https://doi.org/10.1016/j.jclinepi.2024.111395.

[44] T. J. VanderWeele. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21(4):540–51, 2010. doi: https://doi.org/10.1097/EDE.0b013e3181df191c.

[45] T. J. VanderWeele and O. A. Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22(1):42–52, 2011. doi: https://doi.org/10.1097/EDE.0b013e3181f74493.

[46] J. Cornfield, W. Haenszel, E.C. Hammond, A.M. Lilienfeld, M.B. Shimkin, and E.L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *JNCI: Journal of the National Cancer Institute*, 22(1):173—-203, 1959. doi: https://doi.org/10.1093/jnci/22.1.173.

[47] K. Imai and T. Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–161, 2013. doi: https://doi.org/10.1093/pan/mps040.

[48] M. Taguri, J. Featherstone, and J. Cheng. Causal mediation analysis with multiple causally non-ordered mediators. *Statistical Methods in Medical Research*, 27(1):3–19, 2018. doi: https://doi.org/10.1177/0962280215615899.

[49] C. Kim, M. J. Daniels, J. W. Hogan, C. Choirat, and C. M. Zigler. Bayesian methods for multiple mediators: relating principal stratification and causal mediation in the analysis of power plant

emission controls. *Annals of Applied Statistics*, 13(3):1927–1956, 2019. doi: https://doi.org/10.1214/19-AOAS1260.

[50] W. Wang, S. Nelson, and J. M. Albert. Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in Medicine*, 32 (24):4211–4228, 2013. doi: https://doi.org/10.1002/sim.5830.

[51] T. J. VanderWeele and S. Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2014. doi: https://doi.org/10.1515/em-2012-0010.

[52] R. M. Daniel, B. L. De Stavola, S. N. Cousens, and S. Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1:14, 2015. doi: https://doi.org/10.1111/biom.12248.

[53] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[54] V. K Rakyan, T. A. Down, D. J Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nature Review Genetics*, 12(8):529–541, 2011. doi: https://doi.org/.

[55] P. A. Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13:484–492, 2012. doi: https://doi.org/10.1038/nrg3230.

[56] M. G. B. Blum, L. Valeri, O. François, S. Cadiou, V. Siroux, J. Lepeule, and R. Slama. Challenges raised by mediation analysis in a high-dimension setting. *Environmental Health Perspectives*, 128(5):1–8, 2020. ISSN 15529924. doi: https://doi.org/10.1289/EHP6240.

[57] H. Zhang, Y. Zheng, Z. Zhang, T. Gao, B. Joyce, G. Yoon, W. Zhang, J. Schwartz, A. Just, E. Colicino, P. Vokonas, L. Zhao, J. Lv, A. Baccarelli, L. Hou, and L. Liu. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 06 2016. ISSN 1367-4803. doi: https://doi.org/10.1093/bioinformatics/btw351.

[58] J. Jin and T. T. Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007. doi: https://doi.org/10.1198/016214507000000167.

[59] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, 2005. doi: https://doi.org/10.1198/016214504000000089.

[60] J. J. Egozcue, J. Graffelman, M. I. Ortego, and V. Pawlowsky-Glahn. Some thoughts on counts in sequencing studies. *NAR Genomics and Bioinformatics*, 2(4):lqaa094, 11 2020. ISSN 2631-9268. doi: https://doi.org/10.1093/nargab/lqaa094. URL `https://doi.org/10.1093/nargab/lqaa094`.

[61] J. Bacon-Shone. A short history of compositional data analysis. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, first edition, 2011.

[62] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. Principles of compositional analysis. In *Modelling and Analysis of Compositional Data*, pages 32–64. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, 2015. ISBN 9781119003144.

[63] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 00359246. URL `http://www.jstor.org/stable/2345821`.

[64] J. Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, Ltd., London (UK), 1986. ISBN 0412280604.

[65] G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue. The principle of working on coordinates. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, first edition, 2011.

[66] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003. ISSN 08828121. doi: https://doi.org/10.1023/A:1023818214614.

[67] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. Coordinate representation. In *Modelling and Analysis of Compositional Data*, pages 32–64. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, 2015. ISBN 9781119003144.

[68] J. J. Egozcue and V. Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005. ISSN 1573-8868. doi: https://doi.org/10.1007/s11004-005-7381-9.

[69] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:1–20, 2017. ISSN 2050084X. doi: https://doi.org/10.7554/eLife.21887.

[70] P. Filzmoser, K. Hron, and M. Templ, editors. *Applied compositional data analysis*. Springer Nature Switzerland AG, Cham, Switzerland, first edition, 2018.

[71] E. Gordon-Rodriguez, T. P. Quinn, and J. P. Cunningham. Learning sparse log-ratios for high-throughput sequencing data. *bioRxiv*, page 2021.02.11.430695, 2021. doi: https://doi.org/10.1101/2021.02.11.430695.

[72] H. Zhang, J. Chen, Z. Li, and L. Liu. Testing for mediation effect with application to human microbiome Data. *Statistics in Biosciences*, 13:313–328, 2021. ISSN 18671772. doi: https://doi.org/10.1007/s12561-019-09253-3.

[73] M. B. Sohn and H. Li. Compositional mediation analysis for microbiome studies. *Annals of Applied Statistics*, 13(1):661–681, 2019. ISSN 19417330. doi: https://doi.org/10.1214/18-AOAS1210.

[74] C. Wang, J. Hu, M. J. Blaser, H. Li, and I. Birol. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*, 36 (2):347–355, 2020. ISSN 14602059. doi: https://doi.org/10.1093/bioinformatics/btz565.

[75] J. Fu, M. D. Koslovsky, A. M. Neophytou, and M. Vannucci. A Bayesian joint model for compositional mediation effect selection in microbiome data. *Statistics in Medicine*, 42(17):2999–3015, 2023. doi: https://doi.org/10.1002/sim.9764.

[76] K. F. Arnold, L. Berrie, P. W. G. Tennant, and M. S. Gilthorpe. A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology*, 49(4):1307–1313, 2020. doi: https://doi.org/10.1093/ije/dyaa021.

[77] J. Graffelman. A parametric test for HWE based on isometric logratio coordinates. pages 1–5, 2011.

[78] J. Graffelman, M. I. Ortego, and J. J. Egozcue. On the asymptotic distribution of proportions of multinomial count data. *Proceedings of the 6th International Workshop on Compositional Data Analysis S.*, (May), 2015.

[79] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[80] B. A. Frigyik, A. Kapila, and M. R. Gupta. Introduction to the dirichlet distribution and related processes. Technical Report UWEETR-2010-0006, Department of Electrical Engineering, University of Washington, Seattle, WA, USA, 2010. URL `https://mayagupta.org/publications/FrigyikKapilaGuptaIntroToDirichlet.pdf`.

[81] T. P. Minka. Estimating a Dirichlet distribution. 2000. URL `https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf`.

[82] V. Pawlowsky-Glahn and A. Buccianti, editors. *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, first edition, 2011.

[83] J. Pelto, K. Auranen, J. Janne V. Kujala, and L. Lahti. Elementary methods provide more replicable results in microbial differential abundance analysis. *Briefings in Bioinformatics*, 26(2): bbaf130, 03 2025. ISSN 1477-4054. doi: https://doi.org/10.1093/bib/bbaf130.

[84] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35:253–278, 2003. doi: https://doi.org/10.1023/A:1023866030544.

[85] J. Palarea-Albaladejo and J. A. Martín-Fernández. zcompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2015.02.019.

[86] J. M. Fry, T. R. L. Fry, and K. R. McLaren. Compositional data analysis and zeros in micro data. *Applied Economics*, 32(8):953–959, 2000. doi: https://doi.org/10.1080/000368400322002.

[87] J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2014. doi: https://doi.org/10.1177/1471082X14535524.

[88] R. Zhou, S. K. Ng, J. J. Y. Sung, W. W. B. Goh, and S. H. Wong. Data pre-processing for analyzing microbiome data - a mini review. *Comput Struct Biotechnol J.*, (21):4804–4815, 2023. doi: https://doi.org/10.1016/j.csbj.2023.10.001.

[89] T. M. Barber, S. Kabisch, A. F.H. Pfeiffer, and M. O. Weickert. The health benefits of dietary fibre. *Nutrients*, 12(10):1–17, 2020. ISSN 20726643. doi: https://doi.org/10.3390/nu12103209.

[90] J. Ahn and R. B. Hayes. Environmental influences on the human microbiome and implications for noncommunicable disease. *Annual Review of Public Health*, 42:277–292, 2021. doi: https://doi.org/10.1146/annurev-publhealth-012420-105020.

[91] O. Simell, H. Niinikoski, T. Rönnemaa, O. T. Raitakari, H. Lagström, M. Laurinen, M. Aromaa, P. Hakala, A. Jula, E. Jokinen, I. Välimäki, J. Viikari, and for the STRIP Study Group. Cohort profile: The strip study (special turku coronary risk factor intervention project), an infancy-onset dietary and life-style intervention trial. *International Journal of Epidemiology*, 38(3):650–655, 04 2008. ISSN 0300-5771. doi: https://doi.org/10.1093/ije/dyn072.

[92] K. Pahkala, T. T. Laitinen, H. Niinikoski, N. Kartiosuo, S. P. Rovio, H. Lagström, B. M. Loo, P. Salo, E. Jokinen, C. G. Magnussen, M. Juonala, O. Simell, A. Jula, T. Rönnemaa, J. Viikari, and O. T. Raitakari. Effects of 20-year infancy-onset dietary counselling on cardiometabolic risk factors in the Special Turku Coronary Risk Factor Intervention Project (STRIP): 6-year post-intervention follow-up. *The Lancet Child and Adolescent Health*, 4(5):359–369, 2020. ISSN 23524642. doi: https://doi.org/10.1016/S2352-4642(20)30059-6.

[93] P. Hakala, J. Marniemi, L. R. Knuts, J. Kumpulainen, R. Tahvonen, and S. Plaami. Calculated vs analysed nutrient composition of weight reduction diets. *Food Chemistry*, 57(1):71–75, 1996. ISSN 03088146. doi: https://doi.org/10.1016/0308-8146(96)00077-5.

[94] Valtion Ravitsemusneuvottelukunta. *Terveyttä ruoasta - Suomalaiset ravitsemussuositukset 2014 (Finnish Nutrition Recommendations 2014) Valtion ravitsemusneuvottelukunta: Tampere.* 2014. ISBN 978-952-453-801-5.

[95] A. Keskitalo, E. Munukka, A. Aatsinki, W. Saleem, N. Kartiosuo, L. Lahti, P. Huovinen, L. L. Elo, S. Pietilä, H. Rovio, S. P. annd Niinikoski, J. Viikari, T. Rönnemaa, H. Lagström, A. Jula, O. Raitakari, and K. Pahkala. An infancy-onset 20-year dietary counselling intervention and gut microbiota composition in adulthood. *Nutrients*, 14(13):2667, 2022. doi: https://doi.org/10.3390/nu14132667.

[96] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–583, 2016. ISSN 15487105. doi: https://doi.org/10.1038/nmeth.3869.

[97] L. D. Moore, T. Le, and F. Guoping. Dna methylation and its basic function. *Neuropsychopharmacology Reviews*, 38:23–38, 2013. doi: https://doi.org/10.1038/npp.2012.112.

[98] Duk-Hee L., M. Porta, Jr. Jacobs, D. R., and L. N. Vandenberg. Chlorinated persistent organic pollutants, obesity, and type 2 diabetes. *Endocrine Reviews*, 35(4):557–601, 08 2014. ISSN 0163-769X. doi: https://doi.org/10.1210/er.2013-1084.

[99] Y. Song, E. L. Chou, A. Baecker, N.-C. Y. You, Y. Song, Q. Sun, and S. Liu. Endocrine-disrupting chemicals, risk of type 2 diabetes, and diabetes-related metabolic traits: A systematic review and meta-analysis. *Journal of Diabetes*, 8(4):516–532, 2016. doi: https://doi.org/10.1111/1753-0407.12325.

[100] S. W. Curtis, D. O. Cobb, V. K., M. L. Terrell, M. E. Marder, D. B. Barr, C. J. Marsit, M. Marcus, K. N. Conneely, and A. K. Smith. Genome-wide dna methylation differences and polychlorinated biphenyl (pcb) exposure in a us population. *Epigenetics*, 16(3):338–352, 2021. doi: https://doi.org/10.1080/15592294.2020.1795605.

[101] A. Cardona, F. R. Day, J. R.B. Perry, M. Loh, A. Y. Chu, B. Lehne, D. S. Paul, L. A. Lotta, I. D. Stewart, N. D. Kerrison, R. A. Scott, K.-T. Khaw, N. G. Forouhi, C. Langenberg, C. Liu, M. M. Mendelson, D. Levy, S. Beck, R. D. Leslie, J. Dupuis, J. B. Meigs, J. S. Kooner, J. Pihlajamäki, A. Vaag, A. Perfilyev, C. Ling, M. F. Hivert, J. C. Chambers, N. J. Wareham, and K. K. Ong. Epigenome-wide association study of incident type 2 diabetes in a british population: Epic-norfolk study. *Diabetes*, 68(12):2315–2326, 09 2019. ISSN 0012-1797. doi: https://doi.org/10.2337/db18-0290.

[102] D. L. Juvinao-Quintero, G. C. Sharp, E. C. M. Sanderson, C. L. Relton, and H.h R. E. Investigating causality in the association between dna methylation and type 2 diabetes using bidirectional two-sample mendelian randomisation. *Diabetologia*, 66(7):1247–1259, 2023. doi: https://doi.org/10.1007/s00125-023-05914-7.

[103] O. T. Raitakari, M. Juonala, T. Rönnemaa, L. Keltikangas-Järvinen, L. Räsänen, M. Pietikäinen, N. Hutri-Kähönen, L. Taittonen, E. Jokinen, J. Marniemi, A. Jula, R. Telama, M. Kähönen, T. Lehtimäki, H. K. Åkerblom, and J. S. A. Viikari. Cohort profile: The cardiovascular risk in young finns study. *International Journal of Epidemiology*, 37(6):1220–1226, 02 2008. ISSN 0300-5771. doi: https://doi.org/10.1093/ije/dym225.

[104] P. Du, X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, 2009. doi: https://doi.org/10.1186/1471-2105-11-587.

[105] S. Lacagnina. The developmental origins of health and disease (dohad). *Am J Lifestyle Med.*, 14 (1):47—-50, 2019. doi: https://doi.org/10.1177/1559827619879694.

[106] C. N. Hales and D. J. P. Barker. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia*, 37(7):595–601, 1992. doi: https://doi.org/10.1007/BF00400248.

[107] T. Roseboom, S. de Rooij, and R. Painter. The dutch famine and its long-term consequences for adult health. *Early Human Development*, 82(8):485–491, 2006. ISSN 0378-3782. doi: https://doi.org/10.1016/j.earlhumdev.2006.07.001. Special Abstract Issue.

[108] E. W. Tobi, L. H. Lumey, R. P. Talens, D. Kremer, H. Putter, A. D. Stein, P. E. Slagboom, and B. T. Heijmans. Dna methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Human Molecular Genetics*, 18(21):4046–4053, 08 2009. ISSN 0964-6906. doi: https://doi.org/10.1093/hmg/ddp353.

[109] R. Yehuda, N. P. Daskalakis, L. M. Bierer, H. N. Bader, T. Klenger, F. Holsboer, and E. B. Binder. Holocaust exposure induced intergenerational effects on fkbp5 methylation. *Biological Psychiatry*, 80(5):372—-380, 2016. doi: https://doi.org/10.1016/j.biopsych.2015.08.005.

[110] E. W. Tobi, R. C. Slieker, R. Luijk, K. F. Dekkers, A. D. Stein, K. M. Xu, Biobank based Integrative Omics Studies Consortium, P. E. Slagboom, E. W. van Zwet, L. H. Lumey, and B. T. Heijmans. Dna methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci. Adv*, 4(1), 2018. doi: https://doi.org/10.1126/sciadv.aao4364.

[111] S. Lapehn and A. G. Paquette. The placental epigenome as a molecular link between prenatal exposures and fetal health outcomes through the dohad hypothesis. *Current Environmental Health Reports*, 9:490—501, 2022. doi: https://doi.org/10.1007/s40572-022-00354-8.

[112] X.-Y. Fan, X. S. Lin, B.-R. Yang, H. W. Zhang, F. Tang, J.-J. Tang, H. Chi, T. Mansell, N. Kartiosuo, Y.-Y. Xia, T.-L. Han, H. Zhang, P. Baker, and R. Saffery. Relationship between prenatal metals exposure and neurodevelopment in one-year-old infants in the climb study. *Ecotoxicology and Environmental Safety*, 291:117860, 2025. doi: https://doi.org/10.1016/j.ecoenv.2025.117860.

[113] R. C. Painter, C. Osmond, P. Gluckman, M. Hanson, D. I. W. Phillips, and Roseboom T. J. Transgenerational effects of prenatal exposure to the dutch famine on neonatal adiposity and

health in later life. *BJOG*, 115(10):1243–9, 2008. doi: https://doi.org/10.1111/j.1471-0528.2008.01822.x.

[114] M. V. Veenendaal, R. C. Painter, S. R. de Rooij, P. M. Bossuyt, J. A. van der Post, M. A. Gluckman, P. D. andd Hanson, and T. J. Roseboom. Transgenerational effects of prenatal exposure to the 1944-45 dutch famine. *BJOG*, 120(5):548–533, 2013. doi: https://doi.org/10.1111/1471-0528.12136.

[115] J. Santiago, J. V. Silva, J. Howl, M. A. S. Santos, and M. Fardilha. All you need to know about sperm rnas. *Human Reproduction Update*, 28(1):67–91, 10 2021. ISSN 1355-4786. doi: https://doi.org/10.1093/humupd/dmab034.

[116] Q. Chen, W. Yan, and E. Duan. Epigenetic inheritance of acquired traits through sperm rnas and sperm rna modifications. *Nat Rev Genet*, 17(12):733–743, 2016. doi: https://doi.org/10.1038/nrg.2016.106.

[117] Y. Wei, C.-R. Yang, Y.-P. Wei, Z.-A. Zhao, Y. Hou, H. Schatten, and Q.-Y. Sun. Paternally induced transgenerational inheritance of susceptibility to diabetes in mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 111(5):1873–1878, 2014. doi: https://doi.org/10.1073/pnas.1321195111.

[118] M. K. Skinner, M. Ben Maamar, I. Sadler-Riggleman, D. Beck, E. Nilsson, M. McBirney, R. Klukovich, and W. Yan. Alterations in sperm dna methylation, non-coding rna and histone retention associate with ddt-induced epigenetic transgenerational inheritance of disease. *Epigenetics & Chromatin*, 11(8), 2018. doi: https://doi.org/10.1186/s13072-018-0178-0.

[119] Q. Chen, M. Yan, Z. Cao, X. Li, Y. Zhang, J. Shi, G.-H. Feng, H. Peng, X. Zhang, Y. Zhang, J. Qian, E. Duan, Q. Zhai, and Q. Zhou. Sperm tsrnas contribute to intergenerational inheritance of an acquired metabolic disorder. *Science*, 351(6271):397–400, 2016. doi: https://doi.org/10.1126/science.aad7977.

[120] D. L. Maxwell, O. A. Oluwayiose, E. Houle, K. Roth, K. Nowak, S. Sawant, A. L. Paskavitz, W. Liu, K. Gurdziel, M. C. Petriello, and J. R. Pilsner. Mixtures of per- and polyfluoroalkyl substances (pfas) alter sperm methylation and long-term reprogramming of offspring liver and fat transcriptome. *Environment International*, 186:108577, 2024. ISSN 0160-4120. doi: https://doi.org/10.1016/j.envint.2024.108577.

[121] G. Kaati, L. O. Bygren, and S. Edvinsson. Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *European Journal of Human Genetics*, 10(11):682–688, 2002. doi: https://doi.org/10.1038/sj.ejhg.5200859.

[122] K. Northstone, J. Golding, G. Davey Smith, L. L. Miller, and M. Pembrey. Prepubertal start of father's smoking and increased body fat in his sons: further characterisation of paternal transgenerational responses. *European Journal of Human Genetics*, 22(12):1382–1386, 2014. doi: https://doi.org/10.1038/ejhg.2014.31.

[123] J. M. Jaakkola, S. P. Rovio, K. Pahkala, J. Viikari, T. Rönnemaa, A. Jula, H. Niinikoski, J. Mykkänen, M. Juonala, N. Hutri-Kähönen, M. Kähönen, T. Lehtimäki, and O. T. Raitakari. Childhood exposure to parental smoking and life-course overweight and central obesity. *Annals of Medicine*, 53(1):208–216, 2021. doi: https://doi.org/10.1080/07853890.2020.1853215.

[124] F. Wu, K. Pahkala, M. Juonala, J. Jaakkola, S. P. Rovio, T. Lehtimäki, M. A. Sabin, A. Jula, N. Hutri-Kähönen, T. Laitinen, J. S.A. Viikari, C. G. Magnussen, and O. T. Raitakari. Childhood and adulthood passive smoking and nonalcoholic fatty liver in midlife: a 31-year cohort study. *American Journal of Gastroenterology*, 116(6):1256–1263, 2021. doi: https://doi.org/10.14309/ajg.0000000000001141.

[125] Pahkala K., Rovio S., Kartiosuo N., Auranen K., Bourgery M., Elovainio M., Fogelholm M., Haapala J., Hirvensalo M., Hutri N., Jokinen E., Jula A., Juonala M., Kaikkonen J., Kiviranta H., Koskinen JS., Kotaja N., Kähönen M., Laitinen TP., Lehtimäki T., Lisinen I., Loo BM., Lyytikäinen LP., Magnussen CG., Mishra PP., Mykkänen J., Mäkelä JA., Männistö S., Nevalainen J., Pulkki-Råback L., Raitoharju E., Rantakokko P., Rönnemaa T., Stenbacka S., Taittonen L., Tammelin TH., Toppari J., Tossavainen P., Viikari J., and Raitakari O. Cohort profile update: Expanding the cardiovascular risk in young finns study into a multigen-

erational cohort. *International Journal of Epidemiology*, 55(1):dyaf206, 2026. doi: https://doi.org10.1093/ije/dyaf206/.

[126] M. Naveed, Z. Shen, and J. Bao. Sperm-borne small non-coding rnas: potential functions and mechanisms as epigenetic carriers. *Cell & Bioscience*, 15(1):5. doi: https://doi.org/10.1186/s13578-025-01347-4.

[127] S. A. Krawetz. Paternal contribution: new insights and future challenges. *Nature Reviews Genetics*, 6(8):633–642, 2005. doi: https://doi.org/10.1038/nrg1654.

[128] D. Nätt, U. Kugelberg, E. Casas, E. Nedstrand, S. Zalavary, P. Henriksson, C. Nijm, J. Jäderquist, J. Sandborg, E. Flinke, R. Ramesh, L. Örkenby, F. Appelkvist, T. Lingg, N. Guzzi, C. Bellodi, M. Löf, T. Vavouri, and A. Öst. Human sperm displays rapid responses to diet. *PLoS Biology*, 17(12):e3000559, 2019. doi: https://doi.org/10.1371/journal.pbio.3000559.

[129] A. Tomar, M. Gomez-Velazquez, R. Gerlini, G. Comas-Armangué, L. Makharadze, T. Kolbe, A. Boersma, M. Dahlhoff, J. P. Burgstaller, M. Lassi, J. Darr, J. Toppari, H. Virtanen, A. Kühnapfel, M. Scholz, K. Landgraf, W. Kiess, M. Vogel, V. Gailus-Durner, H. Fuchs, S. Marschall, M. Hrabě de Angelis, N. Kotaja, A. Körner, and R. Teperino. Epigenetic inheritance of diet-induced and sperm-borne mitochondrial rnas. *Nature*, 630:720–727, 2024. doi: https://doi.org/10.1038/s41586-024-07472-3.

[130] E. W. Harville and N. Kartiosuo. Transgenerational studies: How do we investigate multigenerational effects? *Obesity*, 28(3):482–483, 2020. doi: https://doi.org/10.1002/oby.22723.

[131] G. McGee, N. J. Perkins, S. L. Mumford, M.-A. Kioumourtzoglou, M. G. Weisskopf, J. S. Schildcrout, B. A. Coull, E. F. Schisterman, and S. Haneuse. Methodological issues in population-based studies of multigenerational associations. *American Journal of Epidemiology*, 189(12):1600–1609, 2020. doi: https://doi.org/10.1093/aje/kwaa125.

[132] J. M. Williamson, S. Datta, and G. A. Satten. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1):36–42, 03 2003. ISSN 0006-341X. doi: https://doi.org/10.1111/1541-0420.00005.

[133] S. R. Seaman, M. Pavlou, and A. J. Copas. Methods for observed-cluster inference when cluster size is informative: a review and clarifications. *Biometrics*, 70(2):449–456, 01 2014. ISSN 0006-341X. doi: https://doi.org/10.1111/biom.12151.

[134] Z. Niu, S. Mohazzab-Hosseinian, and .C. V. Breton. Transgenerational epigenetic inheritance: Perspectives and challenges. *Journal of Allergy and Clinical Immunology*, 151:1474–1476, 2023. doi: https://doi.org/10.1016/j.jaci.2023.02.027.

[135] A. Kong, G. Thorleifsson, M. L. Frigge, B. J. Vilhjálmsson, A. I. Young, T. E. Thorgeirsson, S. Benonisdóttir, A. Oddsson, B. V. Halldórsson, G. Masson, D. F. Gudbjartsson, A. Helgason, G. Björnsdóttir, U. Thorsteinsdóttir, and K. Stefánsson. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, 2018. doi: https://doi.org/10.1126/science.aan6877.

[136] C.L Relton and G. Davey Smith. Two-step epigenetic mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology*, 41(1):161–176, 2012. doi: https://doi.org/10.1093/ije/dyr233.

[137] Z. Wei, D. Chen, L. Li, J. Yang, and Y. Wang. Mendelian randomization reveals the causal links between mirnas and rheumatoid arthritis. *Medicine*, 104(44), 2025. doi: https://doi.org/10.1097/MD.0000000000045527.

[138] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022. URL `https://arxiv.org/abs/2206.15475`.

[139] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, and M. van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30:958–968, 2024. doi: https://doi.org/10.1038/s41591-024-02902-1.

[140] E. C. Matthay, D. B. Neill, A. R. Titus, S. Desai, A. B. Troxel, M. Cerdá, I Díaz, M. Santacatterina, and L. E. Thorpe. Integrating artificial intelligence into causal research in epidemiology. *Current Epidemiology Reports*, 12(1):6, 2025. doi: https://doi.org/10.1007/s40471-025-00359-5.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2026

ISBN 978-952-02-0553-9 (PRINT)
ISBN 978-952-02-0554-6 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)