



**POLITECNICO**  
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

# DNA methylation as mediator in the association of dietary nutrient intakes with the risk of CardioVascular diseases: a systematic comparison and evaluation of Meet-in-the-Middle approaches

TESI DI LAUREA MAGISTRALE IN  
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: **India Ermacora**

Student ID: 10608359

Advisor: Prof. Francesca Ieva

Co-advisors: Solène Cadiou, Giovanni Fiorito

Academic Year: 2022-23

## Abstract

DNA methylation (DNAm) is a biomolecular mechanism of gene regulation involving the addition of methyl groups to DNA molecules, often playing a crucial role in various disease mechanisms, including CardioVascular diseases (CVD). Contrary to the genetic sequence, DNAm is influenced by internal and external exposures, and it is widely modifiable. Recent evidence shows that DNAm is linked to CVD risk and food intake, but its role in mediating the association of different types of nutrients with cardiovascular risk factors has not been fully investigated.

In this study, we analyse DNAm profiles as an intermediate biological layer between the independent variables (dietary nutrient intakes) and the dependent variable (CVD risk), representing an essential but complex additional information to understand the molecular mechanisms linking dietary habits with CVD risk. We conduct a systematic comparison and evaluation of three distinct methods. The core of the analyses is represented by two developments of the Meet-in-the-Middle (MITM) method, an innovative implementation of high-dimensional mediation framework. The first application employs the methylome layer to identify potential new exposures likely to be causally associated with CVD, while the second application focuses on selecting potential mediators for exposures associated with CVD and to assess their significance. Finally, we compare the results of the MITM approaches with a third method consisting of a stability selection approach, which sought to identify causal associations between nutrients and CVD without considering the intermediate layer. Through a comprehensive comparison, we discuss strengths and weaknesses of each method, and we identify two nutrients showing statistical evidence of a causal association with CVD risk, with the methylome layer playing a mediating role in this process.

While our work acknowledges certain limitations, our results contribute to a rapidly advancing research domain, for the understanding of the molecular mechanisms linking dietary habits with the risk of CVD.

**Keywords:** DNA methylation; DNAm; nutrients; CardioVascular disease; CVD; Meet-in-the-Middle; mediation analysis; causal inference; stability selection; expsome

## Abstract in lingua italiana

La metilazione del DNA (DNAm) è un meccanismo biomolecolare di regolazione genica che comporta l'aggiunta di gruppi metilici alle molecole di DNA e che spesso svolge un ruolo cruciale in vari meccanismi patologici, tra cui le malattie cardiovascolari (CVD). A differenza della sequenza genetica, DNAm è influenzata da esposizioni interne ed esterne ed è ampiamente modificabile. Recenti evidenze dimostrano che DNAm è legata al rischio di CVD e all'assunzione di cibo, ma il suo ruolo nel mediare l'associazione di diversi tipi di nutrienti con i fattori di rischio cardiovascolare non è stato completamente indagato.

In questo studio, analizziamo i profili della DNAm come strato biologico intermedio tra le variabili indipendenti (assunzione di nutrienti con la dieta) e la variabile dipendente (rischio di CVD), rappresentando un'informazione aggiuntiva essenziale ma complessa per comprendere i meccanismi molecolari che collegano le abitudini alimentari al rischio di CVD. Abbiamo condotto un confronto sistematico e una valutazione di tre metodi distinti. Il nucleo delle analisi è rappresentato da due sviluppi del metodo Meet-in-the-Middle (MITM), un'implementazione innovativa del quadro di mediazione ad alta dimensionalità. La prima applicazione impiega lo strato del metiloma per identificare potenziali nuove esposizioni che potrebbero essere causalmente associate alla CVD, mentre la seconda applicazione si concentra sulla selezione di potenziali mediatori per le esposizioni associate alla CVD e sulla valutazione della loro significatività. Infine, confrontiamo i risultati degli approcci MITM con un terzo metodo consistente in un approccio di selezione di stabilità, che ha cercato di identificare associazioni causali tra nutrienti e CVD senza considerare lo strato intermedio. Attraverso un confronto completo, discutiamo i punti di forza e di debolezza di ciascun metodo e identifichiamo due nutrienti che mostrano evidenza statistica di un'associazione causale con il rischio di CVD, con lo strato metilico che svolge un ruolo di mediazione in questo processo.

Pur riconoscendo alcune limitazioni, i nostri risultati contribuiscono a un settore di ricerca in rapida evoluzione, per la comprensione dei meccanismi molecolari che collegano le abitudini alimentari al rischio di CVD.

**Parole chiave:** Metilazione del DNA; nutrienti; malattie cardiovascolari; CVD; Meet-in-the-Middle; analisi di mediazione; inferenza causale; selezione di stabilità; esposoma

# Contents

<b>Abstract</b>	<b>i</b>
<b>Abstract in lingua italiana</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>Introduction</b>	<b>1</b>
<b>1 General context</b>	<b>5</b>
1.1 General medical context . . . . .	5
1.1.1 DNA methylation . . . . .	5
1.1.2 Cardiovascular Disease . . . . .	6
1.1.3 Nutrients intake . . . . .	7
1.2 General methodological context . . . . .	8
1.2.1 Exposome-related issues . . . . .	8
1.2.2 Reduce the rate of false positive and false negative signals in expo- some studies . . . . .	9
1.2.3 Causal inference . . . . .	10
1.2.4 Mediation analysis . . . . .	12
1.2.5 Meet-in-the-Middle method . . . . .	15
1.2.6 Stability selection . . . . .	16
<b>2 Data</b>	<b>18</b>
2.1 Data presentation . . . . .	18
2.1.1 EPIC . . . . .	18
2.1.2 Dataset . . . . .	20

<b>3</b>	<b>Preliminary and descriptive analyses of the dataset</b>	<b>25</b>
3.1	Sample selection . . . . .	25
3.2	Exploratory analysis . . . . .	27
3.2.1	Sample characteristics . . . . .	27
3.2.2	Nutrients' characteristics . . . . .	30
3.2.3	DNA methylation levels' characteristics . . . . .	31
<b>4</b>	<b>Meet-in-the-Middle and mediation analysis</b>	<b>33</b>
4.1	Methods for the first MITM approach . . . . .	33
4.1.1	Overall strategy . . . . .	34
4.1.2	Step a): a priori selection of CVD-relevant CpG sites . . . . .	35
4.1.3	Step b): Model and assumptions . . . . .	37
4.1.4	Step c): Model and assumptions . . . . .	40
4.1.5	Visualization tools: Volcano plot, inflation plot, dose-response relationships plot . . . . .	43
4.2	Results for the first MITM approach . . . . .	45
4.2.1	Results step b) . . . . .	45
4.2.2	Results step c) . . . . .	47
4.3	Methods for the second MITM approach . . . . .	52
4.3.1	Overall strategy . . . . .	52
4.3.2	Step a) : Models and assumptions . . . . .	53
4.3.3	Step b) : A priori selection of CVD-relevant CpG sites . . . . .	55
4.3.4	Step c) : Model and assumptions . . . . .	55
4.3.5	Step d) : Models and assumptions . . . . .	56
4.3.6	Step e): Assess the mediation proportion . . . . .	57
4.3.7	Visualization tools: Manhattan Plot . . . . .	59
4.4	Results for the second MITM approach . . . . .	60
4.4.1	Results step a) . . . . .	60
4.4.2	Results step c) . . . . .	66
4.4.3	Results step d) . . . . .	67
4.4.4	Results step e) . . . . .	68
<b>5</b>	<b>Stability selection</b>	<b>71</b>
5.1	Methods . . . . .	71
5.2	Results . . . . .	74
<b>6</b>	<b>Comparison</b>	<b>75</b>

<b>7</b>	<b>Conclusions and future developments</b>	<b>80</b>
	<b>Bibliography</b>	<b>83</b>
<b>A</b>	<b>Appendix A</b>	<b>90</b>
<b>B</b>	<b>Appendix B</b>	<b>93</b>
B.1	Preliminary and descriptive analysis of the dataset . . . . .	93
B.2	MITM 1 . . . . .	99
B.3	MITM 2 . . . . .	107
B.4	Stability selection . . . . .	120

## List of Figures

1	Schematic representation of the link between exposome, intermediate biological layers and health outcome, with a particular highlight on our case study. . . . .	2
1.1	DNA methylation occurs at the C-5 or N-4 positions of cytosine and at the N-6 position of adenine, catalyzed by enzymes DNA MethylTransferases (MTases) . . . . .	6
1.2	On the left a representation of a CpG site (yellow strand) and a GpC site (blue strand), on the right a base pairing between a cytosine and guanine basis. . . . .	6
1.3	Worldwide causes of death in 2019, according to IHME. . . . .	7
1.4	Causal diagram illustrating the structure of confounding [? ]. . . . .	11
1.5	Causal diagram illustrating the structure of collider bias [? ]. . . . .	12
1.6	The total effect between the Exposure (E) and the Outcome (Y), represented in Figure 1.6a can be decomposed in direct and indirect effect through the mediation of M, as in Figure 1.6b . . . . .	13
1.7	Important steps in the development of the MITM approach . . . . .	16
2.1	Italian EPIC cohorts . . . . .	19
2.2	Structure of the EPIC Italy study design regarding the process of gathering information. . . . .	20
2.3	Design of a BeadChip from Illumina 450k platform . . . . .	23
3.1	Distribution and frequencies of the three personal covariates: <i>sex</i> , <i>age</i> and <i>center</i> . . . . .	28
3.2	Frequencies of the four medical covariates: <i>smoking status</i> , <i>diabetes</i> , <i>energy</i> and <i>BMI</i> . . . . .	29
3.3	Distribution of the follow-up covariate ( <i>CVD</i> ) in general and divided by center . . . . .	30
3.4	Spearman correlation matrix of the 43 nutrients divided by the personal daily energy intake . . . . .	31

3.5	Spearman correlation matrix of 10 CpG sites taken randomly among the 399,957 CpG sites composing the DNA methylation dataset . . . . .	32
4.1	Schematic pipeline of the first application of the MITM method . . . . .	35
4.2	Schematic representation of the multiple linear models fitted in step b) . . .	38
4.3	Schematic representation of the multiple generalized linear models fitted in step c) . . . . .	41
4.4	Iron intake histogram with quintile divisions highlighted in red. Each value falls within one of the specified ranges (Q1, Q2, Q3, Q4, Q5), determining its categorization into one of the five levels, forming in this way a categorical variable. . . . .	42
4.5	Volcano Plot representing the significant associations between the CpG sites within the Restricted Methylome and the nutrients within the Whole Exposome. . . . .	46
4.6	QQ-plot showing the inflation of the p-value of the tests conducted at step b), i.e. the deviation of the distribution of the observed tests statistics to the distribution of the expected tests statistics (bisector of the first-fourth quadrant). The inflation factor is $\lambda = 1.03$ . . . . .	47
4.7	Volcano Plot representing all the associations between the 8 nutrients within the Reduced Exposome and the CVD as outcome. . . . .	48
4.8	Dose-response relationships for each of the 8 nutrients within the Reduced Exposome, representing the (exponent of the) estimates for the 2 <sup>nd</sup> , 3 <sup>rd</sup> , 4 <sup>th</sup> and 5 <sup>th</sup> quintiles with refer to the baseline (1 <sup>st</sup> quintile), for each nutrient. In red we highlight the threshold $\beta = 0$ , which allow to distinguish between protective factors (below the line) and risk factor (above the line). . . . .	49
4.9	Schematic pipeline of the second application of the MITM method . . . . .	53
4.10	Schematic representation of the multiple generalized linear models fitted in step a) . . . . .	54
4.11	Schematic representation of the multiple generalized linear models fitted in step c) . . . . .	56
4.12	Schematic representation of the multiple linear models fitted in step d) . . .	57
4.13	Volcano Plot representing all the associations between the nutrients within the Restrincted Methylome and the CVD. . . . .	61

4.14	Dose-response relationships for each of the 5 nutrients within the Reduced Exposome, representing the (exponent of the) estimates for the 2 <sup>nd</sup> , 3 <sup>rd</sup> , 4 <sup>th</sup> and 5 <sup>th</sup> quintiles with refer to the baseline (1 <sup>st</sup> quintile), for each nutrient. In red we highlight the threshold $\beta = 0$ , which allow to distinguish between protective factors (below the line) and risk factor (above the line). . . . .	62
4.15	QQ-plot showing the inflation of the p-value of the tests conducted at step a), i.e. the deviation of the distribution of the observed tests statistics to the distribution of the expected tests statistics (bisector of the first-fourth quadrant). The inflation factor is $\lambda = 1.04$ . . . . .	65
4.16	Manhattan Plot representing the associations between the CpG sites within the Restricted Methylome and the CVD. . . . .	66
4.17	QQ-plot showing the inflation of the p-value of the tests conducted at step b), i.e. the deviation of the distribution of the observed tests statistics to the distribution of the expected tests statistics (bisector of the first-fourth quadrant). The inflation factor is $\lambda = 3.33$ . . . . .	67
4.18	Graphical representations among the variables within the fitted SEM model for Iron. . . . .	69
5.1	Pseudo algorithm used for the stability selection process . . . . .	73
6.1	Synthetic representation of the different strategies used for each method . .	76

## List of Tables

2.1	Overview of the personal covariates used in the analysis for 4 patients in the dataset . . . . .	21
2.2	Overview of the technical covariates used in the analysis to account for the batch-effect . . . . .	23
2.3	Mean and Simultaneous 99% Confidence Intervals of the 43 nutrients used in the analysis . . . . .	24
3.1	Sample selection . . . . .	26
3.2	DNA methylation selection . . . . .	27
4.1	Summary of the process we use to create the Restricted Methylome as the union of the significant CpG sites according to the EWAS catalogue and a systematic review of existing studies. . . . .	37
4.2	Summary of the obtained classification as protective/risk factors and some features of the dose-response relations. . . . .	51
4.3	Summary of the comparison between our findings and the ones coming from existing studies. Green color indicated that our results align with the literature, red means that they are in strong contrast, orange means that the comparison is uncertain. . . . .	51
4.4	Recap Table of the obtained classification as protective/risk factors and some features of the dose-response relations. . . . .	64
4.5	Summary of the comparison between our findings and the ones coming from existing studies. Green color indicated that our results align with the literature, red means that they are in strong contrast, orange means that the comparison is uncertain. . . . .	65
4.6	Number of potential mediators (CpG sites) for each nutrient belonging to the Reduced Exposome. . . . .	68
4.7	P-value assessing the statistical significance of the indirect influence of each nutrient (through the set of potential mediators) on CVD . . . . .	69
5.1	Results of the stability selection algorithm . . . . .	74

6.1	Summary of the main difference among the three methods implemented. . .	77
6.2	Nutrients causally linked to CVD through the methylome layer are highlighted in green, while those establishing an independent causal connection with the outcome are represented in blue. . . . .	78

# Introduction

The concept of exposome was introduced over 15 years ago to reflect the important role that the environment exerts on health and disease [1]. It is used to describe all the simultaneous environmental exposures that an individual encounters throughout life, and how these exposures impact biology and health. It encompasses both external and internal factors, with the former including general external factors, such as air pollution, diet and socio-economic factors, as well as specific external factors like chemicals and radiation, and the latter comprising endogenous factors, such as hormones, inflammation, oxidative stress, and gut microbiota.

Exploiting the exposome is crucial as it holds the key to discover the significant impact the environment has in chronic disease development: understanding the environmental factors that contribute to a disease has significant implications for public health and for the development of more effective strategies for prevention and treatment of the disease [2].

In this thesis the focus will be on a study considering a single disease and an exposome restricted to the dietary exposures (e.g. estimates of intake of specific nutrients) in relation to the CardioVascular disease (CVD).

Moreover, it is widely known that DNA methylation is linked to both CVD and nutrient intake [3], but only a limited amount of researches has investigated the role of DNA methylation in mediating the association of different type of nutrients with CardioVascular risk factors, which is the goal of this thesis project. In this work DNA methylation will be used as an intermediate biological layer measured between the independent variable (nutrients intake) and the dependent one (CVD), representing an essential but complex additional information about the link between the two. A schematic representation of the relation between exposome, intermediate biological layers and health outcome is reported in Figure 1.

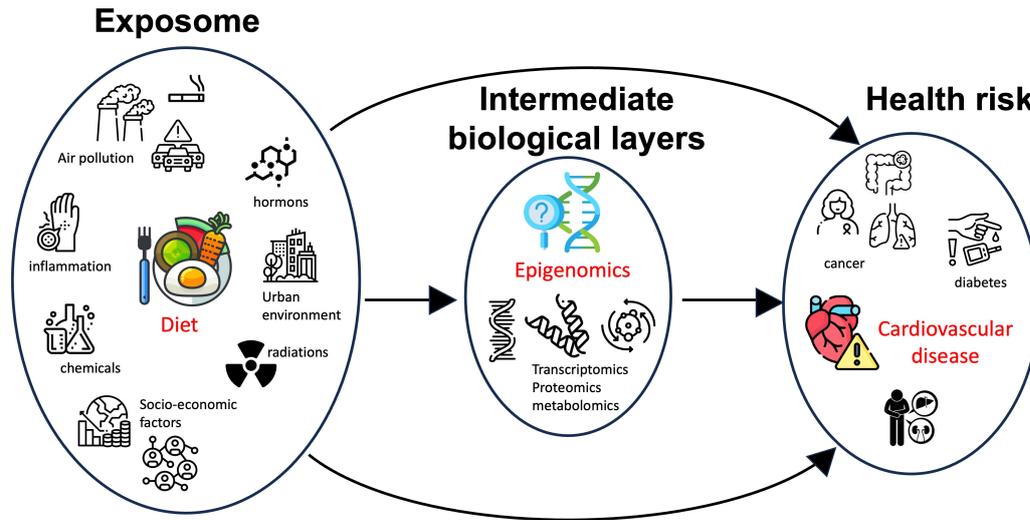


Figure 1: Schematic representation of the link between exposome, intermediate biological layers and health outcome, with a particular highlight on our case study.

Causal inference methods can be implemented to help answer these questions, by offering evidence regarding the presence or absence of a causal effect from the exposure to the disease and by identifying its direction and estimating its magnitude. We focus on a particular category of such methods, the mediation framework. The goal is indeed to identify modifiable factors (here nutrients intake) that, when changed, should lead to a change in the outcome of interest (here CVD).

In this context of causal inference, several methods have been proposed to assess whether DNA methylation's mediation is a significant intermediary and to quantify its magnitude. Among them, a Meet-in-the-Middle (MITM) framework [4] has been developed in the context of studies considering a single exposure and single outcome.

In this work we illustrate two different ways to use a MITM approach to relate a set of dietary exposures to an outcome, here CVD:

- one uses the methylome layer to help pointing potential new exposures likely to be causally associated with CVD,
- the other helps to select potential mediators for exposures associated with CVD.

Moreover, we will compare the results derived from both methodologies with the exposures identified using a variable selection method, specifically stability selection. Notably, this method does not take DNA methylation levels into consideration, allowing us to discern the actual impact of introducing the intermediate biological layer under a MITM framework.

**Aim and structure of the thesis:**

This work is designed to serve as a systematic comparison and evaluation of three methodologies, all converging toward the ultimate goal of establishing causal associations between a set of nutrients and the development of CardioVascular Disease (CVD). The distinctive feature of our approach is the use of the methylome layer as an informative mediator, a critical biological intermediate layer offering precious information on the complex interplay between nutrients and CVD. More in details, our work is structured in the following way:

- Chapter 1: General context.

In this chapter, we furnish a contextual foundation on both the medical and methodological contexts. Regarding the medical background, we offer an exhaustive exploration of three pivotal concepts to this thesis: DNA methylation, CardioVascular disease, and nutrients intake. On the methodological front, we navigate through the methodologies and concepts employed in the study. This includes an exploration of exposome-related issues, causal inference, mediation analysis, Meet-in-the-Middle (MITM) and stability selection methods. It is important to emphasize that a comprehensive understanding of the underlying concepts and methodologies is crucial, given the strong theoretical component of this thesis.

- Chapter 2: Data.

In this chapter, we delve into the analysis of the datasets. We start by introducing the origin of the data and then provide a detailed and comprehensive description of all the factors that will be considered in our analysis: sample characteristics, follow-up information, nutrients, methylation levels and technical covariates. The goal is to offer a clear and thorough understanding of the components that constitute the dataset for the upcoming analysis.

- Chapter 3: Preliminary and descriptive analysis of the dataset.

In this chapter, we conduct the initial exploratory analysis of the dataset. Our journey begins with the careful selection of the definitive samples to be employed in our analysis. Subsequently, we analyse the distributions of the chosen covariates to gain insights into their behavior. Given the pivotal role of correlation among the nutrients constituting the exposome and the CpG sites forming the intermediate layer, we conclude this section with a correlation analysis for both the nutrients and the CpG sites.

- Chapter 4: Meet-in-the-Middle and mediation analysis.

This is the main chapter of the thesis. It presents in details two distinct applications

of the Meet-in-the-Middle approach to relate a set of dietary exposures, the nutrients intakes, to the CVD. The first approach employs the methylome layer to pinpoint new exposures for subsequent testing of their correlation with CVD. The second application aids in the identification of potential mediators linked to CVD exposures and quantifies their impact.

- Chapter 5: Stability selection.

In this chapter we implement a stability selection algorithm with the final aim of selecting potential causal predictor for CVD without considering the intermediate layer.

- Chapter 6: Comparison.

In this chapter we present an extensive comparison of the results obtained with the three methods, highlighting weaknesses and strengths of all of them.

- Chapter 7: Conclusions and future developments.

With this chapter we summarize our results and discuss limitations as well as possible future developments.

# 1 | General context

## 1.1. General medical context

In this section, our aim is to establish clear definitions for key concepts related to DNA methylation, cardiovascular disease, and nutrient intake, while also highlighting the established connections between them.

### 1.1.1. DNA methylation

Genetics is the study of heritable changes in gene activity or function due to the direct alteration of the DNA sequence. Such alterations include point mutations, deletions, insertions, and translocation. In contrast, epigenetics is the study of heritable changes in gene activity or function that is not associated with any change of the DNA sequence itself [5]. One of the major mechanisms that produce such changes is DNA methylation. DNA methylation is an epigenetic mechanism involving the addition of a methyl ( $CH_3$ ) group to the DNA strand itself. It is considered as the major response regulation mechanism for the cell response to environmental changes. DNA contains combinations of four nucleotides which include cytosine, guanine, thymine and adenine [5]; theoretically, each of the DNA bases can be modified, however, only modifications of cytosine and adenine only are known so far. Cytosine methylation is widespread in both eukaryotes and prokaryotes and is the only well-studied DNA modification with established maintenance mechanisms, while adenine methylation has been observed in bacterial, plant, and recently in mammalian DNA, but has received considerably less attention. A visual representation of the adenine and cytosine methylations are shown in Figure 1.1.

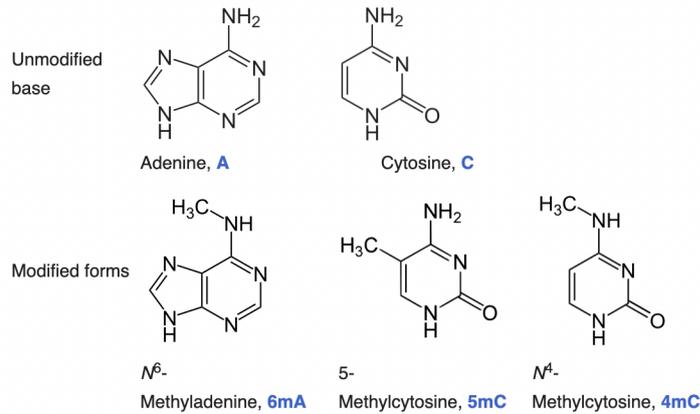


Figure 1.1: DNA methylation occurs at the C-5 or N-4 positions of cytosine and at the N-6 position of adenine, catalyzed by enzymes DNA MethylTransferases (MTases)

DNA methylation is predominantly found at CpG sites. A CpG site refers to a specific DNA sequence where a cytosine (C) nucleotide is followed by a guanine (G) nucleotide in the linear sequence of bases along the DNA molecule. The "p" in CpG stands for the phosphodiester bond that links the cytosine and guanine nucleotides in the DNA backbone.

In mammals, 70% to 80% of CpG cytosines are methylated.



Figure 1.2: On the left a representation of a CpG site (yellow strand) and a GpC site (blue strand), on the right a base pairing between a cytosine and guanine basis.

### 1.1.2. Cardiovascular Disease

CardioVascular diseases (CVDs) are the leading cause of death globally [6]. With this term we indicate any disease involving the heart or blood vessels; the four main types of CVDs are coronary heart disease, stroke, peripheral arterial disease and aortic disease [8]. According to the Global Burden of Disease (GBD) study by the Institute for Health Metrics and Evaluations (IHME) an estimated 18.5 million people died from CVDs in 2019, representing 33% of all global deaths [7].

The World Health Organisation (WHO) estimate that over 75% of premature CVDs is preventable and for this reason understanding the specific risk factors is essential to reduce the growing CVD burden [6].

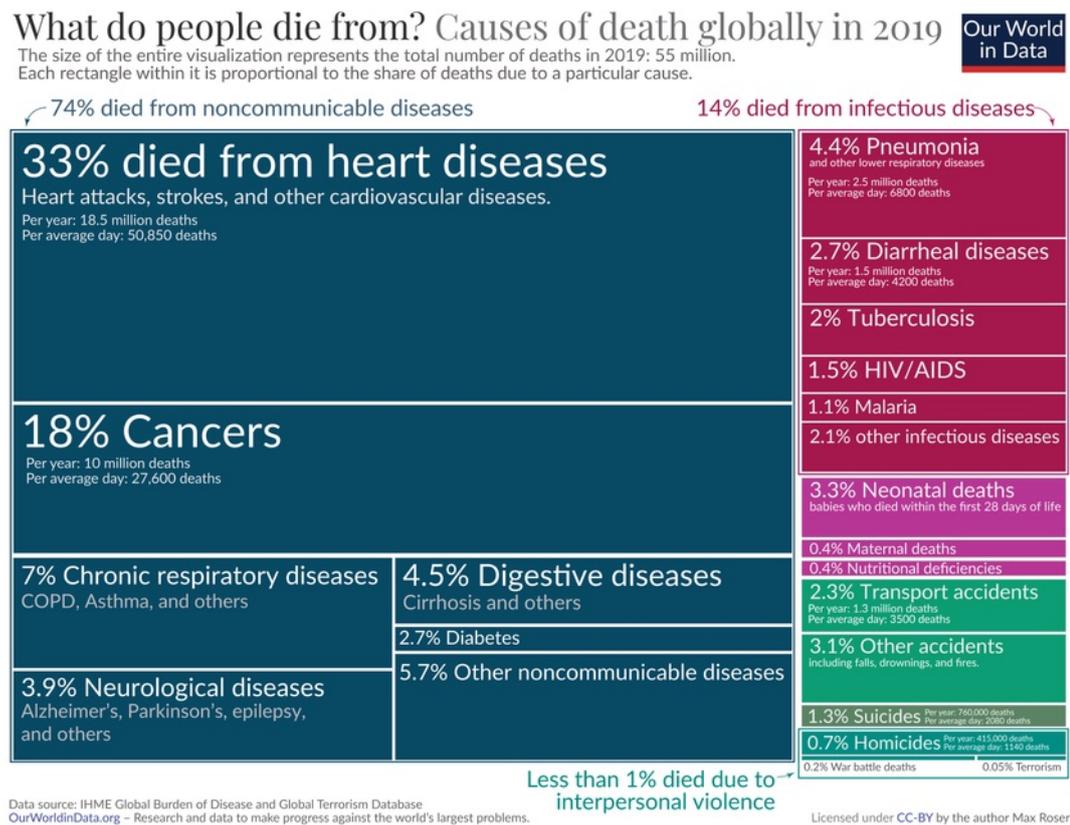


Figure 1.3: Worldwide causes of death in 2019, according to IHME.

It is likely that epigenetic changes mediate, at least in part, the environmental risk for developing or progressing CVD; one prominent factor that is thought to play an important role is DNA methylation. In recent years, numerous studies have solidified the significant association between DNA methylation as a global process and CVD. These findings not only validate this link but also open the way for more focused investigations aimed at yielding specific and refined insights into this established relationship [9].

### 1.1.3. Nutrients intake

Among all the various exposures individuals encounter during every-day life, dietary exposure is one of the most significant and controllable factors. Indeed, unlike air pollution or chemical exposures, managing our dietary choices is more accessible and offers immediate impact. The term "dietary exposure" refers to the measurement of the amount of

a substance consumed by a person or animal in their diet that is intentionally added or unintentionally present (e.g. a nutrient, additive or pesticide) [10]. In our case we are interested to the daily ingested quantity of nutrients, since several studies have elucidated the roles of dietary risk factors on CVD burden [11]. Some nutrients, assumed in specific quantities, can indeed be proven to be harmful or beneficial in the cardiovascular disease matter, so that identifying and regulating them can significantly contribute to enhancing overall health [12]. Moreover, literature regarding experimental studies conducted on animals and humans also supports the idea that diet-induced DNA methylation changes likely contribute to CVD [13].

## 1.2. General methodological context

Now that we have clarified the key components of our analysis, it is necessary to show which role they will have in the development of the analysis. In this section, we will first address certain challenges associated with the exposome (1.2.1), which forms the foundation of our investigation, and outline the strategies we employ to overcome them (1.2.2). Subsequently, we will delve into the theory behind causal inference (1.2.3) and mediation analysis (1.2.4). Finally, we will introduce the two primary methodologies we will utilize and whose outcomes will be compared at the end, namely Meet-in-the-Middle (1.2.5) and stability selection (1.2.6) methods.

### 1.2.1. Exposome-related issues

Exposome studies, while offering invaluable insights into the complex relation between environmental exposures and health outcomes, face significant challenges related to different factors.

- The first one lies is the assessment phase. Exposure may be a single event, but more often it is prolonged and varies over time, with one or more exposure periods. Especially in the latter case, the collected data are often fragmentary and exposure is defined only as a baseline covariate. This approximation might be more relevant when considering some specific kind of exposures (for example chemical or radiation ones), but could be easily overcome when dealing with dietary exposure, like in our case, considering the daily intake [14].
- In exposome studies, one of the main goal is to figure out which causal predictors actually influence the outcome among all the exposures we face, and eventually to

assess their effect [15].

The first method used for this purpose was the Exposome - Wide Association Studies (ExWAS), which looks at how all measured environmental factors in relation to various health outcomes. It uses univariate regressions (adjusted for confounders) relating independently each exposure to the health outcome of interest and possibly corrected for multiple testing. However, a significant downside associated with ExWAS studies is the likely limitation in statistical power. The extensive evaluation of numerous exposures entails substantial costs, rendering it challenging to conduct these studies on a scale large enough to achieve adequate statistical power. Insufficient sample sizes may result in spurious correlations between exposures, contributing to both false negatives (low sensitivity) and false positive findings. Indeed, as for the latter, relying on traditional techniques from basic epidemiology, like univariate linear regression, within ExWAS, may substantially raise the chances of encountering false positives. [15].

### 1.2.2. Reduce the rate of false positive and false negative signals in exposome studies

Using refined statistical methods and/or reducing the dimensionality by adding biological information is helpful in order to reduce the rate of false positive and false negative signals of the classical ExWAS studies [15]. Here, we present both the approaches to address these issues.

- In studies involving numerous comparisons, the likelihood of observing statistically significant results by chance alone is high. Multiple testing corrections reduce the occurrence of false positives by adjusting the p-value to account for the number of comparisons made, thereby controlling the overall Type I error rate. Without correction, the risk of falsely detecting significant results increases as more tests are performed. Two commonly used methods for multiple testing correction are the Benjamini-Hochberg (BH) method and the Bonferroni correction [16]. The Benjamini-Hochberg (BH) method is a widely used procedure for controlling the False Discovery Rate (FDR), which is the expected proportion of false positives among all rejected hypotheses. The Bonferroni correction is a conservative method for controlling the FamilyWise Error Rate (FWER), which is the probability of making one or more Type I errors in a set of tests.
- Another way to cope with the burden of false positive and false negative rate could

be dimension reduction. One way to do dimension reduction is to rely on existing knowledge, using a priori information [15].

A good starting point would be to focus a priori on one exposure (or one set of exposures), using progressed knowledge about the relation with the health outcome of interest [15]. Focusing on a single exposure in research could allow for more precise and targeted results, avoiding the complexity of the wider set of different (and often strongly correlated) factors that make up the exposome.

A further step would be to use the information coming by the introduction of an intermediate biological layer. In general, there are several biological mediators that can be assessed between the exposures and the considered health outcome: the epigenome (DNA methylation), transcriptome (RNA), proteome, and metabolome. These markers can reveal physiological responses to external exposures, effectively serving as internal indicators of health outcomes. These 'omics data, potentially serving as biomarkers for early exposure effects or markers of disease risk, offer valuable but complex additional information regarding the connection between exposures and health. Analyzing the relationship between these biomarkers and the variations in exposure levels or health outcomes may shed light on how the health impact of a given exposure or multiple exposures is biologically mediated, but also to help to select exposures of interest, thus reducing the dimension [15].

Moreover, when dealing with an intermediate layer of high dimension, testing the associations between these potential biomarkers, exposures, and health outcomes can lead to an increased risk of false positive results. In these cases it becomes essential to reduce the dimension of the intermediate layer [17]. Once again, relying on existing knowledge, as for instance comprehensive catalogues and reviews of studies on the same subject, can be a pertinent instrument in achieving the necessary reduction.

### 1.2.3. Causal inference

The task of association analysis, which involves finding interesting connections in large datasets, is not very effective in uncovering the workings of complex diseases [18].

In recent years, there has been a growing realization of the need to transition genomic analysis from association-based approaches to causal inference-based methods and indeed, as a result, a significant portion of recent studies has been moving in this direction. There are big but subtle differences between association and causation. A statistical association between two variables indicates that having knowledge of the value of one variable offers information about the value of the other. However, it doesn't automatically imply a

causal relationship where one variable causes changes in the other. [19]. On the other hand causation means that one variable produces the effect on the other, meaning that the value of one variable changes if measured before and after being subjected to the effect of the other variable. Generally, association does not imply causation (hence the mantra: “association is not causation”), but it could under some specific conditions. To assert that this correlation signifies a causal effect, we must initially address and eliminate two potential issues that can result in a non-causal association: confounding and collider bias. As our aim is to establish a causal relationship between two specific variables, the exposure and health status, we will henceforth refer to the involved variables using this precise terminology, even if causation theory can be applied to more general frameworks.

- Confounding arises when there is a common cause (the confounder) shared between an exposure and an outcome. Failing to account for the confounder may create the appearance of an association between the exposure and the outcome. In reality, both may be influenced by the confounder and may not be directly related (or not as strongly). Only by thoroughly considering and adjusting for all potential confounders can we confidently assert that the exposure causes the outcome. However, identifying and measuring all conceivable confounders in an observational dataset may prove impractical. Consequently, making robust causal claims becomes challenging in practice, especially when dealing with unknown and unmeasured confounders.

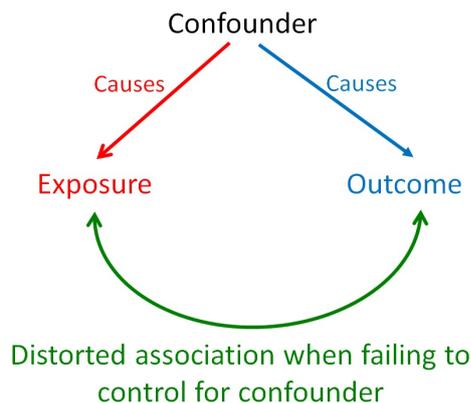


Figure 1.4: Causal diagram illustrating the structure of confounding [? ].

- Collider bias arises when an exposure and outcome have a shared effect, the collider. When we control for this collider, it results in a distorted association between the exposure and outcome. It is possible that both the exposure and the outcome influence a common factor. In such instances, controlling for the collider can lead

to an altered association between the exposure and outcome.

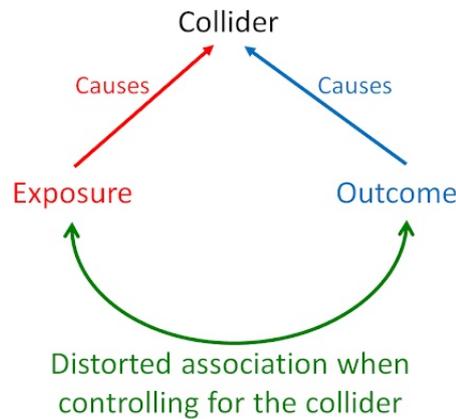


Figure 1.5: Causal diagram illustrating the structure of collider bias [? ].

Colliders introduce bias they are controlled for, while confounders introduce bias when left uncontrolled. To mitigate both biases, it is essential to identify and control for as many confounders as possible during the analysis. Simultaneously, recognizing all colliders and leaving them uncontrolled is crucial for minimizing bias in research outcomes.

Associations may represent causal effects, but this is only potentially valid when we effectively control for all confounders, refrain from controlling for colliders, and establish the temporal precedence of the exposure and outcome. Despite these efforts, the presence of unknown confounders, colliders, and other biases can compromise our conclusions. While excluding confounders and colliders among the measured variables is a crucial step in establishing causation, it alone is insufficient for ensuring robust causal inferences.

Causality has revealed to be more important than a merely association because it unveils the underlying biological mechanisms driving genetic associations between exposures and outcome. Understanding causal relationships allows for targeted interventions and personalized treatments based on a solid understanding of how specific certain exposures directly influence diseases.

#### 1.2.4. Mediation analysis

Mediation analysis is a method from the field of causal inference which, in the last years, is increasingly being applied in many research fields, including the epidemiology one [20]. In epidemiology, it is used to explore the mechanism through which an exposure exposure or a treatment influences the outcome. Assuming a causal relationship from an exposure to an outcome, the mediation analysis aims at decomposing the total exposure-outcome effect into a direct effect and an indirect effect through a mediator variable [20] and

quantifying each of them.

Mediation analysis was developed as path analysis in the genetic field, and later in the area of social sciences, and then further formalized in biomedical research in connection with regression modeling in the counterfactual outcome framework of causal inference [21].

If we assume that part of the effect of the Exposure (E) is mediated by a Mediator (M) then the proportion of the association between E and the outcome (Y) that occurs through M is termed the indirect (or mediated) effect of E. The fraction of the effect of E that occurs independently of M is called direct effect.

It is important to specify that at the basis of all the theory there is the strong assumption of knowing that the effect from E to Y is causal. Rigorous causal inference methods and study designs are necessary to establish a causal relationship and confirm the mediator as true mediators in the pathway between exposure and outcome.

In Figure 1.6 we see how the total exposure-outcome effect can be decomposed into direct and indirect effect through a mediation variable.

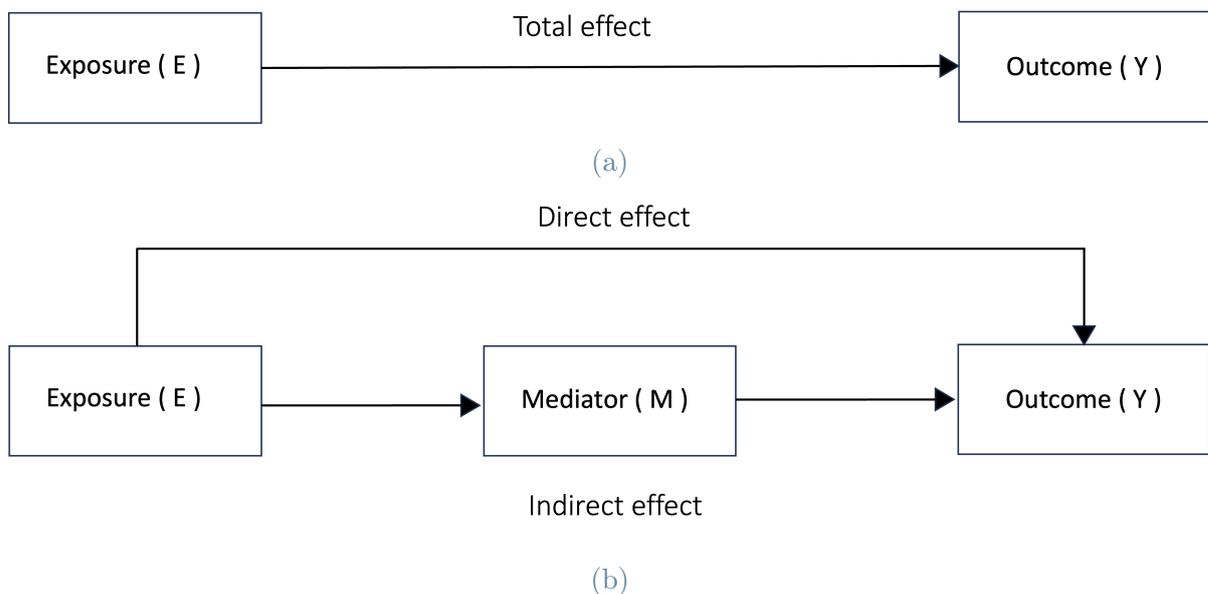


Figure 1.6: The total effect between the Exposure (E) and the Outcome (Y), represented in Figure 1.6a can be decomposed in direct and indirect effect through the mediation of M, as in Figure 1.6b

The more popular guidelines for mediation analysis have been developed assuming that both the mediator and the outcome are continuous variables, but this is not our case since the outcome of interest (which is the development or non-development of the cardiovascular disease) is a binary variable indicating if the cardiovascular event had happened

or not. For this reason we will present here the adapted relative equations under the hypothesis that the mediator is a continuous variable, the outcome is a binary variable and the residuals are normally distributed. We can fit two regression models regarding the two indirect effect contributions, then the direct effect can then be estimated as the sum of the two contributions. [21] [22]

- **Exposure-outcome model:**

$$\text{logit}(\mathbb{P}(Y = 1|E, M)) = \beta_0 + \beta_1 E + \beta_2 M + \beta_3 C$$

Here we use a logistic regression since the output is binary.  $\mathbb{P}$  is the probability of obtaining an event,  $E$  is the exposure variable,  $C$  is a matrix including all the possible confounders of the exposure-outcome associations. The betas are the unknown parameters estimate.

- **Exposure-mediator model:**

$$\mathbb{E}(M|E) = \theta_0 + \theta_1 E + \theta_2 C'$$

Here we use a normal regression since the mediation is a continuous variable.  $\mathbb{E}$  is the mathematical expectation,  $C'$  is a matrix including all the possible confounders of the exposure-mediator associations.

In the setting of binary outcomes, one can define the direct effect using odds ratios. The direct effect corresponds to the change in the odds of the outcome for an increase by one in exposure, assuming that the mediator remains fixed at a specific level.

In this context we are hypothesizing that there is no interaction between the exposure  $E$  and the mediator  $M$ , and indeed the term comprising the combined effect of the two is put to zero in the equation related to the exposure-outcome model.

The estimation of direct and indirect effects requires two major assumptions: a lack of exposure–outcome confounding (i.e. in the equation efficient adjusting for  $C$ ) and a lack of mediator–outcome confounding (efficient control for  $C'$ ) [21].

Mediation analysis can prove to be efficient in revealing hidden mechanisms, but it needs to face some big challenges. First of all, the causal relationship between the exposure and the outcome must be known. Secondly, the extension of mediation analysis to the case of several mediators  $M_1, \dots, M_p$ , where  $p$  is much higher than the number of observations  $n$  (as it happens in our case when the mediator is the DNA methylation, which has a dimension of  $\sim 4 * 10^6$  CpG sites) is challenging and need to be threaten with the proper accuracy [21].

Changes in DNA methylation at CpG sites caused by exposures can influence gene expression patterns and, subsequently, biological processes relevant to health outcomes. This is

the reason why, generally CpG sites are suitable and good candidates as mediators, even if it depends on the specific exposure, outcome, and biological pathways involved.

### 1.2.5. Meet-in-the-Middle method

In recent years, there has been a remarkable increase in the incorporation of endogenous ('omic) biomarkers of effect in applied environmental health research. This trend has paved the way for the need to develop new and more targeted methods, including the Meet-in-the-Middle (MITM) approach, which is one implementation of high-dimensional mediation frameworks [23].

The concept of identifying the overlap between markers of exposure and predictive markers of disease has been defined as "Meet-in-the-Middle" by Vineis and Perera in 2007 [24]. According to this original idea, finding out that some biomarkers are both related to specific exposures and a specific disease would strengthen the causal links between exposures and disease, even without a formal mediation analysis. This approach, according to Vineis and Perera, has potential for prevention by identifying the specific environmental factors involved in the disease process.

A few years later, in 2010, Chadeau-Hyam and colleagues [4] developed a possible implementation of the MITM approach, with the objective of identifying a list of putative intermediate biomarkers that link exposure and disease outcome. This MITM approach consist in measuring intermediate biomarkers and studying how they respond to external exposures and how these responses are linked to subsequent health outcomes. If the same set of markers is robustly associated with both ends of the exposure-to-disease continuum, this is a validation of a causal hypothesis according to the pathway perturbation paradigm. [25]. This means that, in this scenario the MITM approach has been conceived as a research strategy to identify biomarkers that are related to specific exposures and that are, at the same time, predictive of disease outcome. Finding this overlap between exposure and disease of 'intermediate' biomarkers can potentially disclose useful information on the exposure-to-disease pathway [26].

In 2020 Cadiou and colleagues developed a tailored MITM approach [17], which was named oriented Meet-in-the-Middle (oMITM) in a later work in 2021 ([15]) to underline the differences between this new design and the classical Meet-in-the-Middle. The oMITM relies on an intermediary biological layer to restrict the exposures associated with relevant intermediary features, whose association with health is then tested. The idea is that oMITM could 1) allow lowering the high false discovery proportion (FDP) reported for agnostic ExWAS, and 2) could be less sensitive to reverse causality than agnostic dimension reduction methods. This might be obtained at a cost of a decreased sensitivity, in particular as

the proportion of exposures whose health effect is not mediated by the considered layer increases.

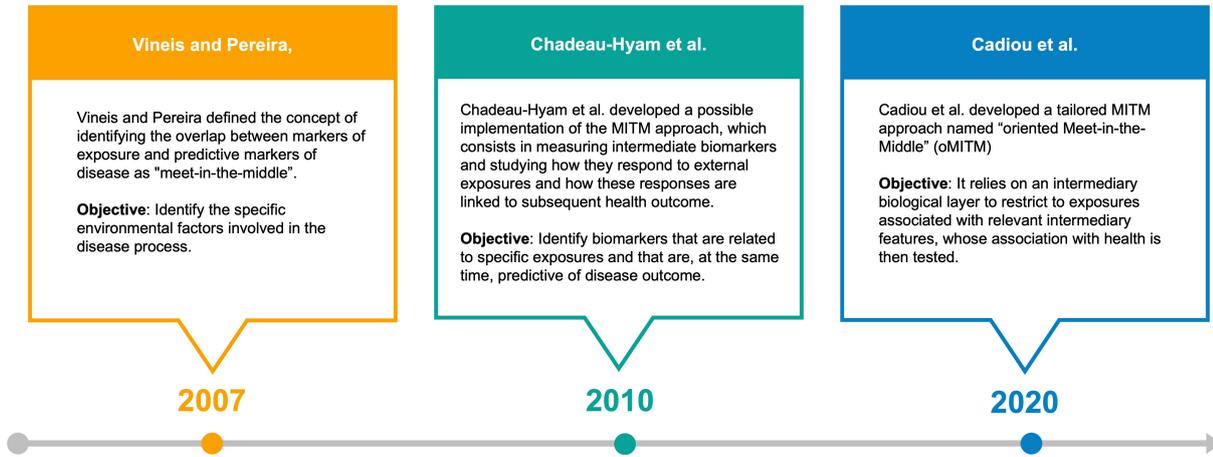


Figure 1.7: Important steps in the development of the MITM approach

Since the Meet-in-the-Middle method has been developed and deepened in the very last years, the existing literature and research on this subject remain somewhat limited. This presents both an opportunity for the exploration of groundbreaking techniques and applications and a challenge when it comes to validating these applications and their outcomes.

### 1.2.6. Stability selection

Stability selection is a variable selection method based on subsampling in combination with (high dimensional) selection algorithms [63]. It is not a new variable selection technique, but instead its aim is rather to enhance and improve existing methods. It uses an existing selection algorithm and complement it with resampling techniques to estimate the probability of selection of each variable using its selection proportion over the resampling iterations. Stability selection ensures reliability of the findings through error control.

In our regression framework, the variable selection algorithm we use is the least absolute shrinkage and selection operator (LASSO). LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

LASSO regression starts with the standard linear regression model and then introduces an additional penalty term based on the absolute values of the coefficients. The goal of the algorithm is to minimize the following:

$$\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

where:

- $y$  is the dependent variable (target)
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  are the coefficients (parameters) to be estimated
- $x_1, x_2, \dots, x_n$  are the independent variables
- $\lambda$  is the regularization parameter that controls the amount of regularization applied.

One might opt for a stability selection method because stability is a pivotal prerequisite for generalizability [27]. The ability of the predictor set to generalize is indispensable for establishing causality within the chosen predictors. Consequently, techniques focused on optimizing "estimation stability," emphasizing the stability of both the predictor set and their estimates, are more inclined to identify causal predictors compared to approaches centered on "prediction stability" [28].

# 2 | Data

## 2.1. Data presentation

Having outlined both the necessary medical and methodological framework, we can now delve into a more detailed and aware articulation of the thesis's objectives.

In this section, we will provide an overview of the datasets that we used in our analysis, with a specific focus on the variables we selected to use in our study.

### 2.1.1. EPIC

The data we use in our analysis were collected from the European Prospective Investigation into Cancer and Nutrition (EPIC) study [29], which is a large-scale study that aims to investigate the relationship between diet, lifestyle, environmental factors, and the development of chronic diseases, including cancer, cardiovascular disease, and diabetes. It is one of the largest cohort studies in the world, including over 521,000 participants enrolled from 23 centres in 10 western European countries. Recruitment of study participants and collection of data and biological samples started in 1993 in four countries (Spain, Italy, France, and the United Kingdom) and was extended between 1994 and 1998 to include six more countries (Greece, Germany, the Netherlands, Denmark, Sweden, and Norway). During the recruitment phase of the study, EPIC collected extensive information on various aspects of participants' lifestyles, including their diet, physical activity levels, medical history, and anthropometric measurements. Additionally, biological samples were taken from a subset of 387,889 individuals at baseline, from which DNA methylation levels were extracted. These samples are currently stored at the International Agency for Research on Cancer (IARC), which is a part of the World Health Organization (WHO).

The EPIC cohort in Italy was initially established in four regions based on cancer registries: the provinces of Florence, Ragusa, Varese and Turin. Subsequently, an additional location, the city of Naples, was included as part of the EPIC study through the Progetto ATENA research program.

The recruitment of participants in each of these regions was motivated by different reasons

and targeting different groups of people, having different habits and health conditions. As a result, the samples obtained from each location exhibit a significant level of diversity. In particular:

- Recruitment of the EPIC-Florence cohort took place between 1993 and 1998. With 13,597 participants (3514 males and 10,083 females), it is the largest Italian EPIC cohort. Moreover it is the only Italian EPIC centre located in central Italy, an area characterized by a high consumption of fresh and cured meats, wine, and olive oil.
- In Naples all the samples were collected through the Progetto ATENA at the beginning of the 1990s. It is a study on the etiology of major chronic diseases in women, recently transformed for incorporation into the EPIC database.
- In Turin the recruitment took place in 1993 - 1998 and involved blood donors and other healthy volunteers.
- In Ragusa the recruitment began between 1992 and 1993, and it was finished in 1997. It was conducted by personnel of the local population-based cancer registry.
- Recruitment was carried out at two hospital centres in the province of Varese between 1992 and 1997. Volunteers were recruited through general practitioners' records, factories, and schools or through letters sent to their homes.



Figure 2.1: Italian EPIC cohorts

### 2.1.2. Dataset

We have access to a subset of the EPIC Italy dataset, which includes 1,780 participants from all the 5 centers. For each of them we are provided 136 variables, which include personal information, daily nutritional and dietary habits, health event follow-ups, as well as methylation levels measured at almost 400,000 sites. In particular blood samples and initial information regarding lifestyle and dietary habits were gathered at the time of recruitment when all individuals were in good health. Over the subsequent 20 years, regular follow-ups were conducted to monitor the development of CVDs or diabetes among the participants. A visual representation of this process is available in Figure 2.2.

#### EPIC Italy study design

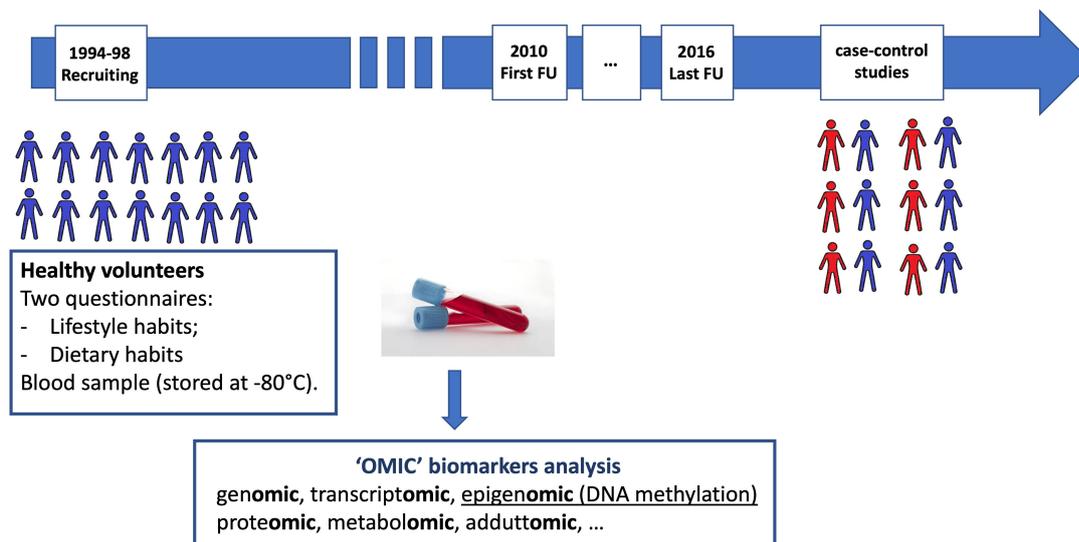


Figure 2.2: Structure of the EPIC Italy study design regarding the process of gathering information.

Let's have a better insight in the variables we will use in our analysis:

- **Sample characteristics:** During the recruitment phase of our study, all the participants (identified by a unique anonymized *key*) were asked to provide a range of information that is essential for a comprehensive characterization of our cohort. This information includes personal covariates such as *sex* (male or female), *age*, and *center* (Florence, Turin, Ragusa, Naples or Varese), as well as medical covariates such as *diabetes status* (categorized as yes or no), daily *energy* intake (expressed in *kcal/day*) and Body Mass Index (*BMI*), which is a measure of body fat based on a person's weight and height. In addition to these factors, participants were

also asked to provide data on lifestyle factors such as *smoking habits* (categorized into three classes: Current, Former or Never). An overview of these variables for 4 participants is reported in the Table 2.1

anonymized key	sex	age	center	diabetes status	energy	BMI	smoking habits
HaGIu4N1SC	Female	53.50034	Turin	No	2289.396	20.51784	Former
41xsoUYL79	Female	53.92471	Ragusa	No	1996.164	31.47599	Never
BDxoY6eg4s	Female	67.08830	Naples	No	2870.708	41.73268	Current
N3m77TNARd	Male	53.90281	Varese	No	2869.963	26.17796	Former

Table 2.1: Overview of the personal covariates used in the analysis for 4 patients in the dataset

- **Nutrients:** In addition to providing personal and medical information, each participant was also asked to complete a questionnaire about their dietary habits, which included a detailed list of 67 different foods and the quantities consumed. Using this information, daily quantities of 67 different foods and 43 nutrients-related data were extracted for each participant. The latter includes daily quantities of nutrients, dietary indexes and total amount of energy; apart from energy, which will be used as a covariate in the analysis, the other quantities here will be referred to for the rest of the analysis with the general term nutrient, being a collection of measurements related to that category. A more specific reference with some statistical quantities is provided in the Table 2.3
- **Follow-up:** Each patient participated in various follow-up sessions, during which any discovered diseases were recorded, along with their respective dates. In this way, we have access to several binary variables that indicate the presence or absence of specific conditions, including cardiovascular disease, cancer (with the specific site of occurrence), and diabetes development, along with the respective diagnostic dates. Moreover we also have access to the current condition of the patient (Alive, Dead, Moved or Emigrated) at the last check-up. In our analysis we decided to focus on the cardiovascular disease (CVD) and we didn't keep into consideration the time line.
- **Methylation levels:** For each patient we also have access to methylation data, which were extracted from blood samples and analysed by the HumanMethylation450 BeadChip platform [30]. It is a platform which offers a unique combination of comprehensive, expert-selected coverage and high throughput at a low price, mak-

ing it ideal for screening large sample populations, like in our case. It covers more than 450,000 methylation sites, with over than 98% reproducibility for technical replicates.

In general, DNA methylation levels can be quantified by using  $\beta$  values or M values, which are connected by the formula:  $M = \log_2\left(\frac{\beta}{1-\beta}\right)$ . As done in other EPIC Italy publications, in our analysis we use  $\beta$  values, which represent the percentage of methylation at a specific CpG site, ranging from 0 (unmethylated) to 1 (fully methylated). They are calculated as the ratio of the intensity of the methylated probe signal to the sum of the intensities of both methylated and unmethylated probe signals.  $\beta$  values are commonly used in DNA methylation studies because they are easy to interpret and compare across samples. For each participant, after the selection explained in section 3.1, we have methylation levels measured at 399,957 different CpG sites.

- **Technical information:** The design of the Illumina 450k platform allows for the simultaneous testing of multiple loci on a fixed array size, which increases the efficiency of the analysis. As shown in Figure 2.3, each BeadChip is made up of 12 arrays arranged in a six rows by two columns layout, with each array capable of processing one sample. Therefore, up to 12 samples can be processed simultaneously on a single BeadChip.

The categorical variable "chip position", assigned to each participant, indicates the location of their sample on the 2D array format of the BeadChip. For example, a sample processed in the third row and first column will have the label "R03C03".

Furthermore, the technology used in the Illumina 450k platform has evolved during the years, resulting in the development of new versions of BeadChips. As a result, each BeadChip has a unique ID, which is stored in the categorical variable "chip", which has 196 levels.

It is crucial to consider these two variables as batch variables to account for any systematic differences in sample processing or measurement that may be associated with the plate position.

An overview of these variables for 4 participants is reported in the Table 2.2

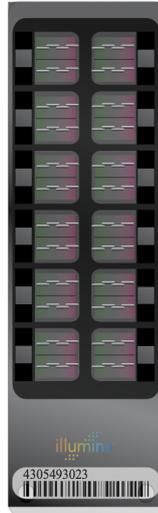


Figure 2.3: Design of a BeadChip from Illumina 450k platform

key	chip position	chip
HaGIu4N1SC	R01C01	200109360008
41xsoUYL79	R02C02	200118310025
BDxoY6eg4s	R01C01	200118310004
N3m77TNARd	R04C01	7668610110

Table 2.2: Overview of the technical covariates used in the analysis to account for the batch-effect

Nutrients	Mean	99% CI
Edible portion	1674.2018 <i>gr/day</i>	[1565.4244, 1782.9792] <i>gr/day</i>
Water	1288.1499 <i>gr/day</i>	[1199.9808, 1376.3191] <i>gr/day</i>
All proteins	90.8508 <i>gr/day</i>	[84.6387, 97.0629] <i>gr/day</i>
Animal protein	58.5629 <i>gr/day</i>	[53.9248, 63.2010] <i>gr/day</i>
Vegetable protein	29.8911 <i>gr/day</i>	[27.1879, 32.5944] <i>gr/day</i>
Total fat	86.7818 <i>gr/day</i>	[80.3001, 93.2636] <i>gr/day</i>
Animal fat	48.4183 <i>gr/day</i>	[43.8258, 53.0107] <i>gr/day</i>
Vegetable fat	38.3402 <i>gr/day</i>	[35.0023, 41.6781] <i>gr/day</i>
Sfa (saturated fatty acids)	30.3134 <i>gr/day</i>	[27.6917, 32.9352] <i>gr/day</i>
Oleic acid	38.3083 <i>gr/day</i>	[35.3510, 41.2656] <i>gr/day</i>
Mufa (monounsaturated fatty acids)	40.8712 <i>gr/day</i>	[37.7490, 43.9933] <i>gr/day</i>
Linoleic acid	8.6843 <i>gr/day</i>	[7.8126, 9.5560] <i>gr/day</i>
Linolenic acid	1.2905 <i>gr/day</i>	[1.1885, 1.3924] <i>gr/day</i>
Pufa (polyunsaturated fatty acids)	10.6213 <i>gr/day</i>	[9.6441, 11.5985] <i>gr/day</i>
Other pufa	0.6369 <i>gr/day</i>	[0.5574, 0.7165] <i>gr/day</i>
Cholesterol	358.3156 <i>gr/day</i>	[328.3018, 388.3294] <i>gr/day</i>
Avail carbohydrates	265.7768 <i>gr/day</i>	[244.5078, 287.0457] <i>gr/day</i>
Starch	161.1243 <i>gr/day</i>	[144.7785, 177.4701] <i>gr/day</i>
Soluble carbohydrates	104.4301 <i>gr/day</i>	[94.7598, 114.1003] <i>gr/day</i>
Fiber	22.5469 <i>gr/day</i>	[20.2905, 24.8032] <i>gr/day</i>
Alcohol	15.1718 <i>gr/day</i>	[11.0550, 19.2886] <i>gr/day</i>
Iron	0.0144 <i>gr/day</i>	[0.0134, 0.0154] <i>gr/day</i>
Calcium	1.0392 <i>gr/day</i>	[0.9430, 1.1355] <i>gr/day</i>
Sodium	2.3888 <i>gr/day</i>	[2.1619, 2.6157] <i>gr/day</i>
Potassium	3.3643 <i>gr/day</i>	[3.1488, 3.5797] <i>gr/day</i>
Phosphorus	1.4684 <i>gr/day</i>	[1.3672, 1.5696] <i>gr/day</i>
Zinc	0.0129 <i>gr/day</i>	[0.0119, 0.0138] <i>gr/day</i>
Vitamin B1	0.0010 <i>gr/day</i>	[0.0010, 0.0011] <i>gr/day</i>
Vitamin B2	0.0016 <i>gr/day</i>	[0.0015, 0.0017] <i>gr/day</i>
Vitamin B3	0.0189 <i>gr/day</i>	[0.0176, 0.0203] <i>gr/day</i>
Vitamin C	0.1406 <i>gr/day</i>	[0.1246, 0.1565] <i>gr/day</i>
Vitamin B6	0.0020 <i>gr/day</i>	[0.0018, 0.0021] <i>gr/day</i>
Folic acid	0.000284 <i>gr/day</i>	[0.000262, 0.000307] <i>gr/day</i>
Retinol equivalents	0.0012 <i>gr/day</i>	[0.0010, 0.0014] <i>gr/day</i>
Retinol	0.0006 <i>gr/day</i>	[0.0005, 0.0008] <i>gr/day</i>
Beta carotene	0.0033 <i>gr/day</i>	[0.0029, 0.0038] <i>gr/day</i>
Vitamin E	0.0084 <i>gr/day</i>	[0.0077, 0.0090] <i>gr/day</i>
Vitamin D	0.0016 <i>gr/day</i>	[0.0013, 0.0018] <i>gr/day</i>
TEAC (trolox equivalent antioxidant capacity)	6.6100	[6.0350, 7.1850]
TRAP (total radical-trapping antioxidant parameter)	9.9563	[9.0654, 10.8472]
FRAP (ferric reducing antioxidant power)	19.8273	[18.1154, 21.5392]
Glycemic load	141.9545 <i>gr/day</i>	[130.2802, 153.6288] <i>gr/day</i>
Glycemic index	0.5331	[0.5269, 0.5394]

Table 2.3: Mean and Simultaneous 99% Confidence Intervals of the 43 nutrients used in the analysis

# 3 | Preliminary and descriptive analyses of the dataset

This chapter will present a preliminary and descriptive analysis that will serve two important purposes. First, it will enable us to select the final samples that we will use for our analysis, based on specific selection rules that will be described in detail (3.1). Second, it will help us to gain a deeper understanding of the underlying structure of the dataset (3.2).

## 3.1. Sample selection

To ensure the validity and reliability of our research findings, it is essential to carefully select a suitable sample for the analysis, based on our specific case. For this reason we begin our inspection by presenting some sample selections we have done on our dataset. Since our response variable of interest is the cardiovascular outcome, we start our inspection by examining the distribution of this variable and its relationship with other covariates in the dataset. As previously mentioned, belonging to different cohorts can induce substantial differences in the distribution of the variables, due to different recruitment groups, medical protocol rules or simply different lifestyle habits. In our specific case, all the 257 people belonging to the Florence cohort don't have any follow-up to establish if they have developed the CVD, and for this reason all the center is excluded, leaving data regarding four centers: Varese, Turin, Ragusa and Naples.

We also consider the potential impact of shared risk factors between cancer and CVD on our study results. Cancer survivors are at an increased risk of developing CVD due to lifestyle factors, cancer-associated inflammation, and iatrogenic effects of cancer therapy [31]. To account for this potential confounding factor, we exclude 15 individuals who developed cancer before experiencing CVD from the 59 people who had both conditions. We conduct an additional check to ensure that the dataset we use for our analysis is clean and does not contain an excessive number of missing values that could potentially invalidate our results. We find that the dataset is, in fact, quite clean, with no missing

values in the covariates *sex*, *age*, *center*, *smoking status*, *chip*, *chip position* and in all the nutrients, while just a few values were found in the variables *bmi* and *diabetes*.

The final dataset we use thus includes 1,508 subjects, with 9 covariates (listed just before) and 43 nutrients.

In the same way, we proceed to inspect the methylation data to identify the CpG sites that are suitable for our analysis.

We first proceed by excluding all the non-CpG methylation sites, which are indicated with the notation "ch" instead of "cg" in the dataset. We identify 1,592 sites to remove, retaining in this way 409,126 remaining CpG sites.

While recent studies have shown that human disease phenotypes can manifest differently by sex [32], we decid to remove X and Y chromosomes from our analysis for two main reasons. Firstly, the data we have prove to be unreliable due to the presence of the Y chromosome in data regarding women, which should not be present. Secondly, including these two chromosomes would complicated the analysis and increase the impact of confounding factors, making it challenging to distinguish between true DNA methylation differences and X-inactivation patterns. Therefore, we deem it best to remove the X and Y chromosomes to improve the accuracy and validity of our study findings, as it is done in most EPIC studies using methylation data. We remove 9'117 CpG sites located in the X chromosome and 52 CpG sites located in the Y one. In this way we remain with a total number of CpG sites equal to 399957.

A recap of the removed data, along with the removal reasons, is shown in Tables 3.1 and 3.2

Number of people removed	Number of people retrieved	Reason of removal
257	1,523	Belonging to a center (Florence) in which there were no follow-ups to establish the development or not of CVD.
15	1,508	Having fallen ill with cancer before experiencing the CVD.

Table 3.1: Sample selection

Number of methylation sites removed	Number of methylation sites retrieved	Reason of removal
1,592	409,126	Non-CpG methylation sites
$9,117 + 52 = 9,169$	399,957	CpG sites located in X and Y chromosomes

Table 3.2: DNA methylation selection

## 3.2. Exploratory analysis

Now that we have the final set of individuals and CpG sites, we can proceed with a preliminary analysis. The aim of this section is to get a general idea of the dataset, look for anomalies or unbalances which could create issues in the future steps and start looking at the correlation among nutrients and CpG sites, which is an important factor to keep under consideration in the later analysis.

### 3.2.1. Sample characteristics

We start by looking at the distribution of the three personal covariates: *sex*, *age* and *center*. The histogram of the age at the recruitment time shows that the values cover the range 34-78 years old, with a median value of 53.8 years. The barplot of the covariate *sex* indicates a slight imbalance between men and women, with women comprising approximately 60% of the sample and men comprising approximately 40%; the difference is limited and it is probably increased by the presence of only women in the "Naples" cohort, as already mentioned before. The distribution of patients across the four centers - Varese, Ragusa, Turin, and Naples - varies significantly, with proportions of 0.34, 0.08, 0.53, and 0.06, respectively, but we expect it not to be an issue.

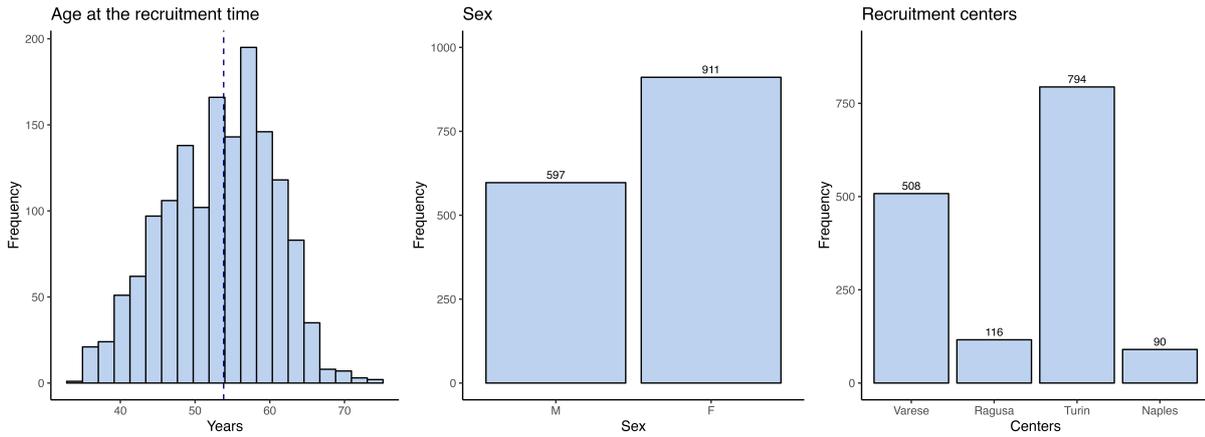


Figure 3.1: Distribution and frequencies of the three personal covariates: *sex*, *age* and *center*

We now examine the distributions of the other four medical and lifestyle covariates. The *smoking status* covariate has three levels - *Never*, *Former*, and *Current* - with proportions of 0.45, 0.28, and 0.27, respectively, indicating a generally balanced distribution. The *diabetes* covariate is constructed as a binary variable, with a value of 1 indicating a patient who had diabetes either at recruitment or developed it later, and 0 indicating no diabetes. Only 6.9% of patients in the sample had diabetes, a percentage consistent with the National Institute of Statistics (ISTAT) estimate of 5.6% in Italy. The *energy* covariate is expressed in units of *kcal/day* and represents the daily caloric intake of each patient. Its mean and median values are 2,251.6 and 2,172.7 *kcal/day*, which is in line with the established mean intake for people in this age range. The *BMI* covariate has mean and median values of 26.3 and 25.9  $kg/m^2$  which fall within the overweight category (characterized by a *BMI* of 25 – 30  $kg/m^2$ ). As for now all the covariates show a balanced behaviour which is indicative of a well-constructed dataset.

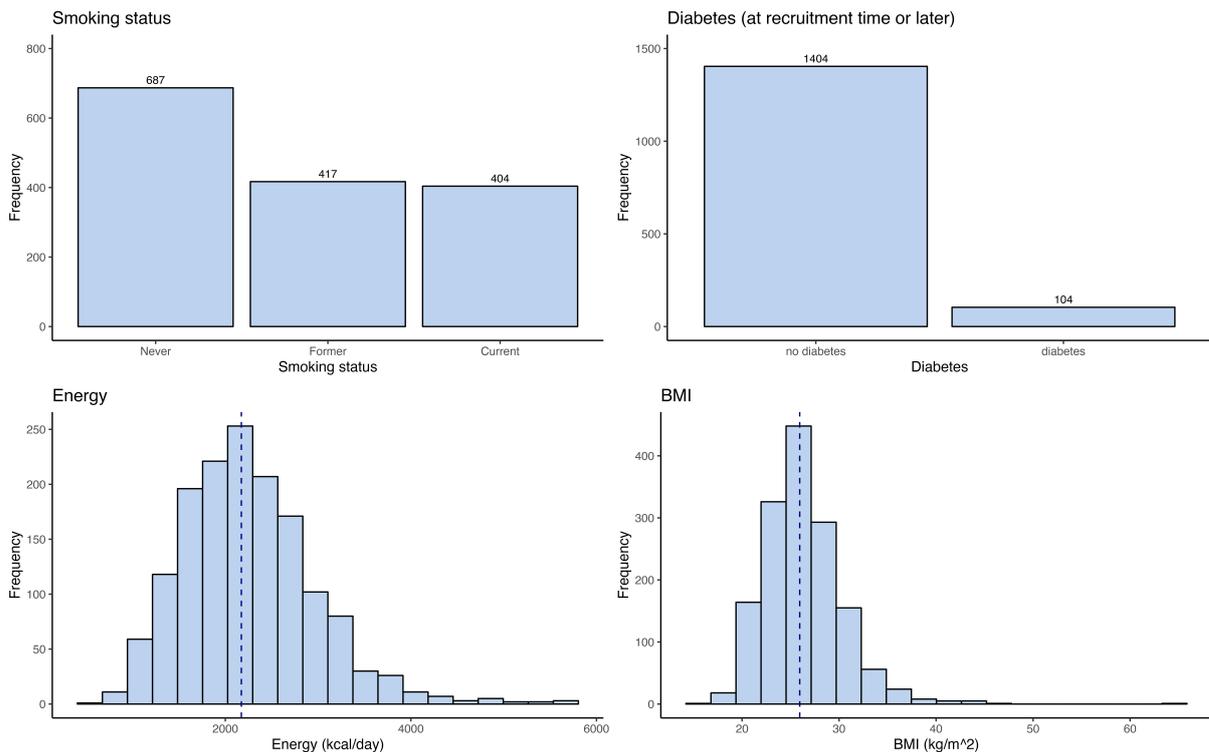


Figure 3.2: Frequencies of the four medical covariates: *smoking status*, *diabetes*, *energy* and *BMI*

We now proceed to analyze the variable describing the outcome of interest: the presence or absence of cardiovascular disease. In this phase we need to ensure that there are no substantial imbalances which could alter the results, given that this outcome variable will serve as the output for numerous regression models utilized throughout the future analysis. For this reason, we also look at its distribution across different centers and calculate the respective proportions within the four cohorts. The division among the centers turns out to be reasonably balanced, with the occurrence of CVD ranging from 13.9% (Naples) to 28.3% (Varese), which is comparable to the 18.6% of people falling ill with CVD in the whole dataset. It's worth noting that the observed proportion is slightly higher than what might be expected in a random population, and this can be attributed to two reasons. Firstly, the participant group has been chosen according to a case-control study design. In this context, the group is formed by joining individuals with the specific disease considered (cases) and a group of individuals without the disease but similar to the cases in other relevant characteristics (controls). Case-control studies are particularly useful when studying rare diseases with long latency periods, as they allow for a more efficient use of resources compared to prospective cohort studies. However, they are also susceptible to biases, as illustrated in this case where the observed proportion of CVD

occurrences deviates from the global expected rate [33]. A second factor contributing to the potential decrease in the observed proportion of CVD cases is the exclusive inclusion of women in Naples. Women typically exhibit a higher life expectancy, which, in turn, results in a reduced risk of mortality.

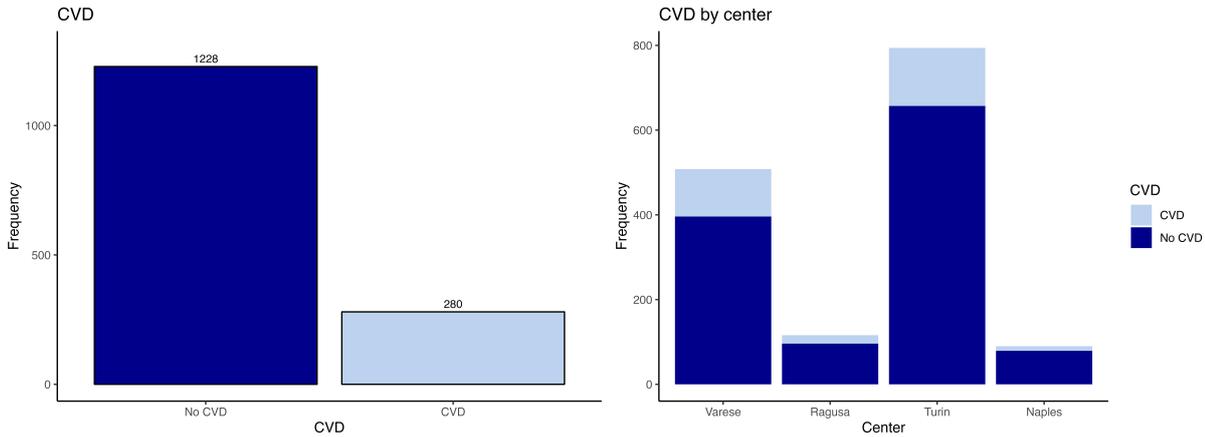


Figure 3.3: Distribution of the follow-up covariate (*CVD*) in general and divided by center

### 3.2.2. Nutrients' characteristics

The purpose of this section is to quantify the correlation among the 43 nutrients, which serve as the exposure variables. The importance of this step will become evident in later stages of the analysis when we construct the necessary models.

Initially, we visually inspect the covariates and construct a Spearman correlation matrix. This matrix provides a representation of the Spearman pair correlation between all the nutrients. The Spearman correlation coefficient, ranging from -1 to 1, measures the strength and direction of association between two ranked variables. In contrast to the Pearson correlation coefficient, which evaluates linear relationships, the Spearman correlation focuses on monotonic relationships, whether linear or not. As a result, it places greater emphasis on the presence or absence of a correlation rather than its specific form. When dealing with nutrients intake, it is crucial to take into account the total energy intake (kcal/day) of each participant in order to ensure accurate directions of association. This is particularly important as the total energy intake can vary significantly across different categories of individuals and our primary interest lies not in the absolute value of nutrient intake, but rather in its proportion relative to the daily intake. Hence, in this phase of the analysis, it appears appropriate to divide all the nutrients of each individual by the corresponding personal daily total intake. This process yields to results that can be compared with each other in order to obtain the relative correlations.

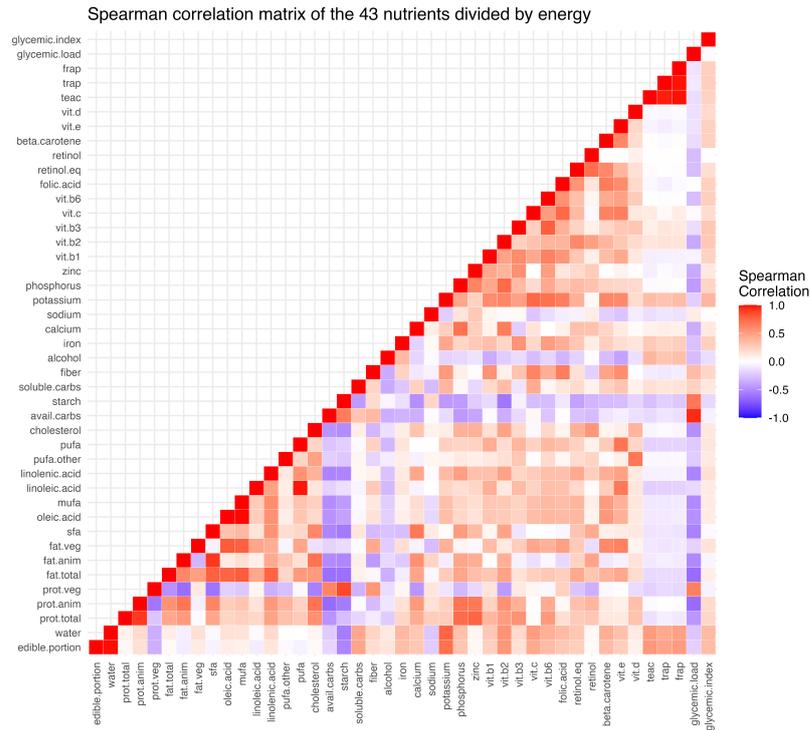


Figure 3.4: Spearman correlation matrix of the 43 nutrients divided by the personal daily energy intake

As shown in the heatmap illustrating the Spearman correlation matrix in Figure 3.4, we observe a mixture of positive and negative associations. Specifically, there are only 9 pairs exhibiting correlations surpassing 90%, and no pairs demonstrate a negative correlation exceeding 90% in absolute value. Within these 9 pairs, we encounter associations that are already established, such as the relationship between glycemic index and available carbohydrates which are linearly dependent, as the glycemic index is derived from the available carbohydrates by division by 100.

### 3.2.3. DNA methylation levels' characteristics

In this section we briefly want to show that also CpG sites can exhibit strong correlations among themselves. These correlations arise because DNA methylation patterns are not entirely random, and nearby CpG sites can be influenced by similar biological factors. The extent and nature of these correlations can vary depending on various factors. For example CpG sites that are physically close to each other on the DNA strand, belong to the same genomic region or cell type may be more likely to have correlated methylation levels. In Figure 3.5 we show the Spearman correlation matrix of 10 CpG sites taken randomly from the 399,957 CpG sites available. As we can see there are some sites which

show very high positive correlation while some others do not, as we would expect from a random choice in a dataset with high within correlations.

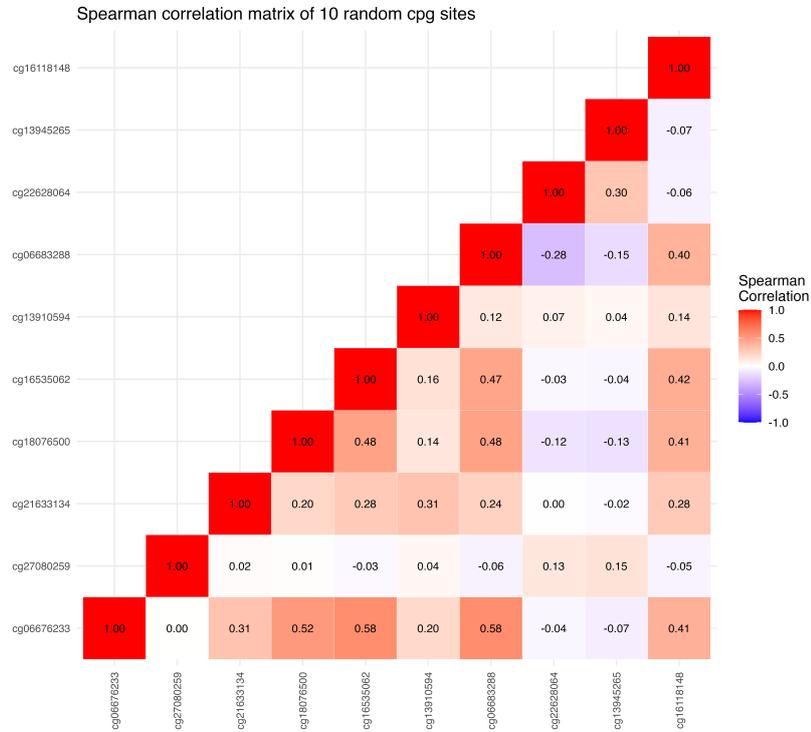


Figure 3.5: Spearman correlation matrix of 10 CpG sites taken randomly among the 399,957 CpG sites composing the DNA methylation dataset

# 4 | Meet-in-the-Middle and mediation analysis

Up to this point, we have established the theoretical foundation of our work and conducted a descriptive and exploratory analysis of the dataset. As previously mentioned, at the basis of our analysis there is the hypothesis that there exists a mediator that connects the exposure and the outcome, sitting on the causal pathway between the exposure and the outcome. In observational studies, identifying a plausible ideally pre-specified mediator can strengthen the causal inference of the findings [46]. In this chapter we are going to explain our first and more accurate methods to assess whether DNA methylation's mediation is a significant intermediary and to assess its extent. The stability selection method to be compared to will be presented in Section 5.

In this Section, we introduce two distinct applications of the Meet-in-the-Middle approach to relate a set of dietary exposures, the nutrients intakes, to the CardioVascular Disease (CVD). The first application, detailed in Sections 4.1 and 4.2, utilizes the methylome layer to identify potential new exposures to be tested for their association with CVD. The second application, discussed in Sections 4.3 and 4.4, helps to select potential mediators for exposures associated with CVD. It's important to highlight that the objectives of these two MITM applications differ. While both employ the methylome layer as a mediator, the ultimate selection criteria vary: the former method aims to identify a specific set of nutrients establishing a causal link with the outcome, whereas the latter focuses on selecting which CpG sites serve as actual mediators and assessing their effect.

## 4.1. Methods for the first MITM approach

In this section, we will present the methodologies and findings of the first Meet-in-the-Middle approach, following the “oriented Meet-in the-Middle” design from Cadiou et al., 2020. Our final objective is to identify potential causal exposures associated with the development of CVD, with DNA methylation acting as the mediating factor in this process.

### 4.1.1. Overall strategy

Here, we provide an overview of the general workflow of our first implementation of the MITM approach (“oriented MITM”), which comprises three sequential steps:

(a) **Dimension reduction based on a prior knowledge:**

We employ existing literature to reduce the dimension of the intermediate methylome layer and thus obtain a restricted methylome. Further details can be found in Section 4.1.2

(b) **Relation between the whole exposome and the restricted methylome:**

We explore the relationship between each nutrient and the restricted set of CpG sites obtained in the previous step through the implementation of a linear model. The specifics and assumptions of this model are detailed in Section 4.1.3.

(c) **Relation between the reduced exposome and the CVD:**

In this stage, we fit a generalized linear model to investigate the association between the Reduced Exposome (i.e., the significant nutrients associated with the restricted set of CpG sites from step (b) ) and CVD. The objective of this step is to identify the final hits (the nutrients) whose causal link with the health outcome (CVD) has been confirmed and strengthened by our implemented approach. Further elaboration on this model can be found in Section 4.1.4.

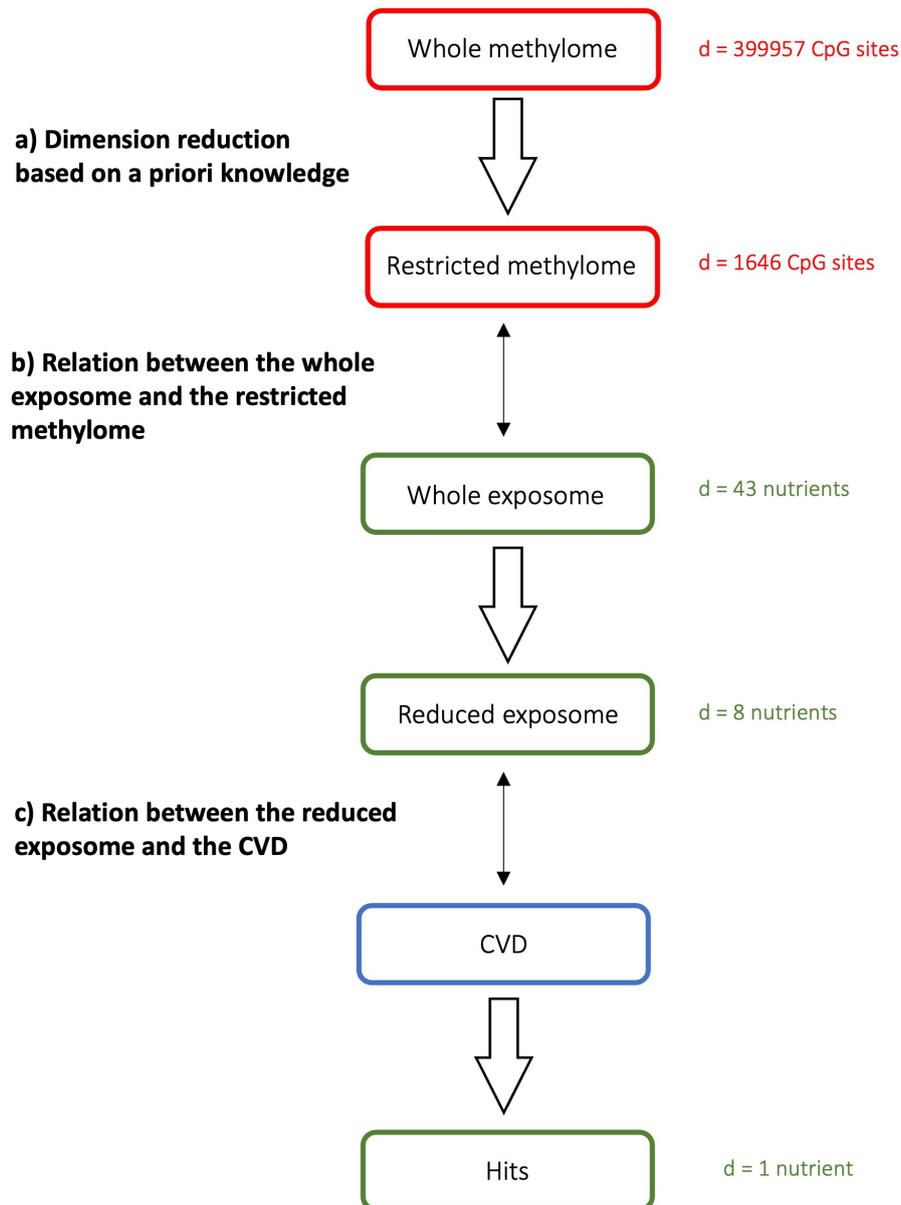


Figure 4.1: Schematic pipeline of the first application of the MITM method

#### 4.1.2. Step a): a priori selection of CVD-relevant CpG sites

As mentioned in Sub-section 1.2.2, studies involving high-dimensional frameworks can lead to a higher rate of false positive findings. Using the information coming from the introduction of an intermediate layer, such as the methylome layer in this case, provides a promising approach to overcome this issue, at least to some extent. However, when the biological intermediate layer has high dimensions, it could be useful to reduce its size before proceeding with the analysis, to prevent introducing an additional source of error. Our hypothesis is that a smaller intermediate layer corresponds to higher specificity (and

possibly lower sensitivity) compared to an agnostic approach. Specificity and sensitivity in binary classification are supposed to have an inverse relationships. Indeed if we increase the specificity (i.e. the number of true negatives) we are making it more conservative in declaring a positive, so in general it tends to classify fewer cases as positive, which reduces the chance of capturing all true positives (i.e. lower the sensitivity). Relying on existing literature and reviews is a good strategy for this reduction task, but the effectiveness of the reduction will highly depend on the quality of the available evidence. Therefore, it is crucial to select sources that align most closely with our objectives. We opted to restrict the initial set of 399,957 CpG sites by considering information from two distinct sources.

- The first source is the EWAS catalogue [34]. It was developed and is maintained by the Integrative Epidemiology Unit (IEU) at the University of Bristol. The Catalog was founded by the IEU in October 2017 by James Staley and was developed to enable the scientific community to search results from epigenome-wide association studies of DNA methylation. The Catalog currently contains results at a significance level of  $p < 10^{-4}$  from three sources. Firstly, results were extracted from the literature. Secondly, EWAS were performed in the Accessible Resource for Epigenomics Studies (ARIES) sub-section of the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort. Thirdly, publicly available data from the Gene Expression Omnibus (GEO) database were extracted using the *geograbi* R package [35] and EWAS were performed on these data after quality control. To ensure we have the latest EWAS are considered, the data extraction team mine the literature each month for new published results.
- The second source is a systematic review of articles examining measurements of CpG methylation levels in CVD [36], conducted in accordance with PRISMA (preferred reporting items for systematic reviews and meta-analyses) guidelines. It has the title "DNA methylation and cardiovascular disease in humans: a systematic review and database of known CpG methylation sites" and is from March 30<sup>th</sup> 2023. The search yielded 5,563 articles from PubMed and CENTRAL databases. From 99 studies with a total of 87,827 individuals eligible for analysis, a database was created combining all CpG-, gene- and study-related information.

Within these sources, we apply our own selection criteria to further tailor the process to our specific problem. Here's a synthesis of our selection process:

- From the EWAS catalogue, we identify CpG sites linked to various heart-related categories. In particular we use the following keywords for the selection: *cardiovascular disease*, *cardiovascular disease risk*, *incident myocardial infarction occurrence*, *my-*

*ocardial infarction, ischemic stroke, risk of future coronary heart disease, non-stroke cardiovascular disease.* This process yield a selection of 448 CpG sites.

- The systematic review, which synthesized findings from studies exploring the relationship between DNA methylation and CVD, provides a comprehensive and easily searchable database. We extract from it the CpG sites that are deemed significant in multiple studies. Upon intersecting these with the CpG sites in our dataset, we discover that 1,227, 424, and 91 are uniquely associated with two or more, three or more, and six or more studies, respectively. We decid to select the 1,227 CpG sites mentioned in two or more studies.

By combining these two sets, which have 29 elements in common, we arrive at a total of 1,646 CpG sites, which represent our Restricted Methylome.

<b>Set 1:</b> number of significant CpG sites extracted from the EWAS catalogue [34]	<b>Set 2:</b> Number of significant CpG sites extracted from a systematic review [36]	<b>Restricted Methylome:</b> number of selected significant CpG sites (union of Set 1 and Set 2)
448	1,227	1,646

Table 4.1: Summary of the process we use to create the Restricted Methylome as the union of the significant CpG sites according to the EWAS catalogue and a systematic review of existing studies.

### 4.1.3. Step b): Model and assumptions

After the initial dimension reduction of the intermediate layer in step a), we continue the development of our MITM approach with the implementation of step b), whose aim is to select those nutrients which are strongly associated with the Restricted Methylome, thus obtaining a reduced set of nutrients (Reduce Exposome).

#### Model:

Here we present the mathematical expression of the 1643\*43 models we want to fit at this stage. The details and underlying motivations are elaborated further in the "Model Fit" section below.

$$Y_i = \beta_0 + \beta_1 X_j + \beta_2 Z + \vec{\gamma} C + \epsilon_{i,j}$$

where:

- $Y_i$  is a vector containing methylation levels at the  $i^{th}$  CpG site belonging to the Restricted Methylome, for each individual,  $i = 1, \dots, 1646$ .
- $X_j$  is a vector representing the daily intakes of the  $j^{th}$  nutrient belonging to the Whole Exposome, for each individual,  $j = 1, \dots, 43$ .
- $Z$  is a vector containing the values for the variable CVD.
- $C$  is a matrix containing all the confounding factors, for each individual.
- $\beta_0$  is the intercept .
- $\beta_1, \beta_2$  and  $\vec{\gamma}$  are the coefficient of the independent variables  $X_j, Z$  and  $C$ .
- $\epsilon_{i,j} \stackrel{iid}{\sim} N(0, 1)$  is the error term for the model with the  $i^{th}$  CpG site and the  $j^{th}$  nutrient.

#### Step b)

**Linear regression model to study the relation between the Whole Exposome and the Restricted Methylome**

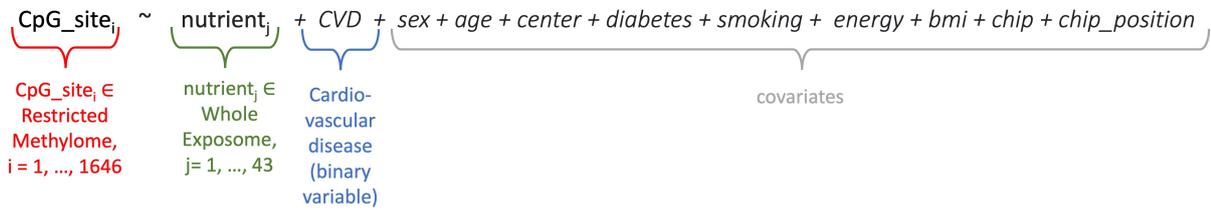


Figure 4.2: Schematic representation of the multiple linear models fitted in step b)

#### Assumptions:

Before fitting the suitable linear models, it is essential to assess whether the assumptions of linear regression hold or not, and how to proceed in the latter case. The model we intend to employ (depicted schematically in Figure 4.2) has the  $i^{th}$  CpG site as the response variable, the  $j^{th}$  nutrient as the independent variable, and all confounding factors incorporated as linear variables. To assess the assumptions of linear regression, we conduct four statistical tests:

1. Shapiro test to assess the normality of the residuals
2. Breusch Pagan Test [37] to check for homoscedasticity of the residuals
3. Agostino test [38] to examine the skewness (asymmetry in the distribution) of the residuals

Regarding the linear regression model implemented in step b), we discover that, at a significance level of  $\alpha = 1\%$ , nearly 96% of the sites rejects the null hypothesis for at least one assumption. Specifically, 93%, 71%, 79%, and 87% of the 1,646 fitted models do not meet the assumptions of normality, homoscedasticity, skewness, and kurtosis of the residuals, respectively. The first concern in using a model who violates the hypothesis is the risk of reporting either false positive or false negative findings in the tests of association. The paper "Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array" [41] addresses three critical analytical challenges that may arise in studies involving DNA methylation associations with complex phenotypes. Among these challenges is the question of whether linear regression, which is the chosen statistical tool for most studies, is appropriate and whether it is biased by the underlying distribution of DNA methylation data. The authors explore if sites that violate assumptions are more likely to be significant in a DNA methylation analysis using a linear regression model, and they determine that this is not the case. They show that the use of linear regression with beta values (as in our case) in DNA methylation studies, even if the data do not satisfy the standard assumptions of the test, does not seem to lead to biased results. Therefore, we proceed with fitting the required models.

#### **Model fit:**

In this step, we conduct an analysis to assess the association of the entire set of nutrients (referred to as the Whole Exposome) with the methylome levels of the preselected CpG sites (known as the Restricted Methylome). This process leads to the identification of a Reduced Exposome. We employ univariate linear regression models and applied p-value correction for multiple testing using a False Discovery Rate (FDR) procedure, specifically the Benjamin-Hochberg method. In all regression models, we include adjustment factors as linear variables. These factors included *sex*, *age*, *recruitment center*, *diabetes status*, *smoking habits*, *energy intake*, *BMI*, *chip* and *chip position*. As previously mentioned in the context of causal inference, identifying and correcting for as many confounding variables as possible is crucial to eliminate bias and establish accurate causal relationships. Among the adjustment factors we choose to integrate the daily *energy* absorption, which is essential when examining nutrient intake to ensure accurate directions of associations, given the potential significant variation in total energy intake across different categories of individuals, which is linked to the nutrients intakes and can have an effect on CVD. Furthermore, we significantly reduce the batch-effect by correcting for *chip* and *chip position*. Batch-effects refer to systematic differences between measurements of different batches of experiments, which can artificially inflate within-group variances, potentially leading to reduced experimental power and creating false positive results [40]. These effects can

arise from various factors, including differences in laboratory environments and operating procedures. In the first simulations, when not correcting for *chip* and *chip position* all the results are severely inflated, due to the presence of a serious batch-effect. When speaking about inflation we refer to the deviation of the distribution of the observed test statistic compared to the distribution of the expected test statistic [42]. In our case, without adjustment, the inflation become evident as it shows an unexpectedly high and unrealistic number of associations.

#### 4.1.4. Step c): Model and assumptions

In this step, our aim is to investigate the relationship of the Reduced Exposome (obtained at the previous step) with the outcome (CVD) to ultimately identify the final hits of the process.

##### Model:

In this section, we provide the mathematical expressions for the 8 models that we aim to apply in this phase. Further insights into the specifics and rationale can be found in the "Model Fit" section below.

$$\text{logit}(p_j) = \beta_0 + \beta_1 X_j + \vec{\gamma} C$$

where:

- $Y_j \sim Be(p_j)$  is the CVD variable.
- $\text{logit}(p_j)$  represents the log-odds of the probability that the dependent variable  $Y_j$  takes on the value 1.
- $X_j$  is a vector representing the daily intakes of the  $j^{\text{th}}$  nutrient belonging to the Reduced Exposome, for each individual,  $j = 1, \dots, 8$ .
- $C$  is a matrix containing all the confounding factors, for each individual.
- $\beta_0$  is the intercept .
- $\beta_1$  and  $\vec{\gamma}$  are the coefficient of the independent variables  $X_j$  and  $C$ .

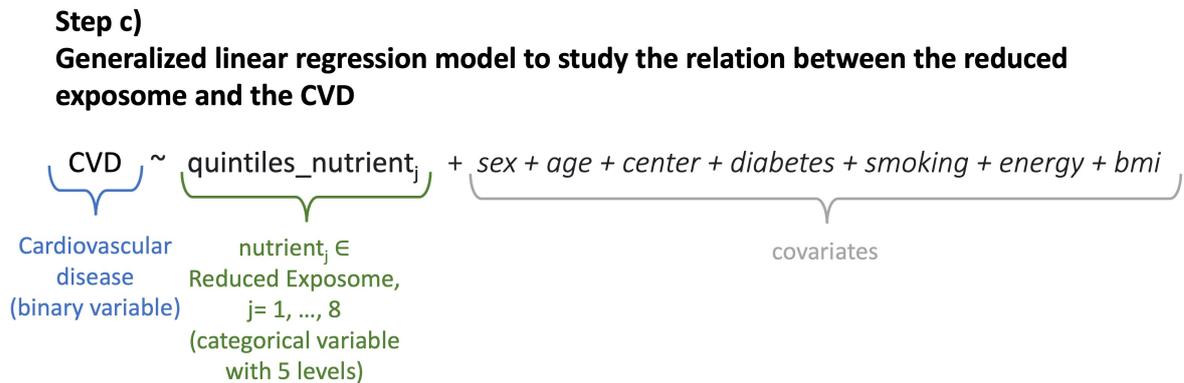


Figure 4.3: Schematic representation of the multiple generalized linear models fitted in step c)

### Division of the distribution into quantiles:

Given that we are constructing a logistic regression model, it's not necessary to validate the assumptions as in previous cases with the linear model.

However, before proceeding, we make some adjustments to the model. Dividing a distribution into quantiles means partitioning the data into equal-sized groups based on their rank or position within the dataset. In dietary epidemiology, tertiles, quartiles, and quintiles are commonly employed to categorize individuals according to their different levels of dietary intake. This practice is particularly common in the EPIC cohort, as it helps address non-linearity and provides a deeper consideration of the significant influence of extremes in the distribution on Cardiovascular Disease. In fact, the association between nutrients and the outcome frequently exhibits non-linear patterns, such as U-shaped or inverted U-shaped relationships. It is common for both low and high intakes of a specific nutrient to carry risks (or benefits), while moderate values may have no discernible impact. Given that extreme values tend to exert the most influence on the outcome, segmenting the nutrient distribution into quantiles and evaluating the significance level of the highest quantile relative to the baseline (first quantile) proves to be an effective strategy for addressing non-linearity.

In this context, we refer to three papers ([43], [44], [45]) in which tertiles, quartiles, and quintiles were respectively used in the EPIC Italy framework when examining the relationship between nutrients and health outcomes.

For our study, we opt to utilize quintiles. We transform the continuous variables representing the nutrients' intake in discrete ones, using the 20<sup>th</sup>, 40<sup>th</sup>, 60<sup>th</sup>, and 80<sup>th</sup> percentiles as dividers. Each value is then assigned to one of these categories. We report in Figure 4.4 the example for one nutrient, iron, of how the division into quintiles would transform

the variable. Furthermore, it is essential to note that when assessing the significance of the association between a nutrient (categorized into quintiles) and the outcome, we specifically extract the p-value and beta coefficients of the 5<sup>th</sup> quintile in comparison to the baseline (1<sup>st</sup> quintile). This selection is made based on the consideration that more extreme values often exert a greater influence on the outcome.

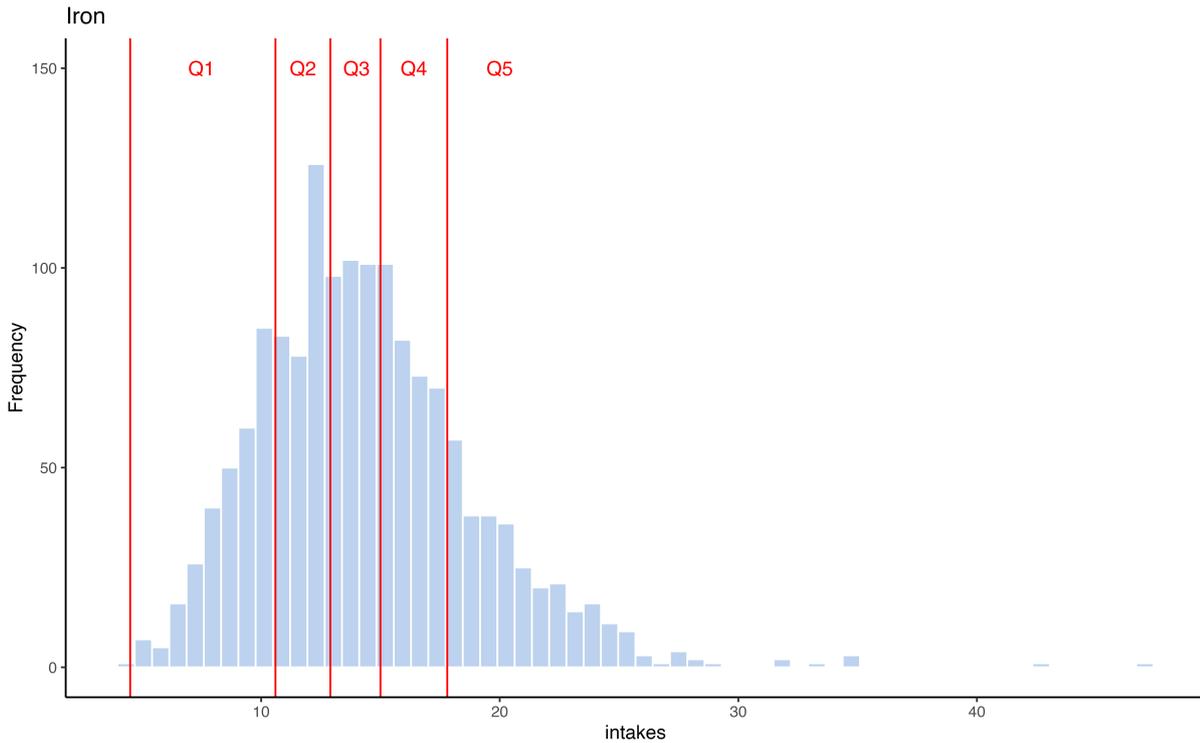


Figure 4.4: Iron intake histogram with quintile divisions highlighted in red. Each value falls within one of the specified ranges (Q1, Q2, Q3, Q4, Q5), determining its categorization into one of the five levels, forming in this way a categorical variable.

### Model fit:

In this step we examine the associations between the Reduced Exposome and the occurrence of CVD as the outcome. Here we employ univariate generalized linear regression models, being the outcome (development or non-development of the cardiovascular disease) a binary variable. Once again all p-values are adjusted using Benjamin-Hockberg procedure. Differently from before, here we reduce the number of adjustment factors. Indeed, since now we are not dealing with methylation levels, it doesn't make sense to correct for differences in DNA methylation sample processing, as indicated by the variables *chip* and *chip position*. We therefore used as covariates: *sex*, *age*, *recruitment center*, *diabetes status*, *smoking habits*, *energy intake* and *BMI*.

### 4.1.5. Visualization tools: Volcano plot, inflation plot, dose-response relationships plot

Here we introduce three essential plots that will enhance the visualization of our results in the next section:

#### Volcano Plot:

The volcano Plot [47] is a type of scatter-plot that is used to quickly identify direction and magnitude of changes. In particular it shows statistical significance ( $-\log_{10}(p - value)$  on the y-axis) versus magnitude of change ( $\log_2$  of the fold change ( $\log_2 FC$ ) on the x-axis). Higher points on the y-axis indicate stronger evidence of association. The fold change is a measure describing how much a quantity changes between an original and a subsequent measurement [48]. The specifics vary from case to case, but in a linear regression framework like ours, it represents the ratio of the value of the outcome (Y) after one unit increase of the considered covariate (X) to the value before the increase. The mathematical expression is the following:

$$\log_2 FC = \text{sign}(\beta_X) * |\log_2(|\frac{\beta_X + \beta_0}{\beta_0}|)|$$

where  $\beta_X$  denotes the regression coefficient of the covariate X and  $\beta_0$  represents the intercept value obtained from the regression output.

In our analysis, the covariate X will consistently refer to the  $j^{th}$  nutrient, while the outcome might vary between the  $i^{th}$  CpG site and the CVD.

Regarding the plot, it is important to highlight that:

- The sign of the  $\log_2 FC$  indicates the direction of change induced by the covariate X on the outcome. A positive  $\log_2 FC$  implies that covariate X contributes to an increase in the outcome, while a negative value suggests a decrease.
- The absolute value of the  $\log_2 FC$  reflects the magnitude of this contribution.
- The horizontal line represents significant p-value of 0.05, while the vertical line in 0 separates positive and negative  $\beta$ -values
- Features with a p-value below 0.05, are positioned in the upper-right and upper-left parts of the plot. They are color-coded for emphasis: positively associated features appear in red, negatively associated features in blue.

#### Inflation plot:

The inflation factor ( $\lambda$ ) is a statistical measure used to assess the extent of inflation in

test statistics. It helps identify if there is an excess of significant associations compared to what would be expected by chance. The inflation factor is calculated by dividing the observed median of the distribution of test statistics by the expected median under the null hypothesis of no true associations:

$$\lambda = \frac{\text{median}(\text{observed test statistics})}{\text{median}(\text{expected test statistics})}$$

This detachment can occur due to various reasons, such as unaccounted sources of variation or biases in the data. For this reason at each step of our analysis we will control the inflation related to the multiple tests conducted, to ensure reliability and validity of the statistical analysis.

To control the inflation we use the R package "bacon" [49]. It executes the Gibbs Sampler algorithm and, among other things, can generate a QQ-plot representing the observed vs expected  $-\log_{10}(p - \text{value})$  values for each test, starting from the z-scores ( $\frac{X-\mu}{\sigma}$ ).

We can also compute an inflation factor, which estimate the total amount of inflation:

- If the inflation factor ( $\lambda$ ) is equal to 1, it indicates that there is no inflation, and the test statistics follow the expected null distribution. This suggests that the assumptions of the regression model are met, and there is no evidence of inflation or deflation of test statistics.
- If the inflation factor ( $\lambda$ ) is greater than 1, it suggests that there is inflation of test statistics. This means that the observed p-values are smaller than expected, indicating a higher rate of false positives. In this case, it may indicate issues like unaccounted-for confounding or other sources of bias.
- If the inflation factor ( $\lambda$ ) is less than 1, it suggests that there is deflation of test statistics. This means that the observed p-values are larger than expected, indicating a lower rate of false positives. While rare, this can sometimes occur due to factors like over-correction for confounding.

#### **Dose-response relationships plot:**

As previously mentioned, in step c), we will examine the relationship between nutrients and CVD, with the former's distributions divided into quintiles. This entails that the output of the linear model will include estimates ( $\beta$  coefficients) for the 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup>, and 5<sup>th</sup> quintiles in reference to the baseline (1<sup>st</sup>). In a linear model, each  $\beta$  coefficient represents the estimated change in the dependent variable (CVD) for a one-unit change in a predictor variable ( $j^{\text{th}}$  nutrient). Subsequently, for each nutrient, we can ascertain whether it acts as a risk or protective factor ( $\beta > 0$  or  $\beta < 0$ ) and within which ranges of

each nutrient's intake the odds of having CVD increase or decrease, and to what extent. The most effective way to assess these changes is by plotting the values and analyzing the shape of the resulting curve. We observe a so-called dose-response relationship when increasing levels of exposure are associated with either an increasing or decreasing risk of the outcome, signifying that the curve exhibits a monotonic increase or decrease. Conversely, if the curve exhibits a U-shape, it indicates that both lower and higher intakes of that nutrient contribute to an increased probability of developing CVD. Lastly, if the curve displays an inverted U-shape, it implies that the central ranges pose the greatest risk.

Furthermore, examining the absolute value of the  $\beta$  coefficient for the 5<sup>th</sup> quintile (in relation to the 1<sup>st</sup>) allows us to determine if it exhibits more pronounced behavior compared to the other  $\beta$  coefficients. This provides confirmation as to whether it contributes more significantly to the increase or decrease in the odds of developing CVD, as expected.

## 4.2. Results for the first MITM approach

In this section we present, step by step, the results obtained by the first application of the MITM approach.

### 4.2.1. Results step b)

The aim of this step is to select those nutrients which are strongly associated with the Restricted Methylome, thus obtaining a reduced set of nutrients (Reduce Exposome).

From the test of association of the 1,646 CpG sites (composing the Restricted Methylome) with each of the 43 nutrients (composing the Whole Exposome) we identify 8 nutrients significantly associated with at least one CpG site (in order of significance): *alcohol*, *available carbohydrates*, *iron*, *glycemic load*, *TRAP* ("*total radical trapping antioxidant potential*"), *Vitamin D*, *potassium* and *FRAP* ("*ferric reducing antioxidant power*").

An association occurs when the p-value related to the  $j^{\text{th}}$  nutrient (in the output of the regression model:  $CpG\_site_i \sim nutrient_j + sex + age + center + diabetes + smoking + energy + bmi + chip + chip\_position$ ), after the Benjamin-Hockberg correction for multiple testing, is lower than a threshold  $\alpha = 0.05$ . These 8 nutrients significantly associated with the intermediate methylome layer compose the Reduced Methylome, which will be the starting point of the following step.

Moreover in total 5 different CpG sites are associated with at least one nutrient.

We show the 10 significant associations through a Volcano Plot in order to better visualize the different level of significance of the nutrients as well as the different directions of

association. As illustrated in Figure 4.5:

- Alcohol emerges as the most remarkably significant nutrient, showing a corrected p-value of approximately  $10^{-21}$
- Among these associations, iron stands out as the only statistically significant negatively associated nutrient, indicating its role in reducing the methylation level of the corresponding CpG site. Conversely, vitamin D exhibits a positive association, leading to an increase in the methylation levels of the respective CpG site. It's worth noting that, at this stage, these observations lack of interpretation since the outcome of each model is a different CpG site, which makes a general comparison difficult.

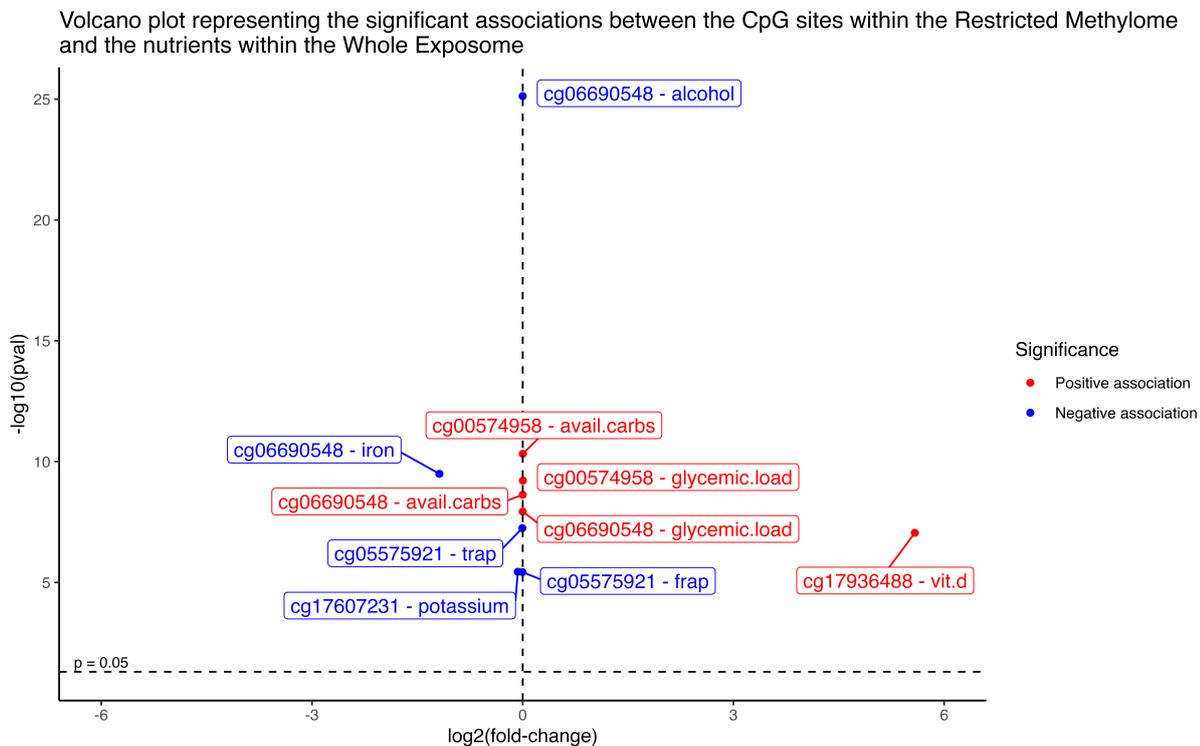


Figure 4.5: Volcano Plot representing the significant associations between the CpG sites within the Restricted Methylome and the nutrients within the Whole Exposome.

We also check if our results are reliable or show too much inflation. We both plot the QQ-plot of observed vs expected  $-\log_{10}(p - \text{value})$  for each model (Figure 4.6) and compute the general inflation factor, which results to be 1.03. Analysing both we can affirm that the inflation is minimal and does not affect the results.

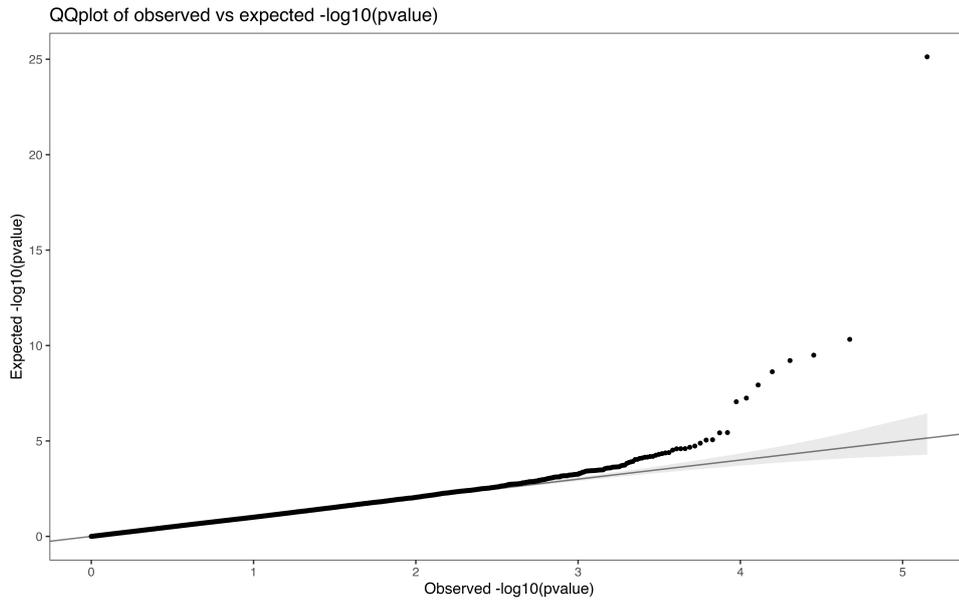


Figure 4.6: QQ-plot showing the inflation of the p-value of the tests conducted at step b), i.e. the deviation of the distribution of the observed tests statistics to the distribution of the expected tests statistics (bisector of the first-fourth quadrant). The inflation factor is  $\lambda = 1.03$

#### 4.2.2. Results step c)

In this last step we study the association between the Reduced Exposome, composed by the 8 nutrients found at the previous step, with the CVD as outcome.

We identify that *Iron* can be considered as directly associated with CVD, after the BH correction for multiple testing, with a p-value of 5.2%, which is slightly above the threshold 5%.

Again we visualize the results using a Volcano plot, which we report below in Figure 4.7. As regression coefficients for each nutrient we use the  $\beta$  associated with the 5<sup>th</sup> quintile with refer to the baseline (1<sup>st</sup> quintile)

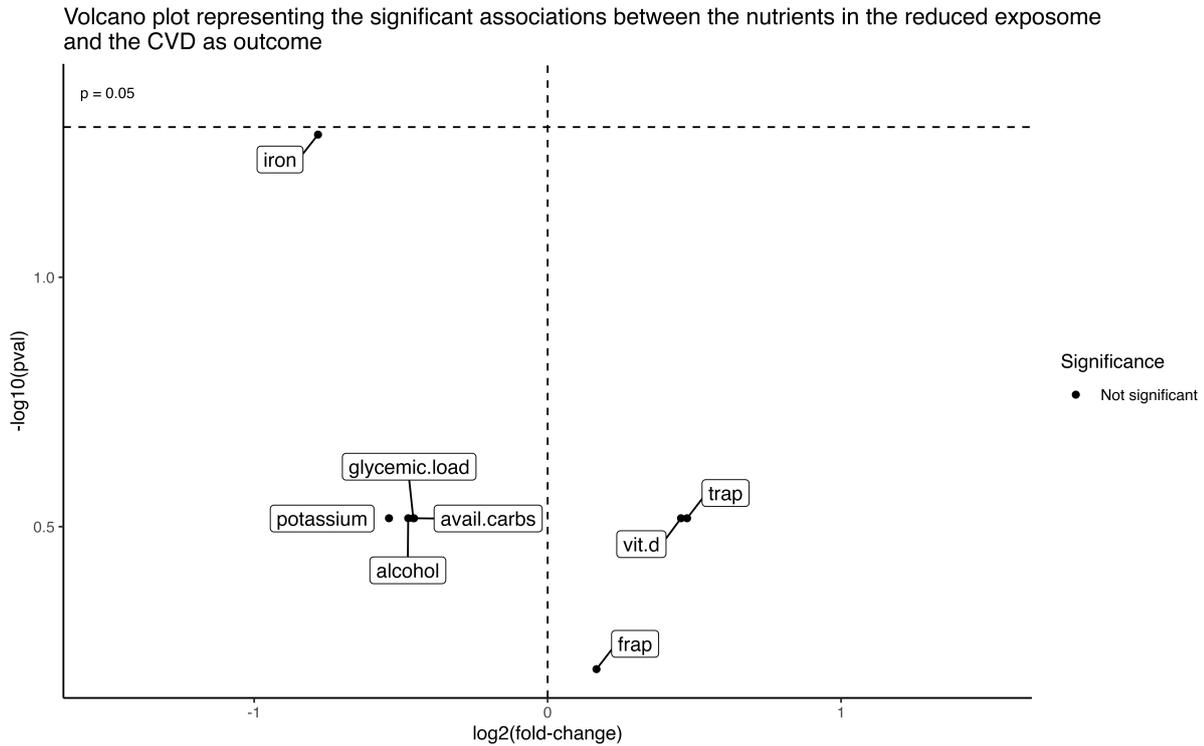


Figure 4.7: Volcano Plot representing all the associations between the 8 nutrients within the Reduced Exposome and the CVD as outcome.

While Iron is the only nutrient showing almost significant results, we apply all the following considerations to all the 8 nutrients. It's worth noting that these observations are only indicative, and the lack of significance does not provide certainties regarding their reliability. Indeed, as we will discover, the less significant a nutrient is, the more unreliable the findings regarding it tend to be. We proceed in this way:

- Firstly, by examining the sign of the  $\text{Log}_2(FC)$  (which corresponds to the sign of the beta associated with the 5<sup>th</sup> quintile with refer to the 1<sup>st</sup> for each nutrient), we can infer whether a nutrient might be considered a protective factor (negative sign) or a risk factor (positive sign) based on our analysis. Subsequently, we validate this hypothesis by verifying the consistency of the signs of the 4<sup>th</sup>, 3<sup>rd</sup>, and 2<sup>nd</sup> quintiles (relative to the baseline) through the dose-response plot. Using the relative regression coefficients ( $\beta$ ) we can quantify the reduction/increase in the risk of developing a CVD.
- Secondly, we explore the dose-response relationship to understand the relation between each nutrient and the outcome. About this second analysis, it is worth noting that, in general, lower estimates ( $\beta$ ) tend to have a greater impact on decreasing the

odds (in the case of protective factors) or increasing them less (in the case of risk factors) for CVD. Therefore, lower estimates are typically associated with a more favorable outcome for CVD. It's worth noting that for a protective factor, assuming a specific quantity of a nutrient associated to the lowest beta is beneficial for preventing CVD. On the other hand, for a risk factor, assuming a certain quantity of a nutrient related to the lowest beta may be better than another quantity, but it is still less favorable than not assuming it at all. We plot (in Figure 4.8) the exponential of each  $\beta$  coefficient (y-axis) associated with each quintile from the regression model.

- Finally, in order to check the obtained results, we compare the obtained classification in risk-protective factor for CVD with the notions known in literature.

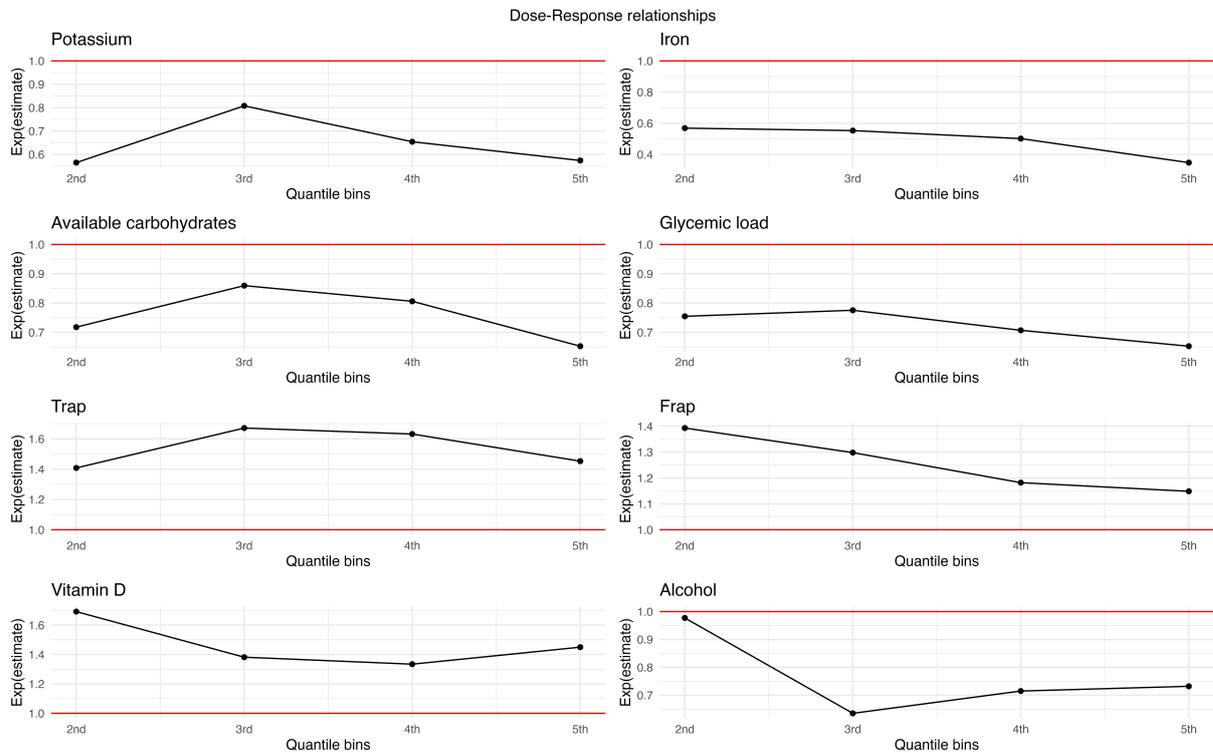


Figure 4.8: Dose-response relationships for each of the 8 nutrients within the Reduced Exposome, representing the (exponent of the) estimates for the 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup> and 5<sup>th</sup> quintiles with refer to the baseline (1<sup>st</sup> quintile), for each nutrient. In red we highlight the threshold  $\beta = 0$ , which allow to distinguish between protective factors (below the line) and risk factor (above the line).

We report here the results for iron and in Appendix A the considerations for the other 7 non significant nutrients.

**Iron:**

According to our findings, iron emerges as a protective factor, as evidenced by the negative regression coefficient of  $\beta = -1.05$ . This indicates that individual in the 5<sup>th</sup> quintile for iron consumption have a decreased risk in the odds of the event (development of CVD) compared with the reference group (individuals in the 1<sup>st</sup> quintile for iron consumption). This means that people whose iron intakes belong to the 5<sup>th</sup> quintile have 65% ( $1 - e^{-1.05}$ ) less probabilities of developing CVD than people assuming iron values within the 1<sup>st</sup> quintile.

By looking at the shape of the dose-response curve for iron, we observe that it exhibits a monotonic decreasing curve, meaning that the effect of reducing the CVD increases by increasing the consumption range (i.e. the quintile considered), and we have the best outcome at the 5<sup>th</sup> quintile, which corresponds to a range of (17.8, 47.2) mg/day.

Before presenting the findings based on the literature, it is crucial to underscore that our conclusion, suggesting that higher levels of iron are associated with a reduced likelihood of CVD, is based on a study analysing assumed iron quantities within the range of 4.5 to 47.2 mg/day. We do not investigate the effects of much higher iron levels originating from alternate sources, such as genetic disorder, blood transfusions or drug-induced toxicities. Therefore, from our analysis, we cannot draw any conclusions on the influence of such higher iron doses on CVD. By reviewing the literature [51][52] we found out that:

- Iron deficiency is the most prevalent malnutrition-related condition and affects up to 75% of patients with heart failure.
- Conversely, both primary and secondary forms of iron overload can cause heart disease via oxidative damage. Iron over-load occurs when the body stores excess iron. The heart is particularly susceptible to damage induced by accumulation of iron. Primary iron overload is caused by genetic disorders whereas secondary iron overload develops as a result of repeated blood transfusions, drug-induced toxicity or excess consumption of iron.

These assessments confirm our results because, as highlighted above, in our analysis we just focus on normal doses coming from daily intakes, not on external extraordinary causes leading to much higher levels.

We summarize all the findings for each nutrient in Table 4.2 and the comparison with literature notions in 4.3

Nutrients	Risk or protective factors	shape of the curve	Quantile related to the best quantities to assume	Ranges of best quantities to assume ( <i>gr/day</i> )
Iron	Protective	decreasing	5 <sup>th</sup>	[0.0178, 0.0472]
Potassium	Protective	Inverted U-shape	2 <sup>nd</sup> or 5 <sup>th</sup>	[0.00258, 0.00305], [0.00405, 0.0095]
Available Carbohydrates	Protective	Inverted U-shape	2 <sup>nd</sup> or 5 <sup>th</sup>	[184, 232], [336, 780]
Glycemic load	Protective	decreasing	4 <sup>th</sup>	[149, 180]
TRAP	Risk	Inverted U-shape	2 <sup>nd</sup> or 5 <sup>th</sup>	[6.48, 8.67], [13.2, 30.8]
FRAP	Risk	decreasing	4 <sup>th</sup>	[21, 25.7]
Vitamin D	Risk	U-shape	3 <sup>rd</sup> or 4 <sup>th</sup>	[1.13, 1.53], [1.53, 2.13]
Alcohol	Protective	U-shape	3 <sup>rd</sup>	[3.49, 11.8]

Table 4.2: Summary of the obtained classification as protective/risk factors and some features of the dose-response relations.

Nutrients	Associated p-value of significance (after BH correction)	Classification in Risk/Protective factor according to our result	Classification in Risk/Protective factor according to the literature	References to the source
Iron	0.052	Protective	Protective	[51], [52]
Potassium	0.304	Protective	Protective	[50]
Available carbohydrates	0.304	Protective	Protective	[53]
Glycemic load	0.304	Protective	Protective/Risk	[54]
TRAP	0.304	Risk	Protective	[55]
FRAP	0.610	Risk	Protective	[55]
Vitamin D	0.304	Risk	Risk	[56]
Alcohol	0.304	Protective	Risk	[57]

Table 4.3: Summary of the comparison between our findings and the ones coming from existing studies. Green color indicated that our results align with the literature, red means that they are in strong contrast, orange means that the comparison is uncertain.

### 4.3. Methods for the second MITM approach

In this section we present the methods and results of the second application of the Meet-in-the-Middle approach. Here, our aim is to identify with a MITM approach those CpG sites that could potentially mediate the effect of exposures on CardioVascular disease and then quantify the extent of this mediation.

In contrast to the first application, we relax the hypothesis and do not correct for multiple testing. Indeed, given that the objective here is to test and quantify the mediation of all selected CpG sites, it is deemed more acceptable to be less stringent in the process. This distinction arises from the differing objectives between the two applications of the MITM. In the first case (4.1), our focus was on identifying specific nutrients likely to be causal of CVD, while in this application our primary goal is to confirm the impact of all selected CpG sites as potential mediators. While adhering strictly to hypotheses is generally preferred, we find it acceptable to relax them in this framework.

Given that many of the underlying reasoning, assumptions, and pre-selections remain consistent with those explained in the first implementation of the MITM approach (4.1), we will refrain from reiterating them and instead refer back to the specific explanation.

#### 4.3.1. Overall strategy

We begin by outlining the general pipeline for the second implementation of the MITM approach, followed by a detailed explanation of each step in the subsequent sections. This implementation comprises five sequential steps:

(a) **Relation between the Whole Exposome and the CVD:**

We assess the significance of each of the 43 nutrients that constitute the Whole Exposome with the health outcome, the CVD. We obtain a restricted set referred to as Reduced Exposome.

(b) **Dimension reduction based on a prior knowledge:**

Once again, we leverage biological information from genetic databases to priorly reduce the methylome's dimension. This process yields a restricted set, known as the Restricted Methylome.

(c) **Relation between the Restricted Methylome and the CVD:**

In this step, we explore the association between the Restricted Methylome and CVD, obtaining an even smaller set of significant CpG sites, the Reduced Methylome.

(d) **Relation between the Reduced Exposome and the Reduced Methylome:**

Here, we employ the reduced sets obtained from steps a) and c). We conduct asso-

ciation studies of each CpG site within the Reduced Methylome with each nutrient within the Reduced Exposome. This yields a set of potential mediators (CpG sites) for each nutrient in the Reduced Exposome.

(e) **Assess the mediation for each nutrient:**

For each nutrient within the Reduced Exposome, we investigate the presence and statistical significance of mediation effects of the identified mediators from the previous step.

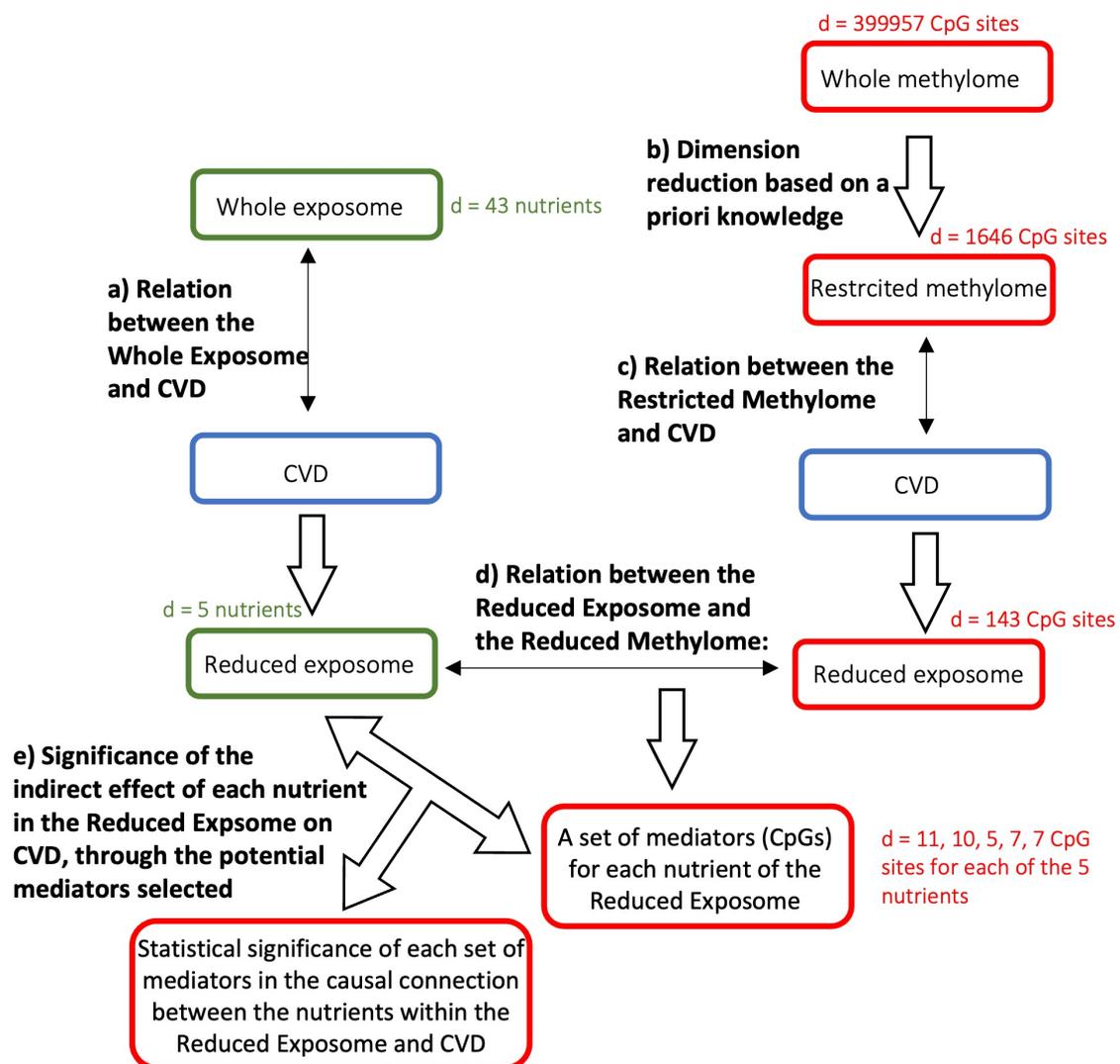


Figure 4.9: Schematic pipeline of the second application of the MITM method

#### 4.3.2. Step a) : Models and assumptions

**Model:**

Here, we outline the mathematical representations of the 43 models we intend to apply

in this stage. For a deeper understanding, please refer to the "Model Fit" section below.

$$\text{logit}(p_j) = \beta_0 + \beta_1 X_j + \vec{\gamma} C$$

where:

- $Y_j \sim Be(p_j)$  is the CVD variable
- $\text{logit}(p_j)$  represents the log-odds of the probability that the dependent variable  $Y_j$  takes on the value 1.
- $X_j$  is a vector representing the daily intakes of the  $j^{\text{th}}$  nutrient belonging to the Whole Exposome, for each individual,  $j = 1, \dots, 43$ .
- $C$  is a matrix containing all the confounding factors, for each individual.
- $\beta_0$  is the intercept .
- $\beta_1$  and  $\vec{\gamma}$  are the coefficient of the independent variables  $X_j$  and  $C$ .

#### Step a)

#### Generalized linear regression model to study the relation between the Whole Exposome and the Cardiovascular disease

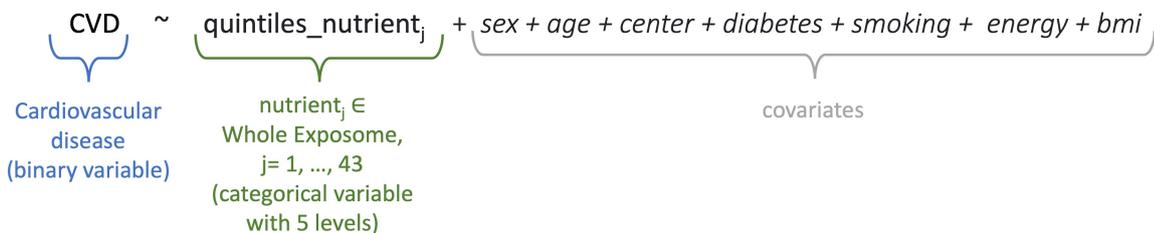


Figure 4.10: Schematic representation of the multiple generalized linear models fitted in step a)

#### Division of the distribution into quintiles:

In the first step of the second MITM approach, we employ multiple generalized linear models. Here, the binary variable CVD serves as the outcome, the nutrient acts as the dependent variable, and all confounding factors are included as linear variables. Once again, as previously detailed in Section 4.1.4, for each nutrient, we utilize the categorical variable denoting the division into quintiles instead of the continuous one. This decision is motivated by the same considerations as earlier: the improved modeling approach.

**Model fit:**

In this phase, we examine the association between all the nutrients and CVD, treated as the outcome variable. We utilize univariate generalized linear models for this purpose. As mentioned earlier we refrain from applying multiple testing corrections. The standard set of adjustment factors (in the case in which we are not dealing with methylation levels) are introduced as linear variables in the model: *sex, age, recruitment center, diabetes status, smoking habits, energy intake* and *BMI*.

**4.3.3. Step b) : A priori selection of CVD-relevant CpG sites**

Here again, we employ biological data from genetic databases to a priori reduce the methylome dimension. This selection process is the same of the one detailed in Section 4.1.2, that indeed leads us to the same set of 1646 preselected CpG sites, forming the Restricted Methylome.

**4.3.4. Step c) : Model and assumptions****Model :**

Here we present the mathematical expression of the 1646 models we want to fit at this stage. The details and underlying motivations are elaborated further in the "Model Fit" section below.

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + \vec{\gamma} C$$

where:

- $Y_i \sim Be(p_i)$  is the CVD variable
- $\text{logit}(p_i)$  represents the log-odds of the probability that the dependent variable  $Y_i$  takes on the value 1.
- $X_i$  is a vector representing the  $i^{\text{th}}$  CpG site belonging to the Restricted Methylome, for each individual,  $i = 1, \dots, 1646$ .
- $C$  is a matrix containing all the confounding factors, for each individual.
- $\beta_0$  is the intercept .
- $\beta_1$  and  $\vec{\gamma}$  are the coefficient of the independent variables  $X_i$  and  $C$ .

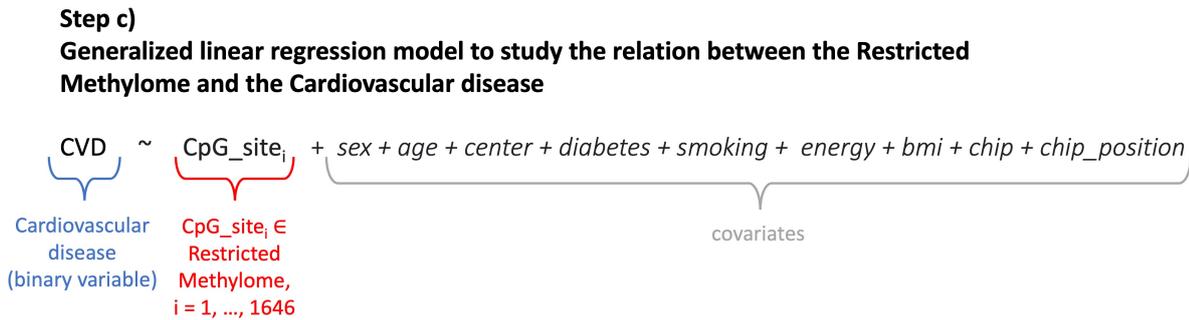


Figure 4.11: Schematic representation of the multiple generalized linear models fitted in step c)

### Model fit:

We test the association between the methylation levels of the preselected CpG sites (Restricted Methyome) and CVD as the outcome. Once again, we refrain from adjusting the p-values for multiple testing. Additionally, we incorporate the standard set of adjustment factors (in presence of methylation levels) as linear variables in the model: *sex*, *age*, *recruitment center*, *diabetes status*, *smoking habits*, *energy intake BMI*, *chip* and *chip position*. This step is crucial in mediation analysis, as it ensures that the mediating CpG sites are indeed associated with the outcome (CVD). The CpG sites selected in this step will constitute the Reduced Methyome.

### 4.3.5. Step d) : Models and assumptions

#### Model:

Here we present the mathematical expression of the 143\*5 models fitted at this stage. The details are elaborated in the "Model Fit" section below.

$$Y_i = \beta_0 + \beta_1 X_j + \vec{\gamma} C + \epsilon_{i,j}$$

where:

- $Y_i$  is a vector containing methylation levels at the  $i^{th}$  CpG site belonging to the Reduced Methyome, for each individual,  $i = 1, \dots, 143$ .
- $X_j$  is a vector representing the daily intakes of the  $j^{th}$  nutrient belonging to the Reduced Exposome, for each individual,  $j = 1, \dots, 5$
- $C$  is a matrix containing all the confounding factors, for each individual.

- $\beta_0$  is the intercept
- $\beta_1$  and  $\vec{\gamma}$  are the coefficient of the independent variables  $X_j$  and  $C$
- $\epsilon_{i,j} \stackrel{iid}{\sim} N(0, 1)$  is the error term for the model with the  $i^{th}$  CpG site and the  $j^{th}$  nutrient.

**Step d)**

**Linear regression model to study the relation between the Reduced Exposome and the Reduced Methylome**

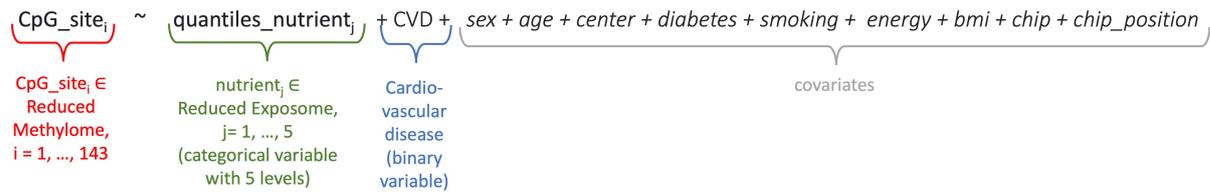


Figure 4.12: Schematic representation of the multiple linear models fitted in step d)

**Assumptions:**

As explained in detail in section 4.1.3), the use of linear regression with beta values in DNAm studies is accepted as unbiased even if the data do not satisfy the assumptions of the test. Following this theoretical basis, we proceed to construct the necessary model.

**Model fit:**

In this step, our aim is to investigate the association between each nutrient from the Reduced Exposome (identified in step a)) and the CpG sites composing the Reduced Methylome (identified in step c)). We employ various linear regression models, without applying multiple testing corrections. Additionally, we incorporate the typical confounding factors (*sex, age, recruitment center, diabetes status, smoking habits, energy intake BMI, chip and chip position*) as linear variables. The outcome of this step will yield a set of potential mediators for each nutrient within the Reduced Methylome.

**4.3.6. Step e): Assess the mediation proportion**

At this stage, we aim to assess the mediation of the CpG sites selected as potential mediators in the previous step for each nutrient within the Reduced Exposome. It's important to note that while some methylation sites may mediate for multiple nutrients, each nutrient typically has its own unique set of mediators. Using mediation analysis, we explore the extent to which the effect of a factor (a specific nutrient) on an outcome (CVD) is mediated by an intermediate layer (DNA methylation levels).

There are different ways to assess mediation [62]:

- One widely used method for estimating this proportion is the difference method, which quantifies the change in estimates obtained from separate linear regression models examining the exposure-outcome relationship, with and without the mediator. However, within this framework, evaluating the mediation of multiple CpG sites becomes a challenge, as they are incorporated individually as distinct covariates rather than being collectively considered as a whole.
- An alternative approach, as we employ here, involves the use of Structural Equation Modeling (SEM), offering a more appropriate inference framework for mediation analyses and various causal investigations. SEM is a versatile and potent multivariate technique that utilizes a conceptual model, path diagram, and a system of interconnected regression-style equations to capture complex and dynamic relationships within a network of observed and unobserved variables. Unlike regression, SEM doesn't strictly adhere to a clear division between dependent and independent variables; instead, these distinctions hold in relative terms, as a variable considered dependent in one equation may function as an independent variable in other parts of the SEM system. The SEM framework brings several advantages to mediation analysis, including:
  - SEM simplifies the examination of mediation hypotheses by being specifically designed, in part, to assess these complex mediation models within a single comprehensive analysis.
  - The SEM analysis approach returns model fit information, offering insights into the consistency of the hypothesized mediational model with the data. It also provides evidence regarding the plausibility of the causality assumptions made during the construction of the mediation model.

In particular in our case, for each nutrient belonging to the Reduced Exposome, we build a SEM model with the following specifications:

- We build a latent variable which is constructed from the set of potential mediators (CpG sites).
- We model the outcome variable CVD as influenced by a set of predictors (the nutrient under consideration and all the confounders) and mediators (the latent variable just built).
- We model the influence of the nutrient on the latent variable in the mediator model.

- We compute indirect effect as the product of the coefficients related to the influence of the nutrient on the latent variable and of the latent variable on the CVD, in the previously built models.
- The total effect of the predictors (nutrient) on CVD is calculated as the sum of the direct and indirect effect.

For greater clarity we report a representation of the SEM model built for one nutrient (folic acid in this case). The symbol ”  $\sim$  ” is commonly used to denote the relationships between latent variables and their indicators.

#### Latent variable:

$$dnam \sim cg11088672 + cg05575921 + cg12065531 + cg01353448 + cg27510066 + cg00247963 + cg24613083 + cg03636183 + cg09958192 + cg19866478 + cg27585074$$

#### Outcome model:

$$cvd \sim c * folic.acid + c_1 * age.recr + c_2 * sex + c_3 * energy + c_4 * bmi + c_5 * diabetes + c_6 * smoking_c + c_7 * smoking_f + c_8 * center_V + c_9 * center_R + c_{10} * center_T + m1 * dnam$$

#### Mediator model:

$$dnam \sim a_1 * folic.acid$$

#### Indirect effect (IE):

$$dnam\_IE := a_1 * m_1$$

$$sumIE := (a_1 * m_1)$$

#### Total effect (IE):

$$total := c + (a_1 * m_1)$$

### 4.3.7. Visualization tools: Manhattan Plot

Here we introduce another essential plots that, along with the three already presented, will enhance the visualization of our results in the next section: **Manhattan Plot**:

A Manhattan plot is a graphical representation commonly used in genome-wide association studies (GWAS) to visualize the results of statistical tests conducted across the genome. In our case, each data point represents a CpG site, with its position on the x-axis indicating its chromosomal location, and the y-axis representing the statistical significance of its association with a particular outcome. The significance is represented as

the negative logarithm (base 10) of the p-value obtained from the statistical test. Therefore, higher points on the y-axis indicate stronger evidence of association. We also use horizontal lines to highlight the desired thresholds. It provides a concise overview of the genetic variants most likely to be associated with the trait or disease under investigation, making it a valuable tool for identifying potential sites of interest.

## 4.4. Results for the second MITM approach

### 4.4.1. Results step a)

In this section we will present, step by step, all the results obtained.

From the tests of associations of the nutrients (treated with the distribution divided into quantiles) with the health outcome (CVD) we obtain 5 significant nutrients, which will form the Reduced Exposome: *folic acid*, *iron*, *water*, *edible portion* and *vitamin B6*. Here we establish an association when the p-value associated to the 5<sup>th</sup> quintile of a nutrient (with refer to the 1<sup>st</sup>) is lower than 0.05, without correction for multiple testing. Since in this phase the number of regression models fitted is feasible (43) we can show, using a Volcano plot, all the significance levels (p-values) of each nutrient with refer to the outcome. As we see in Figure 4.13 there are no nutrients which result particularly significant, but even the 5 ones which have a p-value lower than 0.05 are just slightly below the threshold (higher in the plot since on the y-axis there is  $-\log_{10}(p - value)$ ). As regression coefficients for each nutrient we used the associated with the 5<sup>th</sup> quintile with refer to the baseline (1<sup>st</sup> quintile)

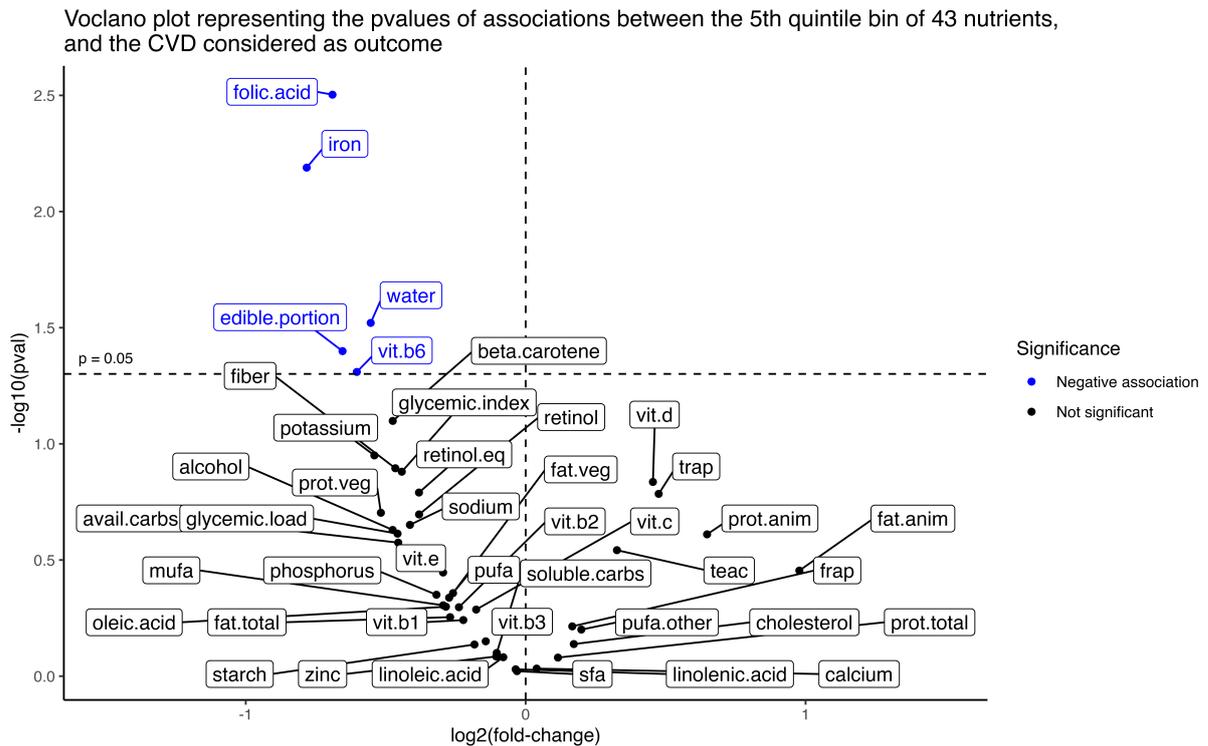


Figure 4.13: Volcano Plot representing all the associations between the nutrients within the Restricted Methylome and the CVD.

Furthermore, utilizing a Volcano Plot facilitates the assessment of the direction of association. The x-axis represents the  $\log_2FC$ , which reflect the sign of the beta coefficient in the regression model for each nutrient. Even if all the nutrients have a  $\log_2FC$  not deviating significantly from 0, we can still make tentative inferences about the sign of the beta coefficient.

In this section, as previously conducted in Section 4.2.2, we categorize each of the five significant nutrients as either a protective or a risk factor. We subsequently examine the dose-response curves (reported in Figure 4.14) and finally conduct a comparative analysis with existing literature.

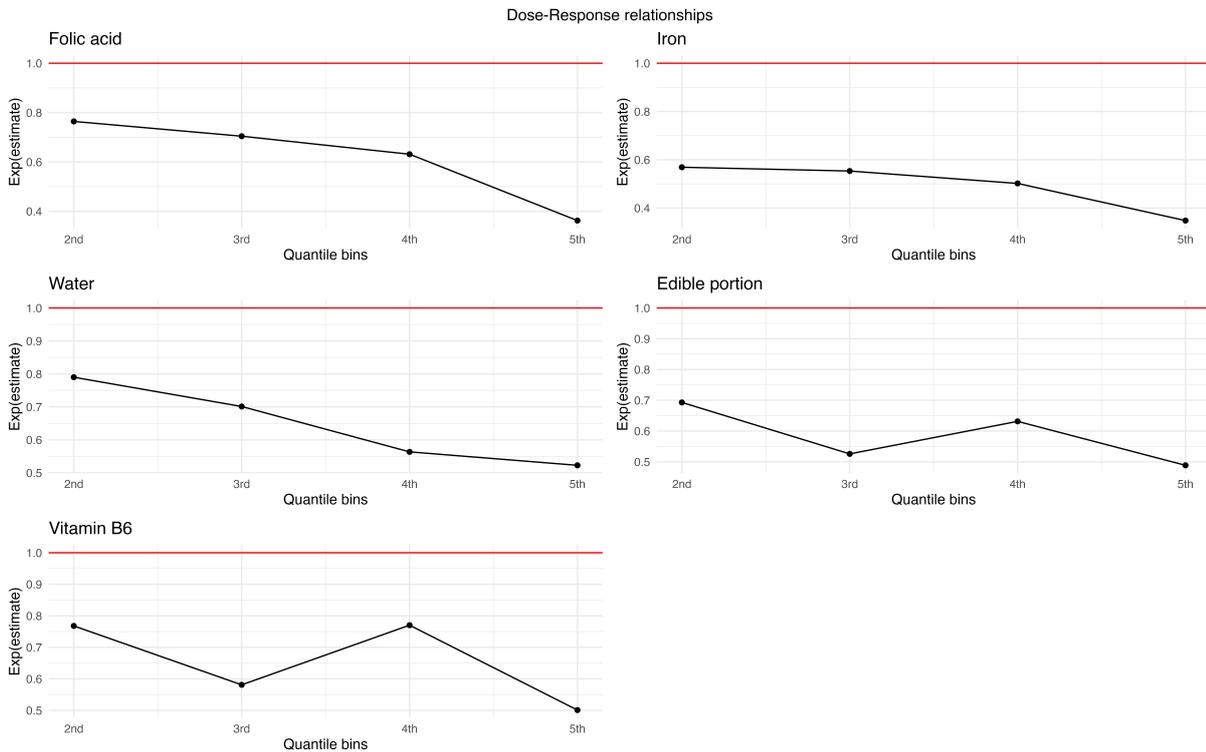


Figure 4.14: Dose-response relationships for each of the 5 nutrients within the Reduced Exposome, representing the (exponent of the) estimates for the 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup> and 5<sup>th</sup> quintiles with refer to the baseline (1<sup>st</sup> quintile), for each nutrient. In red we highlight the threshold  $\beta = 0$ , which allow to distinguish between protective factors (below the line) and risk factor (above the line).

This is what we obtained for each of the 5 nutrients:

#### **Folic acid:**

According to our findings, folic acid emerges as the most significant nutrient and functions as a protective factor, demonstrated by a negative regression coefficient of  $\beta = -1.02$ . This indicates that individual in the 5<sup>th</sup> quintile for folic acid consumption have a decreased risk in the odds of the event (development of CVD) compared with the reference group (individuals in the 1<sup>st</sup> quintile for folic acid consumption)

Upon examining the dose-response curve, we observe a consistently decreasing trend. This indicates that administering more folic acid leads to greater benefits for CVD, within the analyzed value range. The optimal outcome is observed at the 5<sup>th</sup> quintile, corresponding to a range of (0.348, 1.02) mg/day.

A review of the literature [59][60], drawing from the most comprehensive and current

evidence, supports a modest benefit of folic acid supplementation in reducing both CVD risk factors and overall risks. Our results strongly align with these findings.

**Iron:**

Within this framework (as also observed in the results from the previous method, detailed in Section 4.2.2), iron proves to be a protective factor, with a regression coefficient of  $\beta = -1.05$ , suggesting that individuals in the highest quintile of iron consumption experience a reduced likelihood of the event (onset of CVD) when compared to the reference group (those in the lowest quintile for iron consumption).

As the results are the same of those reported in the previous method, detailed information concerning the dose-response relationship and the comparison with the literature can be found in subsection 4.2.2.

**Water:**

Water intake emerges as a protective factor, evidenced by an estimated log ratio ( $\beta$ ) of -0.65.

The dose-response curve consistently displays a decreasing pattern, indicating that the optimal dose lies within the highest range, corresponding to the 5<sup>th</sup> quintile.

Literature [58] supports these findings, affirming that higher total water intake is associated with a reduced risk of CVD mortality. A comprehensive review indicates that adequate hydration lowers the risk of conditions like hypertension, fatal coronary heart disease, venous thromboembolism, and cerebral infarct. The reference of a decreasing curve in the dose-response relationship with existing research further strengthens the reliability of our findings.

**Edible portion:**

The evaluation of results for the total edible portion yields a somewhat distinct perspective. Given that this factor encompasses the entirety of food consumed in a day, it comprehend both healthy and unhealthy forms of nutrients. While it generally seems to act as a protective factor here, there is no point in reasoning on the meaning of this results, neither in quantifying its contribution to reduce/increase the risk of having a CVD. Additionally, the oscillating pattern observed in the dose-response relationship stresses the absence of a definitive safest quantity (in grams) of daily food intake.

**Vitamin B6:**

Vitamin B6 emerges as a protective factor, with the safest doses observed in the 5<sup>th</sup> quintile, followed by the 3<sup>rd</sup>, while the higher risk levels are associated with the 2<sup>nd</sup> and 4<sup>th</sup> quintiles.

Similarly, the dose-response curve exhibits an oscillating pattern, indicating that the optimal dosage isn't discernible through a straightforward pattern (such as extremes in a U-shape or highest/lowest values in increasing/decreasing curves), making it a challenging value to pinpoint.

This observation aligns with existing knowledge about vitamin B6 [61]. A brief review of the literature related to the function of vitamin B6 in metabolic changes, and thus its participation in maintaining health, proves that it is a molecule necessary for the proper functioning of the entire body, and its role cannot be overestimated. However, vitamin B6 deficiency was observed to be connected with Cardiovascular disease, confirming in some extents the reliability of our results.

We summarize all the findings for each nutrient in Table 4.4 and the comparison with literature notions in 4.5

Nutrients	Risk or protective factors	shape of the curve	Quintile related to the best quantities to assume	Ranges of best quantities to assume ( <i>gr/day</i> )
Folic acid	Protective	decreasing	5 <sup>th</sup>	[0.000348, 0.00102],
Iron	Protective	decreasing	5 <sup>th</sup>	[0.0178, 0.0472]
Water	Protective	decreasing	5 <sup>th</sup>	[1580, 3560]
Edible portion	Protective	oscillating	5 <sup>th</sup>	[2050, 4070]
Vitamin B6	Protective	oscillating	5 <sup>th</sup>	[0.00241, 0.00575]

**Table 4.4:** Recap Table of the obtained classification as protective/risk factors and some features of the dose-response relations.

Nutrients	Classification in Risk/Protective factor according to our result	Classification in Risk/Protective factor according to the literature	References to the source
Folic acid	Protective	Protective	[59], [60]
Iron	Protective	Protective	[51], [52]
Water	Protective	Protective	[58]
Edible portion	Protective	-	-
Vitamin B6	Protective	Protective	[61]

Table 4.5: Summary of the comparison between our findings and the ones coming from existing studies. Green color indicated that our results align with the literature, red means that they are in strong contrast, orange means that the comparison is uncertain.

We again check the reliability of our results by computing the inflation factor, which resulted to be 1.04, and by plotting the QQ-plot of observed vs expected  $\log_{10}(p - value)$  for each of the 43 models fitted. Analysing both we can affirm that the inflation is minimal and does not affect the results.

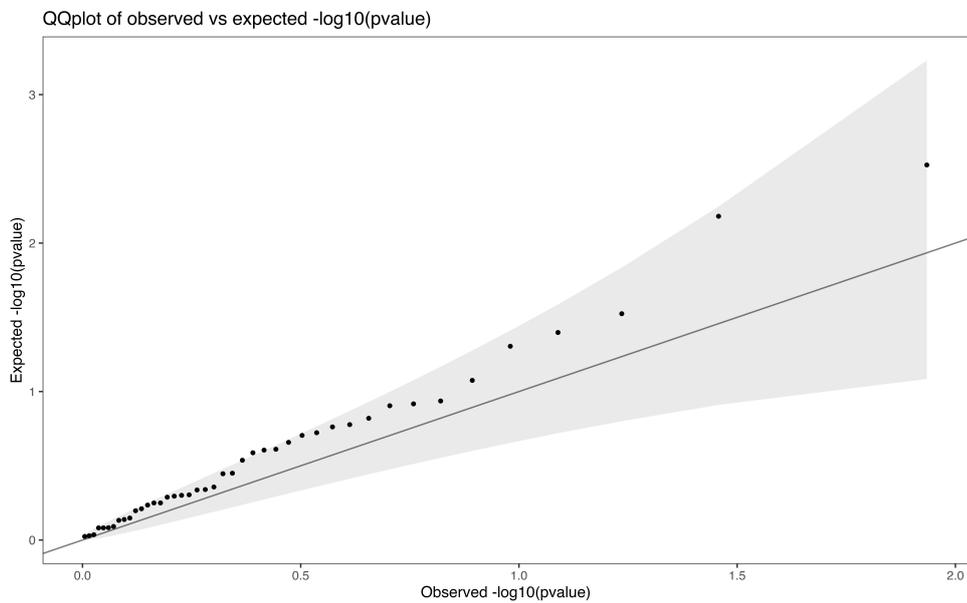


Figure 4.15: QQ-plot showing the inflation of the p-value of the tests conducted at step a), i.e. the deviation of the distribution of the observed tests statistics to the distribution of the expected tests statistics (bisector of the first-fourth quadrant). The inflation factor is  $\lambda = 1.04$

#### 4.4.2. Results step c)

We want to select the potential mediators for exposures associated with CVD, hence in this step we studied the association of each CpG site belonging to the Restricted Methylome with the health outcome (CVD). Without correction we find 143 significant CpG sites, which will form the Reduced Methylome. To visualize the results of the tests conducted across the genome we use a Manhattan Plot (Figure 4.16), where we plot the p-values related to each CpG site in the different regression models. This step is essential because the mediating sites need to be associated with the outcome.

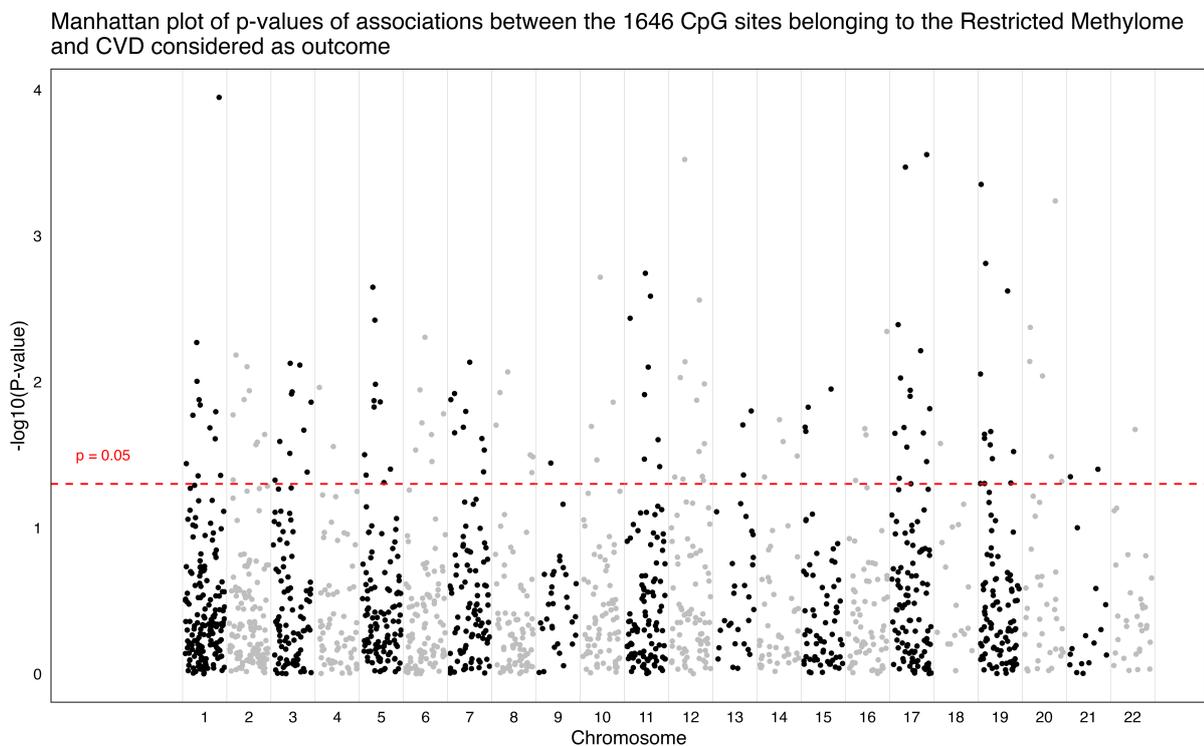


Figure 4.16: Manhattan Plot representing the associations between the CpG sites withing the Restricted Methylome and the CVD.

When examining the data for inflation, it becomes apparent that there is a slight inflationary trend, as evidenced by an inflation factor of  $\lambda = 3.33$  and the QQ-plot comparing observed and expected  $-\log_{10}(p\text{-value})$  (in Figure 4.17). However, this inflation is expected and justifiable in this context. Indeed the analysis includes all CpG sites within the Restricted Methylome, a selection derived from sites that have been previously reported as significantly associated with CVD in both the EWAS catalogue and an exhaustive literature review. So, if the selection is accurate, one can expect the Restricted Methylome

to contain an higher proportion of true predictors of CVD than the Whole Methylome.

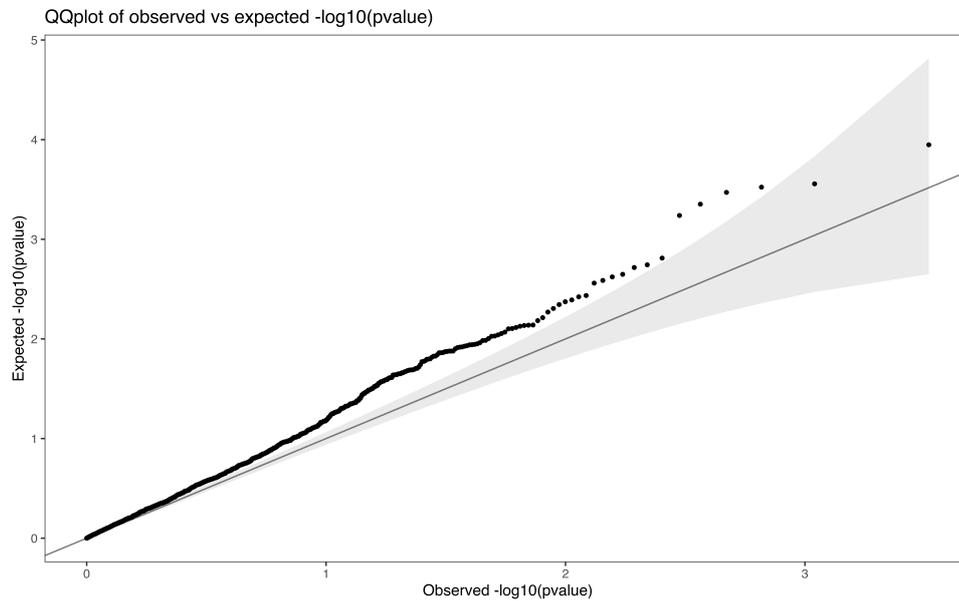


Figure 4.17: QQ-plot showing the inflation of the p-value of the tests conducted at step b), i.e. the deviation of the distribution of the observed tests statistics to the distribution of the expected tests statistics (bisector of the first-fourth quadrant). The inflation factor is  $\lambda = 3.33$

#### 4.4.3. Results step d)

From the test of association of the 143 significant CpG sites (Reduced Methylome) with each of the 5 nutrients composing the Reduced Exposome, for each of the 5 nutrient we identify a set of potential mediators. The latter are chosen by taking all the CpG sites which exhibited a p-value lower than the threshold 0.05 in the association test with each nutrient. No corrections for multiple testing are applied here.

Nutrients belonging to the Reduced Exposome	Number of CpG sites belonging to the Reduced Exposome, associated with each nutrient (potential mediators)
Folic Acid	11
Iron	10
Water	7
Edible portion	7
Vitamin B6	7

Table 4.6: Number of potential mediators (CpG sites) for each nutrient belonging to the Reduced Exposome.

#### 4.4.4. Results step e)

Through Structural Equation Models (SEMs), we assess the extent to which the significant CpG sites identified in the previous step mediate the impact of each significant nutrient on CVD.

For each of the five nutrients, we employ a graphical representation of the relationships among the variables within the fitted SEM model. In Figure 4.18 we report this graph for iron, as an example. In this visualization, variables are depicted as nodes, and directed edges (arrows) signify unilateral relationships. Latent variables are denoted by ellipses, while others are represented as rectangles. Along the edges, both the estimate (numeric value) and the significance of the influence (following the standard convention used in linear model outputs—with more asterisks indicating higher p-values of significance) are displayed .

As we can see, the outcome variable (CVD) receives both direct and indirect contributions. All confounding factors provide direct contributions, while iron contributes both directly and indirectly (via the latent variable *dnam*).

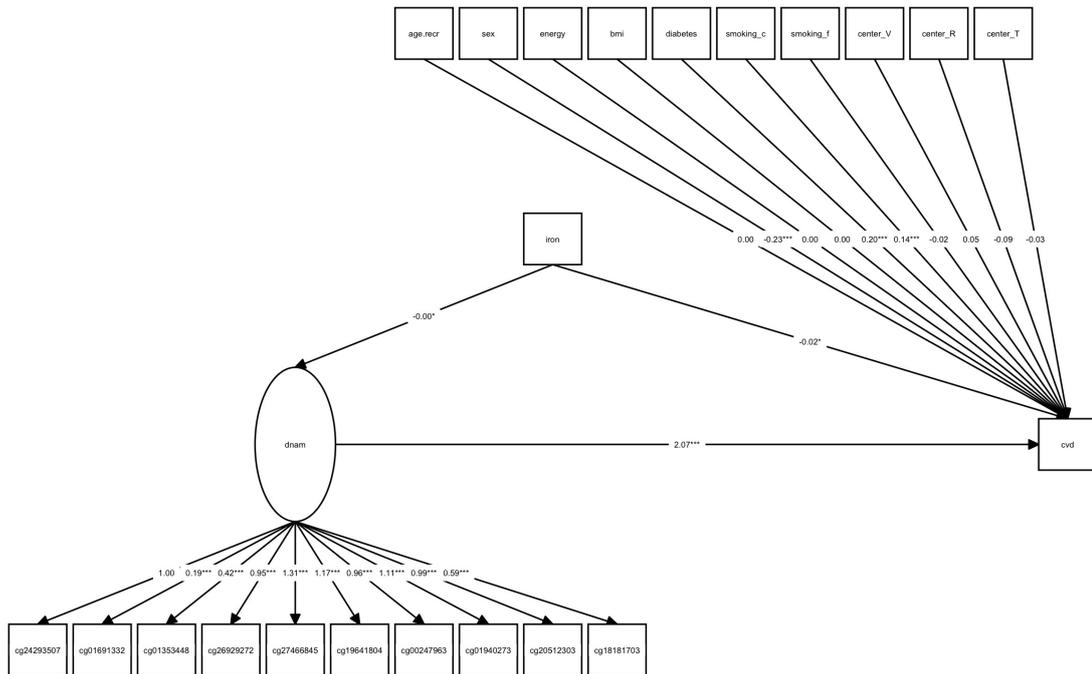


Figure 4.18: Graphical representations among the variables within the fitted SEM model for Iron.

The constructed model structure enables us to evaluate the statistical significance of this indirect contribution. This assessment helps determine whether there is statistical evidence that alterations in nutrient intakes influence CVD through changes in DNA methylation levels. The associated p-values of significance for the indirect effects of each nutrient on CVD are presented in Table 4.7.

Nutrients	p-value
Iron	0.042
Folic Acid	0.083
Water	0.247
Vitamin B6	0.535
Edible portion	0.811

Table 4.7: P-value assessing the statistical significance of the indirect influence of each nutrient (through the set of potential mediators) on CVD

These results indicate that the protective effect of iron and folic acid on CVD risk may go through changes in DNAm profile, although the result for folic acid is only borderline

statistically significant. On the contrary, the protective effect of water, vitamin B6 and edible portion is more likely to act through alternative biological mechanisms independent from DNAm.

# 5 | Stability selection

The aim of this chapter is to select causal predictor for CVD without considering the intermediate layer, implementing a stability selection algorithm. In this way we will be able to discuss, by comparison, if the introduction of the methylome as intermediary produces different results or not.

Before proceeding it's important to emphasize that from now on the parameter  $\lambda$  will represent the penalization parameter in the Lasso regression, and not the inflation factor as in the previous chapters.

## 5.1. Methods

The implemented stability selection algorithm follows the guidelines described in the original paper by Meinshausen and Bühlmann [63]. Here we present a general overview of the algorithm. To ensure the robustness of the results, a series of steps is averaged over a set number of iterations (in this case, 1,000).

Initially, in each iteration, a random 50% subset of the data (i.e. participants), is selected. For each chosen participant, we retrieve relevant information including the outcome (CVD), the variables of interest (nutrients, used with their continuous distribution, and not with the division into quintiles), and the explanatory variables (confounders: *sex*, *age*, *recruitment center*, *diabetes status*, *smoking habits*, *energy intake* and *BMI*).

Subsequently, we apply a Lasso regression model to this subset. The selection of an optimal  $\lambda$  value (tuning parameter in the Lasso regression), which crucially balances the trade-off between bias and variance, is pivotal. To identify this optimal value, we employ a Cross-Validation procedure. Cross-validated Lasso performs multiple rounds of training and testing on different subsets of the current data, using all the  $\lambda$  values contained in a predefined grid (here 100 equally separated values, from 0.00001 to 0.1). Ultimately, the  $\lambda$  value associated with the lowest mean squared error across the validation sets is chosen. This  $\lambda$  value corresponds to the model that strikes the most effective balance between complexity and accuracy, yielding an optimal and robust predictive model.

From the model using the best penalization parameter we extract the selected variables

(those with non-zero regression coefficients) and the signs of the corresponding regression coefficients (beta).

We repeat this process for 1,000 iterations.

At this stage, as our focus is solely on the significance of nutrients in relation to CVD, we retain information related to the nutrients and discard all other variables, i.e. the confounders which were used for adjusting each regression model.

We now calculate the probability of each variable being selected. This is done by computing the ratio of the total number of times it was selected to the number of iterations. We also calculate the probability for the coefficient associated with each selected nutrient at each iteration to be positive. This is achieved by determining the ratio of the number of times it exhibited a positive sign to the total number of times it was selected. An average value of 1 suggests a consistently positive sign, while an average of 0 indicates a consistently negative sign. Additionally, values of 1.0 and 0.0 signify that the sign remains stable across iterations.

Finally, a chosen threshold (here 0.5) is applied to filter out nutrients that surpass a desired level of importance. This culminates in a final selection of the most relevant variables. Of course the choice of the threshold is arbitrary as well, so there is not a specific value, but it must be chosen according to the specific situation.

We stress that with stability selection we do not simply select one model in the list. Instead the data are perturbed (e.g. by subsampling) many times and we choose all structures or variables that occur in a large fraction of the resulting selection sets.

Below, in Figure 5.1 we present a pseudo-algorithm of the method just implemented.

```

stability_selection_lasso <- function(x, y, z, prop = 0.5, n_iter = 1000) {
  # Initialization of needed variables
  # x dataframe containing the predictors (nutrients)
  # y vector containing the outcome(CVD)
  # z dataframe containing the adjustment factors
  n <- nrow(x)
  p <- ncol(x)
  count_variables_selected = rep(0,54) # how many times each variable is selected
  betas = data.frame(matrix(ncol = 0, nrow = 54)) # beta for each variable at each iteration (for best_lambda)
  lambda.grid <- seq(0.00001,0.1,length=100) # to use for CV
  # Repeat the process for a certain number of times:
  for (iter in 1:n_iter) {
    # 1. randomly select 50% of the data:
    sample_idx <- sample(1:n, floor(0.5 * n)) # Randomly select 50% of the data
    x_sample <- x[sample_idx, , drop = FALSE]
    y_sample <- y[sample_idx]
    z_sample <- z[sample_idx, , drop = FALSE]
    # Remove lines with NAs (there are 2 in z_sample)
    rows_no_na = which(complete.cases(z_sample))
    x_sample = x_sample[rows_no_na,]
    y_sample = y_sample[rows_no_na]
    z_sample = z_sample[rows_no_na,]
    # Create necessary structures
    d = data.frame(y_sample,x_sample,z_sample)
    x_new <- model.matrix(y_sample ~ .,data = d)[-1] #Build the matrix of predictors. (-1 to remove intercept)
    y_new = as.numeric(y_sample)
    # 2. Fit a Lasso regression model on the current samples,
    # using cross validation to select the best lambda value (penalization) on a fixed grid of 100 values.
    fit = cv.glmnet(x_new,y_new,lambda=lambda.grid, family = 'binomial')
    best_lambda <- fit$lambda.min
    #Only at the first iteration give the name at the rows
    if (iter == 1) {
      rownames(betas) = rownames(betas_iter)
      names(count_variables_selected) = rownames(betas_iter)
    }
    # 3. Extract, at each iteration, the variables (nutrients + confounders) which are selected
    betas_iter <- predict(fit, s=best_lambda, type = 'coefficients')
    count_variables_selected = count_variables_selected + !betas_iter==0 #sum 1 to selected nutrients
    # 4. Extract, at each iteration, the beta (regression coefficient) associated with each variable selected
    col_name = paste0("iter", iter)
    betas[,col_name] = as.numeric(betas_iter)
  }
  # 5. Select only the information regarding the nutrient
  count_variables_selected_nutr = count_variables_selected[2:44,]
  betas_nutr = betas[2:44,]
  # 6. Compute the probability for each nutrient to be selected
  selection_prob = count_variables_selected_nutr/n_iter
  # 7. Select those nutrients which have a selection proportion above a chosen threshold
  final_selected <- selection_prob[selection_prob >= prop]
  # 8. For the selected nutrients compute the proportion of positive betas across all the iterations.
  proportion_positive_betas = rep(0,length(final_selected))
  names(proportion_positive_betas) = names(final_selected)
  proportion_positive_betas =
    rowSums(betas[names(final_selected),]>0)/count_variables_selected_nutr[names(final_selected)]
  return(list(final_selected,proportion_positive_betas)) }

```

Figure 5.1: Pseudo algorithm used for the stability selection process

## 5.2. Results

Using the stability selection algorithm with a selection proportion threshold of 50%, we identify 3 significant nutrients, presented in decreasing order of selection proportion: glycemic index, TRAP, and folic acid.

Each of these nutrients exhibits a unique sign, serving as an indicator of an unequivocally positive or negative contribution. Specifically, glycemic index and folic acid emerge as protective factors, while TRAP appear to be a risk factor. These associations align with the findings from the two previous applications of the MITM and are partially consistent with those reported in the literature. Indeed:

- Glycemic index's findings vary in the literature, possibly influenced by different factors at play. Its classification as a protective or risk factor is not universally recognized.
- In contrast to the literature, TRAP is identified as a risk factor in this phase, presenting a discrepancy with its classification as a protective factor in existing research.
- Folic Acid is the only nutrient that entirely mirrors previously established results, consistently reported as a protective factor.

However, it is important to keep in mind that while Lasso is effective for variable selection, it operates under the assumption of a linear relationship between the selected variables and the outcome. Nevertheless, our observation reveals that most nutrients do not conform to a linear relationship with the outcome. Consequently, Lasso might not capture non-linear relationships effectively. Despite this, it remains a valuable selection tool. It's essential to acknowledge that certain significant variables with non-linear relations to the outcome may not be selected by Lasso but still be significant for CVD.

We summarize all the findings in Table 5.1

Selected nutrients	Selection proportion	Probability of having a positive sign	Sign
Glycemic index	0.785	0.000	-
TRAP	0.573	1.000	+
Folic acid	0.519	0.000	-

Table 5.1: Results of the stability selection algorithm

## 6 | Comparison

In this thesis we implement three methods to relate a set of dietary exposures, the nutrients intakes, to CardioVascular Disease:

- **First Application of MITM:**

Discussed in Sections 4.1 and 4.2, this application utilizes the methylome layer to pinpoint potential causal exposures for testing their association with CVD. The primary goal is to identify a specific set of nutrients establishing a causal link with the outcome.

- **Second Application of MITM:**

Developed in Sections 4.3 and 4.4, this application aids in the selection of potential mediators for exposures associated with CVD. Here, the focus is on pinpointing which CpG sites act as actual mediators for significant nutrients and assessing their effect.

- **Stability Selection Method:**

This method, independent of the intermediate methylome layer, aims to select causal predictors for CVD.

We report a synthetic visual representation of the structures of the three implemented methods in Figure 6.1 and a summary of the main difference among the three methods in Table 6.1.

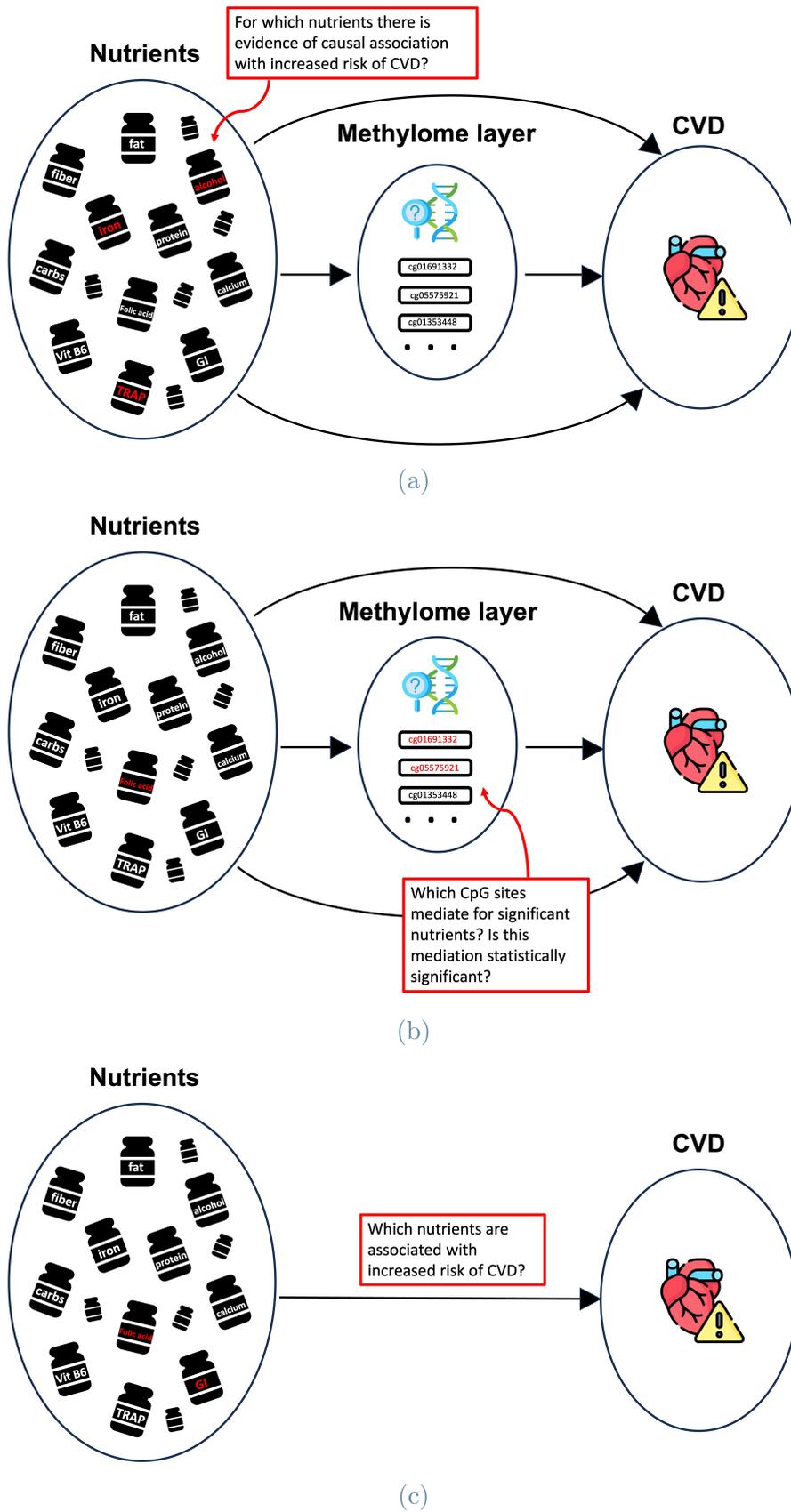


Figure 6.1: Synthetic representation of the different strategies used for each method

	<b>First MITM method</b>	<b>Second MITM method</b>	<b>Stability selection algorithm</b>
<b>Aim</b>	Select exposures likely to be causally associated with CVD	Select CpG sites that mediate the effect on CVD of significant nutrients and assess their statistical significance	Select causal predictors for CVD among nutrients
<b>Use of the methylome</b>	Yes	Yes	No
<b>Method used to reduce the size of the methylome</b>	Literature	Literature + test of association	-
<b>Correction for multiple testings</b>	Yes	No	No
<b>Use of quantiles for the exposures</b>	Yes	Yes (depending on the step)	No

Table 6.1: Summary of the main difference among the three methods implemented.

While all the methods have different structures and ultimate aims, we can still draw comparisons in terms of the selection of exposures that are likely to be causally associated with CVD. This comparative analysis serves as a critical lens through which we evaluate the robustness and consistency of our findings.

Before proceeding with the due considerations, we report in Table 6.2 a summary of all the nutrients selected as significant for the development of CVD from each method.

First MITM method	Second MITM method	Stability selection algorithm
Iron	Iron	Glycemic index
	Folic acid	TRAP
	Water	Folic acid
	Vitamin B6	
	Edible portion	

Table 6.2: Nutrients causally linked to CVD through the methylome layer are highlighted in green, while those establishing an independent causal connection with the outcome are represented in blue.

Here are some considerations about the different methods and some comparison between them:

- **Sensitivity vs specificity:**

It is worth to emphasize that the design of the first MITM, which includes corrections for multiple testings at each step, results in higher specificity (obtained by reducing the likelihood of false positives) but the trade-off is a potential decrease in sensitivity (as the correction may lead to the exclusion of some true positives) when compared to the results of the first step of the second MITM, where we do not correct for multiple testing. As a result, the first MITM produces a considerably narrowed set of effective significant nutrients (ultimately identifying only one nutrient: iron), while the first step of the second MITM identifies five distinct nutrient (iron, folic acid, water, vitamin B6 and edible portion), which could however include both false positives or true negatives.

- **The role of the methylome layer:**

The effectiveness of the MITM method in its first application heavily relies on the intermediate methylome layer for the purpose of selecting nutrients causally linked to the outcome (eventually only iron is identified). The last step of second implementation of the MITM identifies which of the five selected nutrients are causally connected to the outcome passing through the methylome layer (eventually identified in iron and folic acid), assuming that all the 5 nutrients significantly associated with CVD are causal predictors. Both methods identify very small sets of exposures potentially causally connected to the outcome through the methylome layer, consisting in only one nutrient for the first MITM (iron) and two nutrients for the second MITM (iron and folic acid). However, these sets also exhibit p-values that are not significantly high, indicating a weak statistical significance. These observations sug-

gest that the methylation layer might not exert a sufficiently strong influence in mediating the effects of nutrients on CardioVascular disease.

- **Limitations of the stability selection algorithm:**

The use of the stability selection algorithm for comparison with the first and second MITM methods is subject to two significant limitations. Firstly, as previously emphasized, it exclusively identifies nutrients with a linear relationship with the outcome. Secondly, as it does not consider the methylome layer, it potentially identifies a bigger set of nutrients, since it points out both the nutrient which are causally connected through changes in DNA methylation levels and those who establish an independent causal connection. Consequently, it stands as a valuable tool for confirming the significance of selected nutrients, but it is not useful in the opposite direction: it lacks the capability to deny the validity of the obtained results.

If we want to quantify the results by pointing out some specific nutrients, we can say that the more likely to be causally related to the development of CVD are iron and folic acid. Indeed:

- **Iron:** It is the only nutrient identified as significant in the first application of the MITM and ranks as the one with higher statistical evidence that alterations in its intakes influence CVD through changes in DNA methylation levels. It is not selected by the stability selection method, possibly owing to its non-linear U-shape relationship with the outcome. The selection of iron is notably influenced by the methylome layer, a pivotal factor in the first and second MITM, as stated in preceding reasoning.
- **Folic acid:** Folic acid is selected by the last step of the second MITM for showing statistical evidence that alterations in its intakes influence CVD through changes in DNA methylation levels. It also presents a selection proportion exceeding 50% in the stability selection algorithm. This dual significance suggests a potential causal relationship with CVD. Its selection however appears to be less influenced by the presence of the methylome layer compared to iron.

# 7 | Conclusions and future developments

The primary objective of this thesis is to investigate the potential causal relationship between a preselected exposome (nutrients) and a single health outcome (CardioVascular Disease), utilizing DNA methylation as an additional layer of information, possibly mediating an effect.

We start by conducting a comprehensive analysis of two datasets containing information on nutrient intake and DNA methylations for a group of participants belonging to the EPIC cohort. After an initial descriptive analysis of the dataset, we delve into the core of the study. The research develop to be a systematic comparison and evaluation of different methods with the final goal to inform the nutrients-CVD relations. We propose to use a recent approach, the Meet-in-the-Middle (MITM) method, which has been developed and deepened in the very last years. This approach offer an exciting opportunity for exploring groundbreaking techniques while also posing challenges in adapting it to this specific context. Two implementations of the MITM method are presented, both designed to strengthen the causal links between exposures and disease. The first application of the MITM method uses the methylome layer to identify potential new exposures for testing their association with CVD. The second application aims to select potential mediators for exposures associated with CVD. While both methods employ the methylome as intermediate layer, their selection criteria differed: the former focuses on identifying a specific set of nutrients establishing a causal link with the outcome, while the latter concentrates on selecting CpG sites as actual mediators and assessing their effect. As a third method, we propose the development of a stability selection approach, which seeks to identify causal paths between nutrients and CVD without considering the intermediate layer. The objective of this algorithm is to search confirmations or refutations of the importance of the methylome layer in the mediation process, as well as strengthening the evidence of possible causal links between nutrients and CVD. Throughout the research, we consistently reference the literature, both as a starting point (evident in the selection of the restricted set of CpG sites) and as a comparison for the obtained results. Eventually we identify iron

and folic acid as two nutrients that are likely to be causally connected to the development of CVD. The obtained results are important from a public health perspective because they can help develop prevention strategies for high-risk individuals.

However, it's crucial to acknowledge several limitations and opportunities for improvement within our study, suggesting a cautious interpretation of the specific results obtained. Our study, fundamentally, is designed more as a methodological development than a generator of results. The major limitation is that we were not consistent across all three methods, as reported in Table 6.1. For example, we only corrected for multiple testing in the first application of the MITM, we do not use the quintile division and the methylome layer in the stability selection algorithm and we applied different methods to reduced the size of the methylome. This inconsistency arose because we had to adapt the methods to the available data and specific goals. Outlined below are some other recognised limitations, which could serve as a useful starting point for future developments:

- The selected intermediate layer may not be optimal for establishing links between nutrients and CVD. This limitation became particularly evident in the first application of the MITM, where the initial selection of causally related nutrients heavily depended on this layer. A more suitable intermediate biological layer could be one that is more responsive to varying nutrient intakes, such as the microbiome [64]. Many nutrients have indeed the ability to selectively influence the growth of specific microbial species, offering more evident differences that could be considered in a study.
- Another way to overcome the limited impact of nutrients on the variation of methylation levels would be to increase the sample size. Even though 280 recorded case of CVD on 1,508 individuals is not a small number, considering a larger cohort could yield significant benefits for the research, allowing the detection of even small changes.
- To increase the performance, one might consider exploring different adaptations of the conventional linear regression models. For instance, employing Linear Mixed Effect Models with a random intercept on the Center covariate, instead of simple linear regression models, could be beneficial. Also, using Lasso regression during variable selection (e.g., in step b) of the first MITM or step c) of the second MITM) could improve the filtering. While we have tested some of these adjustments on a restricted subset as a trial, it's important to note that implementing such changes would introduce a higher level of complexity and demand increased optimization efforts, especially given the enormous size of the treated data.

- To enhance consistency between methods and facilitate a more meaningful comparison with the stability selection algorithm, it might be beneficial to incorporate the quintile division in the latter as well. This could involve implementing five distinct algorithms, each including one of the five quintiles into which we categorize the distribution of each nutrient. Such an approach could reduce the burden of solely identifying nutrients with a linear relationship with the outcome, allowing a more accurate selection process.

## Bibliography

- [1] Wild, Christopher Paul. «Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology». *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 14, fasc. 8, agosto 2005, pp. 1847–50. PubMed, <https://doi.org/10.1158/1055-9965.EPI-05-0456>.
- [2] Siroux, Valérie, et al. «The Exposome Concept: A Challenge and a Potential Driver for Environmental Health Research». *European Respiratory Review*, vol. 25, fasc. 140, giugno 2016, pp. 124–29. [err.ersjournals.com, https://doi.org/10.1183/16000617.0034-2016](https://doi.org/10.1183/16000617.0034-2016).
- [3] Glier, Melissa B., et al. «Methyl Nutrients, DNA Methylation, and Cardiovascular Disease». *Molecular Nutrition & Food Research*, vol. 58, fasc. 1, gennaio 2014, pp. 172–82. PubMed, <https://doi.org/10.1002/mnfr.201200636>.
- [4] Chadeau-Hyam, Marc, et al. «Meeting-in-the-Middle Using Metabolic Profiling - a Strategy for the Identification of Intermediate Biomarkers in Cohort Studies». *Biomarkers: Biochemical Indicators of Exposure, Response, and Susceptibility to Chemicals*, vol. 16, fasc. 1, febbraio 2011, pp. 83–88. PubMed, <https://doi.org/10.3109/1354750X.2010.533285>.
- [5] Moore, Lisa D., et al. «DNA Methylation and Its Basic Function». *Neuropsychopharmacology*, vol. 38, fasc. 1, gennaio 2013, pp. 23–38. [www.nature.com](http://www.nature.com), <https://doi.org/10.1038/npp.2012.112>.  
«CpG Site». Wikipedia, [https://en.wikipedia.org/w/index.php?title=CpG\\_site&oldid=1175298732](https://en.wikipedia.org/w/index.php?title=CpG_site&oldid=1175298732).
- [6] Cardiovascular Diseases. <https://www.who.int/health-topics/cardiovascular-diseases>.
- [7] Dattani, Saloni, et al. «Causes of Death». *Our World in Data*, settembre 2023. [our-worldindata.org, https://ourworldindata.org/causes-of-death](https://ourworldindata.org/causes-of-death).

- [8] Cardiovascular Disease. <https://www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels/conditions/cardiovascular-disease>.
- [9] Krolevets, Mykhailo, et al. «DNA methylation and cardiovascular disease in humans: a systematic review and database of known CpG methylation sites». *Clinical Epigenetics*, vol. 15, fasc. 1, marzo 2023, p. 56. BioMed Central, <https://doi.org/10.1186/s13148-023-01468-y>.
- [10] Dietary Exposure | EFSA. <https://www.efsa.europa.eu/en/glossary/dietary-exposure>.
- [11] Chareonrungrueangchai, Kridsada, et al. «Dietary Factors and Risks of Cardiovascular Diseases: An Umbrella Review». *Nutrients*, vol. 12, fasc. 4, aprile 2020, p. 1088. PubMed Central, <https://doi.org/10.3390/nu12041088>.
- [12] Dong, Caijuan, et al. «Cardiovascular Disease Burden Attributable to Dietary Risk Factors from 1990 to 2019: A Systematic Analysis of the Global Burden of Disease Study». *Nutrition, Metabolism, and Cardiovascular Diseases: NMCD*, vol. 32, fasc. 4, aprile 2022, pp. 897–907. PubMed, <https://doi.org/10.1016/j.numecd.2021.11.012>.
- [13] Zhong, Jia, et al. «The Role of DNA Methylation in Cardiovascular Risk and Disease: Methodological Aspects, Study Design, and Data Analysis for Epidemiological Studies». *Circulation research*, vol. 118, fasc. 1, gennaio 2016, pp. 119–31. PubMed Central, <https://doi.org/10.1161/CIRCRESAHA.115.305206>.
- [14] Pazzagli, Laura, et al. «Methods for time-varying exposure related problems in pharmacoepidemiology: An overview». *Pharmacoepidemiology and Drug Safety*, vol. 27, fasc. 2, febbraio 2018, pp. 148–60. PubMed Central, <https://doi.org/10.1002/pds.4372>.
- [15] Solène Cadiou. Using dna-methylation to inform the exposome-health relations. Human health and pathology. Université Grenoble Alpes [2020-..], 2020. English. NNT : 2020GRALS038 . tel-03438103
- [16] Vickerstaff, Victoria, et al. «Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes». *BMC Medical Research Methodology*, vol. 19, fasc. 1, giugno 2019, p. 129. BioMed Central, <https://doi.org/10.1186/s12874-019-0754-4>.
- [17] Cadiou, Solène, et al. «Using methylome data to inform exposome-health association

- studies: An application to the identification of environmental drivers of child body mass index». *Environment International*, vol. 138, maggio 2020, p. 105622. ScienceDirect, <https://doi.org/10.1016/j.envint.2020.105622>.
- [18] Hu, Pengfei, et al. «Application of Causal Inference to Genomic Analysis: Advances in Methodology». *Frontiers in Genetics*, vol. 9, 2018. Frontiers, <https://www.frontiersin.org/articles/10.3389/fgene.2018.00238>.
- [19] Hernán, Miguel A. «The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data». *American Journal of Public Health*, vol. 108, fasc. 5, maggio 2018, pp. 616–19. PubMed, <https://doi.org/10.2105/AJPH.2018.304337>.
- [20] Rijnhart, Judith J. M., et al. «Mediation analysis methods used in observational research: a scoping review and recommendations». *BMC Medical Research Methodology*, vol. 21, fasc. 1, ottobre 2021, p. 226. BioMed Central, <https://doi.org/10.1186/s12874-021-01426-3>.
- [21] Blum, Michaël G. B., et al. «Challenges Raised by Mediation Analysis in a High-Dimension Setting». *Environmental Health Perspectives*, vol. 128, fasc. 5, maggio 2020, p. 055001. DOI.org (Crossref), <https://doi.org/10.1289/EHP6240>.
- [22] Rijnhart, Judith J. M., et al. «Statistical Mediation Analysis for Models with a Binary Mediator and a Binary Outcome: The Differences Between Causal and Traditional Mediation Analysis». *Prevention Science*, vol. 24, fasc. 3, aprile 2023, pp. 408–18. Springer Link, <https://doi.org/10.1007/s11121-021-01308-6>.
- [23] Babin, Étienne, et al. «A review of statistical strategies to integrate biomarkers of chemical exposure with biomarkers of effect applied in omic-scale environmental epidemiology». *Environmental Pollution*, vol. 330, agosto 2023, p. 121741. ScienceDirect, <https://doi.org/10.1016/j.envpol.2023.121741>.
- [24] Vineis, Paolo, e Frederica Perera. «Molecular Epidemiology and Biomarkers in Etiologic Cancer Research: The New in Light of the Old». *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 16, fasc. 10, ottobre 2007, pp. 1954–65. PubMed, <https://doi.org/10.1158/1055-9965.EPI-07-0457>.
- [25] Vineis, Paolo, et al. «Long-Term Effects of Air Pollution: An Exposome Meet-in-the-Middle Approach». *International Journal of Public Health*, vol. 65, fasc. 2, marzo 2020, pp. 125–27. PubMed, <https://doi.org/10.1007/s00038-019-01329-7>.
- [26] Assi, Nada, et al. «A Statistical Framework to Model the Meeting-in-the-Middle

- Principle Using Metabolomic Data: Application to Hepatocellular Carcinoma in the EPIC Study». *Mutagenesis*, vol. 30, fasc. 6, novembre 2015, pp. 743–53. PubMed, <https://doi.org/10.1093/mutage/gev045>.
- [27] Poggio, Tomaso, et al. «General Conditions for Predictivity in Learning Theory». *Nature*, vol. 428, fasc. 6981, marzo 2004, pp. 419–22. [www.nature.com](http://www.nature.com), <https://doi.org/10.1038/nature02341>.
- [28] Lazarevic, Nina, et al. «Statistical Methodology in Studies of Prenatal Exposure to Mixtures of Endocrine-Disrupting Chemicals: A Review of Existing Approaches and New Alternatives». *Environmental Health Perspectives*, vol. 127, fasc. 2, febbraio 2019, p. 26001. PubMed, <https://doi.org/10.1289/EHP2207>.
- [29] EPIC Centres - ITALY. <https://epic.iarc.fr/centers/italy.php>.
- [30] Daniel J. Weisenberger, David Van Den Berg, Fei Pan, Benjamin P. Berman, and Peter W. Laird. «Comprehensive DNA Methylation Analysis on the Illumina® Infinium® Assay Platform». [https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote\\_dna\\_methylation\\_analysis\\_infinium.pdf](https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf).
- [31] Truong, Lan-Linh, et al. «Cancer and cardiovascular disease: can understanding the mechanisms of cardiovascular injury guide us to optimise care in cancer survivors?» *ecancermedicalscience*, vol. 16, luglio 2022, p. 1430. PubMed Central, <https://doi.org/10.3332/ecancer.2022.1430>.
- [32] Morrow, Edward H. «The evolution of sex differences in disease». *Biology of Sex Differences*, vol. 6, fasc. 1, marzo 2015, p. 5. BioMed Central, <https://doi.org/10.1186/s13293-015-0023-0>.
- [33] Fiorito, Giovanni, et al. «Socioeconomic Position, Lifestyle Habits and Biomarkers of Epigenetic Aging: A Multi-Cohort Analysis». *Aging*, vol. 11, fasc. 7, aprile 2019, pp. 2045–70. PubMed, <https://doi.org/10.18632/aging.101900>.
- [34] Battram, Thomas, et al. "The EWAS Catalog: a database of epigenome-wide association studies" Wellcome Open Res 2022.
- [35] geograbi. <https://mrcieu.github.io/software/geograbi/>.
- [36] Krolevets, M., Cate, V.t., Prochaska, J.H. et al. DNA methylation and cardiovascular disease in humans: a systematic review and database of known CpG methylation sites. *Clin Epigenet* 15, 56 (2023). <https://doi.org/10.1186/s13148-023-01468-y>.

- [37] «Breusch–Pagan Test». Wikipedia, 23 settembre 2023. Wikipedia, [https://en.wikipedia.org/w/index.php?title=Breusch%E2%80%93Pagan\\_test&oldid=1176713923](https://en.wikipedia.org/w/index.php?title=Breusch%E2%80%93Pagan_test&oldid=1176713923).
- [38] «D’Agostino’s K-Squared Test». Wikipedia, 3 aprile 2023. Wikipedia, [https://en.wikipedia.org/w/index.php?title=D%27Agostino%27s\\_K-squared\\_test&oldid=1147925467](https://en.wikipedia.org/w/index.php?title=D%27Agostino%27s_K-squared_test&oldid=1147925467).
- [39] Anscombe, F. J. «Tests of Goodness of Fit». *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 25, fasc. 1, gennaio 1963, pp. 81–94. DOI.org (Crossref), <https://doi.org/10.1111/j.2517-6161.1963.tb00485.x>.
- [40] Ross, Jason P., et al. «Batch-effect detection, correction and characterisation in Illumina HumanMethylation450 and MethylationEPIC BeadChip array data». *Clinical Epigenetics*, vol. 14, fasc. 1, aprile 2022, p. 58. BioMed Central, <https://doi.org/10.1186/s13148-022-01277-9>.
- [41] Mansell, Georgina, et al. «Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array». *BMC Genomics*, vol. 20, fasc. 1, maggio 2019, p. 366. BioMed Central, <https://doi.org/10.1186/s12864-019-5761-7>.
- [42] Van Den Berg, Sanne, et al. «Significance Testing and Genomic Inflation Factor Using High-density Genotypes or Whole-genome Sequence Data». *Journal of Animal Breeding and Genetics*, vol. 136, fasc. 6, novembre 2019, pp. 418–29. DOI.org (Crossref), <https://doi.org/10.1111/jbg.12419>.
- [43] Pala V, Sieri S, Berrino F, Vineis P, Sacerdote C, Palli D, Masala G, Panico S, Mattiello A, Tumino R, Giurdanella MC, Agnoli C, Grioni S, Krogh V. Yogurt consumption and risk of colorectal cancer in the Italian European prospective investigation into cancer and nutrition cohort. *Int J Cancer*. 2011 Dec 1;129(11):2712-9. doi: 10.1002/ijc.26193. Epub 2011 Aug 5. PMID: 21607947.
- [44] Bendinelli B, Masala G, Saieva C, Salvini S, Calonico C, Sacerdote C, Agnoli C, Grioni S, Frasca G, Mattiello A, Chiodini P, Tumino R, Vineis P, Palli D, Panico S. Fruit, vegetables, and olive oil and risk of coronary heart disease in Italian women: the EPI-COR Study. *Am J Clin Nutr*. 2011 Feb;93(2):275-83. doi: 10.3945/ajcn.110.000521. Epub 2010 Dec 22. PMID: 21177799.
- [45] Fontana, Luigi et al. ‘Dietary Intake of Animal and Plant Proteins and Risk of All Cause and Cause-specific Mortality: The Epic-Italy Cohort’. 1 Jan. 2021 : 257 – 268.
- [46] Nevo D, Liao X, Spiegelman D. Estimation and Inference for the Mediation Propor-

- tion. *Int J Biostat*. 2017 Sep 20;13(2):/j/ijb.2017.13.issue-2/ijb-2017-0006/ijb-2017-0006.xml. doi: 10.1515/ijb-2017-0006. PMID: 28930628; PMCID: PMC6014631.
- [47] «Volcano Plot (Statistics)». Wikipedia, 29 novembre 2022. Wikipedia, [https://en.wikipedia.org/w/index.php?title=Volcano\\_plot\\_\(statistics\)&oldid=1124624468](https://en.wikipedia.org/w/index.php?title=Volcano_plot_(statistics)&oldid=1124624468).
- [48] «Fold Change». Wikipedia, 11 settembre 2023. Wikipedia, [https://en.wikipedia.org/w/index.php?title=Fold\\_change&oldid=1174841074](https://en.wikipedia.org/w/index.php?title=Fold_change&oldid=1174841074).
- [49] Controlling bias and inflation in association studies using the empirical null distribution. <https://bioconductor.org/packages/release/bioc/vignettes/bacon/inst/doc/bacon.html>.
- [50] Aburto NJ, Hanson S, Gutierrez H, Hooper L, Elliott P, Cappuccio FP. Effect of increased potassium intake on cardiovascular risk factors and disease: systematic review and meta-analyses. *BMJ*. 2013 Apr 3;346:f1378. doi: 10.1136/bmj.f1378. PMID: 23558164; PMCID: PMC4816263.
- [51] Naito Y, Masuyama T, Ishihara M. Iron and cardiovascular diseases. *J Cardiol*. 2021 Feb;77(2):160-165. doi: 10.1016/j.jjcc.2020.07.009. Epub 2020 Jul 30. PMID: 32739111.
- [52] Fang, X., Ardehali, H., Min, J. et al. The molecular and metabolic landscape of iron and ferroptosis in cardiovascular disease. *Nat Rev Cardiol* 20, 7–23 (2023). <https://doi.org/10.1038/s41569-022-00735-4>.
- [53] Clemente-Suárez VJ, Mielgo-Ayuso J, Martín-Rodríguez A, Ramos-Campo DJ, Redondo-Flórez L, Tornero-Aguilera JF. The Burden of Carbohydrates in Health and Disease. *Nutrients*. 2022 Sep 15;14(18):3809. doi: 10.3390/nu14183809. PMID: 36145184; PMCID: PMC9505863.
- [54] Vega-López S, Venn BJ, Slavin JL. Relevance of the Glycemic Index and Glycemic Load for Body Weight, Diabetes, and Cardiovascular Disease. *Nutrients*. 2018 Sep 22;10(10):1361. doi: 10.3390/nu10101361. PMID: 30249012; PMCID: PMC6213615.
- [55] Zujko ME, Waśkiewicz A, Witkowska AM, Cicha-Mikołajczyk A, Zujko K, Drygas W. Dietary Total Antioxidant Capacity-A New Indicator of Healthy Diet Quality in Cardiovascular Diseases: A Polish Cross-Sectional Study. *Nutrients*. 2022 Aug 6;14(15):3219. doi: 10.3390/nu14153219. PMID: 35956397; PMCID: PMC9370392.
- [56] Cosentino N, Campodonico J, Milazzo V, De Metrio M, Brambilla M, Camera M, Marenzi G. Vitamin D and Cardiovascular Disease: Current Evidence and Future

- Perspectives. *Nutrients*. 2021 Oct 14;13(10):3603. doi: 10.3390/nu13103603. PMID: 34684604; PMCID: PMC8541123.
- [57] Krittanawong C, Isath A, Rosenson RS, Khawaja M, Wang Z, Fogg SE, Virani SS, Qi L, Cao Y, Long MT, Tangney CC, Lavie CJ. Alcohol Consumption and Cardiovascular Health. *Am J Med*. 2022 Oct;135(10):1213-1230.e3. doi: 10.1016/j.amjmed.2022.04.021. Epub 2022 May 14. PMID: 35580715; PMCID: PMC9529807.
- [58] Majdi M, Hosseini F, Naghshi S, Djafarian K, Shab-Bidar S. Total and drinking water intake and risk of all-cause and cardiovascular mortality: A systematic review and dose-response meta-analysis of prospective cohort studies. *Int J Clin Pract*. 2021 Dec;75(12):e14878. doi: 10.1111/ijcp.14878. Epub 2021 Sep 23. PMID: 34525269.
- [59] Li Y, Huang T, Zheng Y, Muka T, Troup J, Hu FB. Folic Acid Supplementation and the Risk of Cardiovascular Diseases: A Meta-Analysis of Randomized Controlled Trials. *J Am Heart Assoc*. 2016 Aug 15;5(8):e003768. doi: 10.1161/JAHA.116.003768. PMID: 27528407; PMCID: PMC5015297.
- [60] An P, Wan S, Luo Y, Luo J, Zhang X, Zhou S, Xu T, He J, Mechanick JI, Wu WC, Ren F, Liu S. Micronutrient Supplementation to Reduce Cardiovascular Risk. *J Am Coll Cardiol*. 2022 Dec 13;80(24):2269-2285. doi: 10.1016/j.jacc.2022.09.048. PMID: 36480969.
- [61] Stach K, Stach W, Augoff K. Vitamin B6 in Health and Disease. *Nutrients*. 2021 Sep 17;13(9):3229. doi: 10.3390/nu13093229. PMID: 34579110; PMCID: PMC8467949.
- [62] Gunzler, Douglas, et al. «Introduction to mediation analysis with structural equation modeling». *Shanghai Archives of Psychiatry*, vol. 25, fasc. 6, dicembre 2013, pp. 390–94. PubMed Central, <https://doi.org/10.3969/j.issn.1002-0829.2013.06.009>.
- [63] Meinshausen, N. and Bühlmann, P. (2010), Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 417-473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>.
- [64] Dahl, Wendy J., et al. «Diet, Nutrients and the Microbiome». *Progress in Molecular Biology and Translational Science*, vol. 171, 2020, pp. 237–63. PubMed, <https://doi.org/10.1016/bs.pmbts.2020.04.006>.

# A | Appendix A

In this appendix we report results and consideration for the 7 non significant nutrients extracted at step c) of the first application of the MITM, which can be found at subsection 4.2.2.

## **Potassium:**

Potassium also emerges as a protective factor, with a regression coefficient associated with the 5<sup>th</sup> quintile  $\beta = -0.54$ , indicating that individual in the 5<sup>th</sup> quintile for potassium consumption have a decreased risk in the odds of the development of CVD compared with the reference group (individuals in the 1<sup>st</sup> quintile for potassium consumption)

The dose-response curve exhibits an inverted U-shape, with optimal doses at the extremes (either low or high consumption values).

High-quality evidence in the literature [50] supports increased potassium intake in reducing blood pressure, thereby lowering the risk of stroke. Our results are therefore partially confirmed by already known information.

## **Available Carbohydrates:**

Available carbohydrates also demonstrated a protective effect, indicated by an estimated log ratio  $\beta$  of -0.43, which suggests that individuals in the highest (5<sup>th</sup>) quintile of available carbohydrates consumption experience a reduced likelihood of the event (development of CVD) when compared to the reference group (those in the lowest (1<sup>st</sup> quintile for available carbohydrates consumption).

The dose-response relationship exhibits an almost inverted U-shape, with the most favorable quantities found in the 5<sup>th</sup> quintile (with refer to the 1<sup>st</sup>), followed by slightly higher values in the 2<sup>nd</sup> quintile (always with refer to the 1<sup>st</sup>).

The literature [53] strongly emphasizes the role of carbohydrate consumption in the development of major Western diseases in the 21<sup>st</sup> century. While various studies have produced different and sometimes conflicting results, the consensus is that the type of carbohydrate, rather than the quantity, plays a pivotal role in mitigating the risk of car-

diovascular disease and improving cardiovascular risk markers. Carbohydrates derived from certain food sources such as fruits, vegetables, legumes, and whole grains are recommended to lower the risk of CVD, whereas others, particularly those rich in redefined sugars and sweets, are linked to increases in the risk of CVD. Our analysis did not differentiate between different types of carbohydrates, which may account for the observed non-linear decreasing curve.

#### **Glycemic load:**

Our findings regarding glycemic load closely mirror those of available carbohydrates, with very similar regression coefficients and significance p-values. This similarity arises from the linear dependence between these two nutrients, as described by the formula:

$$glycemic\_load = glycemic\_index * available\_carbohydrates$$

where the glycemic index quantifies how rapidly a carbohydrate is metabolized and enters the bloodstream of each individual.

Additionally, the dose-response curve exhibits a comparable, with a more expected (almost completely) decreasing trend. The optimal consumption value definitely lies in the 5<sup>th</sup> quintile, meaning that higher doses are recommended to lower the risk of CVD.

Existing reviews [54] present varying results on this matter. The diverse findings likely stem from a complex interplay of factors related to dietary influences on carbohydrate digestion and metabolism (e.g., dietary fiber or carbohydrate quantity in the diet), variations in study design and participant demographics, as well as limitations inherent to different study methodologies.

#### **TRAP & FRAP:**

In our analysis, both TRAP (total radical-trapping antioxidant parameter) and FRAP (ferric reducing antioxidant power) emerged as risk factors, as indicated by their respective regression coefficients ( $\beta_{TRAP} = 0.37$  and  $\beta_{FRAP} = 0.14$ ).

The dose-response relationship revealed an Inverted U-shape for TRAP and a monotonic decreasing curve for FRAP. The safest intakes, denoting the lowest risk, were observed in the 2<sup>nd</sup>/5<sup>th</sup> quintile for TRAP and in the 5<sup>th</sup> quintile for FRAP.

Both TRAP and FRAP are methods used to evaluate the antioxidant capacity of biological samples, but they measure different aspects of this capacity. High values in these

assays generally indicate a higher level of antioxidants in the sample. However, according to the literature [55], both of them results to be protective factors, which is in contrast with our findings.

#### **Vitamin D:**

According to our results, vitamin D as well appeared to be a risk factor, with a regression coefficient of  $\beta = 0.37$ , which implies that individual in the 5<sup>th</sup> quintile for vitamin D consumption have a higher risk in the odds of the event (development of CVD) compared with the reference group (individuals in the 1<sup>st</sup> quintile for vitamin D consumption)

Its dose-response curve showed an intermediate behaviour between a U-shape and a decreasing curve, with the best intake value at the lowest range (with refer to the 1<sup>st</sup> quintile).

However, according to the literature [56], vitamin D deficiency is emerging as a new risk factor for cardiovascular disease (CVD). In particular, several epidemiological and clinical studies have reported a close association between low vitamin D levels and major CVDs, such as coronary artery disease, heart failure, and atrial fibrillation. According to our results, the worst assumption range belongs to the lowest intake values, meaning that the results appear to be consistent with literature. Based on our findings, the worst consumption range for Vitamin D corresponds to the lowest intake values (2<sup>nd</sup> quintile with refer to the 1<sup>st</sup>), aligning with the trends observed in existing literature.

#### **Alcohol:**

Lastly, our focus is on the results regarding alcohol. The general trend suggests it as a protective factor, with a regression coefficient associated with the 5<sup>th</sup> quintile of  $\beta = -0.21$ .

However, upon examining its dose-response curve, a notable discrepancy arises. While it presents as a risk factor when comparing the coefficient of the 2<sup>nd</sup> quintile against the 1<sup>st</sup>, it emerges as a protective factor for all the other quintiles. This would mean that low doses of alcohol may pose a danger in terms of cardiovascular disease, while higher doses appear to mitigate this risk.

This finding stands in strong contrast to the established understanding of alcohol's impact. Conventionally [57], lower levels of alcohol intakes are linked with a reduced risk of cardiovascular mortality, while heavier daily or weekly consumption is associated with a heightened risk of cardiovascular disease mortality.

# B | Appendix B

In this Appendix we report the essential code implemented for the development of the project, divided by chapters.

## B.1. Preliminary and descriptive analysis of the dataset

In this first part we focus on the R code defining the initial exploratory analysis, regarding in particular the plots of all the covariates and the Spearman correlation matrices of the 43 nutrients (divided by energy) and 10 random CpG sites.

```

1 #EXPLORATORY ANALYSIS
2 #load the dataset:
3 load("datasets/epic_italy_diet_noFlorence_noch_nocancerbefore_noXY.rda")
4 library(ggplot2)
5 library(gridExtra)
6
7 # Analysis of the sample dataset:
8 dim(samples) #1508 individuals (rows), 136 features each (cols)
9 # 1. Sample characteristics:
10 sample_char = samples[,1:17]
11 sample_nutr = samples[,18:61]
12 sample_outcome = samples[,129:136]
13 head(sample_char)
14 colnames(sample_char)
15
16 ##### Histograms and barplots of the 7 covariates #####
17 #####
18
19 # 1. Histogram of the ages at the recruitment time:
20 hist_age = ggplot(data = sample_char, aes(x = age.recr)) +
21   geom_histogram(color = "black", fill = "lightsteelblue2", bins = 20) +
22   ggtitle("Age at the recruitment time") +
23   geom_vline(xintercept = median(sample_char$age.recr), col = "navy",
24     lwd = 0.5, linetype = "dashed") +
25   theme_classic() +

```

```

25   xlab("Years") +
26   ylab("Frequency")
27 # 2. Barplot of the genders:
28 barplot_gender = ggplot(data = sample_char, aes(x = sex)) +
29   geom_bar(color = "black", fill = "lightsteelblue2") +
30   ggtitle("Sex") +
31   theme_classic() +
32   xlab("Sex") +
33   ylab("Frequency") +
34   stat_count(aes(label=..count..), geom="text", vjust=-0.5, size=3,
35     color="black") +
36   ylim(0,1000)
37 # 3. Barplot of the smoking status:
38 barplot_smoking = ggplot(data = sample_char, aes(x = smoking.status)) +
39   geom_bar(color = "black", fill = "lightsteelblue2") +
40   ggtitle("Smoking status") +
41   xlab("Smoking status") +
42   ylab("Frequency") +
43   theme_classic() +
44   stat_count(aes(label=..count..), geom="text", vjust=-0.5, size=3,
45     color="black") +
46   ylim(0,800)
47 # 4. Histogram of the energy:
48 hist_energy = ggplot(data = sample_nutr, aes(x = energy)) +
49   geom_histogram(color = "black", fill = "lightsteelblue2", bins = 20) +
50   geom_vline(xintercept = median(sample_nutr$energy), col = "navy", lwd
51     = 0.5, linetype = "dashed") +
52   ggtitle("Energy") +
53   xlab("Energy (kcal/day)") +
54   ylab("Frequency") +
55   theme_classic()
56 # 5. Barplot of the diabetes:
57 rows_diabetes = union(which(sample_char$diabetes), which(!is.na(
58   sample_outcome$date.diabetes.dx)))
59 diabetes = rep(0,1508)
60 diabetes[rows_diabetes] = 1
61 sample_char$diabetes_ever = as.factor(diabetes)
62 levels(sample_char$diabetes_ever) = c("no diabetes", "diabetes")
63
64 barplot_diabetes = ggplot(data = sample_char, aes(x = diabetes_ever)) +
65   geom_bar(color = "black", fill = "lightsteelblue2") +
66   ggtitle("Diabetes (at recruitment time or later)") +
67   xlab("Diabetes") +
68   ylab("Frequency") +

```

```

65   stat_count(aes(label=..count..), geom="text", vjust=-0.5, size=3,
66     color="black") +
67   ylim(0,1500) +
68   theme_classic()
69 # 6. Barplot of the recruitment centers:
70 center_table <- table(sample_char$center)
71 sample_char$center = droplevels(sample_char$center)
72 barplot_center = ggplot(data = sample_char, aes(x = center))+
73   geom_bar(color = "black", fill = "lightsteelblue2") +
74   ggtitle("Recruitment centers") +
75   xlab("Centers") +
76   ylab("Frequency") +
77   stat_count(aes(label=..count..), geom="text", vjust=-0.5, size=3,
78     color="black") +
79   ylim(0,900) +
80   theme_classic()
81 # 7. Histogram of bmi:
82 hist_bmi = ggplot(data = sample_char, aes(x = bmi)) +
83   geom_histogram(color = "black", fill = "lightsteelblue2", bins = 20) +
84   geom_vline(xintercept = median(sample_char$bmi, na.rm = TRUE), col = "
85     navy", lwd = 0.5, linetype = "dashed") +
86   ggtitle("BMI") +
87   xlab("BMI (kg/m^2)") +
88   ylab("Frequency") +
89   theme_classic()
90 # 8. Barplot of CVD:
91 CVD = as.factor(ifelse(!is.na(sample_outcome$date.cvd.dx),1,0))
92 levels(CVD) = c("No CVD", "CVD")
93 my_data = data.frame(CVD, sample_char, sample_nutr)
94 my_data$CVD = as.factor(CVD)
95 barplot_CVD = ggplot(data = my_data, aes(x = CVD)) +
96   geom_bar(color = "black", fill = c("darkblue", "lightsteelblue2")) +
97   ggtitle("CVD") +
98   xlab("CVD") +
99   ylab("Frequency") +
100  theme_classic() +
101  stat_count(aes(label=..count..), geom="text", vjust=-0.5, size=3,
102    color="black") +
103  ylim(0,1300)
104 # 9. Barplot of CVD by center
105 set.seed(123)
106 center <- sample_char$center
107 data <- data.frame(CVD, center)

```

```

105 # calculate sum of males and females for each center
106 sum_data <- aggregate(data$CVD, by = list(data$center), FUN = function(x
    ) {
107   c(sum(x == "1"), sum(x == "0"))
108 })
109 data= data.frame(center = sum_data$Group.1, CVD = sum_data$x[,1], no_CVD
    = sum_data$x[,2])
110 # convert data from wide to long format
111 data_long <- tidyr::pivot_longer(data, cols = c("CVD", "no_CVD"),
112                                   names_to = "CVD", values_to = "count")
113 # create stacked bar plot
114 quartz()
115 barplot_CVD_by_center = ggplot(data_long, aes(x = center, y = count,
    fill = CVD)) +
116   geom_bar(stat = "identity", position = "stack") +
117   scale_fill_manual(values = c("lightsteelblue2", "darkblue"), name = "
    CVD",
118                     labels = c("CVD", "No CVD")) +
119   labs(title = "CVD by center",
120        x = "Center", y = "Frequency") +
121   theme_classic()
122
123
124 ##### Spearman correlation matrix for the nutrients #####
125 #####
126
127 # Sample nutrients:
128 sample_nutr = samples[,18:61]
129 head(sample_nutr)
130 colnames(sample_nutr)
131 #iron, calcium, sodium, potassium, phosphorus, zinc, vit.b1, vit.b2, vit
    .b3, vit.c, vit.b6, vit.e, vit.d are in milligrams/day
132 sample_nutr[,c(23:33, 38, 39)] = sample_nutr[,c(23:33, 38, 39)]/1e3
133 #folic.acid, retinol.eq, retinol, beta.carotene are in micrograms
134 sample_nutr[, 34:37] = sample_nutr[, 34:37]/1e6
135 #remove energy
136 sample_nutr = sample_nutr[,-22]
137 sample_nutr = sample_nutr/samples$energy
138 #divide the nutrients by energy
139 row_energy = which(colnames(sample_nutr)=="energy")
140 sample_nutr_energy = sample_nutr[,-row_energy]/sample_nutr$energy #each
    individual (row) is divided by the energy of that individual
141 dim(sample_nutr_energy)
142

```

```

143
144
145 #compute correlation matrix
146 corr_matrix = matrix (0, dim(sample_nutr_energy)[2], dim(
      sample_nutr_energy)[2])
147 colnames(corr_matrix) = colnames(sample_nutr_energy)
148 rownames(corr_matrix) = colnames(sample_nutr_energy)
149
150 for (j in 1:dim(sample_nutr_energy)[2]) {
151   for (k in 1:dim(sample_nutr_energy)[2]) {
152     corr_matrix [j,k] = corr_matrix [j,k] + cor(sample_nutr_energy[,j],
      sample_nutr_energy[,k], method = "spearman", use = "complete.obs")
153   }
154 }
155
156 reorder_cormat <- function(cormat){
157   # Use correlation between variables as distance
158   dd <- as.dist((1-cormat)/2)
159   hc <- hclust(dd)
160   cormat <- cormat[hc$order, hc$order]
161 }
162
163 # Get lower triangle of the correlation matrix
164 get_lower_tri <- function(cormat){
165   cormat[upper.tri(cormat)] <- NA
166   return(cormat)
167 }
168 # Get upper triangle of the correlation matrix
169 get_upper_tri <- function(cormat){
170   cormat[lower.tri(cormat)] <- NA
171   return(cormat)
172 }
173
174 # Reorder the correlation matrix
175 cormat <- round(reorder_cormat(corr_matrix), 3)
176 upper_tri <- round(get_upper_tri(corr_matrix), 3)
177 # Melt the correlation matrix
178 library(reshape2)
179 melted_cormat <- melt(upper_tri, na.rm = TRUE)
180 # Create a ggheatmap
181 quartz()
182 ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value)) +
183   geom_tile(color = "white") +
184   scale_fill_gradient2(low = "blue", high = "red", mid = "white",

```

```

185         midpoint = 0, limit = c(-1,1), space = "Lab",
186         name="Spearman\nCorrelation") +
187 theme_minimal() + # minimal theme
188 theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
189         size = 8, hjust = 1), # modify axis.
190         text.x
191         axis.text.y = element_text(size = 8)) + # modify axis.text.y
192 labs(x = "", y = "") + # remove x and y axis labels
193 coord_fixed() +
194 ggtitle("Heatmap for the 43 nutrients, each divided by the total
195         energy for each individual") # add a title
196
197 ##### Spearman correlation matrix for the cpg sites #####
198 #####
199
200 # Set the maximum value 'n'
201 n <- dim(dnam)[1] # Replace with your desired maximum value
202
203 # Generate six random numbers between 1 and 'n'
204 random_numbers <- sample(1:n, 10, replace = TRUE)
205 #random_numbers = 1:8
206
207 dnam_random = t(dnam[random_numbers,])
208 dim(dnam_random)
209
210 #compute correlation matrix
211 corr_matrix = matrix(0, dim(dnam_random)[2], dim(dnam_random)[2])
212 colnames(corr_matrix) = colnames(dnam_random)
213 rownames(corr_matrix) = colnames(dnam_random)
214
215 for (j in 1:dim(dnam_random)[2]) {
216   for (k in 1:dim(dnam_random)[2]) {
217     corr_matrix [j,k] = corr_matrix [j,k] + cor(dnam_random[,j],
218         dnam_random[,k], method = "spearman", use = "complete.obs")
219   }
220 }
221
222 reorder_cormat <- function(cormat){
223   # Use correlation between variables as distance
224   dd <- as.dist((1-cormat)/2)
225   hc <- hclust(dd)
226   cormat <- cormat[hc$order, hc$order]

```

```

226 }
227
228 # Get lower triangle of the correlation matrix
229 get_lower_tri<-function(cormat){
230   cormat[upper.tri(cormat)] <- NA
231   return(cormat)
232 }
233 # Get upper triangle of the correlation matrix
234 get_upper_tri <- function(cormat){
235   cormat[lower.tri(cormat)]<- NA
236   return(cormat)
237 }
238
239 # Reorder the correlation matrix
240 cormat <- round(reorder_cormat(corr_matrix), 3)
241 upper_tri <- round(get_upper_tri(corr_matrix),3)
242 # Melt the correlation matrix
243 library(reshape2)
244 melted_cormat <- melt(upper_tri, na.rm = TRUE)
245 # Create a ggheatmap
246 #quartz()
247 ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value)) +
248   geom_tile(color = "white") +
249   geom_text(aes(label = sprintf("%.2f", value)), vjust = 0.5, hjust =
250     0.5, size = 3) +
251   scale_fill_gradient2(low = "blue", high = "red", mid = "white",
252     midpoint = 0, limit = c(-1,1), space = "Lab",
253     name="Spearman\nCorrelation") +
254   theme_minimal() + # minimal theme
255   theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
256     size = 8, hjust = 1), # modify axis.
257     text.x
258     axis.text.y = element_text(size = 8)) + # modify axis.text.y
259   labs(x = "", y = "") + # remove x and y axis labels
260   coord_fixed() +
261   ggtitle("Spearman correlation matrix of 10 random cpg sites") # add a
262   title

```

## B.2. MITM 1

In this section we report the R code regarding the first application of the MITM method, divided by steps as it is presented in the manuscript.

```

1 #Load dataset and create the appropriate sub-datasets

```

```

2 load("datasets/epic_italy_diet_noFlorence_noch_nocancerbefore_noXY.rda")
3 sample_char = samples[,1:17]
4 sample_nutr = samples[,18:61]
5 sample_food = samples[,62:128]
6 sample_outcome = samples[,129:136]
7
8 #Create a numerical variable CVD:
9 CVD = ifelse(!is.na(sample_outcome$date.cvd.dx),1,0)
10
11 #Create the other covariates:
12 sex = sample_char$sex
13 age = sample_char$age.recr
14 smoking = sample_char$smoking.status
15 center = sample_char$center
16 energy = samples$energy
17 rows_diabetes = union(which(sample_char$diabetes), which(!is.na(
18     sample_outcome$date.diabetes.dx)))
19 length(rows_diabetes) #104 = 39 + 69 - 4
20 diabetes = rep(0,1508)
21 diabetes[rows_diabetes] = 1
22 bmi = sample_char$bmi
23 chip.pos = sample_char$chip.pos
24 chip = as.factor(sample_char$chip)
25 #Macroelements + vitamins are in milligrams/day, folic acid .... beta
26     carotene are in micrograms/day
27 colnames(sample_nutr)
28 #iron, calcium, sodium, potassium, phosphorus, zinc, vit.b1, vit.b2, vit
29     .b3, vit.c, vit.b6, vit.e, vit.d are in milligrams/day
30 sample_nutr[,c(23:33, 38, 39)] = sample_nutr[,c(23:33, 38, 39)]/1e3
31 #folic.acid, retinol.eq, retinol, beta.carotene are in micrograms
32 sample_nutr[, 34:37] = sample_nutr[, 34:37]/1e6
33
34 #Remove energy (to be used as covariate)
35 sample_nutr = sample_nutr[, -22]
36 colnames(sample_nutr)
37
38 #function plots:
39 library(ggplot2)
40 library(ggrepel)
41
42 Volcano_plot = function(pval, log2FC, title, names, num_nutrients,
43     y_annotate, xlim_lower, xlim_upper) {
44     col = rep(3, num_nutrients) #grey

```

```

42 col[log2FC > log2(2) & pval > -log10(0.05)] = 0 #red
43 col[log2FC < -log2(2) & pval > -log10(0.05)] = 1 #blue
44 col[log2FC > -log2(2) & log2FC < log2(2) & pval < -log10(0.05)] = 2 #
   black
45 df = data.frame(x = log2FC, y = pval, col = col)
46 df$name = names
47 colors <- c("0" = "red", "1" = "blue", "2" = "black", "3" = "grey")
48 labels <- c("0" = "Up-regulated", "1" = "Down-regulated", "2" = "Not
   significant", "3" = "Inconclusive")
49 my_plot = ggplot(df, aes(x = x, y = y, color = as.factor(col))) +
50   geom_point() +
51   scale_color_manual(values = colors, labels = labels, drop = FALSE) +
52   geom_hline(yintercept = -log10(0.05), linetype = "dashed") +
53   geom_vline(xintercept = -log2(2), linetype = "dashed") +
54   geom_vline(xintercept = log2(2), linetype = "dashed") +
55   geom_vline(xintercept = 0, linetype = "dashed", col = "grey", size
   = 0.25) +
56   geom_label_repel(aes(label = name),
57     box.padding = 0.5,
58     #data = subset(df, col == 0 | col == 1)
59     data = df, size = 4, show.legend = FALSE, max.
   overlaps = 100) +
60   labs(x = "log2(fold-change)", y = "-log10(pval)", color = "
   Significance", title = title) +
61   annotate("text", x=xlim_lower, y=y_annotate, label= "p = 0.05", col=
   "black", size=3) +
62   theme_classic() +
63   xlim(xlim_lower, xlim_upper)
64 my_plot$guides$color$title <- NULL
65 return (my_plot)
66 }
67
68 inflation_plot <- function(ps, ci = 0.95) {
69   n <- length(ps)
70   df <- data.frame(
71     observed = -log10(sort(ps)),
72     expected = -log10(ppoints(n)),
73     clower = -log10(qbeta(p = (1 - ci) / 2, shape1 = 1:n, shape2 = n
   :1)),
74     cupper = -log10(qbeta(p = (1 + ci) / 2, shape1 = 1:n, shape2 = n
   :1))
75   )
76   ggplot(df) +

```

```

77   geom_ribbon(mapping = aes(x = expected, ymin = clower, ymax = cupper
78   ), alpha = 0.1) +
79   geom_point(aes(expected, observed), size = 1) +
80   geom_abline(intercept = 0, slope = 1, alpha = 0.5) +
81   labs(x = "Observed -log10(pvalue)", y = "Expected -log10(pvalue)",
82   title = "QQplot of observed vs expected -log10(pvalue)")
83 }
84
85 inflation <- function(pval) {
86   chisq <- qchisq(1 - pval, 1)
87   lambda <- median(chisq) / qchisq(0.5, 1)
88   lambda
89   # lambda is defined as the median of the resulting chi-squared test
90   # statistics divided
91   #by the expected median of the chi-squared distribution
92   #The genomic inflation factor expresses the deviation of the
93   #distribution
94   #of the observed test statistic compared to the distribution of the
95   #expected test statistic.
96 }
97
98 inflation_bacon = function(zscore) {
99   library(bacon)
100   bc <- bacon(zscore)
101   estimates(bc)[,"sigma.0"]
102 }
103
104 ##### a. Dimensionality reduction #####
105 #####
106 #Select the restricted set of cpg sites (1646)
107
108 ##### b. Relation between each nutrient and the restricted methylome ##
109 #####
110 #Cpg_i ~ nutr_j + CVD + covariates, i = 1, ,1646 and j = 1, ,43:
111 pvalue_func_dataframe = function (sample_nutr, dnam) {
112   outcome_df = data.frame(matrix(0, nrow = dim(sample_nutr)[2]*dim(dnam)
113   [1], ncol = 4))
114   colnames(outcome_df) = c("pvalue", "beta", "log2FC", "zscore")
115   for (i in 1:dim(dnam)[1]) { #i = 1,...1646

```

```

115   for (j in 1:dim(sample_nutr)[2]) { #j = 1,...,43
116     fit <- lm(dnam[i,]~ sample_nutr[,j] + CVD + sex + age + smoking +
center + bmi + diabetes + energy + chip + chip.pos)
117     s = summary(fit)
118     outcome_df[(i-1)*dim(sample_nutr)[2]+j,1] = s$coefficients[2,4] #
pvalue
119     outcome_df[(i-1)*dim(sample_nutr)[2]+j,2] = s$coefficients[2,1] #
beta
120     fold_change = (coefficients(fit)[2] + coefficients(fit)[1])/
coefficients(fit)[1]
121     outcome_df[(i-1)*dim(sample_nutr)[2]+j,3] = sign(coefficients(fit)
[2]) * abs(log2(abs(fold_change)))
122     outcome_df[(i-1)*dim(sample_nutr)[2]+j,4] = s$coefficients[2,1]/
s$coefficients[2,2]
123     rownames(outcome_df)[(i-1)*dim(sample_nutr)[2]+j] = paste0(i,": ",
rownames(dnam)[i], " - ", colnames(sample_nutr)[j])
124   }
125 }
126 return (outcome_df)
127 }
128
129 outcome_df = pvalue_func_dataframe(sample_nutr, dnam)
130 dim(outcome_df) # 70778, 3 (1646 cpgs*43 nutr)
131 outcome_df$zscore = abs(outcome_df$beta*qnorm(outcome_df$pvalue/2))
132
133 #How many nutrients significant after BH correction:
134 outcome_df$pvalue_BH_corrected = p.adjust(outcome_df$pvalue, method = "
BH")
135 outcome_df$significant = outcome_df[which(outcome_df$pvalue_BH_corrected
<0.05),]
136 #10 significant cpg-nutr
137 #8 significant nutr for at least one cpg (REDUCED EXPOSOME): potassium,
avail.carbs (x2) , alcohol, iron, glycemic.load (x2), trap, frap, vit
.d
138
139 #for IRON, test with and without quantiles:
140 sample_nutr_temp = sample_nutr
141 sample_nutr_temp[, "iron"] = cut(sample_nutr_temp[, "iron"], quantile(
sample_nutr_temp[, "iron"], probs = seq(0, 1, 0.2)))
142
143 #with quantiles, cg06690548:
144 fit <- lm(dnam["cg06690548",] ~ sample_nutr_temp[, "iron"] + CVD + sex +
age + smoking + center + energy + diabetes + bmi + chip + chip.pos)
145 summary(fit)$coef[5,]

```

```

146
147 #without quantiles, cg06690548:
148 fit <- lm(dnam["cg06690548",] ~ sample_nutr["iron"] + CVD + sex + age
149       + smoking + center + energy + diabetes + bmi + chip + chip.pos)
150 summary(fit)$coef[2,]
151
152 # 1.VOLCANO PLOT OF 8 SIGNIFICANT NUTRIENTS IN REDUCED EXPOSOME:
153
154 my_plot = Volcano_plot(-log10(outcome_df_significant$pvalue),
155       outcome_df_significant$log2FC,
156       "Volcano plot representing the significant associations between the CpG
157       sites within the Restricted Methylome
158       and the nutrients within the Whole Exposome",
159       sub("^\\d+:\\s", "", rownames(outcome_df_significant)), 8, 1.7)
160
161 # 2.INFLATION PLOT with Bacon:
162 zscore <- outcome_df$zscore
163 my_plot = inflation_plot(outcome_df$pvalue) +
164       theme_bw() +
165       theme(axis.ticks = element_line(size = 0.5), panel.grid =
166       element_blank()
167       )
168 my_plot
169
170 ##### c. Relation between the reduced exposome and CVD #####
171 #####
172 #CVD ~ nutr_j + covariates, j = 1, 8
173 sample_nutr2 = sample_nutr
174 names_nutr = colnames(sample_nutr)
175
176 #divide distributions of nutrients into quantiles
177 f = function(name_nutr){
178   sample_nutr2[,name_nutr] = cut(sample_nutr2[,name_nutr],quantile(
179     sample_nutr2[,name_nutr],probs = seq(0, 1, 0.2)), include.lowest =
180     TRUE)
181   fit <- glm(CVD ~ sample_nutr2[,name_nutr] + sex + age + smoking +
182     center + energy + diabetes + bmi, family = binomial)
183   out = summary(fit)$coeff[5,]
184   fold_change = (coefficients(fit)[1] + coefficients(fit)[5]) /
185     coefficients(fit)[1]
186   log2fc = sign(coefficients(fit)[5]) * abs(log2(abs(fold_change)))

```

```

182   out = c(out, log2fc)
183   return(out)}
184
185 result = data.frame(do.call("rbind",lapply(names_nutr,f)))
186 colnames(result) = c("beta", "std.error", "z-value", "pvalue", "log2fc")
187 rownames(result) = names_nutr
188 result_ordered <- result[order(result$pvalue), ]
189
190 #CVD ~ nutr_j + covariates, j = 1, 8
191 #select only the 8 nutrients in the reduced exposome
192 significant_nutr = c("potassium", "avail.carbs", "alcohol", "iron", "
    glyceimic.load", "trap", "frap", "vit.d")
193 result_significant = result[significant_nutr,]
194 result_significant$pvalue_BH_corrected = p.adjust(
    result_significant$pvalue, method = "BH")
195 #only significant: Iron, with 5.2%
196
197
198 #1. VOLCANO PLOT OF 8 SIGNIFICANT NUTRIENTS IN REDUCED EXPOSOME:
199 my_plot = Volcano_plot(-log10(result_significant$pvalue_BH_corrected),
    result_significant$log2fc,
200     "Volcano plot representing the significant
    associations between the nutrients in the reduced exposome
201 and the CVD as outcome", rownames(result_significant), 8, 1.37, -1.5,
    1.5)
202 my_plot
203
204 #2. DOSE-RESPONDE RELATIONSHIP:
205
206 f2 = function(name_nutr){
207   sample_nutr2[,name_nutr] = cut(sample_nutr2[,name_nutr],quantile(
    sample_nutr2[,name_nutr],probs = seq(0, 1, 0.2)), include.lowest =
    TRUE)
208   fit <- glm(CVD ~ sample_nutr2[,name_nutr] + sex + age + smoking +
    center + energy + diabetes + bmi, family = binomial)
209   out = summary(fit)$coeff[2:5,]
210   rownames(out) = c("2nd quantile bin", "3rd quantile bin", "4th
    quantile bin", "5th quantile bin")
211   return(out)}
212
213 #potassium:
214 data <- data.frame(x = seq(1, 4), y = exp(f2("potassium")[,1]))
215 names <- c("2nd", "3rd", "4th", "5th")
216 dr_potassium = ggplot(data, aes(x = x, y = y)) +

```

```
217 geom_point() +
218 geom_line() +
219 theme_minimal() +
220 labs(x = "Quantile bins", y = "Exp(estimate)", title = "Potassium") +
221 scale_x_continuous(breaks = data$x, labels = names)
222
223 #iron:
224 data <- data.frame(x = seq(1, 4), y = exp(f2("iron")[,1]))
225 names <- c("2nd", "3rd", "4th", "5th")
226 dr_iron = ggplot(data, aes(x = x, y = y)) +
227     geom_point() +
228     geom_line() +
229     theme_minimal() +
230     labs(x = "Quantile bins", y = "Exp(estimate)", title = "Iron")
231     +
232     scale_x_continuous(breaks = data$x, labels = names)
233
234 #avail.carbs:
235 data <- data.frame(x = seq(1, 4), y = exp(f2("avail.carbs")[,1]))
236 names <- c("2nd", "3rd", "4th", "5th")
237 dr_avail.carbs = ggplot(data, aes(x = x, y = y)) +
238     geom_point() +
239     geom_line() +
240     theme_minimal() +
241     labs(x = "Quantile bins", y = "Exp(estimate)", title = "Available
242         carbohydrates") +
243     scale_x_continuous(breaks = data$x, labels = names)
244
245 #glycemic.load:
246 data <- data.frame(x = seq(1, 4), y = exp(f2("glycemic.load")[,1]))
247 names <- c("2nd", "3rd", "4th", "5th")
248 dr_glycemic.load = ggplot(data, aes(x = x, y = y)) +
249     geom_point() +
250     geom_line() +
251     theme_minimal() +
252     labs(x = "Quantile bins", y = "Exp(estimate)", title = "Glycemic load")
253     +
254     scale_x_continuous(breaks = data$x, labels = names)
255
256 #trap:
257 data <- data.frame(x = seq(1, 4), y = exp(f2("trap")[,1]))
258 names <- c("2nd", "3rd", "4th", "5th")
259 dr_trap = ggplot(data, aes(x = x, y = y)) +
260     geom_point() +
```

```

258 geom_line() +
259 theme_minimal() +
260 labs(x = "Quantile bins", y = "Exp(estimate)", title = "Trap") +
261 scale_x_continuous(breaks = data$x, labels = names)
262
263 #frap:
264 data <- data.frame(x = seq(1, 4), y = exp(f2("frap")[,1]))
265 names <- c("2nd", "3rd", "4th", "5th")
266 dr_frap = ggplot(data, aes(x = x, y = y)) +
267   geom_point() +
268   geom_line() +
269   theme_minimal() +
270   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Frap") +
271   scale_x_continuous(breaks = data$x, labels = names)
272
273
274 #vit.d:
275 data <- data.frame(x = seq(1, 4), y = exp(f2("vit.d")[,1]))
276 names <- c("2nd", "3rd", "4th", "5th")
277 dr_vit.d = ggplot(data, aes(x = x, y = y)) +
278   geom_point() +
279   geom_line() +
280   theme_minimal() +
281   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Vitamin D") +
282   scale_x_continuous(breaks = data$x, labels = names)
283
284 #alcohol:
285 data <- data.frame(x = seq(1, 4), y = exp(f2("alcohol")[,1]))
286 names <- c("2nd", "3rd", "4th", "5th")
287 dr_alcohol= ggplot(data, aes(x = x, y = y)) +
288   geom_point() +
289   geom_line() +
290   theme_minimal() +
291   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Alcohol") +
292   scale_x_continuous(breaks = data$x, labels = names)

```

### B.3. MITM 2

In this section we report the R code regarding the second application of the MITM method, divided by steps as it is presented in the manuscript.

```

1 #same initial setting as MITM1
2
3 Manhattan_plot = function(title, pval) {

```

```

4  chr <- as.integer(gsub('chr', '', annot$chr))
5  pos <- as.integer(annot$pos)
6  data <- data.frame(chr, pos, pval)
7  sig_threshold <- -log10(0.05)
8  set.seed(1)
9  chr_jitter = chr + runif(length(chr), min = -0.45, max = 0.45)
10 df = data.frame(x = chr_jitter, y = -log10(pval), name = rownames(dnam
    ))
11 my_plot = ggplot(df, aes(x = chr_jitter, y = y, color = factor(chr %%
    2))) +
12   geom_point(size = 0.8) +
13   scale_x_continuous(breaks = seq(1, 22, by = 1)) +
14   scale_color_manual(values = c("grey", "black")) +
15   geom_hline(yintercept = sig_threshold, linetype = "dashed", color =
    "red") +
16   geom_vline(xintercept = seq(0.5, 22.5, by=1), color = "grey", size =
    0.1) +
17   labs(title = title,
18         x = "Chromosome", y = "-log10(P-value)") +
19   guides(color = "none") +
20   theme_classic() +
21   theme_linedraw(base_line_size = 0) +
22   annotate("text", x=-1.3, y=1.5, label= "p = 0.05", col="red", size
    =3)
23 }
24
25
26 ##### a. Relation between nutrients and CVD #####
27 #####
28
29
30 #The best way to find robust, reproducible results is to compare
31 #the extremes of the distribution of each nutr (regarding association
    with outcome)
32
33 #find nutrients belonging to fifth quantile associated with outcome CVD:
34 sample_nutr2 = sample_nutr
35 names_nutr = colnames(sample_nutr)
36
37 #divide distributions of nutrients into quantiles
38 f = function(name_nutr){
39   sample_nutr2[,name_nutr] = cut(sample_nutr2[,name_nutr],quantile(
    sample_nutr2[,name_nutr],probs = seq(0, 1, 0.2)), include.lowest =
    TRUE)

```

```

40 fit <- glm(CVD ~ sample_nutr2[,name_nutr] + sex + age + smoking +
  center + energy + diabetes + bmi, family = binomial)
41 out = summary(fit)$coeff[5,]
42 fold_change = (coefficients(fit)[1] + coefficients(fit)[5]) /
  coefficients(fit)[1]
43 log2fc = sign(coefficients(fit)[5]) * abs(log2(abs(fold_change)))
44 out = c(out, log2fc)
45 return(out)}
46
47 result = data.frame(do.call("rbind",lapply(names_nutr,f)))
48 rownames(result) = names_nutr
49 colnames(result) = c("est","se","t","p", "log2fc")
50 result$zscore = result$est/result$se
51
52 result_ordered = result[order(result$p),]
53 result_ordered = round(result_ordered,4)
54 result_ordered
55 #sig: folic.acid, iron, water, edible.portion and vit.b6 have p<0.05
56 significant_nutr = c("folic.acid", "iron", "water", "edible.portion", "
  vit.b6")
57
58 # 1. VOLCANO PLOT
59 #Volcano_plot = function(pval, log2FC, title, names, num_nutrients,
  y_annotate, xlim_lower, xlim_upper)
60 my_plot = Volcano_plot(-log10(result$p), result$log2fc,
61 "Volcano plot representing the pvalues of associations between the 5th
  quintile bin of 43 nutrients,
62 and the CVD considered as outcome", rownames(result), 43, 1.37, -1.5,
  1.5)
63 my_plot
64
65 # 2. INFLATION PLOT with Bacon:
66 zscore <- result$zscore
67 my_plot = inflation_plot(result$p) +
68   theme_bw() +
69   theme( axis.ticks = element_line(size = 0.5), panel.grid =
  element_blank()
70         # panel.grid = element_line(size = 0.5, color = "grey80")
71       )
72 my_plot
73 inflation_bacon(zscore)
74
75 # 3. DOSE-RESPONSE RELATIONSHIP for the significant nutrients
76

```

```
77 f2 = function(name_nutr){
78   sample_nutr2[,name_nutr] = cut(sample_nutr2[,name_nutr],quantile(
      sample_nutr2[,name_nutr],probs = seq(0, 1, 0.2)), include.lowest =
      TRUE)
79   fit <- glm(CVD ~ sample_nutr2[,name_nutr] + sex + age + smoking +
      center + energy + diabetes + bmi, family = binomial)
80   out = summary(fit)$coeff[2:5,]
81   rownames(out) = c("2nd quantile bin", "3rd quantile bin", "4th
      quantile bin", "5th quantile bin")
82   return(out)}
83
84
85 #folic.acid:
86 data <- data.frame(x = seq(1, 4), y = exp(f2("folic.acid")[,1]))
87 names <- c("2nd", "3rd", "4th", "5th")
88 dr_folic.acid = ggplot(data, aes(x = x, y = y)) +
89   geom_point() +
90   geom_line() +
91   theme_minimal() +
92   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Folic acid") +
93   scale_x_continuous(breaks = data$x, labels = names)
94
95 #iron:
96 data <- data.frame(x = seq(1, 4), y = exp(f2("iron")[,1]))
97 names <- c("2nd", "3rd", "4th", "5th")
98 dr_iron = ggplot(data, aes(x = x, y = y)) +
99   geom_point() +
100  geom_line() +
101  theme_minimal() +
102  labs(x = "Quantile bins", y = "Exp(estimate)", title = "Iron") +
103  scale_x_continuous(breaks = data$x, labels = names)
104
105 #water:
106 data <- data.frame(x = seq(1, 4), y = exp(f2("water")[,1]))
107 names <- c("2nd", "3rd", "4th", "5th")
108 dr_water = ggplot(data, aes(x = x, y = y)) +
109   geom_point() +
110   geom_line() +
111   theme_minimal() +
112   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Water") +
113   scale_x_continuous(breaks = data$x, labels = names)
114
115 #edible.portion:
116 data <- data.frame(x = seq(1, 4), y = exp(f2("edible.portion")[,1]))
```

```

117 names <- c("2nd", "3rd", "4th", "5th")
118 dr_edible.portion = ggplot(data, aes(x = x, y = y)) +
119   geom_point() +
120   geom_line() +
121   theme_minimal() +
122   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Edible portion"
123     ) +
124   scale_x_continuous(breaks = data$x, labels = names)
125
126 #vit.b6:
127 data <- data.frame(x = seq(1, 4), y = exp(f2("vit.b6")[,1]))
128 names <- c("2nd", "3rd", "4th", "5th")
129 dr_vit.b6 = ggplot(data, aes(x = x, y = y)) +
130   geom_point() +
131   geom_line() +
132   theme_minimal() +
133   labs(x = "Quantile bins", y = "Exp(estimate)", title = "Vitamin B6") +
134   scale_x_continuous(breaks = data$x, labels = names)
135
136 ##### b. Dimensionality reduction #####
137 #####
138
139 #Select the restricted set of cpg sites (1646)
140
141
142 ##### c. Relation between restricted methylome and CVD #####
143 #####
144
145 #Select the restricted set of cpg sites (1646)
146
147 all_func = function (row_cpg) {
148   fit <- glm(CVD ~ row_cpg + sex + age + smoking + center + diabetes +
149     bmi + energy + chip + chip.pos, family = binomial)
150   summary(fit)
151   pvalue = (summary (fit)$coefficients [2,4])
152   beta = (summary (fit)$coefficients [2,1])
153   z_score <- (summary (fit)$coefficients [2,1]) / (summary (fit)
154     $coefficients [2,2])
155   return( c(pvalue, z_score, beta))
156 }
157
158 df = apply(dnam, 1, all_func)
159 significant_cpg = rownames(df)[which(df$pvalue < 0.05)]

```

```

158 length(significant_cpg) #143
159
160 # 1. MANHATTAN PLOT
161 my_plot = Manhattan_plot("Manhattan plot of p-values of associations
162   between the 1646 CpG sites belonging to the Restricted Methylome
163   and CVD considered as outcome", df$pvalue)
164 my_plot
165
166 # 2. INFLATION PLOT
167 zscore <- df$zscore
168 my_plot = inflation_plot(df$pvalue) +
169   theme_bw() +
170   theme( axis.ticks = element_line(size = 0.5), panel.grid =
171     element_blank()
172   )
173 my_plot
174 inflation_bacon(zscore)
175
176 ##### d.Relation between each nutrient of the reduced exposome (5) #####
177 ##### and the reduced methylome (143) #####
178 #####
179 significant_cpg
180 length(significant_cpg)
181 significant_nutr = c("folic.acid", "iron", "water", "edible.portion", "
182   vit.b6")
183 significant_nutr
184 length(significant_nutr)
185
186 #Cpg_i ~ nutr_j + CVD + covariates, i = 1, ...,143 and j = 1, ...,5
187 sample_nutr_restricted = sample_nutr[,significant_nutr]
188 dnam_restrcited = dnam[significant_cpg,]
189
190 pvalue_func_dataframe = function (sample_nutr_restricted,
191   dnam_restricted) {
192   outcome_df = data.frame(matrix(0, nrow = dim(sample_nutr_restricted)
193     [2]*dim(dnam_restricted)[1], ncol = 3))
194   colnames(outcome_df) = c("pvalue", "beta", "log2FC")
195   for (i in 1:dim(dnam_restricted)[1]) { #i = 1,...1646
196     for (j in 1:dim(sample_nutr_restricted)[2]) { #j = 1,...,43
197       fit <- lm(dnam_restricted[i,]~ sample_nutr_restricted[,j] + CVD +
198         sex + age + smoking + center + bmi + diabetes + energy + chip + chip.
199         pos)

```

```

195     s = summary(fit)
196     outcome_df[(i-1)*dim(sample_nutr_restricted)[2]+j,1] =
s$coefficients[2,4] #pvalue
197     outcome_df[(i-1)*dim(sample_nutr_restricted)[2]+j,2] =
s$coefficients[2,1] #beta
198     fold_change = (coefficients(fit)[2] + coefficients(fit)[1])/
coefficients(fit)[1]
199     outcome_df[(i-1)*dim(sample_nutr_restricted)[2]+j,3] = sign(
coefficients(fit)[2]) * abs(log2(abs(fold_change)))
200     rownames(outcome_df)[(i-1)*dim(sample_nutr_restricted)[2]+j] =
paste0(i,": ", rownames(dnam_restrcited)[i], " - ", colnames(
sample_nutr_restricted)[j])
201   }
202 }
203 return (outcome_df)
204 }
205
206 outcome_df_restricted = pvalue_func_dataframe(sample_nutr_restricted,
dnam_restrcited)
207 dim(outcome_df_restricted) # 715, 3 (143 cpgs*5 nutr)
208
209
210 #For each nutrient, find the cpgs to which it is associated:
211 #Table number of sig cpgs per nutrient
212 sum(outcome_df_restricted$pvalue<0.05) #42 significant:
213 significant_nutrients_df = outcome_df_restricted[which(
outcome_df_restricted$pvalue<0.05),]
214 all_nutr <- sapply(strsplit(rownames(significant_nutrients_df), " "),
tail, n = 1)
215 length(all_nutr) #42 ok
216 unique_nutr = unique(all_nutr)
217 length(unique_nutr) #5, ok all 5 have at least one significant CpG sites
218 t = data.frame(table(all_nutr)) #number of significant CpG sites for
each nutrient
219
220 #11 for folic.acid
221 #10 for iron
222 #7 for water
223 #7 for edible.portion
224 #7 for vit.b6
225
226
227 #significant cpgs folic.acid:
228 rows_folic.acid = which(all_nutr == "folic.acid")

```

```

229 df_folic.acid = significant_nutrients_df[rows_folic.acid,]
230 rownames(df_folic.acid) <- gsub("^\\d+: (\\s+)?(\\w+)\\s+-.*$", "\\2",
    rownames(df_folic.acid))
231
232 #significant cpgs iron:
233 rows_iron = which(all_nutr == "iron")
234 df_iron = significant_nutrients_df[rows_iron,]
235 rownames(df_iron) <- gsub("^\\d+: (\\s+)?(\\w+)\\s+-.*$", "\\2", rownames
    (df_iron))
236
237 #significant cpgs water:
238 rows_water = which(all_nutr == "water")
239 df_water = significant_nutrients_df[rows_water,]
240 rownames(df_water) <- gsub("^\\d+: (\\s+)?(\\w+)\\s+-.*$", "\\2",
    rownames(df_water))
241
242 #significant cpgs edible.portion:
243 rows_edible.portion = which(all_nutr == "edible.portion")
244 df_edible.portion = significant_nutrients_df[rows_edible.portion,]
245 rownames(df_edible.portion) <- gsub("^\\d+: (\\s+)?(\\w+)\\s+-.*$", "\\2"
    , rownames(df_edible.portion))
246
247 #significant cpgs vit.b6:
248 rows_vit.b6 = which(all_nutr == "vit.b6")
249 df_vit.b6 = significant_nutrients_df[rows_vit.b6,]
250 rownames(df_vit.b6) <- gsub("^\\d+: (\\s+)?(\\w+)\\s+-.*$", "\\2",
    rownames(df_vit.b6))
251
252 #list of 5 dataframe, one for each nutr. In each dataframe we have, for
    each sig CpG sites pvalue, beta and log2FC
253 list_nutr = list(df_folic.acid, df_iron, df_water, df_edible.portion,
    df_vit.b6)
254 names(list_nutr) = c("folic.acid", "iron", "water", "edible.portion", "
    vit.b6")
255 saveRDS(list_nutr, file = "tables/final/list_pvalues_cpg_per_nutr.rds")
256
257 #cpgs per nutr:
258 cpg_folic.acid = t(dnam[rownames(list_nutr$folic.acid),])
259 cpgs_iron = t(dnam[rownames(list_nutr$iron),])
260 cpg_water = t(dnam[rownames(list_nutr$water),])
261 cpg_edible.portion = t(dnam[rownames(list_nutr$edible.portion),])
262 cpg_vit.b6 = t(dnam[rownames(list_nutr$vit.b6),])
263
264 #for each nutrient we have the dnam value of each CpG site significant.

```

```

265 list_cpgs = list(cpg_folic.acid, cpgs_iron, cpg_water, cpg_edible.
    portion, cpg_vit.b6)
266 names(list_cpgs) = c("folic.acid", "iron", "water", "edible.portion", "
    vit.b6")
267 saveRDS(list_cpgs, file = "tables/final/list_cpgs_per_nutr.rds")
268
269
270 ##### e. Mediation for each nutrient #####
271 #####
272
273 library(lavaan)
274
275 dummy_smoking <- model.matrix(~ . - 1, data = smoking)
276 colnames(dummy_smoking) = c("smoking_n", "smoking_f", "smoking_c")
277 dummy_center = model.matrix(~ . - 1, data = center)
278 colnames(dummy_center) = c("center_V", "center_R", "center_T", "center_N
    ")
279
280 ## FOLIC ACID:
281 folic.acid = cut(samples$folic.acid,
282     quantile(samples$folic.acid, probs = seq(0, 1, 0.2)),
283     include.lowest=TRUE)
284 folic.acid = as.numeric (folic.acid)
285 data.folic_acid = data.frame(cvd = CVD, folic.acid, age.recr, sex,
286     energy, bmi, diabetes,
287     dummy_smoking, dummy_center,
288     list_cpgs$folic.acid)
289
290 colnames(list_cpgs$folic.acid)
291 model <- '
292 # latent variable
293 dnam =~ cg11088672 + cg05575921 + cg12065531 +
294 cg01353448 + cg27510066 + cg00247963 +
295 cg24613083 + cg03636183 + cg09958192 +
296 cg19866478 + cg27585074
297
298 # outcome model
299 cvd ~ c*folic.acid + c1*age.recr + c2*sex + c3*energy + c4*bmi + c5*
300 diabetes +
301 c6*smoking_c + c7*smoking_f + c8*center_V+ c9*center_R + c10*center_T +
302 m1*dnam
303
304 # mediator models
305 dnam ~ a1*folic.acid

```

```

302
303 # indirect effects (IE)
304 dnam_IE := a1*m1
305
306 sumIE := (a1*m1)
307
308 # total effect
309 total := c + (a1*m1)
310 '
311 fit = sem(model, data=data.folic_acid)
312 summary(fit)
313
314
315 ## IRON:
316 iron = cut(samples$iron,
317           quantile(samples$iron, probs = seq(0, 1, 0.2)), include.lowest=
           TRUE)
318 iron = as.numeric (iron)
319 data.iron= data.frame(cvd = CVD, iron, age.recr, sex, energy, bmi,
           diabetes,
320                   dummy_smoking, dummy_center, list_cpgs$iron)
321 colnames(list_cpgs$iron)
322
323 model <- '
324 # latent variable
325 dnam =~ cg24293507 + cg01691332 + cg01353448 +
326 cg26929272 + cg27466845 + cg19641804 +
327 cg00247963 + cg01940273 + cg20512303 +
328 cg18181703
329
330 # outcome model
331 cvd ~ c*iron + c1*age.recr + c2*sex + c3*energy + c4*bmi + c5*diabetes +
332 c6*smoking_c + c7*smoking_f + c8*center_V + c9*center_R + c10*center_T +
333 m1*dnam
334
335 # mediator models
336 dnam ~ a1*iron
337
338 # indirect effects (IE)
339 dnam_IE := a1*m1
340
341 sumIE := (a1*m1)
342
343 # total effect

```

```

344 total := c + (a1*m1)
345 '
346 fit = sem(model, data=data.iron)
347 summary(fit)
348 #elimintae self loops (error variances) from the visualizations
349 g = get_edges(fit)[c(1:23, 37:41),]
350 #customize layout
351 l2 = get_layout(fit)
352 l3 = matrix(data =NA, nrow = 4, ncol = 17)
353 l3[4,1:10] = l2[3, 1:10]
354 l3[3,5] = "dnam"
355 l3[3,17] = "cvd"
356 conf1 = c("age.recr", "sex", "energy", "bmi","diabetes")
357 conf2 = c("smoking_c", "smoking_f", "center_V", "center_R", "center_T")
358 l3[1,7:11] = conf1
359 l3[1,12:16] = conf2
360 l3[2,9] = "iron"
361 #plot graph of relations
362 library(tidySEM)
363 quartz()
364 graph_sem(fit, edges = g, layout = l3, direction = "down", text_size =
      1.95,
365           rect_width = 1.8, rect_height = 0.1,
366           ellipses_width = 2.5, ellipses_height = 0.3, variance_diameter
      = 0, angle = 180, spacing_y = 0.4,
367           spacing_x = 2)
368
369
370 ## WATER
371 water = cut(samples$water,
372             quantile(samples$water,probs = seq(0, 1, 0.2)),include.
      lowest=TRUE)
373 water = as.numeric (water)
374 data.water= data.frame(cvd = CVD, water, age.recr, sex, energy, bmi,
      diabetes,
375                       dummy_smoking, dummy_center, list_cpgs$water)
376 colnames(list_cpgs$water)
377 cor(list_cpgs$water, use = "complete.obs")
378 #I removed the first CpG sites cg03019000 whcih has correlation of 0.32
      with second and 0.36 with third
379
380 model <- '
381
382 # latent variable

```

```

383 dnam =~ cg12184221 + cg22972055 +
384 cg27631256 + cg07291563 + cg14859204 + cg09678939
385
386 # outcome model
387 cvd ~ c*water + c1*age.recr + c2*sex + c3*energy + c4*bmi + c5*diabetes
      +
388 c6*smoking_c + c7*smoking_f + c8*center_V + c9*center_R + c10*center_T +
389 m1*dnam
390
391 # mediator models
392 dnam ~ a1*water
393
394 # indirect effects (IE)
395 dnam_IE := a1*m1
396
397 sumIE := (a1*m1)
398
399 # total effect
400 total := c + (a1*m1)
401 '
402
403 fit = sem(model, data=data.water)
404 summary(fit)
405
406
407 ## EDIBLE PORTION
408 edible.portion = cut(samples$edible.portion,
409                       quantile(samples$edible.portion,probs = seq(0, 1,
410                               0.2)),include.lowest=TRUE)
411 edible.portion = as.numeric (edible.portion)
412 data.edible.portion= data.frame(cvd = CVD, edible.portion, age.recr, sex
413                               , energy, bmi, diabetes,
414                               dummy_smoking, dummy_center,
415                               list_cpgs$edible.portion)
416 colnames(list_cpgs$edible.portion)
417 cor(list_cpgs$edible.portion, use = "complete.obs")
418
419 #removed first CpG site: cg07191189
420
421 model <- '
422 # latent variable
423 dnam =~ cg03019000 + cg12184221 +
424 cg22972055 + cg27631256 + cg14859204 + cg09678939

```

```

423
424 # outcome model
425 cvd ~ c*edible.portion + c1*age.recr + c2*sex + c3*energy + c4*bmi + c5*
      diabetes +
426 c6*smoking_c + c7*smoking_f + c8*center_V + c9*center_R + c10*center_T +
427 m1*dnam
428
429 # mediator models
430 dnam ~ a1*edible.portion
431
432 # indirect effects (IE)
433 dnam_IE := a1*m1
434
435 sumIE := (a1*m1)
436
437 # total effect
438 total := c + (a1*m1)
439 '
440
441 fit = sem(model, data=data.edible.portion)
442 summary(fit)
443
444
445 ## VIT.B6
446 vit.b6 = cut(samples$vit.b6,
447             quantile(samples$vit.b6,probs = seq(0, 1, 0.2)),include.
             lowest=TRUE)
448 vit.b6 = as.numeric (vit.b6)
449 data.vit.b6= data.frame(cvd = CVD, vit.b6, age.recr, sex, energy, bmi,
             diabetes,
450                       dummy_smoking, dummy_center, list_cpgs$vit.b6)
451 colnames(list_cpgs$vit.b6)
452
453 model <- '
454
455 # latent variable
456 dnam =~ cg01691332 + cg05575921 + cg01353448 +
457 cg26929272 + cg12985418 + cg03636183 + cg11762703
458
459 # outcome model
460 cvd ~ c*vit.b6 + c1*age.recr + c2*sex + c3*energy + c4*bmi + c5*diabetes
      +
461 c6*smoking_current + c7*smoking_former + c8*center_Varese + c9*
      center_Ragusa + c10*center_Turin +

```

```

462 m1*dnam
463
464 # mediator models
465 dnam ~ a1*vit.b6
466
467 # indirect effects (IE)
468 dnam_IE := a1*m1
469
470 sumIE := (a1*m1)
471
472 # total effect
473 total := c + (a1*m1)
474 '
475
476 fit = sem(model, data=data.vit.b6)
477 summary(fit)

```

## B.4. Stability selection

In this section we report the code regarding the implementation of the stability selection algorithm.

```

1 #same initial setting as MITM1 and MITM2
2 library(glmnet)
3 stability_selection_lasso <- function(x, y, z, prop = 0.5, n_iter =
  1000) {
4   #Initialization of needed variables
5   n <- nrow(x)
6   p <- ncol(x)
7   count_variables_selected = rep(0,54) #how many times each variable is
  selected
8   betas = data.frame(matrix(ncol = 0, nrow = 54)) #beta for each
  variable at each iteration (for best_lambda)
9   lambda.grid <- seq(0.00001,0.1,length=100) #to use for CV
10  for (iter in 1:n_iter) {
11    sample_idx <- sample(1:n, floor(0.5 * n)) # Randomly select 50% of
  the data
12    x_sample <- x[sample_idx, , drop = FALSE]
13    y_sample <- y[sample_idx]
14    z_sample <- z[sample_idx, , drop = FALSE]
15    #Remove lines with NAs (there are 2 in z_sample)
16    rows_no_na = which(complete.cases(z_sample))
17    x_sample = x_sample[rows_no_na,]
18    y_sample = y_sample[rows_no_na]

```

```

19   z_sample = z_sample[rows_no_na,]
20   #Create necessary structures
21   d = data.frame(y_sample,x_sample,z_sample)
22   x_new <- model.matrix(y_sample ~ .,data = d)[,-1] #Build the matrix
      of predictors
23   y_new = as.numeric(y_sample)
24   #Fit a Lasso regression model
25   fit = cv.glmnet(x_new,y_new,lambda=lambda.grid, family = 'binomial')
26   best_lambda <- fit$lambda.min
27   #only at the first iteration give the name at the rows
28   if (iter == 1) {
29     rownames(betas) = rownames(betas_iter)
30     names(count_variables_selected) = rownames(betas_iter)
31   }
32   #sum 1 to selected nutrients
33   betas_iter <- predict(fit, s=best_lambda, type = 'coefficients')
34   count_variables_selected = count_variables_selected + !betas_iter==0
      #sum 1 to selected nutrients
35   #report obtained betas for each nutrient at this iteration
36   col_name = paste0("iter", iter)
37   betas[,col_name] = as.numeric(betas_iter)
38 }
39 #only nutrients
40 count_variables_selected_nutr = count_variables_selected[2:44,]
41 betas_nutr = betas[2:44,]
42 #selection nutrients with high selection proportion
43 selection_prob = count_variables_selected_nutr/n_iter
44 #final_selected <- selection_prob[selection_prob >= prop]
45 final_selected = selection_prob
46 #betas for selected nutr
47 proportion_positive_betas = rep(0,length(final_selected))
48 names(proportion_positive_betas) = names(final_selected)
49 proportion_positive_betas =
50   rowSums(betas[names(final_selected),]>0)/
      count_variables_selected_nutr[names(final_selected)]
51
52 return(list(final_selected,proportion_positive_betas))
53 }
54
55 # x is a matrix or dataframe containing the predictors:
56 x = data.frame(sample_nutr)
57 #y vector containing the outcome
58 y = CVD
59 #z adjustment factors

```

```
60 z = data.frame(sex = as.factor(sex), age, center = as.factor(center),
61               smoking = as.factor(smoking), energy, diabetes = as.
        factor(diabetes), bmi)
62
63
64 output <- stability_selection_lasso(x, y, z)
65 final_selected = output[[1]]
66 final_selected
67 #folic.acid      trap      glycemc.index
68 #0.508           0.566           0.765
69
70
71 proportion_pos_betas = output[[2]]
72 proportion_pos_betas
73 #folic.acid      trap glycemc.index
74 #0.00000000     1.00000000     0.00130719
75
76 result <- data.frame(
77   selection_probability = final_selected,
78   proportion_positive_betas = proportion_pos_betas
79 )
80 result_ordered <- result[order(-result$selection_probability),]
```