



**POLITECNICO
MILANO 1863**

**SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE**

EXECUTIVE SUMMARY OF THE THESIS

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

LAUREA MAGISTRALE IN MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: INDIA ERMACORA

Advisor: PROF. FRANCESCA IEVA

Co-advisor: SOLÈNE CADIOU, GIOVANNI FIORITO

Academic year: 2022-2023

1. Introduction

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 (CVDs). CVDs are the leading cause of death, constituting 33% of the total mortality. The *World Health Organisation* estimates 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. 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 nutrient intake, but its role in mediating the association of different type of nutrients with cardiovascular risk factors has not been fully investigated. The primary objective of this study is to investigate the potential causal relationship between a preselected exposome (nutrients) and a single health outcome (CVD), using DNAm as an additional layer of information, possibly mediating an effect.

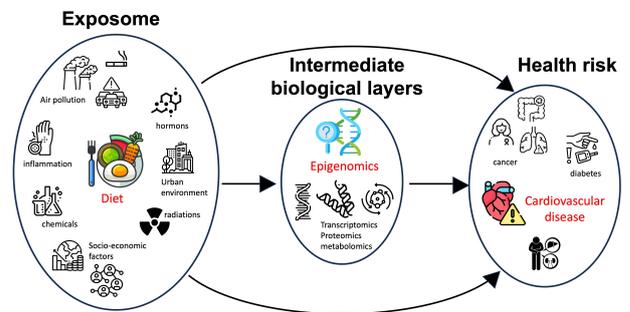


Figure 1: Structure of our case study.

This work is designed to serve as a systematic comparison and evaluation of three methodologies: two applications of the Meet-in-the-Middle (MITM) method [1], an innovative implementation of a high-dimensional mediation framework, and an implementation of a stability selection approach. Our results contribute to a rapidly advancing research domain and are important from a public health perspective because they can help develop effective prevention strategies for high-risk individuals.

2. Data

Our research applies the outlined methodologies to a case-control study within the *European Prospective Investigation into Cancer and Nutrition* (EPIC) project [2], which is an extensive investigation that focuses on exploring the connections between diet, lifestyle, environmental factors, and the onset of chronic diseases. Accessing a subset of the EPIC Italy dataset, our analysis includes 1,580 participants from four centers, obtained through a selection process aimed at extracting a suitable sample. For each participant, we retrieve data on lifestyle habits (*sex, age, recruitment center, smoking status*), medical covariates (*diabetes status, BMI, daily energy intake*), dietary habits (with details on the intake of 43 nutrients), methylation levels (measured with beta values at 399,957 different CpG sites through blood samples), technical covariates (*chip and chip positions*, used to account for systematic differences in DNAm sample processing) and the development of CVD. Initial information and blood samples were obtained at the recruitment stage when all individuals were in good health. Over the subsequent 20 years, routine follow-ups have been conducted to monitor the development of CVDs or diabetes among the participants.

3. Methods

We systematically compared and evaluated three distinct methods, with the ultimate aim of informing the relationship between nutrients and CVD. The core of the analyses involves two developments of the MITM method, while as a third method, we proposed a stability selection approach.

3.1. First application of the MITM

The first application of the MITM, following the “oriented Meet-in the-Middle” design from Cadiou et al., 2020 [3], employs the methylome layer to identify potential new exposures likely to be causally associated with CVD. It consists of three sequential steps:

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

We employ existing literature (EWAS catalogue [4] and a systematic review [5]) to reduce the dimension of the intermediate

methylome layer, thus obtaining a Restricted Methylome.

(b) **Relation between the Whole Exposome and the Restricted Methylome:**

By employing multiple univariate linear models, we identify the nutrients which are strongly associated with the Restricted Methylome, resulting in a reduced set of nutrients, the Reduced Exposome.

(c) **Relation between the Reduced Exposome and the CVD:**

Using univariate logistic models we investigate the relation of the Reduced Exposome with the outcome (CVD) to ultimately identify the nutrients causally linked with CVD.

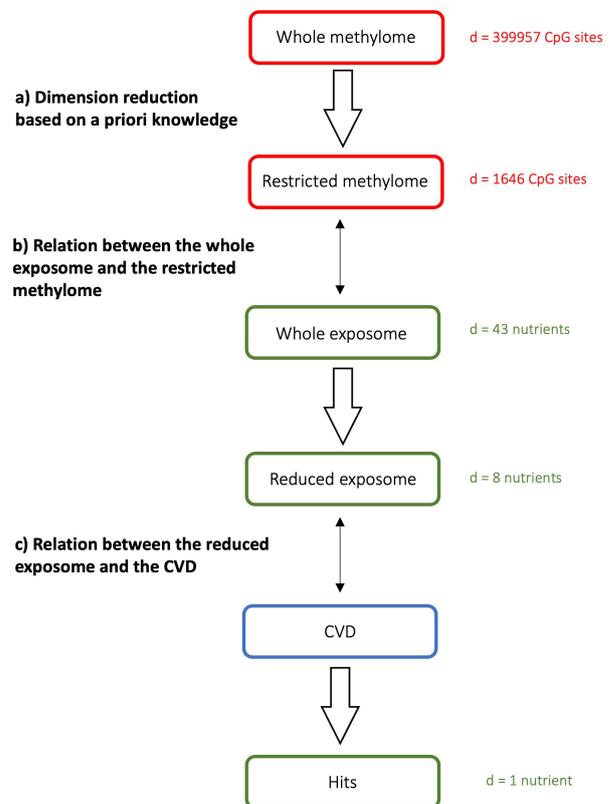


Figure 2: Schematic pipeline of the first MITM.

In all regression models, we incorporate adjustment factors as linear predictors, recalling that 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.

Throughout each step of the process, we implement p-value correction for multiple testing using a False Discovery Rate (FDR) procedure, specifically the Benjamin-Hochberg method. Additionally, we employ quintiles to divide the

nutrient intakes into five equal groups, a common practice in the EPIC cohort and more largely in nutritional epidemiology for addressing non-linearity and evaluating the impact of the most influential intake range. Indeed, given that extreme values tend to exert the most influence on the outcome, by considering p-values and β -coefficients of the 5th quintile with refer to the baseline (1st) in the regression models, we can assess the risks and benefits or the most dangerous or safest range of intakes, excluding those that do not manifest a noticeable impact.

3.2. Second application of the MITM

The second application of the MITM aims at identifying among the methylome potential mediators of the effect of some nutrients associated with CVD, whose mediation effect is then statistically tested. The method is articulated as follows:

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

Through univariate logistic models, we separately evaluate the significance of each nutrient within the Whole Exposome with CVD, resulting in a restricted set, the Reduced Exposome. Hereafter, we assume that the significant nutrients have a causal effect on CVD, and look for mediators among the methylome using a MITM approach.

(b) Dimension reduction based on a prior knowledge:

Once again, leveraging biological information from the mentioned genetic databases, we reduce the methylome's dimension, resulting in the Restricted Methylome.

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

In this step, through linear regression models, 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:

Her we conduct association 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) Significance of the indirect effect of each nutrient in the Reduced Exposome on CVD, through the potential mediators selected:

Through Structural Equation Models (SEMs), we assess the extent to which the subsets of significant CpG sites identified in the previous step mediate the impact of each nutrient from the Reduced Exposome on CVD. We want to determine whether there is statistical evidence that alterations in nutrient intakes influence CVD through changes in DNA methylation levels.

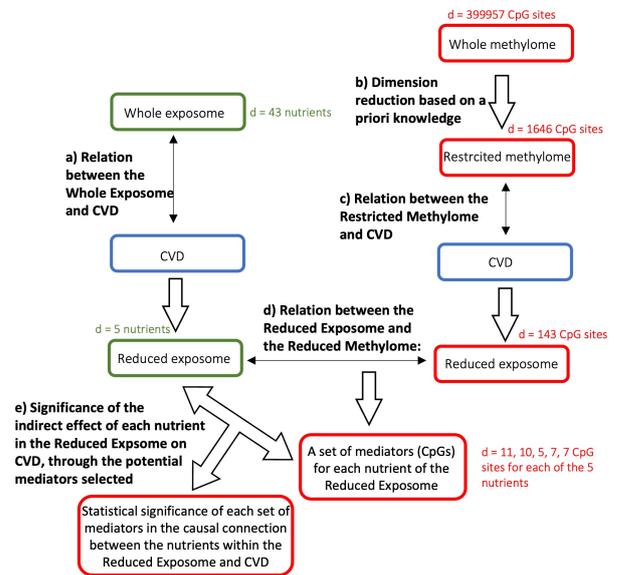


Figure 3: Schematic pipeline of the second MITM.

Given the different objective and structure of the method, in contrast to the first application, here we relax the hypotheses and do not correct for multiple testing. We still incorporate the confounding factors as linear predictors and use quintiles for the division of the distribution of nutrients intake into equal groups, as in the first approach.

3.3. Stability selection algorithm

We implement a stability selection algorithm to select causal predictors for CVD without considering the methylome layer. We opt for a stability selection method because stability is a pivotal prerequisite for generalizability [6], which 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" [7].

The implemented stability selection algorithm follows the guidelines described in the original paper by Meinshausen and Bühlmann [8]. The algorithm uses an existing selection algorithm (LASSO regression here) and complement it with resampling techniques to estimate the probability of selection of each variable using its selection proportion over the resampling iterations.

While LASSO is effective in variable selection, it assumes a linear relationship between selected variables and the outcome. Our observation, however, suggests that many nutrients exhibit non-linear relationships with the outcome, which hence might be not captured by the algorithm.

4. Results

4.1. First application of the MITM

We present, step by step, the results obtained in the first implementation of the MITM:

- (a) By combining the information coming from the two sources considered, we obtain a Restricted Methylome comprising 1,646 CpG sites.
- (b) 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*").
- (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. According to our findings, *iron* emerges as the only significant nutrient, with a with an association p-value of 5.2%. By looking at the shape of its dose-response curve we see that the regression coefficients of the 2nd, 3rd, 4th and 5th quintiles with refer to the baseline (1st) are all negative, indicating that it acts as a protective factor for all consumptions intakes compared to the lowest. Moreover our results mirror the literature sources, confirming the reliability of our findings.

4.2. Second application of the MITM

We outline here the results obtained in each step of the second MITM approach:

- (a) We identify five nutrients that potentially exhibit causal associations with the outcome, forming the Reduced Exposome: *folic acid*, *iron*, *water*, *edible portion*, and *vitamin B6*. According to the dose-response plots, all these nutrients appear to act as protective factors. These findings align with existing literature, with the exception of the *edible portion*, which reasonably lacks a general classification.
- (b) The process leading to the identification of the Restricted Methylome follows the same selection procedure outlined in step a) of the first MITM.
- (c) Analyzing the association of each CpG site within the Restricted Methylome with the health outcome (CVD), we pinpoint 143 significant CpG sites, constituting the Reduced Methylome. This step is crucial as the mediating sites must be linked to the outcome.
- (d) For each of the 5 nutrients within the Reduced Exposome we identify a set of significant CpG sites within the Reduced Methylome, acting as potential mediators.
- (e) The constructed SEM model structure enables us to evaluate the statistical significance of this indirect contribution of each nutrient within the Reduced Exposome through the selected subsets of CpG sites acting as potential mediators. The 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 almost 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.

4.3. Stability selection algorithm

Applying the described stability selection algorithm with a 50% selection proportion threshold,

we identify 3 significant nutrients: *glycemic index*, *TRAP*, and *folic acid*.

5. Comparison and analysis

While all the methods have different structures and ultimate aims, we can still draw comparisons regarding some aspects. Before proceeding with the due considerations, we report in Table 1 the nutrients selected as to be potentially causally connected with the health outcome (CVD).

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

Table 1: Nutrients found 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 among 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 and two nutrients for the second MITM. However, these sets also exhibit p-values that are not significantly high, indicating a weak statistical significance. These observations suggest that the methylation layer might not exert a sufficiently strong influence in mediating the effects of nutrients on CVD.

- **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 which have 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 nutrients which are causally connected through changes in DNAm 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 highest 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 interaction of the methylome layer, a piv-

otal factor in the first and second MITM, as stated in preceding reasonings.

- **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 the methylome layer. It also presents a selection proportion exceeding 50% in the stability selection algorithm. This dual significance suggests a potential causal relationship with CVD, but its selection, however, appears to be less influenced by the presence of the methylome layer if compared to *iron*.

6. Conclusions

The primary objective of this thesis is to investigate the potential causal relationship between a preselected exposome (nutrients) and a single health outcome (CVD), using DNA methylation as an additional layer of information, possibly mediating an effect. After an initial descriptive analysis of the dataset, we delve into the core of the study. The research develops 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 offers 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. 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. 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 devel-

opment 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 results obtained.

References

- [1] Chadeau-Hyam, M. et al., 2011. Meeting-in-the-middle using metabolic profiling-a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers* 16, 83–88. <https://doi.org/10.3109/1354750X.2010.533285>
- [2] EPIC Centres - ITALY. <https://epic.iarc.fr/centers/italy.php>.
- [3] Cadiou, S. et al., 2020. Using methylome data to inform exposome-health association studies: An application to the identification of environmental drivers of child body mass index. *Environ. Int.* 138, 105622. <https://doi.org/10.1016/j.envint.2020.105622>
- [4] Battram, T. et al., The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res* 2022.
- [5] Krolevets, M. et al., 2023. DNA methylation and cardiovascular disease in humans: a systematic review and database of known CpG methylation sites. *Clin Epigenet* 15, 56. <https://doi.org/10.1186/s13148-023-01468-y>
- [6] Poggio, T. et al., 2004. General conditions for predictivity in learning theory. *Nature* 428, 419–422. <https://doi.org/10.1038/nature02341>
- [7] Lazarevic, N. et al., 2019. Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: A review of existing approaches and new alternatives. *Environ. Health Perspect.* 127, 26001. <https://doi.org/10.1289/EHP2207>
- [8] Meinshausen, N. et al., 2010. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72: 417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>.