

Using dna-methylation to inform the exposome-health relations

Solène Cadiou

► To cite this version:

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

HAL Id: tel-03438103

<https://tel.archives-ouvertes.fr/tel-03438103>

Submitted on 21 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Modèles, méthodes et algorithmes en biologie, santé et environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Solène CADIOU

Thèse dirigée par **Rémy SLAMA**, Université Grenoble Alpes

préparée au sein du **Institute for Advanced Biosciences (IAB)**,
Inserm U1209, CNRS UMR 5309, Université Grenoble Alpes
dans l'**École Doctorale Ingénierie pour la Santé la Cognition et l'Environnement**

Influence de l'exposome sur la santé : apport des données de haute dimension de méthylation de l'ADN

Using DNA-methylation to inform the exposome-health relations

Thèse soutenue publiquement le **20 novembre 2020**,
devant le jury composé de :

Monsieur REMY SLAMA

DIRECTEUR DE RECHERCHE, INSERM DELEGATION AUVERGNE-RHÔNE-ALPES, Directeur de thèse

Monsieur ARTHUR TENENHAUS

PROFESSEUR, CENTRALE SUPELEC - PARIS-SACLAY, Rapporteur

Madame JULIE HERBSTMAN

PROFESSEUR ASSOCIE, UNIVERSITE COLUMBIA A NEW YORK - USA, Rapporteure

Monsieur RODOLPHE THIEBAUT

PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE DE BORDEAUX, Président

Monsieur XAVIER BASAGANA

PROFESSEUR ASSOCIE, INS DE BARCELONE POUR LA SANTE MONDIALE, Examinateur



"Chacun a le droit de vivre dans un environnement équilibré et respectueux de la santé."

"Toute personne a le devoir de prendre part à la préservation et à l'amélioration de l'environnement."

Charte de l'environnement, articles 1 et 2

"La fumée fait mal à mon ami."

Marcel Proust, *Le côté de Guermantes*

"Les événements n'appartiennent au hasard que tant qu'on ne connaît pas les lois générales de leur catégorie."

Guy-Ernest Debord, dans *Le mal du Debors*, Rémy Slama

"C'était l'austère simplicité de la fiction plutôt que la trame embrouillée de la réalité."

Raymond Chandler, *Le grand sommeil*

"L'égarement de la vérité entre le beau et le réel"

Mahmoud Darwich, *Comme des fleurs d'amandiers ou plus loin*

"Où t'en vas-tu pensée où t'en vas-tu rebelle"

Louis Aragon, *Cantique à Elsa*

Acknowledgments

Je veux remercier ici les nombreuses personnes qui ont contribué à l'aboutissement de ma thèse.

I would like to thank all the members of my jury: Rodolphe Thiebaut, who did me the honor to chair the jury, Julie Herbstman and Arthur Tenenhaus, who carefully reviewed my manuscript, and Xavier Basagana, also member of my PhD committee, who kindly shared his expertise and his advice during the three years of my PhD. I also thank the members of my PhD committee: Olivier François, Xavier Basagana again, Andrea Baccarelli, and François Mariotti, who all contributed to the progression of my thesis.

Je souhaite également remercier tous les autres chercheurs des laboratoires extérieurs avec qui j'ai collaboré et qui m'ont fait bénéficier de leur expérience et leur expertise au cours de discussions passionnantes ; je pense en particulier à mes collègues de l'ISGlobal à Barcelone, notamment Mariona Bustamante, à Michaël Blum du TIMC ainsi qu'à Simon Barthelme et Florent Chatelain du Gipsa Lab.

Je remercie aussi Alexandre Péry, pour son suivi attentif, et par son intermédiaire, le Ministère de l'Agriculture et de l'Alimentation et le Corps des Ponts, Eaux et Forêts, qui ont financé cette thèse.

J'ai eu la chance d'être accueillie dans l'équipe d'épidémiologie environnementale de l'Institut pour l'Avancée des Biosciences (IAB) de Grenoble, une équipe dont l'excellence scientifique, l'atmosphère conviviale et le dynamisme – mais aussi la localisation avec vue sur la chaîne de Belledonne – m'ont permis d'accomplir ma thèse dans les meilleures conditions. Je veux remercier en premier lieu mon directeur de thèse et directeur d'équipe, Rémy Slama : ma thèse doit énormément à son excellence et sa rigueur scientifique, sa hauteur de vue et son esprit de synthèse face à des problèmes complexes à l'interface de plusieurs disciplines. Plus que cela, son humanité, sa patience pleine de détermination face à mon propre entêtement lors de nos points réguliers, et la grande autonomie, toujours alliée à un encadrement rigoureux, qu'il me laissait (sans oublier ses habituels retards qui me permettaient de finir mes présentations à temps) ont rendu ces trois années de thèse très agréables à vivre, qu'il en soit ici vivement remercié.

Plus largement, je souhaite remercier tous les membres de l'équipe, pour leur apport scientifique comme pour les moments plus conviviaux : les chercheuses de l'équipe, en particulier Valérie Siroux et ses conseils avisés dans la dernière ligne droite et Claire Philippat, pour son expertise gentiment partagée tant sur les perturbateurs endocriniens que sur les spots de grimpe ; tous les autres permanents de l'équipe : en particulier Sarah Lyon-Caen, pour sa gentillesse et sa disponibilité, et Mailys Barbagallo et Karine Supernant si aidantes ; enfin tous les stagiaires, thésards, post-doctorants ou ingénieurs de passage, en particulier Lydiane Agier, qui m'a apporté une aide et une amitié précieuse pendant la première année de thèse et avec qui je partage le souvenir d'un agréable séjour barcelonais, et Matthieu Rolland, pour son aide en R (ggplot !), mais aussi pour son enthousiasme à faire vivre l'équipe et ses talents de brasseur.

Enfin j'ai un mot particulier pour les membres du bureau 103 que j'ai vu défiler durant ces trois ans, qui ont vraiment rendu cette thèse si agréable au quotidien : je remercie Dorothy bien sûr pour son soutien, ses contributions majeures au dialecte endémique du bureau et ses photos de Willa ; Stephan, pilier du bureau 103 avec son grand cœur, son expertise et ses qualités d'esthète, avec qui je déchiffrais les hiéroglyphes de Rémy (et nos verres au Groove avec Hubert, parfois déraisonnables mais toujours passionnantes, resteront sans aucun doute parmi mes meilleurs souvenirs de thèse !) ; Hubert donc, membre honoraire du bureau 103, ses rires un peu trop

sonores, son optimisme réfléchi, sa solidité et son immense gentillesse ; Alicia, avec qui les discussions statistiques étaient toujours éclairantes, les séances de bloc sympathiques (malgré les chevilles tordues et les démotivations de dernier moment) et les discussions sur ses exs passionnantes ; Ian, toujours si gentil et aussi toujours prêt à aller visiter un buffet gratuit, dont l'expertise en Unix a sans doute divisé mon temps passé sur Ciment par 2 ; Pau, with our memorable memories of Utrecht and our (attempt of) record du monde ; enfin André, je garde un très bon souvenir de ces deux semestres où nos horaires étaient variables, de sa générosité, de ses conseils en bloc, et de nos débats aux déjeuners.

Je remercie également Florent Chuffart, si aidant et disponible pour m'empêcher de faire planter Ciment et grâce à qui j'ai pu mener mes études de simulations à terme.

Je tiens aussi à remercier les amis, les nouveaux grenoblois comme les anciens, fidèles et lointains, grâce auxquels ce furent de belles années grenobloises. Au moment où j'achève cette thèse, j'ai une pensée particulière pour mes coloc, Antoine et Claire, compagnons quotidiens aux thèses parallèles.

Enfin, c'est bien sûr ma famille que je veux remercier, mes chers parents, mes petites sœurs Jeanne et Camille, mon frère Erwan, avec Mathilde et Augustine. Vous avez joué un rôle inestimable dans l'achèvement de cette thèse comme dans les choix qui m'y ont conduite, par les valeurs transmises et partagées, par votre joie de vivre, et bien sûr par votre soutien plein d'humour, vos encouragements et votre affection.

Et je termine en remerciant mon Jean-Gabriel, que ma thèse a également amené à Grenoble et qui a rendu bien plus mouvementées et belles ces trois années de vie alpine.

Abstract

Context: The exposome is defined as encompassing all the life-course environmental exposures from the prenatal period onwards. Challenges in the characterization of its effects on health include a limited statistical power and a possibly high rate of false positive signals of current studies. False positives findings may correspond to reverse causality. To cope with such challenges, refinement in statistical methods may be needed. In addition, using (a priori) biological information, e.g. from intermediary layers such as DNA methylation, may help to reduce the problem dimensionality and possibly false positive signals due to reverse causality.

Aims: We aimed to identify strategies to limit the false positive rate in exposome studies, in particular by integrating a priori information from the methylome and to apply these strategies to the question of the exposome's influences on child health. We also aimed to illustrate issues of exposome studies related to models' stability.

Methods: We first implemented two “oriented Meet-in-the-Middle” (oMITM) approaches to characterize the link between the exposome and child health outcomes (Body Mass Index, BMI and lung function) in the Helix cohorts (1173 mother-child pairs); the approach relied on 3 steps: a) identification of CpGs loci (i.e. methylation sites) independently associated with exposures and the outcome using a priori information and/or univariate linear regressions; b) identification by linear regression of the exposures associated with at least one of these CpGs, thus constituting a *reduced exposome*; c) test of their association with the outcome. We then performed a Monte-Carlo simulation study to characterize the performance of the oMITM design under various causal assumptions. We simulated realistic exposome, intermediary layer and outcome relying on data from the Helix BMI study and assuming linear relationships between components of the three layers. The magnitude of links was allowed to vary, leading to 2281 scenarios under 5 different causal structures, including a structure corresponding to reverse causality. For each scenario, we generated 100 datasets and tested 6 methods: 3 ignoring the methylome data (“agnostic approaches”: ExWAS; DSA; LASSO) and 3 using methylome data (two implementations of oMITM and a mediation analysis). Methods’ performance was assessed by sensitivity and specificity. We further performed a two-layer simulation study to assess the instability of some agnostic methods, with a focus on LASSO.

Results: The oMITM approaches performed on Helix data identified one exposure, copper post-natal blood level, associated with higher BMI and with lower lung function. An ExWAS relating exposome to BMI identified in the same data 18 additional (lipophilic) exposures, whose association with BMI could possibly be due to reverse causality. The simulation study showed that, compared to the other approaches, the oMITM design may allow to discard some false positive findings in at least one situation of reverse causality and to increase specificity when the intermediate layer mediates part of the effect of the exposome on the outcome, at a cost in terms of sensitivity loss. The oMITM – DSA implementation showed better performances (sensitivity, specificity) than the oMITM – ExWAS. The second simulation study showed that the stabilization step changes model performance thus illustrating its importance when using agnostic machine learning algorithms such as LASSO.

Discussion and perspectives: The use of complex statistical methods tailored for intermediate or high dimensional data, or the consideration of biological information, could help tackle the question of false positives in exposome studies. We developed a design, oMITM, which was less prone to reverse causation bias than agnostic approaches ignoring intermediate layers between the exposome and health, at a cost in terms of sensitivity.

Résumé

Contexte : L'exposome est défini comme l'ensemble des expositions environnementales reçues au cours de la vie (dont la vie prénatale). La puissance statistique limitée et le taux élevé de faux positifs des études actuelles sont deux défis majeurs pour la caractérisation de ses effets sur la santé. Les faux-positifs peuvent notamment être dus à de la causalité inverse. Pour faire face à ces défis, affiner les méthodes statistiques est utile, mais l'utilisation d'information biologique, par exemple provenant de couches intermédiaires telle la méthylation de l'ADN, peut aussi contribuer à réduire la dimension du problème, et les faux positifs liés à la causalité inverse.

Objectifs : Notre objectif principal est d'identifier des stratégies pour limiter les faux positifs dans les études sur l'exposome, en particulier en intégrant des informations *a priori* provenant du méthylome, et d'appliquer ces stratégies à l'étude de l'influence de l'environnement sur la santé de l'enfant. Nous avons également cherché à illustrer d'autres enjeux des études sur l'exposome liés à l'instabilité des modèles.

Méthodes : Nous avons d'abord mis en œuvre deux approches "Meet-in-the-Middle orientées" (oMITM) pour caractériser le lien entre exposome et santé de l'enfant (indice de masse corporelle, IMC et fonction pulmonaire) dans les cohortes Helix (1173 mères-enfants) ; l'approche comprenait 3 étapes : a) identification de CpG indépendamment associés aux expositions et à la santé en utilisant des connaissances *a priori* et/ou des régressions linéaires univariées ; b) identification par régression linéaire des expositions associées à au moins un de ces CpG, constituant un *exposome réduit* ; c) test de leur association avec la santé. Nous avons ensuite réalisé une simulation de Monte-Carlo pour caractériser la pertinence du design oMITM sous différentes structures causales. Nous avons simulé un exposome, une couche intermédiaire et un événement de santé à partir des données Helix en postulant des relations linéaires entre les couches. La magnitude des liens variait, générant 2281 scénarios sous 5 structures causales différentes, dont une de causalité inverse. Pour chaque scénario, 100 jeux de données étaient générés et 6 méthodes testées : 3 ignorant le méthylome ("approches agnostiques" : ExWAS ; DSA ; LASSO) et 3 l'utilisant (deux implémentations d'oMITM et une analyse de médiation). Les performances étaient évaluées par leur sensibilité et spécificité. Nous avons aussi effectué une étude de simulation pour évaluer l'instabilité de certaines méthodes agnostiques, en particulier le LASSO.

Résultats : Les approches oMITM sur les données Helix ont identifié une exposition, le niveau postnatal de cuivre dans le sang, associé à un IMC accru et à une fonction pulmonaire diminuée. Une ExWAS entre exposome et IMC dans HELIX a identifié 18 autres expositions (lipophiles), dont l'association avec l'IMC pourrait de ce fait être due à de la causalité inverse. L'étude de simulation a montré que, par rapport aux autres approches, le design oMITM peut éviter certains faux positifs dans au moins une situation de causalité inverse et augmenter la spécificité lorsque la couche intermédiaire médie une partie de l'effet de l'exposome sur la santé, ceci à un coût en terme de sensibilité. L'implémentation oMITM-DSA montrait de meilleures performances qu'oMITM-ExWAS. La deuxième simulation a montré que l'étape de stabilisation du modèle est cruciale lors de l'utilisation d'algorithmes d'apprentissage agnostique tels LASSO, car elle en modifie les performances.

Discussion et perspectives : L'utilisation de méthodes statistiques complexes adaptées à des données de dimensions intermédiaires ou élevées, ou la prise en compte de connaissances biologiques, pourraient aider à limiter les faux positifs dans les études sur l'exposome. Nous avons proposé un design, oMITM, qui est moins sujet au biais de causalité inverse que les approches agnostiques avec un coût en termes de sensibilité.

Large audience abstract

The exposome includes all the life-course environmental exposures. However, its application to find causal predictors of a health outcome raise challenges, due to limited statistical power and high false positive rate in current exposome projects. Reducing the exposome dimension, e.g. using a priori knowledge, can be a solution. We proposed a new design, the oriented Meet-in-the-Middle (oMITM): it relied on an intermediary biological layer to restrict to exposures associated with relevant intermediary features, whose association with health is then tested. In Helix data, oMITM with methylome pointed copper post-natal blood level as associated with higher child body mass index. We performed a simulation study to compare oMITM to existing methods: oMITM allowed to discard some associations due to reverse causality and to increase specificity in case of an effect of exposome on health. We also provided new insights on the use of unstable machine learning algorithms in epidemiology.

Résumé grand-public

L'exposome est l'ensemble des expositions environnementales au long de la vie. Y identifier des prédicteurs causaux de la santé pose des défis, de par la puissance statistique limitée et le taux élevé de faux positifs des projets exposome actuels. Réduire a priori la dimension de l'exposome peut être une solution. On propose un nouveau design, le "Meet-in-the-Middle orienté" (oMITM), qui restreint l'exposome aux expositions associées à des variables biologiques intermédiaires pertinentes avant de tester son association avec la santé. Dans les données Helix, oMITM a pointé le niveau sanguin postnatal de cuivre comme associé à un IMC plus élevé de l'enfant. Une simulation a montré qu'oMITM écartait des associations dues à la causalité inverse et augmentait la spécificité par rapport à des méthodes existantes. Ce travail permet de mieux comprendre dans quelles situations certaines approches statistiques ou l'intégration des données intermédiaires rendent les études exposome performantes.

Keywords

Exposome; cohort; methylome; biological a priori; false-discovery; causality; high-dimension; Body-Mass-Index; stability.

Mots-clefs

Exposome ; cohorte ; methylome ; a priori biologique ; faux-positifs ; causalité ; haute dimension ; Indice de Masse Corporelle ; stabilité.

List of acronyms

BMI: Body Mass Index
BPA: bisphenol A
BPF: bisphenol F
BPS: bisphenol S
BUPA: butyl paraben
CI: Confidence Interval
cxMiNP: Mono-4-methyl-7-carboxyoctyl-phthalate
CpG: Cytosine phosphate Guanine
DAG: Directed Acyclic Graph
DDE: 4,4'dichlorodiphenyl dichloroethylene
DMR: Differentially methylated regions
DNA: Deoxyribonucleic acid
DOHaD: Developmental Origins of Health and Disease
DSA: Deletion Substitution Addition
ETPA: ethyl paraben
ExWAS: Exposome wide association study
FDP: False discovery proportion
FDR: False discovery rate
FEV₁: Forced Expiratory Volume in 1 second
HCB: Hexachlorobenzene
HELIX: Human Early Life Exposome
LASSO: Least Absolute Shrinkage and Selection Operator
LOD: Limit of Detection
LOQ: Limit of Quantification
LUR: Land use regression
MBzP: mono benzyl phthalate
MECPP: mono-2-ethyl 5-carboxypentyl phthalate
MEHHP: mono-2-ethyl-5-hydroxyhexyl phthalate
MEHP: Mono(2-éthylhexyl) phthalate
MEOHP: mono-2-ethyl-5-oxohexyl phthalate
MEP: monoethyl phthalate
MEPA: methyl paraben
MiBP: mono-iso-butyl phthalate
MITM: Meet-in-the-Midle
MLR: Multivariate Linear Regression
MMCHP: Mono-2-carboxymethyl hexyl-phthalate
MnBP: mono-n-butyl phthalate
MWAS: Methylome-Wide Association Study
NDVI: Normalized Differential Vegetation Index
ohMiNCH: 2-(((Hydroxy-4-methyoctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid
OHMiNP: mono-4-methyl-7-hydroxyoctyl phthalate
ohMPHP: 6-Hydroxy Monopropylheptyl-phthalate
oMITM: oriented Meet-in-the-Middle
OXBE: oxybenzone (benzophenone-3)
oxoMiNCH: 2-(((4-Methyl-7-oxyoctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid
OXOMiNP: mono-4-methyl-7-oxooctyl phthalate
PBDE: Polybrominated diphenyl ethers
PCA: Principal components analysis
PCB: Polychlorinated biphenyl
PLS: Partial Least Square
PM: Particulate Matter
PRPA: propyl paraben
RMSE: Root Mean Squared Error
RNA: Ribonucleic acid

SD: Standard deviation

VEMS: volume expiratoire maximal en 1 seconde

VIF: Variance inflation factor

WQS: Weighted Quantile Sum

Table of contents

Acknowledgments	1
Abstract	1
Résumé	2
List of acronyms	4
Table of contents	6
Detailed table of contents	7
Lists of figures and tables	11
CHAPTER I: Introduction	17
CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks	45
CHAPTER III: Early-life Exposures and child lung function: a modified Meet-in-the-Middle approach using preselected methylation marks	101
CHAPTER IV: Performance of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health: a simulation study under various causal structures	111
CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology	167
CHAPTER VII: Discussion	203
References	217
APPENDIX I: Oral communications and publications; curriculum vitae	239
APPENDIX II: Prenatal exposures and birthweight in SEPAGES mother-child study, an adapted oriented Meet-in-the-Middle approach	243
APPENDIX III: Large supplementary materials	263

Detailed table of contents

Acknowledgments

Abstract	1
Résumé	2
List of acronyms	4
Table of contents	6
Detailed table of contents	7
Lists of figures and tables	11
Main figures:	11
Supplementary Figures:	12
Main tables:	13
Supplementary tables:	14
Other supplementary materials:	15
CHAPTER I: Introduction	17
I. 1. Challenges of the use of the exposome concept in environmental epidemiology	17
I.1.1. The exposome concept	17
I.1.2. Assessing the exposome	18
I.1.3. Relevance of the exposome concept for public health	19
I.1.4. Relating the exposome to health	20
I.1.5. Considering the false positive rate issue in a structural causal framework	26
I. 2. Statistical techniques to address false positive and false negative challenges in exposome studies	29
I.2.1. Multiple testing correction	29
I.2.2. Curse of dimensionality and dimension reduction techniques	30
I.2.3. Multivariate variable selection methods	31
I. 3. Adding biological information	34
I.3.1. Focusing <i>a priori</i> on a single exposure	34
I.3.2. Information from intermediate biological layers	34
I.3.3. Mediation analysis	37
I.3.4. Using intermediate biological layer to find true predictors of health within the exposome	39

I. 4. Environmental effects on child weight	40
I.4.1. Environmental effects on child Body Mass Index	40
I.4.2. Environmental effects on birth weight	41
I. 5. PhD project's aim	41
CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks	45
II. 1. Abstracts	46
II.1.1. English abstract	46
II.1.2. French abstract	47
II. 2. Published article	49
II. 3. Supplementary materials	65
CHAPTER III: Early-life Exposures and child lung function: a modified Meet-in-the-Middle approach using preselected methylation marks	101
III. 1. Abstracts	102
III.1.1. English abstract	102
III.1.2. French abstract	103
III. 2. Background	104
III. 3. Methods	104
III. 4. Results	107
III. 5. Discussion	108
CHAPTER IV: Performance of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health: a simulation study under various causal structures	111
IV. 1. Abstracts	112
IV.1.1. English abstract	112
IV.1.2. French abstract	113
IV. 2. Introduction	114
IV. 3. Materials and methods	116
IV.3.1. Overview of the simulation	116
IV.3.2. Causal structures considered	116
IV.3.3. Generation of independent realistic exposome, methylome and outcome data, and addition of causal relations within them	117
IV.3.4. Methods to relate the exposome and health compared	121
IV.3.5. Assessing scenarios' characteristics and methods' performances	122
IV.3.6. Comparisons between oMITM, mediation and direct association test using structural causal modelling theory in a three-variable scheme	123

IV. 4. Results	125
IV.4.1. Causal structures assuming an effect of the exposome on health	125
IV.4.2. Causal structures without effect of the exposome on health	134
IV.4.3. Comparisons between methods using causal inference theory	138
IV. 5. Discussion	140
IV.5.1. Strengths and limitations	140
IV.5.2. Summary of methods' performances	141
IV.5.3. Consistency between our structural causal modelling analysis and experimental simulation-based	143
IV.5.4. The need to rely on causal knowledge	144
IV. 6. Supplementary materials	146
CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology	167
V. 1. Abstracts	168
V.1.1. English abstract	168
V.1.2. French abstract	169
V. 2. Introduction	170
V. 3. Methods	173
V.3.1. Simulation study of LASSO, DSA and ExWAS under various correlation structures	173
V.3.2. Simulation study of stabilizations methods for LASSO in a realistic exposome setting	175
V.3.3. Indicators of stability and performance	177
V.3.4. Application: using LASSO to relate the exposome to child body mass index	178
V. 4. Results	180
V.4.1. Stability and performances of the models' default implementations	180
V.4.2. Effectiveness of stabilization methods	181
V.4.3. Relation between stabilization and model performance	182
V.4.4. Application: using LASSO to relate exposome to child body mass index in Helix data	185
V. 5. Discussion	187
V.5.1. Strengths and limitations	188
V.5.2. Stability selection, a relevant approach when selecting true predictors is the aim	189
V.5.3. Importance of the calibration of model averaging approaches	191
V.5.4. Practical consequences and conclusion	191
V. 7. Supplementary Materials	193

CHAPTER VII: Discussion	203
VII. 1. Overview of results	203
VII. 2. Dimension reduction approaches in the context of exposome studies	204
VII.2.1. Which dimension reduction approaches when selection of variable(s) of interest is the aim?	205
VII.2.2. Which dimension reduction approaches when information concentration is the aim?	207
VII. 3. Multilayer designs to identify causal predictors	210
VII.3.1. Infer causality in exposome studies	210
VII.3.2. Multiple layers design as a clue to overcome the challenges of data-driven causal modelling	211
VII. 4. Perspectives and conclusion	213
References	217
APPENDIX I: Oral communications and publications; curriculum vitae	239
APPENDIX II: Prenatal exposures and birthweight in SEPAGES mother-child study, an adapted oriented Meet-in-the-Middle approach	243
APPENDIX III: Large supplementary materials	263

Lists of figures and tables

Main figures:

Figure I.1: Median absolute correlation within exposure groups (diagonal) and between exposure groups (off-diagonal) for the prenatal exposome assessed in the Helix Project (Figure from Tamayo et al., 2019).	20
Figure I.2: Sensitivity and FDP of 6 different statistical methods assessed from a Monte-Carlo simulation assuming a causal relationship between predictors drawn from a realistic exposome and a health outcome.	32
Figure I.3 : Schematic representation of links between exposome, intermediate biological layers and health effects. Figure from (Vrijheid, 2014).	35
Figure I.4: Implementation of Meet-in-the-Middle. Adapted from (Chadeau-Hyam et al., 2011).	36
Figure I.5: Causal graph of mediation.	38
Figure III.1: Manhattan plot of the step b) of the MITM approach: adjusted-values of adjusted association tests between preselected CpG and exposome.	108
Figure IV.1: Causal structures considered in the simulation study of the efficiency of studies relating a layer of predictors E (e.g., the exposome) to a layer of possibly intermediary parameters (e.g. biological parameters such as DNA methylation) M and a health outcome or parameter Y.	117
Figure IV.2: A. 1- FDP and B. Sensitivity under causal structure A (see Figure IV.1) for all compared methods; performances were averaged across scenarios according to categories of variabilities of Y explained by E (x-axis) and by M (color) and categories of mean variability of a covariate from M affected by E by E.	127
Figure IV.3: Comparisons between oMITM-ExWAS and control methods (<i>oMITM-steps 1 and 2 and ExWAS on subsample</i>) performance (1-Average FDP and sensitivity) for causal structures A, B and C.	128
Figure IV.4: A. 1- FDP; B. sensitivity under causal structure B (see Figure IV.1).	132
Figure IV.5: A. 1- FDP; B. Sensitivity under causal structure C.	133
Figure IV.6: A. Proportion of exposures influenced by Y wrongly identified, and B. number of hits under causal structure D.	136
Figure IV.7: Average number of covariates selected per method under causal structure E.	137
Figure V.1: Stability index (mean Sorenson index) of ExWAS, default LASSO, Elastic-Net and DSA for 3 different structures of pairwise correlations within the tested predictors.	181
Figure V.2: Performance and stability of various stabilization methods of LASSO.	182
Figure V.3: Dependence of performance measures on stability (mean Sorenson index).	184
Figure Appendix II.1: Distribution of the weight of CpGs in the selected PLS component.	251

Supplementary Figures:

Supplementary Figure IV.1: Causal structure A, average sensitivity (A), 1- FDP (B) and number of hits (C) according to the variability of Y explained by E, method by method.	157
Supplementary Figure IV.2: Causal structure B, average sensitivity (A), 1- FDP (B) and number of hits (C) according to the variability of Y explained by E methods by methods.	158
Supplementary Figure IV.3: Causal structure C, average sensitivity (A), 1- FDP (B) and number of hits (C) according to the variability of Y explained by E, method by method.	160
Supplementary Figure IV.4: causal structure D, average number of hits (A), and sensitivity to detect exposures affected by Y (B) according to the variability of one exposure affected by Y explained by Y, methods by methods, in causal structure D (reverse causality).	161
Supplementary Figure IV.5: Example of a causal situation in which a classical Meet-in-the-Middle framework without our additional adjustment would conclude to causal associations between E1 and Y3 whereas there is no causal influence of E1 on Y3 through M.	162
Supplementary Figure V.1: Performance and stability of some common algorithms as a function of the total outcome variability explained by the predictors in situations in which 10 predictors explained the outcome.	193
Supplementary Figure V.2: Performance and stability of the LASSO stabilization methods as a function of the total outcome variability explained by the predictors (log scale).	196
Supplementary Figure V.3: Average number of predictors selected in 20% (A.) and 60% (B.) of the runs on a same dataset for unstable methods, as a function of the total variability of the outcome explained by the true predictors.	197
Supplementary Figure V.4: Occurrence of selection for each exposure when applying 10 times default LASSO and all tested stabilized LASSO to relate an exposome of 173 prenatal and postnatal quantitative exposures (A) or only the smaller exposome of 74 prenatal quantitative variables (B), to zBMI in 1301 mother-child pairs of the Helix cohorts.	198

Main tables:

Table I.1: Exposome project funded by the European Commission. All cited text comes from the website of Community Research and Development Information Service (CORDIS) resources.	22
Table III.1: Preselected genes related to FEV1 according to (Li et al., 2013) and corresponding number of enhancers CpGs available in Helix data.	106
Table III.2: Agnostic ExWAS corrected for relevant potential confounders and corrected for multiple testing relating the exposome and child FEV1.	107
Table IV.1: Details of the methods compared in the simulation study.	119
Table IV.2: Performance for every method under each causal structure.	131
Table IV.3: Number of hits (average mean and standard error across scenarios), sensitivity to find the exposures predicted by Y (average mean and standard error across scenarios) under causal structures D and E.	134
Table IV.4: Number of true causal links detected, false causal links detected, true causal links non-detected, false causal links non-detected by different designs among the causal structures considering all possible links between 3 unidimensional layers.	139
Table V.1: Implementation details for ExWAS, Elastic-Net and DSA.	174
Table V.2: Details of the implemented LASSO methods.	176
Table V.3: Results of the application of default LASSO and various LASSO stabilization methods to relate an exposome of 173 prenatal and postnatal quantitative exposures to zBMI in 1301 mother-child pairs of the Helix cohorts.	186
Table VI.1: Possible strategies of dimension reduction for the exposome and methylome layers	209
Table Appendix II.1: Exposome components assessed in Sepages cohort during pregnancy, with mean and standard deviation for quantitative variables and frequency for qualitative variables, and amount of missing data.	245
Table Appendix II.2: Characteristics of the 438 mother-child pairs included in the exposome analysis based on Sepages study.	250
Table Appendix II.3: Step b) of the oMITM approach: estimates, confidence intervals, uncorrected and corrected for multiple comparisons p-values of the tests of association between exposome and PLS component adjusted on relevant covariates and birth weight (434 mother-child pairs from the Sepages cohort).	253
Table Appendix II.4: Agnostic multiple regression relating the exposome to the birth weight: estimates, confidence intervals, uncorrected and corrected for multiple comparisons p-values of the tests of association between exposome and birthweight adjusted on relevant covariates	256

Supplementary tables:

Supplementary Table IV.1: Meaning and ranges of parameters used in each causal structure to simulate the link between layers.	146
Supplementary Table IV.2: Characteristics of scenarios for structure A, B and C.	149
Supplementary Table IV.3: Characteristics of the scenarios simulated for structure D	150
Supplementary Table IV.4: DAG analysis for different designs when considering all possible links between 3 unidimensional layers (e.g. an exposure, a CpG site, and BMI) according to causal inference theory.	151
Supplementary Table IV.5: Details of causal inference analysis for the oMITM design applied to 3 variables (e.g. an exposure, a CpG site, and BMI) according to causal inference theory in all possible causal structures.	154
Supplementary Table V.1: Distribution of stability index, sensitivity and False Discovery Proportion (FDP), across all scenarios and categorized according to the total variability explained by the true predictors (>1 and ≤ 1) for different stabilization methods of LASSO.	200
Supplementary Table V.2: Results of the application of default LASSO and LASSO stabilization methods to relate an exposome of 74 prenatal quantitative exposures to zBMI in 1301 mother-child pairs of the Helix cohorts.	200
Supplementary Table V.3: List of variables selected by default LASSO and all tested stabilized LASSO for each of the 10 runs applied to relate an exposome of 173 prenatal and postnatal quantitative exposures (A) or only the smaller exposome of 74 prenatal quantitative variables (B), to zBMI in 1301 mother-child pairs of the Helix cohorts.	201

Other supplementary materials:

Supplementary Material II.1: Exposure levels in 1,173 mother-child pairs from the HELIX cohort.	65
Supplementary Material II.2: Pathways identified as relevant for zBMI relying on KEGG database, and corresponding numbers of genes and enhancer CpGs	82
Supplementary Material II.3: Boxplot of child zBMI in HELIX data, by cohorts.	83
Supplementary Material II.4: Population characteristics by cohort	84
Supplementary Material II.5: Distribution of pairwise coefficients of correlation within quantitative variables of the full exposome assessed in 1,173 mother-child pairs from HELIX cohort	87
Supplementary Material II.6: Adjusted associations between the reduced methylome (2284 CpGs) and zBMI in 1,173 children from the HELIX cohort (ExWAS model, step b) of the Meet-in-the-Middle approach.	88
Supplementary Material II.7: Adjusted associations between exposures and CpGs associated with childhood zBMI in 1,173 children from HELIX cohort (ExWAS model, step c) of the Meet-in-the-Middle approach. Results are presented only for CpGs with a (FDR-corrected for multiple hypothesis testing) p-value below 0.05 in ExWAS.	91
Supplementary Material II.8: Sensitivity Analysis I: adjusted associations between the whole exposome and zBMI in 1,173 children from the HELIX cohort (2 multivariate agnostic approaches, one prenatal, one postnatal, ignoring the methylome).	94
Supplementary Material II.9: Characteristics of the CpG selected by a methylome wide analysis on the whole methylome (row percentages).	94
Supplementary Material II.10: Sensitivity analysis III - adjusted association between the whole methylome and zBMI in 1,173 children from the HELIX cohort (ExWAS model, step b of the Meet-in-the-Middle approach applied to the whole methylome).	95
Supplementary Material II.11: Sensitivity analysis III, Meet-in-the-Middle without CpGs preselection: adjusted associations between the exposome and CpGs associated with zBMI in 1,173 children from the HELIX cohort (ExWAS model adjusted on zBMI, step c of the Meet-in-the-Middle approach applied on the whole methylome).	96
Supplementary Material II.12: Sensitivity analysis IV: Meet-in-the-Middle approach considering the cell-types as the intermediate layer. Adjusted associations at steps b),c) and d).	98
Supplementary Material IV.1: Detailed simulation methods	163
Supplementary Material IV.2: Commented simulation script	166
Supplementary Material V.1: Commented simulation script	201
Supplementary Material V.2: Application: using LASSO to relate pregnancy exposome to child body mass index in Helix data	201

CHAPTER I: Introduction**I. 1. Challenges of the use of the exposome concept in environmental epidemiology****I.1.1. The exposome concept**

The exposome concept acknowledges that individuals are simultaneously exposed to a multitude of different factors from conceptions onwards (Wild, 2005) and can be defined as the totality of the individual environmental (i.e. non-genetic) factors. Since the 2000s, environmental epidemiology has progressively embraced it and evolved from studies considering the association of one exposure with one specific disease (e.g. (Hill and Doll, 1950)) to studies including various long term or short term measures of different exposures (see for example (Agier et al., 2019; Lenters et al., 2016)).

The exposome was originally defined as consisting in three main categories (Siroix et al., 2016; Wild, 2012): 1. A large set of external individual exposures, including air pollutants, meteorological factors, radiation, chemical exposures, as well as diet, physical activity and tobacco and other life-style factors; 2. A wider general exposome, including the urban-rural environmental, the education and the socio-economical and climate factors; 3. An internal exposome, consisting in endogenous processes internal to the body (such as metabolic factors, gut microflora, inflammation or oxidative stress). The inclusion of the internal exposome in exposome studies which has been sometimes advocated (Rappaport, 2012; Vermeulen et al., 2020) can be discussed: indeed, if biomarkers of exposures (for example levels of a given phthalate in urine) can be useful to assess the individual exposures, components of the internal exposome can be considered to be biomarkers of effects, whose levels and variations result from a wide range of exogenous and genetic factors (Chadeau-Hyam et al., 2013). Last, the infectious factors, the “infectome”, are also sometimes recognized as a part of the exposome (Bogdanos et al., 2015; Damiani et al., 2020).

Since 2010, numerous ambitious studies have been built in order to describe the exposome at the individual level and its links with the health (see for example the 12 exposome projects funded by the European Commission from 2012 to 2024, Table I.1).

I.1.2. Assessing the exposome

Assessing the exposome involves many challenges: the first one is to know what to measure. Exposome studies most often assessed an a priori defined set of exposures (see for example the Helix project (Vrijheid et al., 2014)). An alternative or a complement is to screen for exogenous chemicals using metabolomic biomarkers (Vermeulen et al., 2020) in an untargeted approach (see for example (Bonvallot et al., 2013)). In targeted studies, the individual exposome is sometimes assessed both by environmental models, for outdoors exposures such as air pollutants and meteorological data, which often have the advantages to document the source of exposures, and by biomarkers for individuals exposures to chemicals, such as phthalates or phenols (in urine) or metals (in blood or other matrices), which allowed a more personal assessment (Maitre et al., 2018; Vineis et al., 2017).

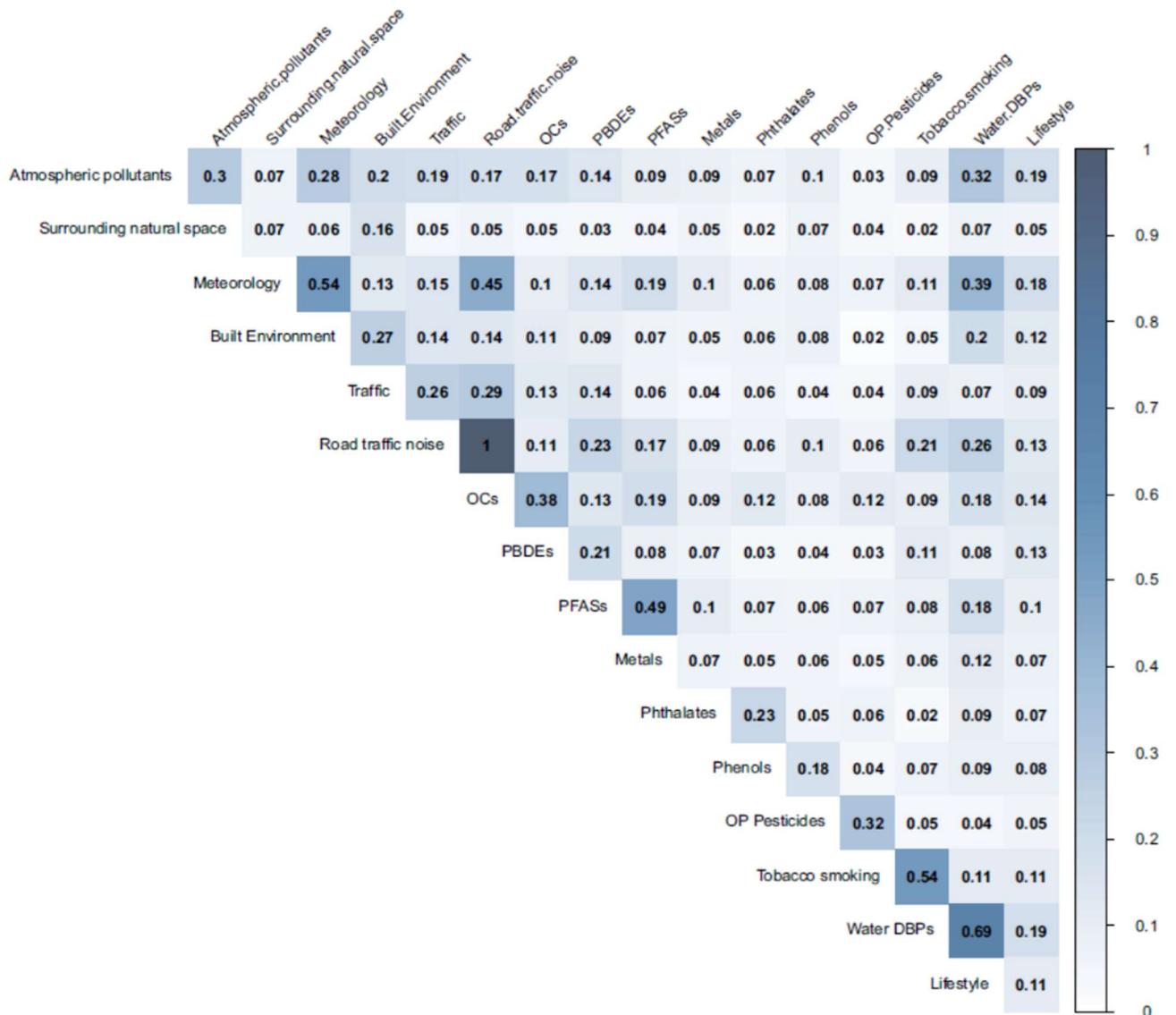
The measurement error (Armstrong, 1998) traditionally encountered in environmental epidemiology is an important challenge for exposome assessment, as the exposome is dynamic throughout time and as some chemicals components of interest have high within-subjects variability (Agier et al., 2020b; Casas et al., 2018; Vernet et al., 2018): this includes in particular short half-life compounds, such as phenols, phthalates and organophosphate pesticides (Casas et al., 2018). Some studies, like the French SEPAGES cohort (Lyon-Caen et al., 2019), aim at decreasing measurement error by increasing the number of measurement points in each subject, e.g. relying on the so-called *within-subject biospecimens pooling* approach (Vernet et al., 2019). Whereas the exposome encompasses all exposures from conceptions, it appears relevant for some exposome studies to reduce measurement time window to focus on early life, as the early-life environment

may be critical for later health, as stated by the Developmental Origins of Health and Disease (DOHaD) paradigm (Heindel et al., 2015); this is done by many exposome projects (see Table I.1).

I.1.3. Relevance of the exposome concept for public health

The relevance of the application of the exposome concept in environmental epidemiology first lies in such ambitious assessments: they allow to describe the correlations between exposures (see for example (Tamayo-Uria et al., 2019) and Figure I.1). Such a description is crucial for environmental justice (Brulle and Pellow, 2006), but also to better relate the environment to health and help assessing the environmental burden of disease: indeed, the exposome may explain an important part of chronic diseases (Manrai et al., 2017; Rappaport, 2016) and deaths (Gakidou et al., 2017), which genetic factors are not sufficient to account for (Rappaport, 2016). Simultaneously considering numerous exposures allows to limit selective reporting and publication bias in etiologic studies (Slama and Vrijheid, 2015), to reckon with multiple testing and to help discarding confounding by co-exposures. Moreover, an exposome approach may, at least in principle, enable to identify *mixture* effects (Slama and Vrijheid, 2015), i.e. combinations and interactions between multiple environmental exposures on their causal path to an health effect (Patel, 2017) (an operational definition of interactions corresponds to departure from additivity in a linear setting (Rothman et al., 2012)).

Figure I.1: Median absolute correlation within exposure groups (diagonal) and between exposure groups (off-diagonal) for the prenatal exposome assessed in the Helix Project (Figure from Tamayo et al., 2019). Between-exposure correlations in a given exposure family can reach 0.8 (Tamayo-Uria et al., 2019).



I.1.4. Relating the exposome to health

The number of environmental factors assessed in such studies are typically in the order of a few hundred (216 in HELIX (Tamayo-Uria et al., 2019)); after the assessment of the exposome, the second main challenge of exposome studies, on which we will focus, is to manage to **identify causal predictors of health outcomes among these exposures, and assessing their effect.**

The first method used to do this was the ExWAS (for exposome-wide association studies), i.e. univariate regressions (adjusted for confounders) relating independently each exposure to the health outcome of interest and possibly corrected for multiple testing (Patel et al., 2010). Patel first assessed the relationship between 266 exposures and type 2 diabetes (Patel et al., 2010). Since then, ExWAS studies have been conducted to relate exposures to birth weight and foetal growth (Agier et al., 2020a; Govarts et al., 2018, 2016; Lenters et al., 2016; Nieuwenhuijsen et al., 2019; Woods et al., 2017), fecundity (Chung et al., 2019; Lenters et al., 2015; Louis et al., 2013), respiratory function (Agier et al., 2019), mortality (Patel et al., 2013), child obesity (Vrijheid et al., 2020) and blood pressure (Warembois et al., 2019). Note that many exposome-type or so-called “mixtures” studies have also been conducted, considering a few dozen exposures, typically for a couple of exposures families (Lenters et al., 2016; Philippat et al., 2019; Woods et al., 2017).

These ExWAS studies probably suffer from a lack of power: indeed, assessing numerous exposures implies costs which make these studies difficult to be performed on a number of individuals sufficient to have enough statistical power (Patel et al., 2017) and to avoid spurious correlation between exposures. This is expected to lead to false negative (low sensitivity) and false positive findings. Moreover, the use in ExWAS of the classical tool of the low-dimension epidemiology, i.e. univariate linear regression, can dramatically increase the “statistical” false-positive rates. This is known as the **‘multiple comparison issue’**: the higher the number of inferences made using an acceptance threshold for type I error (usually 5%), the more likely erroneous inferences are to occur (Benjamini and Hochberg, 1995). Moreover, **some exposures are intrinsically correlated, which makes true predictors of health outcomes difficult to differentiate from correlated exposures** (Slama and Vrijheid, 2015), which are thus predictors but not causal predictors : a simulation study under a realistic exposome setting showed that ExWAS (understood as an exposome-wide study using multiple univariate linear regression) false discovery proportion (FDP) increases with correlation (Agier et al., 2016).

Table I.1: Exposome project funded by the European Commission. All cited text comes from the website of Community Research and Development Information Service (CORDIS) resources.

European Funding Programm	Exposome Project	Aim	Which exposome	Population	Assessing biomarkers
FP7 (2012- 2017)	HELIX - Human Early Life Exposome (Vrijheid et al., 2014)	“To exploit novel tools and methods (remote sensing/GIS-based spatial methods, omics-based approaches, biomarkers of exposure, exposure devices and models [...]), to characterize early-life exposure to a wide range of environmental hazards, and integrate and link these with data on major child health outcomes”.	9 groups of 31 individuals exposures pairs from 6 cohorts assessed by biomarkers; and 5 groups of outdoor exposures : pregnancy exposures, assessed by models (216 exposures in total)	472 mother-child pairs from 6 cohorts : pregnancy and childhood.	Yes, on a subcohort of 1301 individuals: methylome, transcriptome, proteome, metabolome (Vrijheid et al., 2014)
FP7 (2012- 2017)	EXPOSOMICS	“To predict individual disease risk related to the environment, by characterizing the external and internal exposome for common exposures (air and drinking water contaminants) during critical periods of life, including in utero”.	An external exposome focusing on air pollutants and water contaminants, and an “internal exposome”, the metabolome. (Vineis et al., 2017)	Subpopulation of various cohorts, including both prenatal, child and later life cohorts.	Yes: metabolome (“internal exposome”), methylome, transcriptome (Vineis et al., 2017)
FP7 (2012- 2017)	HEALS - Health and Environment-wide Associations based on Large population Surveys	“To [refine] an integrated methodology and [apply] analytical and computational tools for elucidating human exposome through the integrated use of advanced statistical tools for environment-wide association studies in support of EU-wide environment and health assessments”	Both internal biomarkers and external exposures	(135 children cohorts, including twin cohorts (64 exposome) (Steckling et al., 2018)	No intermediate biomarkers

H02020 (2020-2024)	ATHLETE Advancing Tools for Human Early Lifecourse Exposome Research and Translation	“To develop a toolbox of advanced, next-generation, exposome tools and a prospective exposome cohort, which will be used to systematically quantify the effects of a wide range of community-level and individual-level environmental risk factors on mental, cardiometabolic, and respiratory health outcomes and associated biological pathways during the first 2 decades of life, to implement acceptable and feasible exposome interventions, and to translate the resulting evidence to policy recommendations and prevention strategies. »	“multiple environmental risk factors (external/urban, chemical, physical, behavioral, social)”	“15 cohorts in 10 European countries », “metagenomic, during the “20 first years of life”	Yes:
H02020 (2020-2024)	EXPANSE EXposome Powered tools for healthy living in urbAN SEttings	To “study the impact of the Urban Exposome on [...] Cardio-Metabolic and Pulmonary Disease”	Outdoor exposome, diet and individual exposome assessed by untargeted screening for exogenous chemicals	Existing exposome data on “2 million Europeans, and personalized Exposome assessment for 5,000 individuals”	Yes: metabolome to assess individual exposome
H02020 (2020-2024)	HEAP - Human Exposome Assessment Platform	To build an exposome assessment platform	Not yet described	Not yet described	Yes: epigenome and metagenome
H02020 (2020-2024)	REMEDIA	To “expose the impact of environmental factors on debilitating lung disease”	“exposures from the environment, diet, behavior and endogenous processes”	Existing cohorts, not yet described	No

H02020 (2020-2024)	HEDIMED Human Exposomic Determinants of Immune Mediated Diseases	To “identify exposomic determinants which are driving” immune-mediated diseases.	Not yet described	“A combination of data and biological samples from large clinical cohorts, including 350.000 pregnant women, 28.000 children prospectively followed from birth and 6.600 children from cross-sectional studies	Not yet described
H02020 (2020-2024)	EXIMIOUS Mapping Exposure-Induced Immune Effects: Connecting the Exposome and the Immunome	“To construct ‘immune fingerprints’ that reflect a person’s lifetime exposome and identify ‘immune fingerprints’ that are early signs of poor health and predictors of disease at the individual level”	Not yet described	Not described yet	Yes
H02020 (2020-2024)	EPHOR - Exposome project for health and occupational research	To “develop a working-life exposome toolbox”	The working-life exposome, defined as “all occupational and related non-occupational factors (general and socio-economic environment, lifestyle, behavior)”	Not yet described	Not yet described
H02020 (2020-2024)	Equal-Life - Early Environmental quality and life-	“To utilize the exposome concept in an integrated study of the external exposome and its social aspects and of	External exposome and factor	“A combination of birth-cohort data with physiological and physiological factor	Yes (not yet described)

		course mental health effects	measurable internal physiological factors and link those to a child's development and life course mental health"		new sources of data" (N=>250.000)
H02020 (2020-2024)	LONGITOOLS	To "study and measure how exposures to [the environment such as air and noise pollution and the built environment, and individual's lifestyle, psychological and social situation] contribute to the risk of developing [diseases such as obesity, type 2 diabetes and cardiovascular diseases] through a person's life	Not described yet	Not described yet	Yes: DNA methylation, RNA expression and read outs of metabolic pathways.

I.1.5. Considering the false positive rate issue in a structural causal framework

The problem of false positive findings, central in exposome studies, may gain to be considered within a structural causal framework. Indeed, a part of epidemiology is interested in finding the causal predictors of health, i.e. not factors which allow to predict an outcome, but factors which have a causal effect on the outcome. This distinction between risk prediction and causal inference, i.e. between risk predictors and causal factors (or “true predictors”), is crucial for later use of epidemiological findings in public health, as only causal predictors are relevant targets for clinical or public health interventions.

Identifying with accuracy true predictors of an event means **avoiding false-positive as well as false-negative results, i.e. requires to be both specific and sensitive**. When epidemiologists identify statistical associations between factors and an outcome in a sample to try to identify true predictors of this outcome in a population, they are likely to encounter two types of false-positive associations. The first type is the false positive due to sampling variability, the “random error”. The second type occurs **when a structural association which truly exists in the population is identifiable in the sample, but has no causal meaning**. Indeed, as described by Hernán et al. (2004), an association between two variables can occur in five cases:

a. One is cause of the other, i.e.:

a1. The variable of interest (typically the outcome) is influenced by the a priori explanatory variable (typically an exposure): if detected, the association corresponds to a true positive.

a2. The variable of interest causes the a priori explanatory variable: possibly leading to a false-positive finding due to reverse causality.

b. They share a common cause: this leads to a confounding bias, potentially creating false positive (or also false-negative) finding.

- c. They share a common consequence. If one controls for this consequence, a selection bias occurs, and can lead to false positive (or also false-negative) findings.
- d. By chance, due to sampling variability.

All situations a, b, and c correspond to structural associations but only case a1 corresponds to the identification of a causal predictor. Case e corresponds to a false positive “by-chance”: in this case, the observed association, has not only no meaning of causal association, but does not correspond to a structural association. The association then depends on the size of the study sample: chance associations become smaller with increased sample size whereas structural associations remain unchanged (Hernán et al., 2004). In the classical framework of statistical tests, this type of false positive association by chance corresponds to the type I error, i.e. rejecting as false the null hypothesis of independence whereas it is true in the source population. To this classification made by Hernan inside the causal framework, we must add the measurement error (Armstrong, 1998).

Thus, identifying the true predictors of a health outcome without false positive means, for the epidemiologist, to identify structural associations without false-positives (i.e. avoiding case d) and being able to point among these associations the ones which correspond to a causal link (i.e. distinguishing case a1 from the other cases a2, b and c).

Cases a2, b and c are biases and thus cannot be cured by increases in the sample size, but by options related to study design and statistical modeling (e.g., adjustment for confounding factors), supported by information on the causal structure or a priori information. Situation a2 of reverse causality is generally considered to be avoidable by external knowledge on the data generation process (e.g., the underlying biological mechanisms and the study design, including in particular the timing of assessment of E and Y).

Additionally, one should symmetrically avoid also false-negative findings, i.e. be able to identify a causal association when it exists. False negative associations can also occur by random fluctuations,

which, in the framework of statistical tests, correspond to type II error, which is linked to the power of the method of identification.

I. 2. Statistical techniques to address false positive and false negative challenges in exposome studies

I.2.1. Multiple testing correction

To face the above-mentioned challenges of the exposome related to the simultaneous consideration of multiple and possibly correlated exposures, relying on statistical methods more suitable to intermediate and high dimensional data than the classical regression model typically used in “single exposure” analysis (the ExWAS) can be seen as a solution.

Statistical methods have been developed to tackle the problem of by-chance findings. These techniques include correction for multiple testing and dimension reduction, in particular variable selection (Chadeau-Hyam et al., 2013).

Various **multiple testing correction techniques** have been developed to solve the multiple comparison problem, proposing to adapt the significance threshold of 5% most often used in univariate regression, making it stricter to compensate for the number of inferences being made. Two strategies can be distinguished: False-Discoveries-Rate (FDR) controlling procedures and Family-Wise error Rate (FWER) controlling procedures. FDR-controlling procedures control the expected proportion of "discoveries" (rejected null hypotheses) that are false, whereas FWER-controlling procedures control the probability of at least one Type I error, which is more conservative. They are widely used, and all the ExWAS studies that we cited in paragraph I.1.4 applied one of these controlling procedures. Three of the most common strategies of implementation are Benjamini and Hochberg (Benjamini and Hochberg, 1995) and Benjamini and Yekutieli (Benjamini and Yekutieli, 2001) for FDR-controlling techniques and Bonferroni (Dunn, 1961) for FWER-controlling techniques. The simulation by Agier et al. (2016) in realistic exposome settings showed that the addition of a FDR controlling procedure to independent linear regression (which is the most common implementation of the ExWAS method) allowed to decrease the FDP

from always more than 0.89 (without correction) to values between 0.67 and 0.93 (with correction). This decrease in FDP came together a strong decrease in sensitivity: indeed, the Benjamini-Yekutieli procedure used to adjust for multiple testing as well as the other procedures cited above assume that the tests are independent, which is in practical never the case in exposome studies, as exposures are often correlated. Correlation, additionally to create false-positive findings, also makes the number of effective tests performed lower than the number of associations tested, which decreases ExWAS power when a multiple comparisons correction technique is applied. Thus, adaptation of Bonferroni or Benjamini-Hochberg procedures taking into account a computed effective number of tests have been proposed (see for example (Li et al., 2012)) to target the decreased in power, but they also enhance the lack of specificity linked to correlation that we described in I.1.4.

I.2.2. Curse of dimensionality and dimension reduction techniques

To address the issue of false-positive hits linked to correlation and confounding, multiple regression appears as a relevant option: however, the size of current exposome studies prevents its use, as it is expected to be biased in such dimension (Sur and Candès, 2019). Here, an option is to use “dimension reduction techniques”. In fact, some problems encountered when dealing with intermediate or high dimensional data such as the exposome, known as the “curse of dimensionality” (a term first used by Bellman in 1961 (Bellman, 1961)), were the motivation of the development of dimension reduction techniques. Indeed, when the number of variables (and so the dimensionality of the features space) increases, the number of possible configurations increases too, making the configuration covered by an observation smaller compared to all possibilities. In other words, for a defined number of individuals, with more variables, the information may be richer, but it is also more diluted. In practice, this leads to several challenges when trying to extract information from the data: for example, the “vastness” of high-dimensional space often prevents algorithms based on similarity measures (k-neighbors, decision trees...) to work (Houle et al., 2010) and the number of samples needed to estimate an arbitrary function with a given level of accuracy

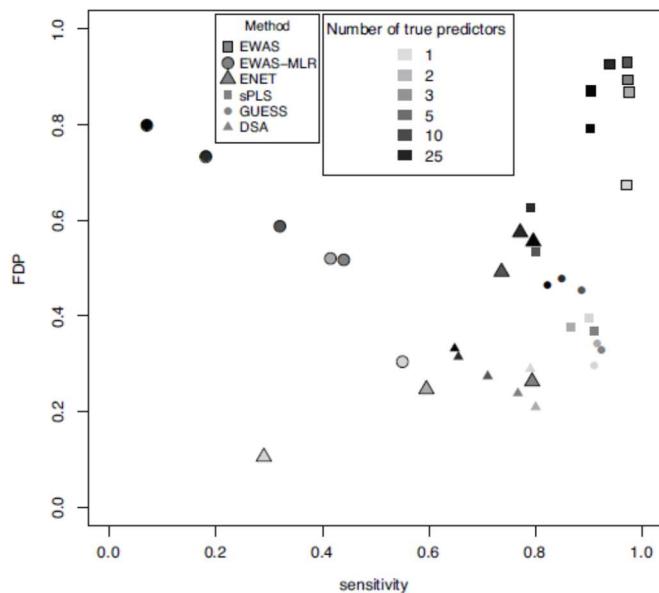
grows with the dimensionality. This explained why the classical maximum-likelihood estimator may be biased when the ratio of the number of independent variables to the number of individuals is typically of 0.2 or more (Sur and Candès, 2019), a ratio which is expected to be even lower if variables are correlated, which explains the difficulty to use multiple regression in current exposome studies. Overall when the number of observations is small compared to the number of features (intermediate or high dimension), dimension reduction may be needed to build good models, in particular for selection of causal predictors. Dimension reduction techniques belong to two major categories: selection techniques, which eliminate some variables while keeping the others, or extraction techniques, which create a set of new variables (Guyon and Elisseeff, 2003). Methods such as sparse Partial Least Squared Regression (sPLS) (Chun and Keleş, 2010) or Regularized Generalized Canonical Correlation Analysis (RGCCA) (Tenenhaus et al., 2017) are for example extraction techniques: they restrict to a small number of new covariates with low or null collinearity whose association with the outcome is assessed. Even if they allow to handle at least partially some of the challenges of false positive association in high dimension, their main drawback is their lack of interpretability (Lazarevic et al., 2019).

I.2.3. Multivariate variable selection methods

Multivariate variable selection methods could be more suitable to the exposome problem, as they may be able to **handle correlation and interactions** while allowing easy interpretability. Some simulation studies (Agier et al., 2016; Barrera-Gómez et al., 2017; Lazarevic et al., 2019; Lenters et al., 2018) have shown that under specific assumptions, methods such as LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996), ElasticNet (Zou and Hastie, 2005), Deletion-Substitution-Addition algorithm (DSA) (Sinisi and van der Laan, 2004), Weighted Quantile Sum regressions (WQS) (Carrico and Gennings, 2013) may allow to identify the true predictors of an health outcome among a set of exposures with good sensitivity and improved specificity (lower FDP) compared to ExWAS (see Figure I.2).

These methods are becoming more common in exposome studies and to some extent tend to replace the ExWAS. Since 2015, (Agier et al., 2020a, 2019; Forns et al., 2016; Lenters et al., 2016; Philippat et al., 2019; Vrijheid et al., 2020) for example, used at least one of the methods cited above.

Figure I.2: Sensitivity and FDP of 6 different statistical methods assessed from a Monte-Carlo simulation assuming a causal relationship between predictors drawn from a realistic exposome and a health outcome. Each predictor explained 3% of the variance of the outcome. DSA, Deletion/substitution/addition; ENET, elastic net; EWAS, environment-wide association study; EWAS-MLR, EWAS- multiple linear regression; GUESS, Graphical Unit Evolutionary Stochastic Search; sPLS, sparse partial least-squares (Agier et al., 2016).



However, these methods have some limits which should be acknowledged. First, none of them allows an accurate balance between sensitivity and specificity when detecting structural associations: the simulation by Agier et al. (2016) showed that, under specific hypotheses, regression-based selection methods have sensitivity that do not exceed 81% and false discovery proportion (FDP) rate which were at least at 34 %. Methods with the best sensitivity were also those with the highest FDP. More generally, in realistic settings, it appears very difficult to reach a FDP of about 5% without having a null sensitivity.

Moreover, stability is also a concern for most of these methods. In machine learning theory, **stability** is the notion that a small perturbation in the training dataset(s) will not change the learned

model, and thus the prediction of the learning algorithm (Poggio et al., 2004). Stability is directly linked to the generalization property of the algorithm (Poggio et al., 2004): intuitively, predictions robust to small perturbations are more likely to be good on a similar dataset. Whereas machine learning often focuses on prediction accuracy, environment epidemiologists, as already stated, are more interested in *feature* selection, as they want to identify causal predictors of health. In our case, stability should therefore be discussed considering the *stability of the subset of selected predictors*, defined in the machine learning field by Nogueira et al. (2017), rather than prediction stability. These may not be equivalent since it is possible that the predicted outcome (or risk) does not always change as the set of selected predictors changes. Instability lowers confidence in results and, as underlined previously (Lee et al., 2013; Nogueira et al., 2017), generalizability. This notion of stability is intrinsically linked to the problem of true predictors: a non-reproducible algorithm which will give non-identical results in term of selection in all different subsamples of a population has necessarily identified false positives, but a stable prediction can in some cases be achieved using different actors correlated with the true predictors, which is often the case in a high dimension setting.

In addition to showing the limits of such algorithms to avoid false positive associations, this also highlights a **fundamental difference between prediction and feature selection**: whereas accurate prediction can be achieved with appropriate data, sample size and learning algorithm, selection of causal predictors (or counterfactual prediction, i.e. prediction using causal predictor) requires additional information. Hernán et al. (2019) underlined that only an expert using a priori knowledge can be able to differentiate a causal predictor from a variate perfectly correlated: “causal analyses typically require not only good data and algorithms, but also domain expert knowledge” (Hernán et al., 2019).

Methods developed in epidemiology to avoid bias and reverse causality all imply a priori knowledge: for example, structural causal modelling (Pearl, 2009) assumes that the epidemiologist knows the

underlying causal structure; the assumption of lack of reverse causality in longitudinal studies comes from the additional information about the causal meaning of the time variable.

However, some methods try to infer causal structure from the data: for example, Bayesian structural learning, also known as causal probabilistic networks (Uusitalo, 2007), which uses Bayesian theory to find the causal structure which better fits the data. However, at least some prior knowledge is most often needed for Bayesian modelling, and as underlined by (Uusitalo, 2007), “theories about causal connections generally result in better models.”

I. 3. Adding biological information

I.3.1. Focusing a priori on a single exposure

An additional way to cope with the challenges of exposome studies related to the high false positive and false negative rate could be **the use of a priori information**. One way to do so would be to focus a priori on an exposure or a set of exposure of interest for example using knowledge from the toxicological field. For many health outcomes relevant to humans, however, a good animal model is lacking (e.g. asthma, autism...). Such studies could be nested in an exposome project, but such an approach is limited by available toxicological knowledge and does not take advantage of biological information at the level of the exposome. One option would be to use a priori information about intermediate biological layer.

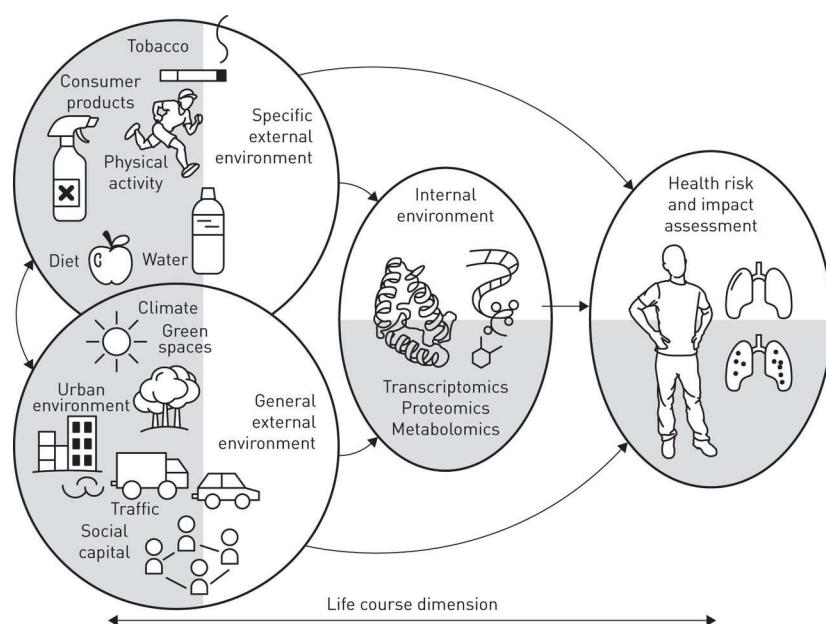
I.3.2. Information from intermediate biological layers

Several intermediary biological layers can be measured between the exposures and the health outcome considered: epigenome (DNA methylation), transcriptome (RNA), proteome and metabolome. They can show physiological responses to external exposures, thus constituting internal signatures of health outcomes. These ‘omics data, **as potential biomarkers of exposures early effect, or biomarkers of disease risk**, are possibly a precious but complex additional information about the link between exposures and health (Chadeau-Hyam et al., 2013; Crews and

Gore, 2011). Assessing the association between these biomarkers and the variation in exposures levels/health outcome may give some insight about how the health effect of one or more given exposure is biologically mediated (Ho et al., 2012).

DNA methylation, which is the addition of a methyl group in cytosine-guanine context (CpG site) on a DNA chain, depends both on genetic and environmental influences (Feil and Fraga, 2012; Marioni et al., 2018). At a biological level, DNA methylation is essential to control DNA transcription and thus cell differentiation, phenotype and functioning (Michalowsky and Jones, 1989). The influence of various early-life environmental factors on interindividual variation in methylation on specific loci (CpG sites) has been demonstrated (Baccarelli et al., 2009; Joubert et al., 2016). These epigenetic alterations can result in modified disease risk (Ho et al., 2012), even if there is so far little convincing estimates of the share of the effect of environmental factors on health mediated by epigenetic changes.

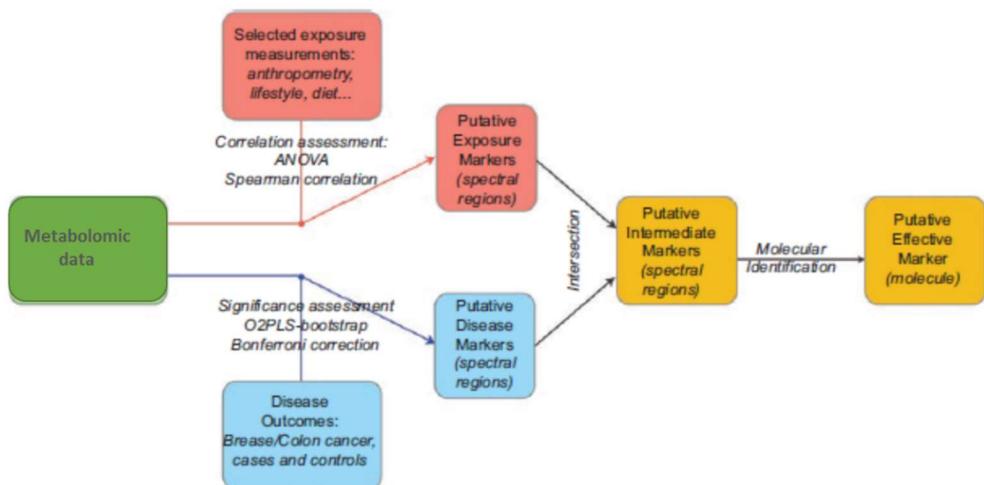
Figure I.3 : Schematic representation of links between exposome, intermediate biological layers and health effects. Figure from (Vrijheid, 2014).



Different strategies were developed to integrate these methylation, and more generally omics, data in epidemiology and use them to better understand mechanisms of environmental effect on health (Blum et al., 2020; Chadeau-Hyam et al., 2013). Due to their high dimension (for one individual, current methylation measures based on commercial arrays assess between 450,000 and 850,000 CpG sites) and their correlation structure, discovering biomarkers of interest for specific exposures and/or diseases is challenging. Dimension reduction techniques and multivariate analyses have, again, been used (see for example (Parkhomenko et al., 2007)).

A Meet-in-the-Middle framework has been developed by Chadeau-Hyam and colleagues in the context of studies considering a single exposure and single outcome to point intermediate biomarkers. The approach is often followed by a mediation analysis for these biomarkers: overlap between omics associated with exposure and outcome are considered as putative mediators (Chadeau-Hyam et al., 2011; Vineis and Perera, 2007).

Figure I.4: Implementation of Meet-in-the-Middle. Adapted from (Chadeau-Hyam et al., 2011).



Nowadays, these complex layers are paradoxically fairly well-known from biological studies and annotated; large database are now available about the functionality of genes/proteins/metabolites and the biological pathways, i.e. the biological network, in which they are involved, such as KEGG (<http://www.genome.ad.jp/kegg/>) (Tanabe and Kanehisa, 2012) or Gene Ontology (<http://www.geneontology.org/>) (Pavlidis et al., 2004).

I.3.3. Mediation analysis

The concept of biological mechanism can be framed in epidemiology with the notion of mediation (Vanderweele and Vansteelandt, 2009).

Mediation analysis aims at identifying the mechanisms (or pathways) through which an exposure E can influence an outcome Y (Figure I.5), and quantifying the importance of these pathways. More precisely, the theory of mediation (VanderWeele, 2011) assumes that there is a causal link between E and Y and that a potential mediator M of this effect is identified. The effect of E on Y which is not mediated by M is called the **direct effect** whereas, if M is indeed a mediator, the proportion of the association between E and Y that occurs through M is called the **indirect (or mediated) effect** of E (see Figure I.5).

In the case where M and Y are quantitative variables and residuals are normally distributed, two linear models representing these effects can be written (Baron and Kenny, 1986):

$$\mathbb{E}(Y) = \theta_0 + \theta_1 E + \theta_2 M + \theta_3 C \quad (\text{Exposure-outcome model})$$

where \mathbb{E} is the mathematical expectation, E the exposure variable and C a vector including all potential confounders of the exposure-outcome, exposure-mediator and mediator-outcome associations.

$$\mathbb{E}(M) = b_0 + b_1 E + b_2 C' \quad (\text{Exposure-mediator model})$$

where C' is a vector including all potential confounders of the exposure mediator association.

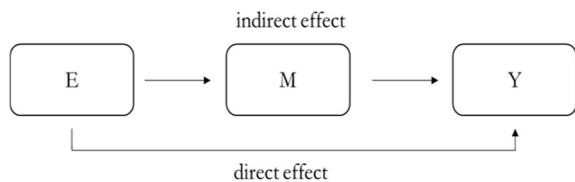
The estimation of direct and indirect effect, requires, beside the postulate that the effect from E to Y is causal, major assumptions (Vanderweele and Vansteelandt, 2009):

- the lack of uncontrolled confounders for all three mediator-outcome / exposure-mediator / exposure-outcome associations,
- the lack of mediator-outcome confounder affected by the exposure and of interaction between exposure and mediator.

If these hypotheses are met, θ_1 is an estimate of the direct effect, and $b_1 \times \theta_2$ is an estimate of the indirect effect (Baron and Kenny, 1986; MacKinnon et al., 2002; Vanderweele and Vansteelandt, 2009).

A well-known test for mediation is the causal inference test, which successively tests the significance of the overall exposome-outcome association not adjusted for M, b_1 and θ_2 (Baron and Kenny, 1986; MacKinnon et al., 2002).

Figure I.5: Causal graph of mediation.



CpGs sites are good candidate mediators between exposures and health. For example, Fasanelli and colleagues, estimated by ExWAS that in adults smoking reversibly caused hypomethylation on two specific CpGs, and that the total effect of smoking on lung cancer risk is mediated by more than 35% by methylation variation at these specific marks (Fasanelli et al., 2015).

However, considering the methylome layer instead of only some specific targeted CpGs complexifies a lot the mediation analysis. Among the challenges of the mediation analysis in high dimension (Barfield et al., 2017; Blum et al., 2020), the a priori causal knowledge on the relation between covariates, which is required for mediation analysis, is expected to be more difficult to decipher, in particular when the relations between potential mediators are complex, which is the case in the methylome layer (Blum et al., 2020).

I.3.4. Using intermediate biological layer to find true predictors of health within the exposome

Omics layers could provide an additional information source to study associations between the exposome and health and to overcome some challenges of the exposome and not only as a way to understand some already known associations. This is one of the key assumptions underpinning this PhD project. In particular, **information from epigenetic marks could be used to reduce the dimension of the exposome** with the hope to overcome some of the limits of purely agnostic exposome approaches: performing informed dimension reduction using the methylome, for example considering only exposures having influence on biological mechanisms relevant for the outcome of interest, could be a way to increase the power of exposome studies or decrease the false discovery proportion as well as to avoid some false positives due to structural association without causal meaning.

At another level, the restriction to candidate genes in studies of the association between the methylome and an outcome (see for example (Richmond et al., 2016)) can be seen as a similar strategy: using information from the genomic layer is expected to increase power as well as to test only the most a priori plausible CpGs.

Due to the complexity of the methylome layer and its high dimension, challenges encountered with the exposome can be even compounded when relying on methylome data to reduce the exposome dimension. Indeed, it could seem paradoxical to aim to use a high dimension layer (the methylome) in order to reduce the dimension of a low dimension layer (the exposome). However, it must be underlined that **part of the difficulties rising for the exposome do not hold when the aim of the use of the methylome is not the identification of causal relevant CpGs but only to inform the relation between exposome and health.** In particular, one can here adopt a risk prediction rather than a causal analysis logic, contrarily to what is sometimes aimed for in a Meet-in-the-Middle approach.

I. 4. Environmental effects on child weight

Such strategies can be used to better understand the causal relationship between early-life exposome and child health outcomes. During this PhD, we focused on child body mass index as our main outcome of interest, and also considered another measure of child growth, birth weight, as well as child lung function.

I.4.1. Environmental effects on child Body Mass Index

Childhood greater Body Mass Index (BMI), defined as the weight in kilogram divided by the squared height in centimeters, is associated with future risk of obesity as well as other risks of diseases, including type 2 diabetes, some cancers and cardiovascular diseases, lack of school achievement, and mental health problems (Han et al., 2010; Park et al., 2012; Quek et al., 2017; Singh et al., 2008).

Childhood obesity and overweight, which have increased rapidly in the three last decades (Finucane et al., 2011), are multifactorial conditions. Changes in the most important risk factors, genetic predisposition and energy imbalance (McAllister et al., 2009), are not sufficient to fully explain the magnitude and speed of the recent increase (Park et al., 2017). Other environmental factors influencing child obesity and adiposity have been identified. They include prenatal exposures, such as maternal smoking (Von Kries et al., 2002) and traffic noise exposure during pregnancy (Weyde et al., 2018) but also exposures occurring during early-life: exposures to some endocrine disruptors during first years of life (Agay-Shay et al., 2015; Holtcamp, 2012; Thayer et al., 2012), exposures to metals (Shao et al., 2017) and life-style factors such as physical activity and thus built environment characteristics like green spaces which contributes to it (Gascon et al., 2016; Lachowycz and Jones, 2011). Overall, the environmental obesogenic hypothesis states that these early exposures play a role in future obesity development by altering metabolic programming (Janesick and Blumberg, 2011; Park et al., 2017).

I.4.2. Environmental effects on birth weight

Birth weight is also a determinant of later health condition and is considered as a marker of the intrauterine environment (Belbasis et al., 2016); low birth weight has been associated with various later comorbidities, including metabolic diseases, cardiovascular diseases and cardiovascular risk factors (Belbasis et al., 2016) and respiratory health (Kindlund et al., 2010). Some evidence exists about the impact of prenatal exposures on birth weight: besides maternal tobacco (Windham et al., 2000) and alcohol consumption (Little, 1977; Mills et al., 1984; Strandberg-Larsen et al., 2017), higher temperature during pregnancy (Strand et al., 2011) and maternal exposure to air pollutants (Bell et al., 2010; Parker et al., 2005; Pedersen et al., 2013; Stieb et al., 2012) have been associated with lower birth weight. An effect of mother undernutrition has also been suggested (Stein and Lumey, 2000). Exposures to some phthalates (MEHHP and MOiNP), a perfluoroalkyl acid (PFOA), and an organochlorine (*p,p'*-DDE) have also been related with decreased birth weight in an exposome study (Lenters et al., 2016). Last, a study on Helix exposome data (Vrijheid et al., 2014) relying on both DSA and ExWAS method pointed the prenatal exposure to green area (NDVI, Normalized Difference Vegetation Index) as significantly associated with an increased birth weight (Nieuwenhuijsen et al., 2019), an association which has already been suggested by Dadvand et al. (2014) .

I. 5. PhD project's aim

Overall, we identified research needs both regarding the understanding of early environmental drivers of child growth and obesity and regarding efficient methods to identify them from the exposome and to improve exposome studies in general. In particular, even if some statistical methods have been pointed as being possibly more efficient than the classical ExWAS, methods currently used most often provide discouraging false-discovery rate when trying to detect the causal predictors of a health outcome (Agier et al., 2016). They are also expected to be prone to reverse

causality, especially when exposures biomarkers are assessed in a cross-sectional design. Last, the possibilities offered by the intermediary biological layers and the a priori knowledge available about them are for the moment not exploited in exposome studies as resources allowing to help pointing new causal predictors of a health outcome.

In this research, we aim at building novel strategies to inform the association between the exposome and a health outcome, and to propose insights about how exposome studies could better tackle the challenges related to high false positive rates and low sensitivity of exposome studies. We aim to do so by considering both “purely statistical” approaches and approaches incorporating (more) a priori biological information, considering specifically biological knowledge related to DNA methylation.

The objectives are both methodological, with the development of specific methods, and applied, aiming at informing the early environmental influences on child birthweight and later BMI.

In the second chapter of this report, we present an exposome study on the environmental determinants of BMI based on Helix data (Vrijheid et al., 2014), which used methylome data using a modified “Meet-in-the-Middle” (oMITM) approach that we developed. In the third chapter, we present a short study where this method is repeated on Helix data considering another child outcome, the lung function. We also provided in an appendix the preliminary results of an ongoing study on Sepages cohort (Lyon-Caen et al., 2019), repeating an oMITM design with a different implementation to study the relationship between prenatal exposome and birthweight, taking advantage of methylome data. After these studies based on real data, we present in the two last results chapters two simulations studies (chapters IV and V) aiming at validating the original approach we proposed in chapter II under various causal structure and to identify the most relevant implementation(s). Chapter IV focuses on the performance of the oMITM design under various causal structures and aims at understanding how the use of methylome data can help limiting the false positive rate. Chapter V presents a simulation study emphasizing the problem of instability

when using complex machine learning algorithms in epidemiology. In our last chapter, we discuss what insights our work can give about how environmental epidemiologists should deal with the dimension and the causality challenges of exposome studies.

A note on terminology: in the article detailed in chapter II (Cadiou et al., 2020), we proposed an innovative design adapted from an existing design usually called “Meet-in-the-Middle” in the literature (Chadeau-Hyam et al., 2011). In our published article reproduced in chapter II, we kept the term “Meet-in-the-Middle” (MITM) to call our adapted design. Later, we chose to rather use the term “oriented Meet-in-the-Middle” (oMITM) to underline the differences between our design and the classical Meet-in-the-Middle, a term which is used in all the other chapters of this thesis.

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

The work presented in this chapter is a study on the relationship between the exposome and child Body Mass Index on the data of the Helix project. To inform this relationship, methylome data are used in an innovative oriented Meet-in-the-Middle design, implemented with ExWAS-type methods. It has been published in Environment International:

Cadiou, S., Bustamante, M., Agier, L., Andrusaityte, S., Basagaña, X., Carracedo, A., Chatzi, L., Grazuleviciene, R., Gonzalez, J.R., Gutzkow, K.B., Maitre, L., Mason, D., Millot, F., Nieuwenhuijsen, M., Papadopoulou, E., Santorelli, G., Saulnier, P.-J., Vives, M., Wright, J., Vrijheid, M., Slama, R., 2020. “Using methylome data to inform exposome-health association studies: An application to the identification of environmental drivers of child body mass index.” Environment International. 138, 105622, doi:10.1016/j.envint.2020.105622

II. 1. Abstracts

II.1.1. English abstract

Background: The exposome is defined as encompassing all environmental exposures one undergoes from conception onwards. Challenges of the application of this concept to environmental-health association studies include a possibly high false-positive rate.

Objectives: We aimed to reduce the dimension of the exposome using information from DNA methylation as a way to more efficiently characterize the relation between exposome and child body mass index (BMI).

Methods: Among 1,173 mother-child pairs from HELIX cohort, 216 exposures (“whole exposome”) were characterized. BMI and DNA methylation from immune cells of peripheral blood were assessed in children at age 6–10 years. A priori reduction of the methylome to preselect BMI-relevant CpGs was performed using biological pathways. We then implemented a tailored Meet-in-the-Middle approach to identify from these CpGs candidate mediators in the exposome-BMI association, using univariate linear regression models corrected for multiple testing: this allowed to point out exposures most likely to be associated with BMI (“reduced exposome”). Associations of this reduced exposome with BMI were finally tested. The approach was compared to an agnostic exposome-wide association study (ExWAS) ignoring the methylome.

Results: Among the 2284 preselected CpGs (0.6% of the assessed CpGs), 62 were associated with BMI. Four factors (3 postnatal and 1 prenatal) of the exposome were associated with at least one of these CpGs, among which postnatal blood level of copper and PFOS were directly associated with BMI, with respectively positive and negative estimated effects. The agnostic ExWAS identified 18 additional postnatal exposures, including many persistent pollutants, generally unexpectedly associated with decreased BMI.

Discussion: Our approach incorporating a priori information identified fewer significant associations than an agnostic approach. We hypothesize that this smaller number corresponds to

a higher specificity (and possibly lower sensitivity), compared to the agnostic approach. Indeed, the latter cannot distinguish causal relations from reverse causation, e.g. for persistent compounds stored in fat, whose circulating level is influenced by BMI.

II.1.2. French abstract

Contexte : L'exposome est défini comme l'ensemble des expositions environnementales auxquelles on est exposé dès la conception. L'application de ce concept à l'étude des liens entre l'environnement et la santé pose des défis, notamment en raison d'un taux de faux positifs potentiellement élevé.

Objectifs : Nous avons cherché à réduire la dimension de l'exposome en utilisant les informations provenant de la méthylation de l'ADN, comme une façon de caractériser plus efficacement la relation entre l'exposome et l'indice de masse corporelle (IMC) de l'enfant.

Méthodes : Parmi 1 173 paires mère-enfant de la cohorte HELIX, 216 expositions ("exposome entier") ont été caractérisées. L'IMC et la méthylation de l'ADN des cellules immunitaires du sang périphérique ont été évalués chez les enfants à l'âge de 6 à 10 ans. Une réduction a priori du méthylome par préselection des CpG pertinents pour l'IMC à partir de banques de données de *pathways* a été effectuée. Nous avons ensuite mis en œuvre une approche *Meet-in-the-Middle* adaptée pour identifier au sein de ces CpG de potentiels médiateurs de relations entre expositions et IMC, en utilisant des modèles de régression linéaire univariés corrigés pour les tests multiples : cela a permis d'identifier les expositions les plus susceptibles d'être liées à l'IMC ("exposome réduit"). L'association de cet exposome réduit avec l'IMC a finalement été testé. L'approche a été comparée à une étude d'association agnostique à l'échelle de l'exposome (ExWAS) ignorant le méthylome.

Résultats : Parmi les 2284 CpG présélectionnés (0,6 % des CpG évalués), 62 étaient associés à l'IMC. Quatre facteurs (3 postnataux et 1 prénatal) de l'exposome étaient associés à au moins un de ces CpG, parmi lesquels les taux sanguins postnataux de cuivre et de PFOS étaient directement associés à l'IMC, avec des effets estimés respectivement positif et négatif. L'ExWAS agnostique a

identifié 18 expositions postnatales supplémentaires, dont de nombreux polluants persistants, généralement associés, de manière non attendue, à une diminution de l'IMC.

Discussion : Notre approche intégrant des informations a priori a identifié moins d'associations significatives qu'une approche agnostique. Nous émettons l'hypothèse que ce nombre plus faible correspond à une spécificité plus élevée (et peut-être à une sensibilité plus faible), par rapport à l'approche agnostique. En effet, cette dernière ne peut distinguer les relations causales de la causalité inverse, par exemple pour les composés persistants stockés dans les graisses, dont le niveau de circulation est influencé par l'IMC.

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

II. 2. Published article

Environment International 138 (2020) 105622



Full length article

Using methylome data to inform exposome-health association studies: An application to the identification of environmental drivers of child body mass index



Solène Cadiou^a, Mariona Bustamante^{b,c,d}, Lydiane Agier^a, Sandra Andrusaityte^e, Xavier Basagaña^{b,c,d}, Angel Carracedo^{f,g}, Leda Chatzi^h, Regina Grazuleviciene^e, Juan R. Gonzalez^{b,c,d}, Kristine B. Gutzkowⁱ, Léa Maitre^{b,c,d}, Dan Mason^j, Frédéric Millot^{k,l}, Mark Nieuwenhuijsen^{b,c,d}, Eleni Papadopoulou^j, Gillian Santorelli^j, Pierre-Jean Saulnier^{k,l,m,n}, Marta Vives^d, John Wright^j, Martine Vrijheid^{b,c,d}, Rémy Slama^{a,*}

^a Team of Environmental Epidemiology, IAB, Institute for Advanced Biosciences, Inserm, CNRS, CHU-Grenoble-Alpes, University Grenoble-Alpes, Grenoble, France

^b ISGlobal, Barcelona Institute for Global Health, Barcelona, Spain

^c Universitat Pompeu Fabra (UPF), Barcelona, Spain

^d CIBER Epidemiología y Salud Pública (CIBERESP), Spain

^e Department of Environmental Sciences, Vytautas Magnus University, Kaunas, Lithuania

^f Fundación Pública Galega de Medicina Xenómica (SERGAS), IDIS, Santiago de Compostela, Spain

^g Centro de Investigación Biomédica Red de Enfermedades Raras (CIBERER), CIBER, CIMUS, Universidad de Santiago de Compostela, Santiago de Compostela, Spain

^h Department of Preventive Medicine, University of Southern California, Los Angeles, USA

ⁱ Norwegian Institute of Public Health, Oslo, Norway

^j Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, United Kingdom

^k CHU Poitiers, Clinical Investigation Centre, CIC 1402, Poitiers, France

^l Poitiers University, Clinical Investigation Centre CIC 1402, Poitiers, France

^m INSERM, CIC 1402, F-86000 Poitiers, France

ⁿ CHU Poitiers, Endocrinology, Diabetology, Nutrition Service, Poitiers, France

ARTICLE INFO

Handling Editor: Shoji Nakayama

Keywords:

Biological a priori

Child body mass index – exposome

dimension reduction – DNA methylation

Reverse causality

ABSTRACT

Background: The exposome is defined as encompassing all environmental exposures one undergoes from conception onwards. Challenges of the application of this concept to environmental-health association studies include a possibly high false-positive rate.

Objectives: We aimed to reduce the dimension of the exposome using information from DNA methylation as a way to more efficiently characterize the relation between exposome and child body mass index (BMI).

Methods: Among 1,173 mother-child pairs from HELIX cohort, 216 exposures (“whole exposome”) were characterized. BMI and DNA methylation from immune cells of peripheral blood were assessed in children at age 6–10 years. A priori reduction of the methylome to preselect BMI-relevant CpGs was performed using biological pathways. We then implemented a tailored Meet-in-the-Middle approach to identify from these CpGs candidate mediators in the exposome-BMI association, using univariate linear regression models corrected for multiple testing: this allowed to point out exposures most likely to be associated with BMI (“reduced exposome”). Associations of this reduced exposome with BMI were finally tested. The approach was compared to an agnostic exposome-wide association study (ExWAS) ignoring the methylome.

Results: Among the 2284 preselected CpGs (0.6% of the assessed CpGs), 62 were associated with BMI. Four factors (3 postnatal and 1 prenatal) of the exposome were associated with at least one of these CpGs, among

Abbreviations: BMI, body mass index; BPA, bisphenol A; DDE, 4,4'dichlorodiphenyl dichloroethylene; DDT, 4,4' dichlorodiphenyltrichloroethane; DNA, Deoxyribonucleic acid; EDTA, ethylenediaminetetraacetic acid; ExWAS, exposome-wide association study; FDP, false discovery proportion; FDR, false discovery rate; HCB, hexachlorobenzene; MWAS, methylome-wide association study; PBDE, polybrominated diphenyl ether; PCB, polychlorinated biphenyl; PFNA, perfluorononanoate; PFOA, perfluorooctanoate; PFOS, perfluorooctane sulfonate; PFUNDA, perfluoroundecanoate; PM, particulate matter; POP, persistent organic pollutants; zBMI, z-score of body mass index

* Corresponding author at: Team of environmental epidemiology, IAB, Institute for Advanced Biosciences, Inserm, CNRS, CHU-Grenoble-Alpes, University Grenoble-Alpes, Allée de Alpes, Grenoble, France.

E-mail address: Remy.slama@univ-grenoble-alpes.fr (R. Slama).

<https://doi.org/10.1016/j.envint.2020.105622>

Received 10 September 2019; Received in revised form 27 February 2020; Accepted 28 February 2020

Available online 14 March 2020

0160-4120/ © 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/BY-NC-ND/4.0/>).

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

which postnatal blood level of copper and PFOS were directly associated with BMI, with respectively positive and negative estimated effects. The agnostic ExWAS identified 18 additional postnatal exposures, including many persistent pollutants, generally unexpectedly associated with decreased BMI.

Discussion: Our approach incorporating a priori information identified fewer significant associations than an agnostic approach. We hypothesize that this smaller number corresponds to a higher specificity (and possibly lower sensitivity), compared to the agnostic approach. Indeed, the latter cannot distinguish causal relations from reverse causation, e.g. for persistent compounds stored in fat, whose circulating level is influenced by BMI.

1. Introduction

The exposome concept recognizes that individuals are simultaneously exposed to a multitude of environmental factors from conception onwards (Wild, 2005). The exposome might explain an important, yet currently not accurately quantified, part of the variability in chronic diseases risk (Manrai et al., 2017). Since the 2010s, environmental epidemiology has progressively embraced the exposome concept and complemented common "single exposure studies" with studies relying on simultaneous measurements of several environmental factors (Agier et al., 2019; Buck Louis et al., 2011; Lenters et al., 2016), which, although not including all possible environmental factors, can be seen as examples of exposome studies. Amongst the many challenges faced by these exposome studies (Agier et al., 2016; Siroux et al., 2016) is a possibly high false discovery rate (Agier et al. 2016). Specifically, a simulation study considering a realistic exposome of 237 exposures assessed in 1200 individuals, with a proportion of the health outcome variability explained by the exposome varying between 3% and 70%, demonstrated that regression-based methods had a suboptimal sensitivity and high false discovery proportion (FDP) (Agier et al. 2016). All of the approaches tested displayed a FDP well above 5% when there was correlation within the exposome. The widely-used ExWAS (Exposome wide association study) approach, consisting in applying independent linear regression models and correcting association p-values for multiple testing, provided the highest sensitivity, at the cost of a very high FDP.

Dimension reduction could be a way to overcome this issue related to false positive findings. Dimension reduction can be performed agnostically, i.e. without relying on external information, with purely statistical techniques, such as variable selection via penalized regression (Lenters et al., 2018; Zou and Hastie, 2005) or Partial-Least-Square (PLS) regressions (Chun & Keles, 2010). However, as Agier et al. (2016) showed, these methods, even if they tend to perform better than ExWAS, are still expected to yield relatively high FDP.

Dimension reduction can also be biologically-driven. Relying on a priori information, one may integrate into statistical models relevant information from, for example, the toxicology and fundamental biology fields. Typically, this could be done by restraining analyses to exposures having biological plausibility, based on existing knowledge on associations with biological layers or on pathways linking exposures and the health outcome of interest. This logic has similarities with the concepts of *Mode of Action* and *Adverse Outcome Pathways* used in toxicology (OECD- Organisation for Economic Co-operation and Development, 2012; Vinken, 2013). DNA methylation can be regarded as such an intermediate informative layer, as it is expected to be under environmental (in addition to genetic) influences (Baccarelli et al., 2009; Feil and Fraga, 2012; Joubert et al., 2016; Marioni et al., 2018) and as these epigenetic alterations can result in modifications of disease risk (Ho et al. 2012, Fasanelli et al., 2015). Epigenetic mechanisms, defined as changes in a chromosome which result in heritable phenotype without alterations in the DNA sequence (Berger et al., 2009), have a key role in regulating transcription and thus cell differentiation, cell functioning, and they can ultimately influence the phenotype.

An option to identify biomarkers associated with both exposures and the health outcome from a single intermediate DNA methylation layer is the Meet-in-the-Middle approach. It has been developed to

point out intermediate biomarkers by considering as putative mediators the overlap between omics signals associated with an exposure and omics signals associated with the outcome (Chadeau-Hyam et al., 2011; Vineis and Perera, 2007).

In the case of an intermediate layer with a high dimension, one would have to test the associations of the intermediate putative biomarkers with exposures and health, which might entail a high false positive rate, in particular in the context of correlated exposures or biomarkers. It would appear relevant here to reduce the dimension of the intermediate layer, focusing on biological pathways (or intermediate biomarkers) a priori relevant for the outcome of interest.

In this study, we aimed to identify, in an exposome context, environmental factors associated with child BMI, by using information from child methylome layer to reduce the exposome dimension. Childhood obesity and overweight, whose prevalence has rapidly increased over the last three decades (Finucane et al., 2011), are multi-factorial conditions, and the most important risk factors, genetic predispositions and energy imbalance, may not suffice to fully explain the magnitude and rapidity of their recent prevalence increase (Park et al., 2017). The effects on BMI of some environmental factors, such as maternal smoking during pregnancy (Oken et al., 2008) or endocrine and metabolic disruptors exposures in early life (Thayer et al., 2012), have already been identified (Agay-Shay et al., 2015; Holtecamp, 2012). The environmental obesogenic hypothesis states that these early exposures play a key role in future obesity risk by altering metabolic programming (Janesick and Blumberg, 2011; Park et al., 2017). Only a few large multi-exposures (Braun, 2017; Fan et al., 2017; Lauritzen et al., 2018) or methylome-wide (Fradin et al., 2017; Rzehak et al., 2017) approaches to the study of child postnatal growth have been conducted.

2. Materials and methods

2.1. Overall strategy

We relied on the HELIX project, in which the exposome (pregnancy and childhood), the DNA methylome (from peripheral blood in childhood) and BMI were assessed in 1173 mother-child pairs (Haug et al., 2018; Tamayo-Uria et al., 2019). Biological information from genetic databases was used to a priori reduce the methylome dimension. We implemented a «Meet-in-the-Middle» approach to identify exposures sharing differentially methylated CpGs (i.e. methylation sites) with BMI, as a way to build a *reduced exposome*. The association of this reduced exposome with BMI was then tested.

More precisely, the approach consisted in 5 steps:

- a) preselection of CpGs located in genes relevant for BMI, using external databases;
- b) test of the associations between the methylation levels of these CpGs and BMI;
- c) test of the association between the methylation levels of the CpGs found to be associated with BMI in b) and each exposure, using child BMI as an adjustment factor, allowing to obtain a *reduced exposome*;
- d) test of the association between BMI and the *reduced exposome*;
- e) comparison with a purely agnostic ExWAS approach ignoring the methylome (i.e. without steps a) to c) allowing exposome dimension reduction, sensitivity analysis I).

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

We also implemented mediation analyses for the significant associations of the main approaches, sensitivity analyses testing the same approach without initial a priori selection of BMI-relevant CpGs, and additionally correcting for cell-type, as well as a sensitivity analysis considering the cell-types as the intermediate layer.

2.2. Study population and outcome

The study was part of the Human Early Life Exposome (HELIX) project (Maitre et al., 2018; Vrijheid et al., 2014), which aimed to describe the early-life exposome and its relations with child development and health.

In HELIX, six population-based European birth cohorts were pooled: BiB (Born in Bradford; United Kingdom) (Wright et al., 2013), EDEN (Étude des Déterminants pré et postnataux du développement et de la santé de l'ENfant; France) (Heude et al., 2016), INMA (Infancia y Medio Ambiente; Spain) (Guxens et al., 2012), KANC (Kaunas Cohort; Lithuania) (Grazuleviciene et al., 2015), MoBa (Norwegian Mother, Father and Child Cohort Study; Norway) (Magnus et al., 2016) and Rhea (Greece) (Chatzi et al., 2017), summing up to 1,301 mother–child pairs from singleton pregnancies for whom external exposures (Tamayo-Uria et al., 2019), health outcomes and confounders were measured and harmonized.

Height and weight were measured according to standardized procedures between 6 and 10 years of age (Maitre et al., 2018). BMI was calculated as the mass in kilograms divided by the squared height in meters. We used age- and sex-standardized z-scores (named hereafter zBMI) according to the international World Health Organization reference curves (de Onis et al., 2007) in order to allow comparison with other studies on child BMI and to take into account the age-related shift in BMI in childhood. Various lifestyle, social and anthropometric factors were additionally assessed (Table 1).

2.3. Exposome assessment

Details of the exposome assessment have been published elsewhere (Haug et al., 2018; Tamayo et al., 2018). Among the 234 exposures assessed in HELIX (Tamayo et al., 2018), we excluded exposures with a time window of one day or one week, which were a priori considered unlikely to influence BMI. This led to 216 prenatal and postnatal exposures (list available in [Supplementary Material 1](#)). Metals, organochlorines, organophosphate pesticides, polybrominated diphenyl ethers (PBDE), perfluorinated alkylated substances (PFAS), phenols and phthalates were assessed by biomarkers in mothers during pregnancy from one urine or blood sample and in children at the time of the clinical examination, from a pool of two urine samples or one blood sample (Casas et al., 2018; Haug et al., 2018). Built environment exposures, indoor air exposures, lifestyle factors, meteorological data, natural spaces quantification, noise, traffic, socio-economic capital and concentrations of disinfection by-products in drinking water were assessed during pregnancy and during the year before child examination by environmental models and questionnaires.

Exposures were transformed to approach normality, using a Box-Cox power transformation approach that chooses among log-transforming or raising the data to the powers $-2, -1, -0.5, 1/3, 0.5, 1$, or 2 . Transformation chosen for each variable is detailed in [Supplementary Material 1](#). Exposures were standardized using their interquartile range after imputing missing data for all exposures using mice R package (Buuren and Groothuis-Oudshoorn, 2011).

2.4. DNA methylation

Peripheral blood was collected in EDTA tubes during the clinical examination that took place when children were between 6 and 10 years old. DNA was extracted from buffy coat; DNA methylation was assessed with the Infinium Human Methylation 450 beadchip

(Illumina), following the manufacturer's protocol. Sample locations on chips were drawn at random balancing chips for cohort and infant sex. Some samples were analysed in duplicate and a control HapMap sample was added in each 96-well plate.

DNA methylation data were pre-processed using the minfi R package (Aryee et al., 2014). A first quality control of the data was done with MethylAid package (van Iterson et al., 2014); probes with low call rate were then filtered following guidelines of Lehne et al. (2015). The functional normalization method was further applied, including Noob background subtraction and dye-bias correction (Triche et al., 2013). Several quality control checks were performed: sex consistency using the shinyMethyl package (Fortin et al., 2014); consistency of duplicates; genetic consistency for the samples that had genome-wide genotypic data. Finally, duplicated samples and control samples were removed as well as probes to measure methylation levels at non-CpG sites (Jang

Table 1
Characteristics of the 1,173 mother–child pairs from HELIX cohort.

Characteristic	Mean (SD)	n (%)
Child BMI (kg/m ²)	16.9 (2.7)	
Child sex		
Female		529 (45)
Male		644 (55)
Child age (years)	7.9 (1.5)	
Cohort		
BiB		203 (17)
EDEN		146 (12)
INMA		215 (18)
KANC		198 (17)
MoBa		212 (18)
RHEA		199 (17)
Maternal education		
Low		176 (15)
Middle		402 (34)
High		595 (51)
Maternal pre-pregnancy BMI (kg/m ²)	25.0 (5.0)	
Parity before index pregnancy		
0		530 (45)
1		430 (37)
2 or more		213 (18)
Trimester of conception		
January–March		368 (31)
April–June		234 (20)
July–September		260 (22)
October–December		311 (27)
Maternal tobacco smoke pregnancy exposure		
None		624 (53)
Only passive exposure		374 (32)
Smoker		175 (15)
Child postnatal tobacco smoke exposure		
Not exposed		745 (64)
Exposed		428 (36)
Maternal age (years)	30.7 (4.9)	
Birthweight		
less than 2500 g		40 (4)
2500 to 3500 g		662 (56)
3500 to 4000 g		357 (30)
≥ 4000 g		114 (10)
Breastfeeding duration		
less than 10.8 weeks		361 (31)
10.8 to 34.9 weeks		419 (36)
greater than 34.9 weeks		393 (34)
Parents born in the country of inclusion		
None		134 (11)
Only one		58 (5)
Both		981 (84)
Ethnicity		
African		7 (1)
Asian		19 (2)
European ancestry		1048 (89)
Native American		2 (0)
Pakistani		79 (7)
Other		18 (2)

et al., 2017). A final filtering was performed to eliminate probes with a single-nucleotide polymorphism (SNP), probes that cross-hybridize and probes on sex chromosomes, restricting to 386,518 CpG probes available for 1,192 subjects. The study was performed on the 1173 subjects among them who also had valid exposures data.

We then used Combat procedure to remove the batch effects supported by the slide. Methylation levels were expressed as Beta values (average methylation levels for an individual, between 0 for a never methylated CpG site and 1 for an always-methylated CpG site).

Cell types were computed according to Houseman et al. (2012) algorithm and Reinius reference panel (Reinius et al., 2012); tests of associations including methylation levels were corrected for cell types only in a sensitivity analysis.

Correlation within the methylome was estimated by averaging the Pearson's correlation within 10 sets of 2284 randomly selected CpGs (same size as the restricted methylome, see next paragraph), to avoid computing all pairwise correlations between the 386,518 CpGs.

2.5. A priori preselection of BMI-relevant CpGs

Biological pathways a priori relevant for BMI were selected using the KEGG database (Tanabe and Kanehisa, 2012), searching with the key words "growth" "obesity" and "fat" in the following categories: "Human energy metabolism", "Human lipid metabolism", "Human endocrine system", "Human digestive system", "Human Excretory system", "Human endocrine and metabolic diseases", "Human genetic Information Processing: 'Transcription – Translation - Folding, sorting and degradation - Replication and repair' ". We thus identified a list of 16 pathways (Supplementary Material 2) and the corresponding list of genes, which were restricted to the CpGs identified as enhancers,

leading to a final dataset of 2284 CpGs belonging to 387 genes and 16 different biological pathways (Supplementary Material 2), which we further refer to as the "restricted methylome". Correspondence between genes and CpGs as well as enhancer annotation and CpGs was based on Illumina annotation (Hansen, 2016). A sensitivity analysis not restricted to enhancers and these 16 pathways was performed.

2.6. Meet-in-the-Middle and ExWAS approaches – Statistical analyses

Our Meet-in-the-Middle design itself consisted in three successive steps, as described in 2.1 (steps b, c and d): in step b), we tested the association of the methylation levels of the preselected CpGs with BMI considered as the outcome; in step c), we tested the associations (adjusted for child BMI) of each exposure with the CpGs found to be associated with BMI in b), leading to the identification of a "reduced exposome"; step d) is the test of the association of this reduced exposome (the exposures found to be associated to some CpGs in step c)) with the outcome.

Univariate linear regression models were applied, and p-values were corrected for multiple testing using a FDR (False-Discovery Rate) control procedure (Benjamini & Hochberg, 1995) at all steps involving regression modelling. Adjustment factors coded linearly in all our regression analyses were maternal pre-pregnancy BMI (additionally coded with a quadratic term in the exposome-outcome associations test), maternal education, maternal age, parental country of birth, maternal smoking during pregnancy, cohort (fixed effect), parity, trimester of conception, child age and child sex (see Table 1 for the categories). We additionally adjusted analyses of postnatal exposures effects for birth weight, breastfeeding duration and passive smoking during childhood (see Table 1) and models including methylation data

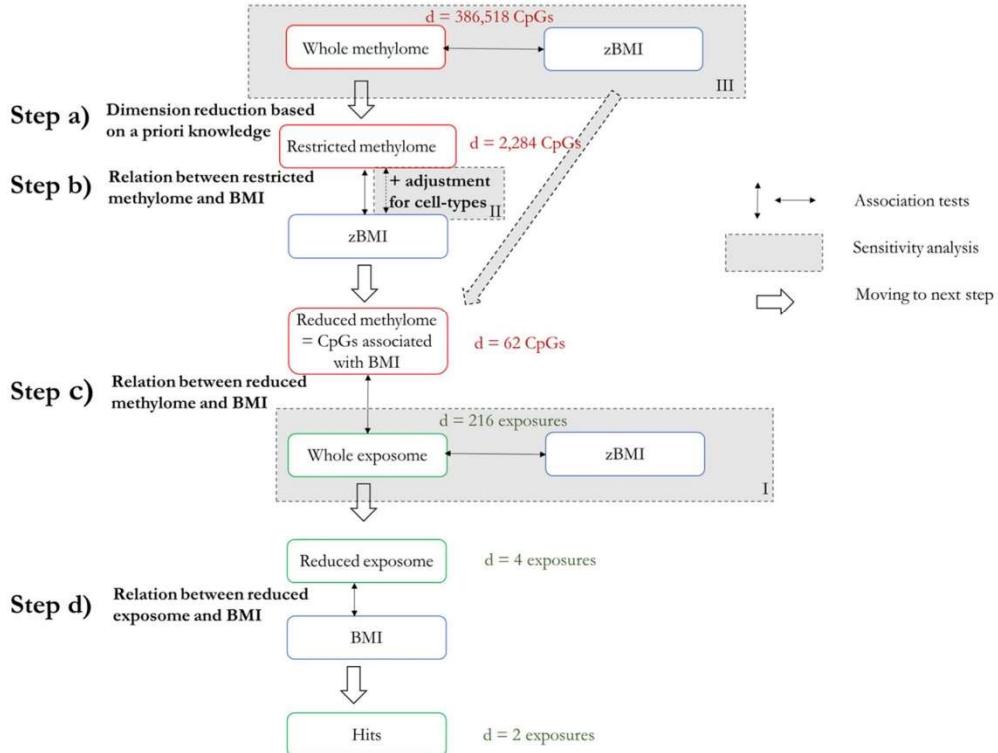


Fig. 1. Workflow of the main statistical analyses. (in color; 1-column fitting image).

for ethnicity (self-reported by parents, with different questions across the cohorts). At step c), correction for multiple testing was done considering together all associations tested between exposures and CpGs, i.e. a number of test equal to the product of the number of exposures

with the number of CpGs associated with zBMI. A mediation analysis using package MMA (Yu and Li, 2017) was performed for exposures found associated with the outcome in step d), considering the CpGs both associated with the exposure and the outcome.

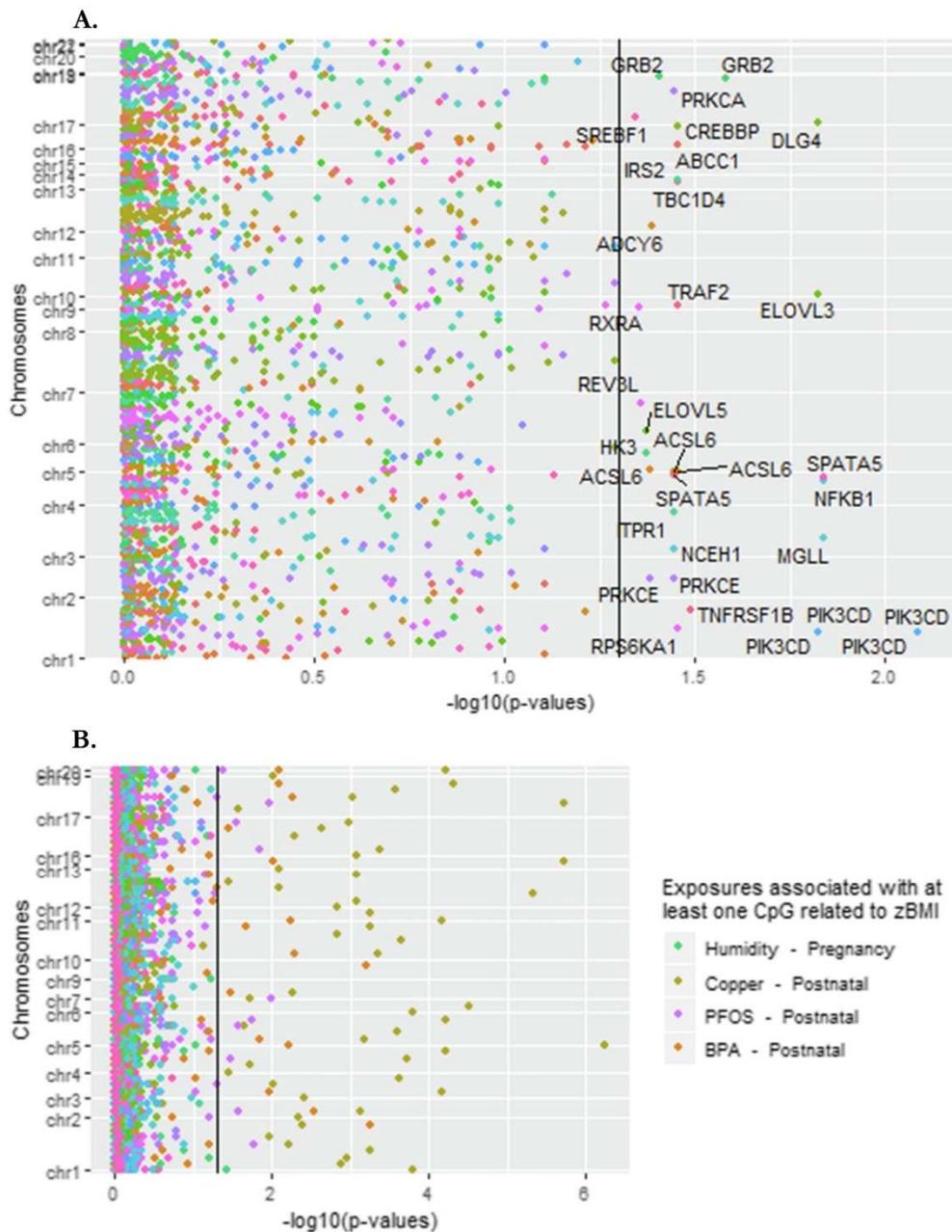


Fig. 2. Manhattan plots of the FDR corrected p-values of adjusted associations obtained with the Meet-in-the-middle approach applied on the reduced methylome at steps b) and c). A: Associations between preselected CpG and zBMI. Each colour corresponds to a gene. The black vertical line shows the (FDR-corrected) 0.05 significance level. Lowest p-value: 3.20×10^{-3} . B: Associations between exposures and CpGs associated with childhood zBMI. Each color corresponds to a different exposure. The black vertical line shows the (FDR-corrected) 0.05 significance level. Lowest p-value: 5.66×10^{-7} . BPA: Bisphenol A; PFOS: Perfluorooctanesulfonic acid.

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

2.7. Sensitivity analyses and test of selection relevance

We compared our results to those obtained with a *totally agnostic* approach, consisting in an ExWAS between the exposome and zBMI, ignoring the methylome, with exactly the same statistical methods as in our step *d*) (sensitivity analysis I). Our agnostic ExWAS has some differences with that performed by Vrijheid et al. (2020) in Helix data (submitted manuscript, available upon request). In particular, we restricted the population of 1301 children used by Vrijheid et al. (2020) to 1173 children with methylome data (see paragraph 2.4.), and we chose to additionally adjust for trimester of conception, ethnicity and pre- and postnatal smoking, which could influence the methylome. We additionally ran two agnostic (prenatal and postnatal) multivariate linear regression models simultaneously adjusted for the whole exposome and potential confounders and corrected for multiple testing.

We performed three other sensitivity analyses. First, we repeated our approach with an additional adjustment on cell-type heterogeneity for all association tests involving the methylome (sensitivity analysis II). Second, we repeated our approach on the unrestricted methylome, i.e. without the first step of a priori selection of CpGs using information from biological pathways database and annotation (i.e. removing step *a*) and starting from step *b*) with 386,518 CpGs). This modified step *b*) corresponds to a methylome-wide analysis, or MWAS (methylome-wide association study) in which we tested the association between methylation levels of the whole unrestricted methylome and zBMI (sensitivity analysis III).

To further inform the relevance of the a priori CpG selection of step *a*), the proportion of CpGs belonging to our candidate list of a priori BMI-relevant CpGs (i.e. our restricted methylome) among CpGs whose methylation levels was found associated with zBMI by MWAS was compared to the corresponding proportion in the whole methylome.

A workflow of the statistical analyses is shown Fig. 1.

We additionally performed a fourth sensitivity analysis (sensitivity analysis IV) by repeating the whole Meet-in-the-Middle approach considering the cell-types instead of methylation data as the intermediate layer.

3. Results

3.1. Population characteristics

At the time of BMI measurement, mean age was 7.9 years. Mean child BMI was 16.9 kg/m² (5th and 95th percentiles: 13.8; 22.4), with substantial differences between cohorts (Supplementary Material 3), INMA and RHEA showing higher zBMI compared to the other cohorts. The other characteristics of the study population are given Table 1 and by cohorts in Supplementary Material 4.

The mean levels of the 216 exposures considered are displayed in Supplementary Material 1. Mean absolute correlation between quantitative exposures was 0.11 (5th and 95th percentiles: 0.00, 0.35); the distribution of the coefficients of correlation is given in Supplementary Material 5.

Within the whole methylome, the estimated mean correlation was 0.09 while it was 0.12 for the 2284 CpGs of the reduced methylome

(5th and 95th percentiles: 0.00; 0.37).

3.2. Meet-in-the-Middle approach

The analysis testing the association between the restricted methylome and zBMI (step *b*) identified 62 CpGs belonging to 43 different genes (FDR adjusted p-values ≤ 0.05 ; Supplementary Material 6). The mean correlation among these CpGs was 0.59. Fig. 2A shows a Manhattan plot of the FDR-adjusted p-values.

The test of association of these 62 CpGs with each of the 216 environmental factors adjusted for child BMI (step *c*) identified 4 exposures associated with at least one CpG (Fig. 2B, Table 2, Supplementary Material 7): copper (postnatal level), BPA (Bisphenol A, postnatal level), PFOS (Perfluorooctanesulfonic acid, postnatal level) and one meteorological variable (humidity, pregnancy average); this constituted our *reduced exposome*. In total, 53 CpGs were associated with at least one exposure.

The last step (step *d*) identified that within the reduced exposome, postnatal blood copper and PFOS levels were directly associated with zBMI. The corresponding estimated parameters were respectively 0.22 (95% CI: 0.14; 0.30; adjusted p-value, 1.43×10^{-6}) and -0.13 (95% CI: -0.23 ; -0.04 ; adjusted p-value, 0.02) (see Table 3 for the other components of the reduced exposome). A mediation analysis quantified that for copper, the 52 CpGs mediated 29% of the total effect of postnatal blood copper level on zBMI, while for PFOS, the 12 selected CpGs mediated 28% of the total effect of postnatal blood PFOS level.

3.3. Agnostic exposome-wide approach

An agnostic ExWAS using FDR correction for multiple testing between the whole (not reduced) exposome and zBMI identified 20 postnatal exposures significantly associated with zBMI (Sensitivity analysis I, Table 4). These included postnatal copper and PFOS level (also identified at step *d*) of the main approach). In addition to metals and perfluorinated alkylated substances (PFAS), these exposures belonged to the organochlorines, polybrominated diphenyl ethers (PBDE), lifestyle and indoor air pollution families. Organochlorines, PBDE and PFAS compounds, as well as postnatal cobalt levels showed negative regression coefficients, corresponding to a decreased zBMI with increasing exposure levels. The most significant associations were observed for 5 of the postnatal PCB levels, which formed a group with higher correlation (mean absolute correlation, 0.50) than the rest of the quantitative exposome (mean absolute correlation, 0.11). When applying a multiple linear regression model simultaneously adjusted for the whole exposome and potential confounders, 2 (postnatal) variables were selected after multiple testing correction: copper (positive parameter) and HCB (negative parameter, Supplementary material 8).

3.4. Other sensitivity analyses

In order to determine if our a priori CpGs selection led to a concentration of information, we quantified the overrepresentation of our preselected BMI-relevant CpGs among the discoveries of a methylome-wide analysis linking the whole methylome to zBMI. As expected, the

Table 2

Number of CpGs associated with both exposures and zBMI in the adjusted associations between the exposome and CpGs associated with zBMI in 1,173 children from the HELIX cohort (ExWAS model adjusted on zBMI, step *c*) of the Meet-in-the-Middle approach applied on the reduced methylome. Results are presented only for exposures associated with a (stringently corrected for multiple hypothesis testing) p-value of less than 0.05 in exposure-CpGs ExWAS-type analyses, with CpGs being previously selected in a CpGs-zBMI ExWAS-type analysis. *Details of the CpGs and genes are given in Supplementary Material 6.

Exposure	Number of CpGs associated both with the exposure and zBMI	Number of corresponding genes*
Copper (postnatal)	52	37
Bisphenol A (BPA) (postnatal)	15	14
Perfluorooctanesulfonic acid (PFOS) (postnatal)	12	12
Humidity average (pregnancy)	1	1

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Table 3

Adjusted associations between the reduced exposome and zBMI in 1,173 children from HELIX cohort (ExWAS model, step d) of the Meet-in-the-Middle approach applied to the reduced methylome. "Significant" associations are indicated in bold.

Group	Label	Unit	Transformation	Effect estimate*	95% CI	Uncorrected p-Value	FDR-corrected p-Value
Meteorological	Humidity (pregnancy)	%	None	0.05	-0.34; 0.44	0.81	0.81
Phenols	BPA (postnatal)	µg/g	Log2	-0.07	-0.14; 2.8x10 ⁻⁴	0.05	0.07
Metals	Copper (postnatal)	µg/L	Log2	0.22	0.14; 0.30	3.57x10⁻⁷	1.43x10⁻⁶
Perfluorinated alkylated substances (PFAS)	PFOS (postnatal)	µg/L	Log2	-0.13	-0.23; -0.04	7.69x10 ⁻³	0.02

* Adjusted change in mean zBMI for each unit increase in transformed exposure level. Models were adjusted for maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for postnatal exposures birth weight, passive smoking during childhood and breastfeeding duration.

CpGs associated with zBMI in an MWAS considering the whole exposome were enriched in enhancers CpGs selected as being relevant for BMI from KEGG database (1.22%) compared to the other CpGs (0.46%, a ratio of 2.6 to 1). However, most significant associations found by the MWAS were not part of the a priori selected list of CpGs (1760 out of 1788, [Supplementary Material 9](#)).

When the whole Meet-in-the-Middle was repeated without the step of CpGs preselection using external biological information (sensitivity analysis III), final results differed from those obtained using the restricted methylome: additionally to blood postnatal copper and PFOS levels (which were also found significant in this analysis), blood postnatal hexachlorobenzene (HCB), Pentabromodiphenyl ether (PBDE) 153 and dichlorodiphenyltrichloroethane (DDT) levels (with negative associations with zBMI) and blood postnatal caesium level (with a positive association with zBMI) were identified in step d) ([Table 5](#)). These 5 exposures were also associated with zBMI in the agnostic ExWAS approach (sensitivity analysis I). To give more details, 1788 out of 386,518 CpGs were associated with zBMI in step b) ([Fig. 3A](#) and [Supplementary Material 10](#)). In step c) of the analysis, 28 exposures were significantly associated with at least one of these 1788 CpGs. Among them, postnatal blood levels of copper, BPA and PFOS ([Supplementary Material 11](#) and [Fig. 3B](#)) and prenatal humidity exposure, which had all been previously found in the main analysis, were associated with respectively 110, 449, 180 and 47 CpGs. All the other exposures were associated with less than 10 CpGs.

When we repeated our analysis adding a correction for blood cell-types (sensitivity analysis II), no association was significant at step b) (lowest p-value with Benjamini-Hochberg correction: 0.72). When we repeated the analysis corrected for blood cell-types without the pre-selection step, we found one association between the whole methylome and zBMI, but the corresponding CpG (cg02032125) was not associated with any exposure at step c) so that no exposure was eventually selected as associated with BMI.

When we considered the cell-types instead of the methylation data as our intermediate layer, (sensitivity analysis IV), results were very similar to those of the main analysis: the reduced exposome consisted in three exposures, postnatal blood copper and BPA levels and average pregnancy humidity exposure. In the last step, only copper level was associated with zBMI. The three cell-types associated with both copper and zBMI mediated 13% of the effect of copper on zBMI. Detailed results of sensitivity analysis IV are available in [Supplementary Material 12](#).

4. Discussion

We implemented a modified Meet-in-the-Middle approach among 1,173 mother-child pairs to identify components of the exposome influencing child BMI through DNA methylation changes. The analysis highlighted postnatal copper blood level as being positively associated with zBMI, an association supported by changes with copper levels in the methylation levels of 52 CpGs from genes that are relevant for BMI

based on a priori knowledge. Blood perfluorooctanesulfonic acid postnatal level was also found related to zBMI in our Meet-in-the-Middle approach, an association likely due to reverse causality.

Our work is one of the first studying the link of an exposome including both chemical and nonchemical stressors during the prenatal and postnatal time windows with child BMI. Beside Helix studies, the largest studies considering multiple chemical exposures and child BMI considered up to 27 components ([Agay-Shay et al., 2015](#); [Fan et al., 2017](#); [Lauritzen et al., 2018](#)).

The efficiency of our Meet-in-the-Middle approach to detect true predictors of zBMI within the exposome relies on three main assumptions: 1) that part of the effects of the exposome on child BMI are mediated by changes in methylation levels that can be observed from peripheral blood; 2) that these methylation changes are strong enough to be detectable and that they can be used to select plausible exposures and thus reduce the exposome dimension; 3) that existing databases of biological pathways and regulatory regions (enhancers) allow to relevantly reduce the dimension of the methylome a priori to study its association with BMI. We discuss here the relevance of these three assumptions, as well as our choice not to correct for cell-types heterogeneity.

4.1. Are some effects of the exposome on child BMI likely to be mediated by the methylome?

Methylation has been reported to mediate part of the effect of specific exposures on health. This has been suggested for example for smoking effects on health: [Fasanelli et al. \(2015\)](#) pointed methylation mediation for smoking effects on lung cancer and [Wahl et al. \(2018\)](#) showed that site-specific methylation can mediate the effect of smoking on the expression of inflammatory proteins. For BMI, evidence of mediation arises from animal toxicological studies, which showed that long-term obesity risk can result from effects of early overfeeding/underfeeding mediated by methylation changes on specific regulatory CpG sites in different tissues ([Carone et al., 2010](#); [Lillycrop et al., 2008, 2005](#); [Plagemann et al., 2009](#)). In humans, studies based on subjects who experienced famine during intra-uterine life suggested that blood cells methylation could mediate effects of prenatal undernutrition on later overweight: early-life exposures to famine had an effect on CpGs regulating growth and metabolic mechanisms involved in obesity ([Heijmans et al., 2008](#); [Tobi et al., 2009](#)). In addition, such a mediation between prenatal exposure to famine and adult metabolic traits has been statistically demonstrated; indeed, studies ([Tobi et al., 2018a;2018b](#)) pointed out that even if, from a biological point of view, DNA methylation measured in peripheral blood was not likely to be a causal mediator of BMI change, it could be a proxy of epigenetic regulation changes in specific tissues, and thus allow to detect mediation. For early exposures other than nutrition, little evidence is currently available regarding an effect on BMI or growth mediated by epigenetic changes ([Richmond et al., 2015](#)): however, [Cao-Lei et al. \(2015\)](#) suggested that some gene-specific methylation could mediate part of the

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Table 4
Sensitivity analysis I: adjusted associations between the whole exposome and zBMI in 1,173 children from the HELIX cohort (ExWAS agnostic approach, ignoring the methylome). Results are presented only for exposures with a (FDR-corrected for multiple hypothesis testing) p-value lower than 0.05.

Exposure group	Exposure variable	Unit	Transform-matn	Effect estimate*	95%CI	Uncorrected p-value	FDR-corrected p-value
Organochlorines	PCB 180 - Postnatal	ng/g lipid	Log ₂	-0.92	-1.05	-0.78	2.14x10 ⁻³⁸
Organochlorines	HCB - Postnatal	ng/g lipid	Log ₂	-0.66	-0.76	-0.57	5.08x10 ⁻³⁸
Organochlorines	PCB 170 - Postnatal	ng/g lipid	Log ₂	-0.82	-0.95	-0.70	9.56x10 ⁻³⁷
Organochlorines	Sum of PCBs - Postnatal	ng/g lipid	Log ₂	-0.80	-0.92	-0.67	7.15x10 ⁻³⁴
Organochlorines	PCB 138 - Postnatal	ng/g lipid	Log ₂	-0.67	-0.78	-0.55	5.56x10 ⁻³⁸
Organochlorines	PCB 153 - Postnatal	ng/g lipid	Log ₂	-0.70	-0.82	-0.58	1.29x10 ⁻²⁷
Organochlorines	DDF - Postnatal	ng/g lipid	Log ₂	-0.54	-0.64	-0.43	3.42x10 ⁻²²
Organochlorines	DDDE 153 - Postnatal	ng/g lipid	Log ₂	-0.40	-0.52	-0.28	8.70x10 ⁻¹¹
Organochlorines	PPDE 118 - Postnatal	ng/g lipid	Log ₂	-0.28	-0.38	-0.19	1.39x10 ⁻⁹
Organochlorines	PCB 118 - Postnatal	ng/g lipid	Log ₂	-0.21	0.13	0.30	3.33x10 ⁻⁸
Metals	Copper - Postnatal	µg/L	Log ₂	-0.24	-0.33	-0.15	5.67x10 ⁻⁷
Perfluorinated alkylated substances	PFQA - Postnatal	µg/L	Log ₂	-0.27	-0.38	-0.16	5.29x10 ⁻⁷
Organochlorines	DDT - Postnatal	ng/g lipid	Log ₂	-0.19	-0.28	-0.10	1.11x10 ⁻⁶
Perfluorinated alkylated substances	PFNA - Postnatal	ng/L	Log ₂	-0.21	-0.32	-0.09	3.65x10 ⁻⁵
Perfluorinated alkylated substances	PFUNDA - Postnatal	ng/L	Log ₂	-0.20	0.09	0.31	4.06x10 ⁻⁴
Metals	Cesium - Postnatal	µg/L	Log ₂	-0.12	-0.19	-0.05	6.27x10 ⁻³
Metals	Cobalt - Postnatal	µg/L	Log ₂	-0.12	-0.12	-0.05	7.52x10 ⁻⁴
Tobacco Smoke	Active smoking - Pregnancy	—	—	0.34	0.13	0.55	1.45x10 ⁻³
Indoor air	Indoor PM _{2.5} - Postnatal	µg/m ³	Log	0.13	0.04	0.21	2.95x10 ⁻³
Indoor air	Indoor PM _{2.5} - Postnatal	µg/m ³	Log	0.11	0.04	0.19	2.94x10 ⁻³
Perfluorinated alkylated substances	PFOS - Postnatal	µg/L	Log ₂	-0.14	-0.24	-0.04	4.59x10 ⁻³

* Adjusted change in mean zBMI for each unit increase in transformed exposure level. Models were adjusted for maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for postnatal exposures birth weight, passive smoking during childhood and breastfeeding duration.

effect of prenatal maternal stress child BMI and central adiposity. Less directly, several exposures were identified as possibly influencing epigenetic marks, and some of these alterations may occur on genes involved in signaling pathways controlling growth and adipose tissue development (Richmond et al., 2015). Importantly, effects in the opposite causal directions are also likely, in that changes in BMI could influence methylation levels on specific loci, as suggested by Dekkers et al. (2016) and Richmond et al. (2016).

Our approach, as well as the agnostic ExWAS, identified postnatal blood copper level as positively associated with zBMI and additionally with changes in BMI-relevant CpGs. The mediating effect that we estimated was of 29% of the estimated total copper effect. Copper is an essential trace element involved via oxidoreduction reactions in a broad range of processes, including energy expenditure, mitochondrial respiration, antioxidant defences and inflammation (Tisato et al., 2010). Human copper intake is most often due to presence of copper in drinking water, food or vitamin supplement (Brewer, 2010; Pal et al., 2014) and is known to influence blood copper level (Silverio Amancio et al., 2003; Uauy et al., 1998). Elevated copper concentrations have been observed in many diseases, including cancer, Alzheimer and metabolic diseases (Brewer, 2010; Salustri et al., 2010; Squitti et al., 2009; Tisato et al., 2010). Specifically, a positive link between copper level and high BMI or obesity in children has been previously described in the same data from HELIX (Vrijheid et al., 2020) and elsewhere (Fan et al., 2017; Lima et al., 2006; Yakinci et al., 1997). These studies in children are cross-sectional and an important question relates to the direction of any causal link between copper level and obesity. Overweight might disrupt copper level, for example due to a higher food intake linked to an increased appetite, as hypothesized by Yakinci et al. (1997), or due to metabolic changes. Other arguments exist in favour of copper being a proximal cause of overweight. Nutrition studies in human showed that changes in copper intake (depletion or supplementation) can have adverse health effects such as metabolic and cardiovascular abnormalities (Klevay, 2018; Milne and Weswig, 2018). The toxicity of copper (Brewer, 2010) and its ability to induce oxidative stress are well-known in humans (Brewer, 2010; Uriu-Adams and Keen, 2005) and from animal models (Galhardi et al., 2004; Pereira et al., 2016). Part of this process can occur via methylation changes, as shown in zebrafish, in which stress-related gene expression can be modified by early-life copper exposure (Dorts et al., 2016). Although we cannot formally exclude a situation in which copper levels are influenced by the child's overweight status (e.g. as in Fig. 4D), in particular due to the cross-sectional design of our study of DNA methylation-BMI links, the above-mentioned experimental and prospective studies make copper a plausible causal biomarker or predictor of BMI, with clues for effects possibly mediated by methylation changes (as in Fig. 4A).

On the contrary, the negative and less strong association of PFOS level with zBMI may correspond to reverse causality (see below). An influence of PFOS levels on blood or serum methylation change has some plausibility (Ruiz-Hernandez et al., 2015; Watkins et al., 2014).

4.2. Can information be borrowed from the blood methylome to reduce the dimension of the exposome?

Dimension reduction is one of the possible cures of the curse of dimensionality (Jimenez and Landgrebe, 1998). Assuming that part of the effects of the exposome on BMI are mediated by the methylome, and that blood methylome constitutes a proxy of methylation levels in other target organs, identifying exposures associated with methylation changes on CpGs relevant for BMI is a way to restrict the analysis to a subgroup of exposures with higher likelihood of having an effect on BMI. If the a priori CpG selection is accurate, one can expect the reduced exposome to contain a higher proportion of true predictors of BMI than the full exposome. In a situation of expected limited power, testing only the association of this reduced exposome with BMI could lead to a better specificity (fewer false positives) than an agnostic

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Table 5
Sensitivity analysis III - Meet-in-the-Middle without CpGs preselection based on KEGG database: adjusted association between the reduced exposure and zBMI in 1,173 children from the HELIX cohort (ExWAS model, step d) of the Meet-in-the-Middle approach applied on the whole methylome). Bold lines indicate significant associations.

Exposure group	Exposure variable	Unit	Transform-mation	Effect estimate*	95%CI	Unadjusted p-value	FDR adjusted p-value
Organochlorines	HCB - Postnatal	ng/g lipids	Log2	-0.56	-0.76; -0.56	7.87×10^{-38}	2.20×10^{-36}
Polybrominated diphenyl ethers	PBDE 153 - Postnatal	ng/g lipids	Log2	-0.40	-0.52; -0.28	2.26×10^{-9}	3.57×10^{-6}
Metals	Copper - Postnatal	µg/L	Log2	0.22	0.14; 0.30	3.33×10^{-7}	4.24×10^{-6}
Organochlorines	DDT - Postnatal	ng/g lipids	Log2	-0.28	-0.39; -0.17	6.06×10^{-7}	4.76×10^{-3}
Metals	Caesium - Postnatal	µg/L	Log2	0.19	0.08; 0.30	8.50×10^{-4}	0.04
Perfluorinated alkylated substances	PFOS - Postnatal	µg/L	Log2	-0.13	-0.23; -0.04	7.69×10^{-3}	
Phenols	BPA - Postnatal	µg/g	Log2	-0.07	-0.14; 0.33	0.05	0.20
Phthalates	OH-1-MNP - Pregnancy	µg/g	Log2	-0.07	-0.14; 0.04	0.06	0.23
Phthalates	MHP - Postnatal	µg/g	Log2	-0.07	-0.16; 0.01	0.10	0.31
Perfluorinated alkylated substances	PFHxS - Postnatal	µg/L	Log2	-0.10	-0.22; 0.02	0.12	0.33
Built Environment	Population density - Postnatal	people / km ²	Square root	0.06	-0.02; 0.15	0.15	0.38
Lifestyle	Soda intake - Postnatal	Times/ week	Tertiles	-0.10	-0.26; 0.05	0.18	0.42
Socio-eco capital	Social participation - Postnatal	-	None	-0.08	-0.24; 0.07	0.28	0.60
Lifestyle	Fastfood intake - Pregnancy	Times/ week	Tertiles	0.13	-0.14; 0.39	0.34	0.66
Phthalates	MIBP - Postnatal	µg/g	Log2	-0.05	-0.16; 0.06	0.35	0.66
Phthalates	MRP - Postnatal	µg/g	Log2	-0.03	-0.11; 0.04	0.40	0.70
Organochlorines	PCB 138 - Pregnancy	ng/g lipids	Log2	-0.05	-0.18; 0.08	0.48	0.78
Organochlorines	Sum of PCBs - Pregnancy	ng/g lipids	Log2	-0.05	-0.20; 0.11	0.55	0.80
Air Pollution	PM2.5 - Pregnancy	µg / m ³	None	-0.04	-0.17; 0.10	0.57	0.80
Noise	Traffic noise (24 h) - Postnatal	dB(A)	None	-0.06	-0.26; 0.13	0.54	0.80
Meteorological	Humidity - Pregnancy	%	None	0.05	-0.34; 0.44	0.81	0.85
Organochlorines	PCB 170 - Pregnancy	ng/g lipids	Log2	-0.03	-0.16; 0.10	0.69	0.85
OP Pesticides	DETP - Pregnancy	µg/g	Log2	0.02	-0.09; 0.12	0.77	0.85
Lifestyle	Vegetables intake - Pregnancy	Times/ week	Tertiles	-0.02	-0.18; 0.15	0.85	
Natural Spaces	Green spaces (300 m) - Pregnancy	-	None	-0.02	-0.17; 0.14	0.84	0.85
Phenols	PRPA - Pregnancy	µg/g	Log2	0.01	-0.09; 0.12	0.82	0.85
Metals	Thallium - Postnatal	Times/ week	None	-0.04	-0.29; 0.21	0.78	0.85
Lifestyle	Yogurt intake - Postnatal	Times/ week	Tertiles	0.04	-0.14; 0.21	0.69	0.85

* Adjusted change in mean zBMI for each increase by 1 in transformed exposure level. Models were adjusted for maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for postnatal exposures birth weight, passive smoking during childhood, breastfeeding duration.

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

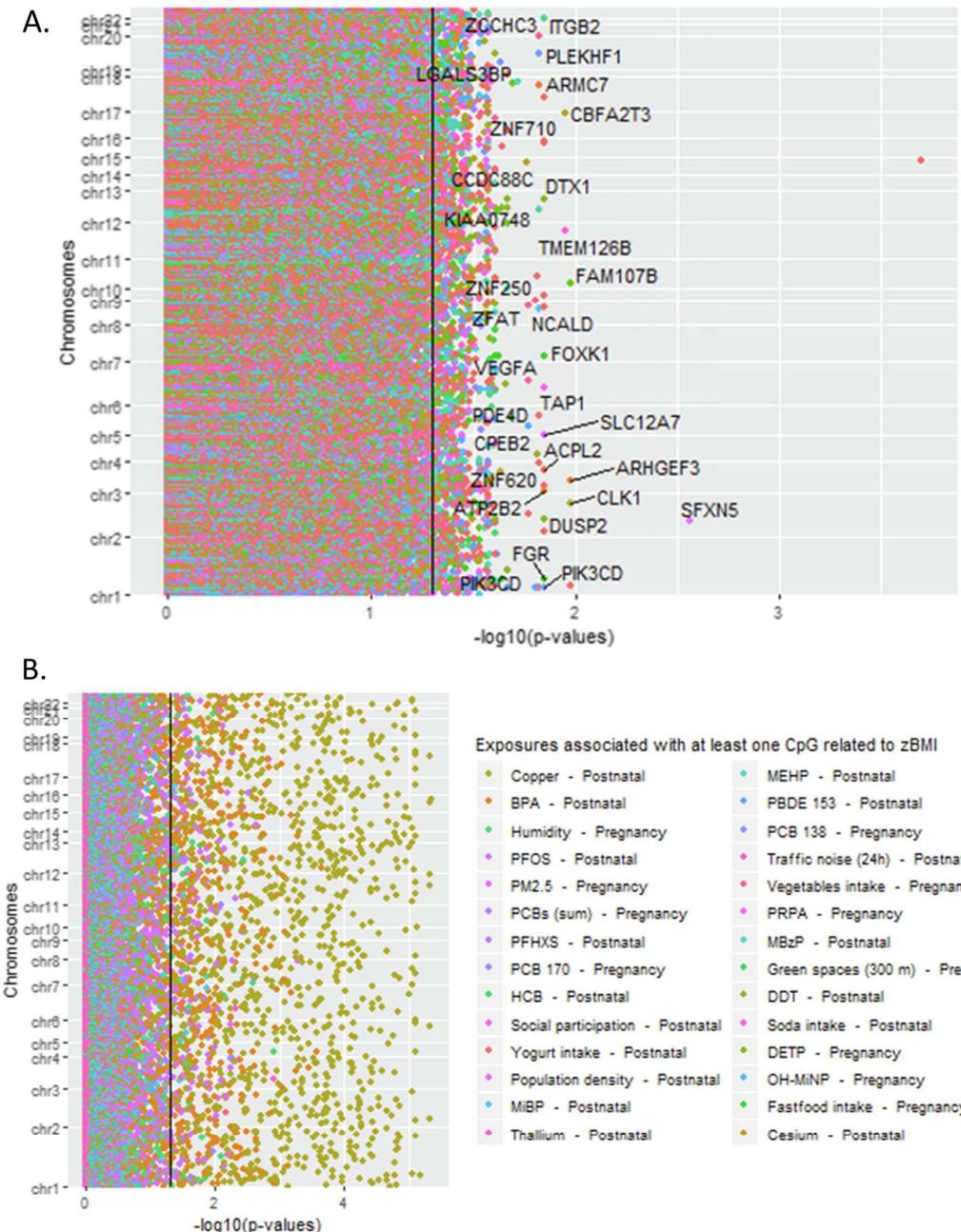


Fig. 3. Sensitivity analysis III - Meet-in-the-Middle approach without a priori preselection of CpGs: Manhattan plots of the FDR corrected p-values of adjusted associations obtained with the Meet-in-the-middle approach applied on the whole methylome at steps b) and c). A: Associations between all CpG and zBMI. Each colour corresponds to genes. The black line is the (FDR-corrected) 0.05 significance threshold. Lowest p-value: 2.04×10^{-4} . B: Associations between exposures and CpGs associated with childhood zBMI. Each colour corresponds to a different exposure. Lowest p-value: 4.51×10^{-6} .

ExWAS performed on the whole exposome, which may suffer from high FDP because of the correlation within the exposome (Agier et al., 2016). This approach however comes at the cost of possibly excluding true BMI predictors within the exposome whose effect on BMI are not identifiable

from the blood methylome. Yet, we considered a high FDP to be of greater concern for exposome studies than low sensitivity.

Compared to the classical Meet-in-the-Middle framework (Vineis et al., 2013), we additionally adjusted for the outcome when testing for

associations between exposures and CpGs. This adjustment was meant to exclude some cases of spurious association between the exposome and the methylome. Thus, we used this tailored Meet-in-the-Middle approach with a different goal than mediators identification, to focus on a subset of the exposome relevant for the considered outcome, with the ultimate goal to increase specificity.

We used ExWAS-type methods to identify our reduced exposome, which has a possibly high sensitivity and an expected high false positive rate (Agier et al. 2016). This could make our reduced exposome possibly inaccurate, containing exposures selected by chance even if the test is adapted to the underlying causal structure. However, our reduced exposome was considerably smaller than the full exposome (4 exposures vs. 216) and our results provided far fewer discoveries (2 vs. 20) than an agnostic ExWAS ignoring the methylome data. This lower number of discoveries is not consistent with our approach having a higher rate of spuriously mediated exposures. Rather, it could be explained by our approach having a lower FDP and/or a lower sensitivity. To discuss these hypotheses, we compare the plausibility of the results of the Meet-in-the-Middle approach and of the agnostic ExWAS obtained in the same population.

As discussed before, copper, identified by both approaches, is a plausible causal predictor of BMI mediated by methylation. Among the 19 other exposures significantly associated with zBMI in the agnostic approach, four associations corresponded to positive slopes: postnatal blood caesium level, prenatal maternal active smoking and postnatal indoor particulate matter ($PM_{2.5}$) concentration and absorbance. The last three associations may be (at least partly) due to the well-known effect of prenatal and postnatal smoking on the obesity (Oken et al., 2008; Vázquez Nava et al., 2006): indeed, postnatal smoking variables were used to compute indoor particulate matter levels. An odd ratio greater than 1 was also reported for the influence of high urinary caesium levels on diabetes, an obesity-related outcome (Menke et al., 2016). The remaining 15 associations corresponded to negative slopes (that is, a lower BMI with increasing exposure levels assessed in child

blood): it was the case for postnatal blood levels of perfluorinated compounds, cobalt and of persistent organochlorine compounds (some PCBs, DDT (an insecticide), its metabolite DDE, and HCB). For some of these exposures, associations of prenatal levels with overweight or obesity have been reported in the literature. PFAS, including PFOS, prenatal exposures have been associated with *higher* BMI in childhood and adulthood (Braun, 2017; Lauritzen et al., 2018; Saikat et al., 2013); obesogenic effects of early-life PCBs levels have also been reported (Heindel and vom Saal, 2009; Thayer et al., 2012). However, for childhood exposure, several studies found *negative* associations of PCBs (Rönn et al., 2011) and PFAS (Nelson et al., 2010) with BMI, similarly to our results. These negative associations may be indicative of reverse causality. This is supported by the facts that 1) the postnatal exposome and outcome were assessed simultaneously; 2) lipophilic compounds such as PFAS, PCBs and DDT are stored in fat, which makes the blood level a possibly inaccurate marker of exposure: studies on seals showed that, for identical levels of exposures, higher persistent organic pollutants (POP) levels in blood are found in thin compared to fatter animals (Debiec et al., 2006; Lydersen et al., 2002). In humans, Rönn et al. (2011) found positive associations of fat mass with blood levels of lightly chlorinated PCBs with fat mass and negative association for highly chlorinated PCBs, which are more lipophilic and therefore more stored in fat. This is in favour of the negative associations between POPs (persistent organic pollutants) levels and BMI being explained by fat levels influencing the blood POPs levels (causal models C to K, Fig. 4), rather than by POPs influencing adiposity (causal models A or B). Unfortunately, we could not access fat tissue POP concentrations, which may have been a better exposure biomarker.

The Meet-in-the-Middle approach did not select PCBs, DDT, DDE, HCB and other POPs highlighted by the agnostic ExWAS. This may be due to our approach coping more efficiently with reverse causality. A recent simulation study showed that Sobel's test of mediation, under specific hypotheses, has far better detection rates for true mediation effects than in a reverse causality situation (Tobi et al., 2018). Our

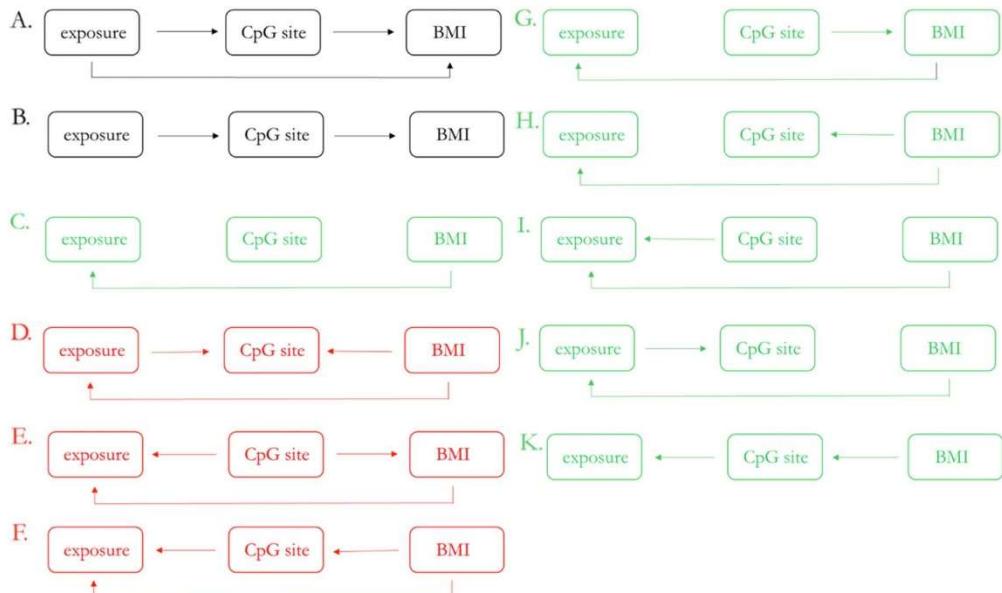


Fig. 4. Different causal models involving one exposure, one CpG site and the outcome (BMI). Among the causal models corresponding to the exposure-BMI link corresponding to reverse causality (C to K), the models in which the Meet-in-the-Middle approach is expected to be able to provide a truly negative result are displayed in green, and causal models in which our Meet-in-the-Middle approach can be expected to provide a false-positive result are displayed in red. (in color; 1.5-column fitting image). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

approach, which has similarities to a mediation test, is moreover specifically designed to efficiently handle some situations of reverse causality.

Fig. 4C to K shows all possible causal situations corresponding to reverse causality between an exposure (E) and zBMI. An ExWAS approach, which cannot distinguish between E influencing zBMI and zBMI influencing E, is expected to detect all of them (assuming perfect power), whereas our Meet-in-the-Middle approach is not expected to conclude that E and zBMI are linked in cases C and I. Moreover, since we adjusted our second test (the exposome-methylome test) for zBMI, cases G and H, which seem very plausible from a biological point of view, should not be erroneously detected by our approach either (indeed, there is no exposure-CpG association conditionally on BMI in models G and H). Finally, as methylation is not likely to influence exposure levels in another way than via its link with the level of fat and as we preselected CpGs in pathways likely to link exposures and BMI, cases E, F and K may not be very frequent situations. Case D however cannot be excluded and may be erroneously detected by our approach, which may be the case for PFOS in our study. We can thus hypothesize that our Meet-in-the-Middle approach is likely to (erroneously) identify a link between E and Y in fewer causal models corresponding to reverse causation than the ExWAS approach. This may limit the number of false-positive findings compared to the agnostic ExWAS.

Regarding now the sensitivity of our approach, it strongly depends on the proportion of exposures truly affecting BMI for which one of the underlying mechanisms relates to changes in blood DNA methylation. If the effect of most exposures is (at least partly) reflected in the blood methylome, then we can expect a high sensitivity, empowered by the smaller dimension of the reduced exposome compared to the whole exposome, which makes the correction for multiple testing less penalizing than in the agnostic ExWAS. If on the contrary most exposures affecting BMI do so by pathways unrelated to the methylome, then our approach is expected to have a low sensitivity. We did not find any association with prenatal exposures to endocrine disruptors (some of which are possible obesogens (Braun, 2017)) or with dietary factors such as soda intake, whose effects on BMI are well documented (Heo et al., 2017; Hu and Malik, 2010; Murakami and Livingstone, 2016) but may not be visible from the blood methylome. Overall, the sensitivity of our approach will depend on the (unknown) proportion of exposures whose effect is mediated by the methylome. Whether the likely lower sensitivity is considered to be compensated by the expected decrease in FDP would depend on the specific nature of the exposome study, with e.g. confirmatory studies putting the emphasis on the limitation of false positive signals.

4.3. Is a preliminary dimension reduction of the methylome necessary to efficiently borrow information from it?

We tried to identify a relevantly reduced exposome with ExWAS-type methods applied to methylome data. ExWAS is known to have a high-false positive rate, particularly if correlation exists among predictors (Agier et al. 2016). To cope with this problem, we modified the original Meet-in-the-Middle framework by testing the exposome-CpGs associations only for CpGs which we had found to be associated with zBMI. This drastically reduced the number of tests done at step c) (10,152 tests versus 83,487,888, a division by 8200) and thus increased power for this step. Moreover, we also performed a preliminary reduction of the size of the methylome, relying on external biological information from the KEGG pathways database (Tanabe and Kanehisa, 2012). Our preselection of CpGs was, again, drastic, as it reduced the methylome from 386,518 to 2284 CpGs, possibly allowing a gain of power to test associations between methylation levels and zBMI, under strong assumptions. These assumptions relate to the quality and completeness of the KEGG database (and of our query) to identify BMI-relevant genes and to the quality of the ILLUMINA annotation about enhancers CpGs and on the link between genes and CpGs. These can be

questioned: in particular the KEGG database is based on publications of various quality, and pathways selected may not be relevant for 6–10-year-old children. Moreover, ILLUMINA annotation is not tissue-specific whereas enhancer characteristic is; it is consequently unclear whether the ILLUMINA enhancer tag is relevant for blood immune cells, on which methylation was assessed. Finally, changes in the methylation level of the enhancer CpGs of one gene may not be linked with the protein level of this same gene, but of a remote gene (Jang et al., 2017).

To try to quantify the impact of our CpGs preselection, we used two approaches. First, we performed a sensitivity analysis, in which the Meet-in-the-Middle approach was performed without step a) of CpGs preselection. This led to quite different results. Although, in the last step, postnatal levels of copper and PFOS were again found to be associated with zBMI, 4 other exposures were additionally significantly negatively associated with zBMI: postnatal blood levels of HCB, PBDE 153, DDT and caesium. Contrarily to copper and PFOS levels, which were associated with respectively 1110 and 180 CpGs at step c), these four additional exposures were each associated with not more than 2 CpGs. Moreover, as discussed above, for the organochlorine compounds, their association with BMI may be due to reverse causality. Thus, the discoveries of this analysis without CpGs preselection contained more compounds little likely to be causative predictors of the zBMI than the discoveries of the main analysis. This might be explained by the high number of false positives expected from an ExWAS-type method in high dimension, at steps b) and c). The dimension reduction of the methylome in our case seemed to help to cope with this problem. Further studies are needed to determine if this impact of CpGs preselection is generally expected or not.

To further test the quality of our a priori CpGs preselection, we tried to establish if our a priori reduction led or not to concentration of information, by quantifying the overrepresentation of our preselected CpGs in the significant associations found by a methylome-wide analysis relating the whole methylome to zBMI. The information was indeed concentrated (apparently higher specificity of the selected CpGs), but at the cost of a considerable loss of information: giving up at least 1760 CpGs associated with zBMI implies a risk not to identify some exposures truly associated with CpGs which are themselves associated with zBMI, and consequently may decrease the sensitivity of our approach (Supplementary Material 9). This illustrates, again, that our approach, which was built to gain in specificity, may in principle have a cost in terms of loss in sensitivity compared to the agnostic approach.

4.4. Correction for cell-type heterogeneity

Correction for the proportion of the cell-types in which DNA methylation is assessed is now applied in most methylome studies, although its relevance is debated (Holbrook et al., 2017). Between-subject differences in DNA methylation may or may not depend on cellular heterogeneity. Generally, it is assumed that differences in DNA methylation not due to cell-types mixture are more likely to be causative of disease, while methylation differences caused by differences in cell-types proportion are considered a likely consequence of disease or at least of the disease process (i.e. to correspond to reverse causality, for example in the case of obesity-induced inflammation leading to changes in leukocytes proportion (Zeyda and Stulnig, 2009)), or to be a confounder whose effect needs to be controlled for. However, diseases are also often associated with the distribution of cell types, and cell type proportion can also in some situations be a cause of disease or a marker for e.g. inflammatory or immune-related diseases (Holbrook et al., 2017). This is particularly relevant for obesity development, which is known to involve inflammatory pathways (Hotamisligil, 2003). The differential methylation driven by cell-type heterogeneity could therefore mediate exposure effects rather than be a consequence of the overweight. Thus, cell types proportion could be 1) a consequence or very close marker of our outcome; in such case, adjusting for it in the model linking methylation and zBMI is irrelevant, as adjustment for

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

consequences of the diseases, which are by definition not confounders, could have harmful consequences (Barton et al., 2019); 2) a cause of the health outcome. In this case, if (option 2a) it is a mediator of an effect of an exposure on zBMI or a proxy of such a mediator, we should not adjust for it: as we want to identify exposures whose effect on zBMI is biologically mediated by the methylome layer, correcting for cell-types could prevent us from selecting potential exposures of interest associated with zBMI and whose effect is mediated by cell-type-dependent methylation. If (2b) it is not a mediator, i.e. if it is a cause of outcome but not a consequence of exposure, it is a potential confounder and as such should be corrected for in the test of association between methylation and zBMI. We chose not to adjust the DNA methylation-BMI model of the main analysis for cell-type heterogeneity because we a priori consider hypotheses 1) and 2a) as more likely than 2b), and because the consequence of erroneously adjusting may, in our study in which identifying the intermediate causal factors was not the main aim, be more harmful (Barton et al., 2019) than adjusting.

In our sensitivity analysis IV, we repeated the whole Meet-in-the-Middle design considering the cell-types instead of the methylation data as the intermediate layer. Final results were very similar to those of the main analysis, making it possible that differential methylation driven by cell-type heterogeneity explained the association between methylome layer and both copper and zBMI, and which therefore mediate copper effects (case 2a). However, the reduced exposome was slightly smaller when considering cell-types and the proportion mediated was lower (13% vs. 29%). This may mean that the cell-type did not convey as much information on the biological effect of exposures than DNA methylation on the path between selected exposures and zBMI.

We acknowledge several limitations to this original work; first, our results are dependent on the quality of a priori information; second, the methylome and the outcome (and a part of the exposome) were assessed simultaneously, a design which makes difficult to exclude reverse causation with an agnostic test of association. Measurement error is also expected for both the methylome and the exposome; finally, only monotonous associations were tested to limit the number of tests. The strengths of our study include a multilayer analysis of the exposome, methylome and BMI in a large and well-characterized population, the consideration of multiple testing in the composite tests and of a priori information on the methylome-BMI relation to restrict the number of tests done. This composite design may allow improving the specificity of exposome-health studies and limit associations due to reverse causality, with a possible cost on sensitivity.

5. Conclusion

This work is to our knowledge the first epidemiological study relying on an intermediate blood methylome layer to try to better characterize the exposome-health association. Purely agnostic exposome studies are expected to suffer from a high false positives rate, and possibly a low sensitivity (due to the correlation within the exposome) (Agier et al. 2016). Our approach may allow reducing the false positive rate by using a modified Meet-in-the-Middle design that permits a biologically driven reduction of the exposome, which may in particular allow discarding some of the associations of the outcome with the exposome due to reverse causality. It comes at a cost of being insensitive to exposures acting on the outcome via pathways not causing changes in the methylome of peripheral blood. Extensions of this approach to other biologically relevant layers might allow avoiding this limitation in the future.

Funding

The study has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 308,333 – the HELIX project for data collection and analyses. The Norwegian Mother and Child Cohort Study is supported by the

Norwegian Ministry of Health and Care Services and the Ministry of Education and Research, NIH/NIEHS (contract no N01-ES-75558), NIH/NINDS (grant n°0.1 U01 NS 047537-01 and grant no.2 U01 NS 047537-06A1). We also received support from Région Auvergne-Rhône-Alpes for collaborations with Catalunya. Dr. Chatzi was supported by NIH P30ES007048, R21ES029681, R01ES029944, R01ES030364, R21ES028903, and by Environmental Protection Agency RD-83544101.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the input of the HELIX consortium. We are grateful to all the participating families in the six countries who took part in this cohort study (BiB, EDEN, INMA, KANC, MoBa and RHEA cohorts). We acknowledge the work of all fieldworkers and health professionals who make the data collection possible. Regarding the EDEN cohort, we further thank Sonia Brisoual, Angélique Serre and Michele Grosdenier (Poitiers Biobank, CRB BB-0033-00068, Poitiers, France) for biological sample management, Elodie Migault, Manuela Boue and Sandy Bertin (Clinical Investigation Centre, Inserm CIC1402, CHU de Poitiers, Poitiers, France) for planification and investigational actions. We are also grateful to Veronique Ferrand-Rigalleau, Céline Leger and Noella Gorry (CHU de Poitiers, Poitiers, France) for administrative assistance (EDEN). We acknowledge the support of Région Auvergne Rhône Alpes for scientific collaboration with Catalogna. Regarding the MoBa study, we like to thank Ingvild Essen and Jorunn Evandt for thorough field work, Heidi Marie Nordheim for biological sample management and the MoBa administrative unit (MoBa). The Norwegian Mother, Father and Child Cohort Study is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. We are grateful to all the participating families in Norway who take part in this on-going cohort study.

Contributions

SC and RS designed the analytical and statistical methods. SC analysed the data. SC, MB, LA, MV and RS interpreted the results and wrote the paper. MV, MB and LM coordinated the HELIX data collection. All authors contributed to the data collection and to the manuscript, and approved the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.envint.2020.105622>.

References

- Agay-Shay, K., Martinez, D., Valvi, D., Garcia-Esteban, R., Basagafia, X., Robinson, O., Casas, M., Sunyer, J., Vrijheid, M., 2015. Exposure to endocrine-disrupting chemicals during pregnancy and weight at 7 years of age: a multi-pollutant approach. *Environ. Health Perspect.* 123, 1030–1037. <https://doi.org/10.1289/ehp.1409049>.
- Agier, L., Basagafia, X., Maitre, L., Granum, B., Bird, P.K., Casas, M., Oftedal, B., Wright, J., Andrusaityte, S., de Castro, M., Cequier, E., Chatzi, L., Donaire-Gonzalez, D., Grazuleviciene, R., Haug, L.S., Sakhi, A.K., Leventakou, V., McEachan, R., Nieuwenhuijsen, M., Petraciene, I., Robinson, O., Roumeliotaki, T., Sunyer, J., Tamayo-Uria, I., Thomsen, C., Urquiza, J., Valentin, A., Slama, R., Vrijheid, M., Sirois, V., 2019. Early-life exposome and lung function in children in Europe: an analysis of data from the longitudinal, population-based HELIX cohort. *Lancet Planet. Health.* 3, e81–e92. [https://doi.org/10.1016/S2542-5196\(19\)30010-5](https://doi.org/10.1016/S2542-5196(19)30010-5).
- Agier, L., Portengen, L., Chadeau-Hyam, M., Basagafia, X., Giorgis-Allemand, L., Sirois, V., Robinson, O., Vlaanderen, J., González, J.R., Nieuwenhuijsen, M.J., Vineis, P., Vrijheid, M., Slama, R., Vermeulen, R., 2016. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ.*

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

- Health Perspect. 124, 1848–1856. <https://doi.org/10.1289/EHP172>.
- Aryee, M.J., Jaffee, A.E., Corradia-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., Irizarry, R.A., 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369. <https://doi.org/10.1093/bioinformatics/btu049>.
- Baccarelli, A., Wright, R.O., Bollati, V., Tarantini, L., Litonjua, A.A., Suh, H.H., Zanobetti, A., Sparrow, D., Vokonas, P.S., Schwartz, J., 2009. Rapid DNA methylation changes after exposure to traffic particles. *Am. J. Respir. Crit. Care Med.* 179, 572–578. <https://doi.org/10.1164/rccm.200807-1097OC>.
- Barton, S.J., Melton, P.E., Titcombe, P., Murray, R., Rauschert, S., Lillycrop, K.A., Huang, R.C., Holbrook, J.D., Godfrey, K.M., 2019. In epigenomic studies, including cell-type adjustments in regression models can introduce multicollinearity, resulting in apparent reversal of direction of association. *Front. Genet.* 10. <https://doi.org/10.3389/fgene.2019.00816>.
- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Stat. Soc. Ser. B J. R.* <https://doi.org/10.2307/2346101>.
- Berger, S.J., Kouzarides, T., Shiekhattar, R., Shilatifard, A., 2009. An operational definition of epigenetics. *Genes Dev.* 23, 781–783. <https://doi.org/10.1101/gad.1787609>.
- Braun, J.M., 2017. Early-life exposure to EDCs: role in childhood obesity and neurodevelopment. *Nat Rev. Endocrinol.* 13, 161–173. <https://doi.org/10.1038/nrendo.2016.186>.
- Brewer, G.J., 2010. Copper toxicity in the general population. *Clin. Neurophysiol.* 121, 459–460. <https://doi.org/10.1016/j.clinph.2009.12.015>.
- Buck Louis, G.M., Schisterman, E.F., Sweeney, A.M., Wilcosky, T.C., Gore-Langton, R.E., Lynch, C.D., Boyd Barr, D., Schrader, S.M., Kim, S., Chen, Z., Sundaram, R., 2011. Designing prospective cohort studies for assessing reproductive and developmental toxicity during sensitive windows of human reproduction and development – the LIFE Study. *Paediatr. Perinat. Epidemiol.* 25, 413–424. <https://doi.org/10.1111/j.1365-3014.2011.02105.x>.
- van Buuren, S., Groothuis-Oudshoorn, K., 2011. *Mice* : multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. <https://doi.org/10.18637/jss.v045.i03>.
- Cao-Lei, L., Dancause, K.N., Elgbeili, G., Massart, R., Szyf, M., Liu, A., Laplante, D.P., King, S., 2015. DNA methylation mediates the impact of exposure to prenatal maternal stress on BMI and central adiposity in children at age 13½ years: Project Ice Storm. *Epidemiology* 10, 749–761. <https://doi.org/10.1080/15352294.2015.1063771>.
- Carone, B.R., Fauquier, L., Habib, N., Shea, J.M., Hart, C.E., Li, R., Bock, C., Li, C., Gu, H., Zamore, P.D., Meissner, A., Weng, Z., Hofmann, H.A., Friedman, N., Rando, O.J., 2010. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell* 143, 1084–1096. <https://doi.org/10.1016/j.cell.2010.12.008>.
- Casas, M., Basagaña, X., Sakhni, A.K., Haug, L.S., Philippat, C., Granum, B., Manzano-Salgado, C.B., Brochot, C., Zeman, F., de Bont, J., Andrusaityte, S., Chatzli, L., Donaire-Gonzalez, D., Giorgis-Allemand, L., Gonzalez, J.R., Gracia-Lavedan, E., Grazuleviciene, R., Kampouri, M., Lyon-Caen, S., Pafella, P., Petraciviciene, I., Robinson, O., Urquiza, J., Vafeiadi, M., Vernet, C., Waiblinger, D., Wright, J., Thomsen, C., Slama, R., Vrijheid, M., 2018. Variability of urinary concentrations of non-persistent chemicals in pregnant women and school-aged children. *Environ. Int.* 121, 561–573. <https://doi.org/10.1016/j.envint.2018.09.046>.
- Chadeau-Hyam, M., Athersuch, T.J., Keun, H.C., De Iorio, M., Ebbels, T.M.D., Jenab, M., Sacerdote, C., Bruce, S.J., Holmes, E., Vineis, P., 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>.
- Chatzli, L., Leventakou, V., Vafeiadi, M., Koutra, K., Roumeliotaki, T., Chalkiadaki, G., Karachaliou, M., Daraki, V., Kyrikaki, A., Kampouri, M., Filionou, E., Sarri, K., Vassilaki, M., Fasouli, M., Bitsios, P., Koutis, A., Stephanou, E.G., Kogevevas, M., 2017. Cohort profile: the mother-child cohort in crete, greece (rhea study). *Int. J. Epidemiol.* 46, 1392–1393. <https://doi.org/10.1093/ije/dyx084>.
- Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72, 3–25. <https://doi.org/10.1111/j.1467-9868.2009.00723.x>.
- de Onis, M., Onyango, A.W., Borghi, E., Siyam, A., Nishida, C., Siekmann, J., 2007. Development of a WHO growth reference for school-aged children and adolescents. *Bull. World Health Organ.* 85, 660–667.
- Dekkers, K.F., van Iterson, M., Slieker, R.C., Moed, M.H., Bonder, M.J., van Galen, M., Mei, H., Zhernakova, D.V., van der Berg, L.H., Deelen, J., van Dongen, J., van Heemst, D., Hofman, A., Hottenga, J.J., van der Kallen, C.J.H., Schalkwijk, C.G., Stehouwer, C.D.A., Tigchelaar, E.F., Uitterlinden, A.G., Witteman, G., Zhernakova, A., Franke, L., 't Hoen, P.A.C., Jansen, R., van Meurs, J., Boomstra, D.I., van Duijn, C.M., van Greevenbroek, M.M.J., Veldink, J.H., Wijmenga, C., van Zwet, E.W., Slagboom, P.E., Jukema, J.W., Heijmans, B.T., 2016. Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* 17, 138. 10.1186/s13059-016-1000-6.
- Debier, C., Chalon, C., Le Beuf, B.J., et al., 2006. Mobilization of PCBs from blubber to blood in northern elephant seals (*Mirounga angustirostris*) during the post-weaning fast. *Aquat. Toxicol.* 80 (2), 149–157. <https://doi.org/10.1016/j.aquatox.2006.08.002>.
- Dorts, J., Falisse, E., Schoofs, E., Flamion, E., Kestemont, P., Silvestre, F., 2016. DNA methyltransferases and stress-related genes expression in zebrafish larvae after exposure to heat and copper during reprogramming of DNA methylation. *Sci. Rep.* 6, 34254. <https://doi.org/10.1038/srep34254>.
- Fan, Y., Zhang, C., Bu, J., 2017. Relationship between selected serum metallic elements and obesity in children and adolescent in the U.S. *Nutrients* 9, 1–12. <https://doi.org/10.3390/nu9020104>.
- Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, Mattias, Grankvist, K., Johansson, Mikael, Assumma, M.B., Naccarati, A., Chadeau-Hyam, M., Ala, U., Faltus, C., Kaaks, R., Risch, A., De Stavola, B., Hodge, A., Giles, G.G., Southey, M.C., Relton, C.L., Haycock, P.C., Lund, E., Polidoro, S., Sandanger, T.M., Severi, G., Vineis, P., 2015. Hypermethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* 6, 10192. <https://doi.org/10.1038/ncomms10192>.
- Feil, R., Fraga, M.F., 2012. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* 13, 97–109. <https://doi.org/10.1038/nrg3142>.
- Finucane, M.M., Stevens, G.A., Cowan, M.J., Danaei, G., Lin, J.K., Paciorek, C.J., Singh, G.M., Gutierrez, H.R., Lu, Y., Bahalim, A.N., Farzadfar, F., Riley, L.M., Ezzati, M., 2011. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* 377, 557–567. [https://doi.org/10.1016/S0140-6736\(10\)62037-5](https://doi.org/10.1016/S0140-6736(10)62037-5).
- Fortin, Jean-Philippe, Fertig, Elana, Hansen, Kasper, 2014. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Res* 3, 175. <https://doi.org/10.12688/f1000research.4680.2>.
- Fradin, D., Boëlle, P.Y., Belot, M.P., Lachaux, F., Tost, J., Besse, C., Deleuze, J.F., De Filippo, G., Bougnères, P., 2017. Genome-wide methylation analysis identifies specific epigenetic marks in severely obese children. *Sci. Rep.* 7. <https://doi.org/10.1038/srep46311>.
- Galhardi, C.M., Diniz, Y.S., Faine, L.A., Rodrigues, H.G., Burneiko, R.C.M., Ribas, B.O., Novelli, E.L.B., 2004. Toxicity of copper intake: Lipid profile, oxidative stress and susceptibility to renal dysfunction. *Food Chem. Toxicol.* <https://doi.org/10.1016/j.fct.2004.07.020>.
- Grazeleviciene, R., Danileviciute, A., Dedele, A., Vencloviene, J., Andrusaityte, S., Uždānaviciute, I., Nieuwenhuijsen, M.J., 2015. Surrounding greenness, proximity to city parks and pregnancy outcome in Kaunas cohort study. *Int. J. Hyg. Environ. Health* 218, 358–363. <https://doi.org/10.1016/j.ijeh.2015.02.004>.
- Guxens, M., Ballester, F., Espada, M., Fernández, M.F., Grimalt, J.O., Ibarluzea, J., Olea, N., Reboliaga, M., Tardón, A., Torrent, M., Vioque, J., Vrijheid, M., Sunyer, J., 2012. Cohort profile: the INMA—INFancia y medio ambiente—(environment and child-hood) project. *Int. J. Epidemiol.* 41, 930–940. <https://doi.org/10.1093/ije/dyr054>.
- Hanssen, K.D., 2016. IlluminaHumanMethylation450anno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays. R Packag. version 0.6.0.
- Haug, I.S., Sakhni, A.K., Cequer, E., Casas, M., Maitre, L., Basagana, X., Andrusaityte, S., Chalkiadaki, G., Chatzli, L., Coen, M., de Bont, J., Dedele, A., Ferrand, J., Grazeleviciene, R., Gonzalez, J.R., Gutzkow, K.B., Keun, H., McEachan, R., Meltzer, H.M., Petraciviciene, I., Robinson, O., Saulnier, P.J., Slama, R., Sunyer, J., Urquiza, J., Vafeiadi, M., Wright, J., Vrijheid, M., Thomsen, C., 2018. In-utero and childhood chemical exposome in six European mother-child cohorts. *Environ. Int.* 121, 751–763. <https://doi.org/10.1016/j.envint.2018.09.056>.
- Heijmans, B.T., Tobi, F.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E., Lumey, L.H., 2008. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci.* 105, 17046–17049. <https://doi.org/10.1073/pnas.0806556105>.
- Heindel, J.J., vom Saal, F.S., 2009. Role of nutrition and environmental endocrine disrupting chemicals during the perinatal period on the aetiology of obesity. *Mol. Cell. Endocrinol.* 304 (1–2), 90–96. <https://doi.org/10.1016/j.mce.2009.02.025>.
- Heo, E.J., Shim, J.E., Yoon, E.Y., 2017. Systematic review on the study of the childhood and adolescent obesity in korea: dietary risk factors. *Korean J. Community Nutr.* 22, 191. <https://doi.org/10.5720/kjcn.2017.22.3.191>.
- Heude, Barbara, Forhan, Anne, Slama, Rémy, Douchaud, L., Bedel, S., Saurel-Cubizolles, M.-J., Hankard, Régis, Thiebautgeorges, Olivier, De Agostini, Maria, Annés-Maesano, Isabella, Kaminski, Monique, Charles, Marie-Aline, Annés-Maesano, I., Bernard, J., Botton, J., Charles, M.-A., Dargent-Molina, P., de Lauzon-Guillain, B., Ducimetière, P., de Agostini, M., Foliguet, B., Forhan, A., Fritel, X., Germa, A., Goua, V., Hankard, R., Heude, B., Kaminski, M., Larroque, B., Lelong, N., Lepeule, J., Magnin, G., Marchand, L., Nabet, C., Pierre, F., Slama, R., Saurel-Cubizolles, M., Schweitzer, M., Thiebautgeorges, O., 2016. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *Int. J. Epidemiol.* 45, 353–363. 10.1093/ije/dyv151.
- Ho, S.M., Johnson, A., Tarapore, P., Janakiraman, V., Zhang, X., Leung, Y.-K., 2012. Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR J.* 53, 289–305. <https://doi.org/10.1093/ilar.53.3-4.289>.
- Holbrook, J.D., Huang, R.-C., Barton, S.J., Saffery, R., Lillycrop, K.A., 2017. Is cellular heterogeneity merely a confounder to be removed from epigenome-wide association studies? *Epigenomics* 9, 1143–1150. <https://doi.org/10.2217/epi-2017-0032>.
- Holtzman, P., 2012. Obesogen: an environmental link to obesity. *Environ. Health Perspect.* 120, a62–a68. <https://doi.org/10.1289/ehp.120-a62>.
- Hotamisligil, G.S., 2003. Inflammatory pathways and insulin action. *Int. J. Obes.* 27, S53–S55. <https://doi.org/10.1038/sj.ijo.0802502>.
- Housman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T., 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinf.* 13, 86. <https://doi.org/10.1186/1471-2105-13-86>.
- Hu, F.B., Malik, V.S., 2010. Sugar-sweetened beverages and risk of obesity and type 2 diabetes: Epidemiologic evidence. *Physiol. Behav.* 100, 47–54. <https://doi.org/10.1016/j.physbeh.2010.01.036>.
- Janesick, A., Blumberg, B., 2011. Endocrine disrupting chemicals and the developmental programming of adipogenesis and obesity. *Birth Defects Res. Part C Embryo Today Rev.* 93, 34–50. <https://doi.org/10.1002/bdrc.20197>.
- Jang, H.S., Shin, W.J., Lee, J.E., Do, J.T., 2017. CpG and Non-CpG methylation in epigenetic gene regulation and brain function. *Genes (Basel).* 8. <https://doi.org/10.3390/GENES8060148>.

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

- Jimenez, L.O., Landgrebe, D.A., 1998. Supervised classification in high-dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev. Inst. Electr. Electron. Eng.* 10, 1109/5326.661089.
- Joubert, B.R., Felix, J.F., Yousefi, P., Bakulski, K.M., Just, A.C., Breton, C., Reese, S.E., Markunas, C.A., Richmond, R.C., Xu, C.J., Küpers, L.K., Oh, S.S., Hoyo, C., Gruzieva, O., Söderhäll, C., Salas, I.A., Baiz, N., Zhang, H., Lepelte, J., Ruiz, C., Ligthart, S., Wang, T., Taylor, J.A., Duijts, L., Sharp, G.C., Jankipersadsing, S.A., Nilsen, R.M., Vaez, A., Fallin, M.D., Hu, D., Litonjua, A.A., Fuemmeler, B.F., Huen, K., Kere, J., Kull, I., Munthe-Kaas, C., Gehring, U., Bustamante, M., Sauré-Coubizolles, M.J., Quraishi, B.M., Ren, J., Tost, J., Gonzalez, J.R., Peters, M.J., Häberg, S.E., Xu, Z., Van Meurs, J.B., Gaunt, T.R., Kerkhof, M., Corpelein, E., Feinberg, A.P., Eng, C., Baccarelli, A.A., Benjamin Neelon, S.E., Bradman, A., Merid, S.K., Bergström, A., Herceg, Z., Hernandez-Vargas, H., Brunekreef, B., Pinart, M., Heude, B., Ewart, S., Yao, J., Lemmonier, N., Franco, O.H., Wu, M.C., Hofman, A., McArdle, W., Van Der Vlies, P., Falahi, F., Gillman, M.W., Barcellos, L.F., Kumar, A., Wickman, M., Guerra, S., Charles, M.A., Holloway, J., Auffray, C., Tielemans, H.W., Smith, G.D., Postma, D., Hirvonen, M.F., Eskenazi, B., Vrijheid, M., Arshad, H., Antó, J.M., Dehghan, A., Karrafaus, W., Annesi-Maesano, I., Sunyer, J., Ghantous, A., Pershagen, G., Holland, N., Murphy, S.K., Demeo, D.L., Burchard, E.G., Ladd-Acosta, C., Snieder, H., Nystad, W., Koppelman, G.H., Relton, C.L., Jaddoe, V.W.V., Wilcox, A., Melén, E., London, S.J., 2016. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *Am. J. Hum. Genet.* 98, 680–696. <https://doi.org/10.1016/j.ajhg.2016.02.019>.
- Klevay, L.M., 2018. Cardiovascular disease from copper deficiency—A history. *J. Nutr.* 130, 489S–492S. <https://doi.org/10.1093/jn/nly130>.
- Lauritzen, H.B., Larose, T.L., Øien, T., Sandanger, T.M., Odland, J.O., Van De Bor, M., Jacobsen, G.W., 2018. Prenatal exposure to persistent organic pollutants and child overweight/obesity at 5-year follow-up: a prospective cohort study. *Environ. Health. Glob. Access Sci. Source* 17, 9. <https://doi.org/10.1186/s12940-017-0338-x>.
- Lehne, B., Drong, A.W., Loh, M., Zhang, W., Scott, W.R., Tan, S.T., Afzal, U., Scott, J., Jarvelin, M.R., Elliott, P., McCarthy, M.I., Kooper, J.S., Chambers, J.C., 2015. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.* 16, 37. <https://doi.org/10.1186/s13059-015-0600-x>.
- Lenters, V., Portengen, L., Rignell-Hydbom, A., Jönsson, B.A.G., Lindh, C.H., Piersma, A.H., Toft, G., Bonde, J.P., Heederik, D., Rylander, L., Vermeulen, R., 2016. Prenatal phthalate, perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: Multi-pollutant models based on elastic net regression. *Environ. Health Perspect.* 124, 365–372. <https://doi.org/10.1289/ehp.1408933>.
- Lenters, V., Vermeulen, R., Portengen, L., 2018. Performance of variable selection methods for assessing the health effects of correlated exposures in case-control studies. *Occup. Environ. Med.* 75, 522–529. <https://doi.org/10.1136/oemed-2016-104231>.
- Lillycrop, Karen A., Phillips, Emma S., Jackson, Alan A., Hanson, Mark A., Burdge, Graham C., 2005. Dietary protein restriction of pregnant rats induces and folic acid supplementation prevents epigenetic modification of hepatic gene expression in the offspring. *J. Nutr.* 135 (6), 1382–1386. <https://doi.org/10.1093/jn/135.6.1382>.
- Lillycrop, K.A., Phillips, E.S., Torrens, C., Hanson, M.A., Jackson, A.A., Burdge, G.C., 2008. Feeding pregnant rats a protein-restricted diet persistently alters the methylation of specific cytosines in the hepatic PPARα promoter of the offspring. *Br. J. Nutr.* 100, 278–282. <https://doi.org/10.1017/S0007114507894438>.
- Lima, S.C.V.C., Arrais, R.F., Sales, C.H., Almeida, M.G., De Sena, K.C.M., Oliveira, V.T.L., De Andrade, A.S., Pedrosa, L.F.C., 2006. Assessment of copper and lipid profile in obese children and adolescents. *Biol. Trace Elem. Res.* 114, 19–30. <https://doi.org/10.1385/BTER:114:1:19>.
- Lydersen, S., Wolkers, H., Severinsen, T., et al., 2002. Blood is a poor substrate for monitoring pollution burdens in phocid seals. *Sci. Tot. Environ.* 292 (3), 193–203. <http://linkinghub.elsevier.com/retrieve/pii/S0048969701011214>.
- Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Dalvæt, A.K., Handal, M., Haugen, M., Heiseth, G., Knudsen, G.P., Paltiel, L., Schreuder, P., Tambs, K., Vold, I., Stoltenberg, C., 2016. Cohort profile update: the norwegian mother and child cohort study (MoBa). *Int. J. Epidemiol.* 45, 382–388. <https://doi.org/10.1093/ije/dyw029>.
- Maitre, L., de Boni, J., Casas, M., Robinson, O., Aasvang, G.M., Agier, L., Andrusaityté, S., Ballester, F., Basagaña, X., Borras, E., Brochot, C., Bustamante, M., Carracedo, A., de Castro, M., Dedeče, A., Donaire-Gonzalez, D., Estivill, X., Evans, J., Fossati, S., Giorgis-Allemand, L., R Gonzalez, J., Granum, B., Grazeviciene, R., Bjerve, K., Gutzkow, K., Småstuen Haug, L., Hernandez-Ferrer, C., Heude, B., Ibarluzea, J., Julvez, J., Karjalainen, M., Keun, H.C., Hjertager, Krog, N., Lau, C.H.E., Leventakou, V., Lyon-Caen, S., Manzano, C., Mason, D., McEachan, R., Meltzer, H.M., Petraciviciene, I., Quenten, J., Roumeliotaki, T., Sabido, E., Saulnier, P.-J., Siskos, A.P., Siroux, P., Sunyer, J., Tamayo, I., Urquiza, J., Vafeiadis, M., van Gent, D., Vives-Usano, M., Waiblinger, D., Warembourg, C., Chatzil, L., Coen, M., van den Hazel, P., Nieuwenhuijsen, M.J., Slama, R., Thomsen, C., Wright, J., Vrijheid, M., 2018. Human Early Life Exposome (HELIx) study: a European population-based exposome cohort. *BMJ Open* 8, e021311. 10.1136/bmjopen-2017-021311.
- Manrai, A.K., Cui, Y., Bushel, P.R., Hall, M., Karakitsios, S., Mattingly, C.J., Ritchie, M., Schmitt, C., Sarigiannis, D.A., Thomas, D.C., Wishart, D., Balshaw, D.M., Patel, C.J., 2017. Informatics and data analytics to support exposome-based discovery for public health. *Annu. Rev. Public Health.* 38, 279–294. <https://doi.org/10.1146/annurev-pubhealth-010116-110607>.
- Marioni, R.E., McAra, A.F., Bressler, J., Colicino, E., Hannon, E., Li, S., Prada, D., Smith, J.A., Trevisi, L., Tsai, P.-C., Vojinovic, D., Simino, J., Levy, D., Liu, C., Mendelson, M., Satizabal, C.L., Yang, Q., Jhun, M.A., Kardia, S.L.R., Zhao, W., Bandinelli, S., Ferrucci, L., Hernandez, D.G., Singleton, A.B., Harris, S.E., Starr, J.M., Kiel, D.P., McLean, R.R., Just, A.C., Schwartz, J., Spiro, A., Vokonas, P., Amin, N., Ikram, M.A., Uitterlinden, A.G., van Meurs, J.B.J., Spector, T.D., Steves, C., Baccarelli, A.A., Bell, J.T., van Duijn, C.M., Fornage, M., Hsu, Y.-H., Mill, J., Mosley, T.H., Seshadri, S., Deary, I.J., 2018. Meta-analysis of epigenome-wide association studies of cognitive abilities. *Mol. Psychiatry* 1. <https://doi.org/10.1038/s41380-017-0008-y>.
- Menke, A., Guallar, E., Cowie, C.C., 2016. Metals in urine and diabetes in U.S. adults. *Diabetes* 65, 164–171. <https://doi.org/10.2337/db15-0316>.
- Milne, D.B., Weswig, P.H., 2018. Effect of supplementary copper on blood and liver copper-containing fractions in rats. *J. Nutr.* <https://doi.org/10.1093/jn/nvz3.429>.
- Murakami, K., Livingstone, M.B.E., 2016. Associations between meal and snack frequency and overweight and abdominal obesity in US children and adolescents from national health and nutrition examination survey (NHANES) 2003–2012. *Br. J. Nutr.* 115, 1819–1829. <https://doi.org/10.1017/njn.2015.00854>.
- Nelson, J.W., Hatch, E.E., Webster, T.F., 2010. Exposure to polyfluoroalkyl chemicals and cholesterol, body weight, and insulin resistance in the general U.S. population. *Environ. Health Perspect.* 118, 197–202. <https://doi.org/10.1289/ehp.0901165>.
- OECD – Organisation for Economic Co-operation and Development, 2012. Proposal for a template, and guidance on developing and assessing the completeness of adverse outcome pathways 1–17.
- Oken, E., Levitan, E.B., Gillman, M.W., 2008. Maternal smoking during pregnancy and child overweight: Systematic review and meta-analysis. *Int. J. Obes.* <https://doi.org/10.1038/sj.ijo.0803760>.
- Pal, A., Jayamani, J., Prasad, R., 2014. An urgent need to reassess the safe levels of copper in the drinking water: lessons from studies on healthy animals harboring no genetic defects. *Neurotoxicology.* <https://doi.org/10.1016/j.neuro.2014.05.005>.
- Park, S.S., Skaar, D.A., Jirtle, R.L., Hoyo, C., 2017. Epigenetics, obesity and early-life cadmium or lead exposure. *Epigenomics* 9, 57–75. <https://doi.org/10.2217/epi-2016-0047>.
- Pereira, T.C.B., Campos, M.M., Bogo, M.R., 2016. Copper toxicology, oxidative stress and inflammation using zebrafish as experimental model. *J. Appl. Toxicol.* 36, 876–885. <https://doi.org/10.1002/jat.3303>.
- Plagmann, A., Harder, T., Brunn, M., Harder, A., Roepeke, K., Wittrock-Staer, M., Ziska, T., Schellong, K., Rodekamp, E., Melchior, K., Dudenhausen, J.W., 2009. Hypothalamic proopiomelanocortin promoter methylation becomes altered by early overfeeding: an epigenetic model of obesity and the metabolic syndrome. *J. Physiol.* 587, 4963–4976. <https://doi.org/10.1113/jphysiol.2009.176156>.
- Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.E., Greco, D., Söderhäll, C., Scheynius, A., Kere, J., 2012. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PLoS ONE* 7, e41361. <https://doi.org/10.1371/journal.pone.0041361>.
- Richmond, R.C., Sharp, G.C., Ward, M.E., Fraser, A., Lytleton, O., McArdle, W.L., Ring, S.M., Gaunt, T.R., Lawlor, D.A., Smith, G.D., Relton, C.L., 2016. DNA Methylation and BMI: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes* 65, 1231–1244. <https://doi.org/10.2337/db15-0996>.
- Richmond, R.C., Timpong, N.J., Sorensen, T.I., 2015. Exploring possible epigenetic mediation of early-life environmental exposures on adiposity and obesity development: Figure 1. *Int. J. Epidemiol.* 44, 1191–1198. <https://doi.org/10.1093/ije/dyv066>.
- Rönn, M., Lind, L., Bavel, B., van, Salihovic, S., Michaelsson, K., Lind, P.M., 2011. Circulating levels of persistent organic pollutants associate in divergent ways to fat mass measured by DXA in humans. *Chemosphere* 85 (3), 335–343. <https://doi.org/10.1016/j.chemosphere.2011.06.095>.
- Ruiz-Hernandez, A., Kuo, C.C., Rentero-Garrido, P., Tang, W.Y., Redon, J., Ordovas, J.M., Navas-Acien, A., Tellier-Plaza, M., 2015. Environmental chemicals and DNA methylation in adults: A systematic review of the epidemiologic evidence. *Clin. Epigenetics* 7. <https://doi.org/10.1186/s13148-015-0055-7>.
- Rzehak, P., Covic, M., Saffery, R., Reischl, E., Wahl, S., Grote, V., Weber, M., Xhonoue, A., Langhendries, J.P., Ferre, N., Closa-Monasterolo, R., Escribano, J., Verdúci, E., Riva, E., Socha, P., Grusfeld, D., Koletzko, B., 2017. DNA-methylation and body composition in preschool children: epigenome-wide-analysis in the european childhood obesity project (CHOP) study. *Epigenet. Rep.* 7. <https://doi.org/10.1038/s41598-017-03099-4>.
- Saihat, S., Kreis, I., Davies, B., Bridgman, S., Kamanyire, R., 2013. The impact of PFOS on health in the general population: a review. *Environ. Sci. Process. Impacts* 15, 329–335. <https://doi.org/10.1039/c3em30698k>.
- Salustri, C., Barbat, G., Ghidoni, R., Quintiliani, L., Ciappina, S., Binetti, G., Squitti, R., 2010. Is cognitive function linked to serum free copper levels? A cohort study in a normal population. *Clin. Neurophysiol.* 121, 502–507. <https://doi.org/10.1016/j.clinph.2009.11.090>.
- Silverio Amancio, O.M., Alves Chaud, D.M., Yanaguibashi, G., Esteves Hilário, M.O., 2003. Copper and zinc intake and serum levels in patients with juvenile rheumatoid arthritis. *Eur. J. Clin. Nutr.* 57, 706–712. <https://doi.org/10.1038/sj.ejcn.1601601>.
- Sirois, V., Agier, L., Slama, R., 2016. The exposome concept: a challenge and a potential driver for environmental health research. *Eur. Respir. Rev.* 25, 124–129. <https://doi.org/10.1183/16000617.0034-2016>.
- Squitti, R., Bressi, F., Pasqualetti, P., Bonomini, C., Ghidoni, R., Binetti, G., Cassetta, E., Moffa, F., Ventriglia, M., Vernieri, F., Rossini, P.M., 2009. Longitudinal prognostic value of serum copper and zinc in patients with Alzheimer disease. *Neurology* 72, 50–55. <https://doi.org/10.1212/01.wnl.0000338568.28960.3f>.
- Tamayo-Uria, I., Maitre, L., Thomsen, C., Nieuwenhuijsen, M.J., Chatzil, L., Sirois, V., Aasvang, G.M., Agier, L., Andrusaityté, S., Casas, M., de Castro, M., Dedeče, A., Haug, L.S., Heude, B., Grazeviciene, R., Gutzkow, K.B., Krog, N.H., Mason, D., McEachan, R.R.C., Meltzer, H.M., Petraciviciene, I., Robinson, O., Roumeliotaki, T., Sakhi, A.K., Urquiza, J., Vafeiadis, M., Waiblinger, D., Warembourg, C., Wright, J., Slama, R., Vrijheid, M., Basagana, X., 2019. The early-life exposome: description and patterns in six European countries. *Environ. Int.* 123, 189–200. <https://doi.org/10.1016/j.envint.2018.11.067>.

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

S. Cadiou, et al.

Environment International 138 (2020) 105622

- Tapamyo, I., Maitre, L., Thomsen, C., Nieuwenhuijsen, M.J., Chatzi, L., 2018. The early-life exposome: description and patterns in six European countries. *Rev. Bioinforma.* 11^{https://doi.org/10.1002/0471250953.bi0112s38}. 1.12.1-1.12.54.
- Tanabe, M., Kanehisa, M., 2012. Using the KEGG database resource. *Curr. Protoc.* 11^{https://doi.org/10.1002/0471250953.bi0112s38}. 1.12.1-1.12.54.
- Thayer, K.A., Heindel, J.J., Bucher, J.R., Gallo, M.A., 2012. Role of environmental chemicals in diabetes and obesity: a national toxicology program workshop review. *Environ. Health Perspect.* 120, 779. ^{https://doi.org/10.1289/EHP.1104597}.
- Tisato, F., Marzano, C., Porchia, M., Pellei, M., Santini, C., 2010. Copper in diseases and treatments, and copper-based anticancer strategies. *Med. Res. Rev.* 30, 705-749. ^{https://doi.org/10.1002/med.20174}.
- Tobi, E.W., Lumey, L.H., Talsens, R.P., Kremer, D., Putter, H., Stein, A.D., Slagboom, P.E., Heijmans, B.T., 2009. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum. Mol. Genet.* 18, 4046-4053. ^{https://doi.org/10.1093/hmg/ddp353}.
- Tobi, E.W., Sleijker, R.C., Luijk, R., Dekkers, K.F., Stein, A.D., Xu, K.M., Biobank-based Integrative Omics Studies Consortium, B.I.O.S., Slagboom, P.E., van Zwet, E.W., Lumey, L.H., Heijmans, B.T., 2018a. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci. Adv.* 4, eaao4364. 10.1126/sciadv.aao4364.
- Tobi, E.W., Zwet, E.W., van Lumey, L., Heijmans, B.T., 2018b. Why mediation analysis trumps Mendelian randomization in population epigenomics studies of the Dutch Famine. *bioRxiv* 362392. 10.1101/362392.
- Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., Siegmund, K.D., 2013. Low-level processing of Illumina Infinium DNA methylation beadarrays. *Nucleic Acids Res.* 41.
- Uauy, R., Olivares, M., Gonzalez, M., 1998. Essentiality of copper in humans. *Am. J. Clin. Nutr.* ^{https://doi.org/10.1093/ajcn/67.5.952s}.
- Uriu-Adams, J.Y., Keen, C.L., 2005. Copper, oxidative stress, and human health. *Mol. Aspects Med.* 26, 268-298. ^{https://doi.org/10.1016/j.mam.2005.07.015}.
- van Iterson, M., Tobi, E.W., Sleijker, R.C., den Hollander, W., Luijk, R., Slagboom, P.E., Heijmans, B.T., 2014. MethylAid: Visual and interactive quality control of large Illumina 450k data sets. *Bioinformatics* 30, 3435-3437. ^{https://doi.org/10.1093/bioinformatics/btu566}.
- Vázquez Navá, F., Saldivar González, A.H., Perales, G.M., Ochoa, D.L., del Carmen Barrientos Gómez, M., Rodríguez, E.M.V., Rodríguez, C.F.V., Guzmán, F.J.B., 2006. Associations Between Family History of Allergy, Exposure to Tobacco Smoke, Active Smoking, Obesity, and Asthma in Adolescents. *Arch. Bronconeumol.* (English Ed. 42, 621-626. 10.1016/s1579-2129(07)60003-2.
- Vineis, P., Perera, F., 2007. Molecular epidemiology and biomarkers in etiologic cancer research: The new in light of the old. *Cancer Epidemiol. Biomarkers Prev.* 10.1158/1055-9965.EPI-07-0457.
- Vineis, P., van Veldhoven, K., Chadeau-Hyam, M., Athersuch, T.J., 2013. Advancing the application of omics-based biomarkers in environmental epidemiology. *Environ. Mol. Mutagen.* 54, 461-467. ^{https://doi.org/10.1002/em.21764}.
- Vinken, M., 2013. The adverse outcome pathway concept: A pragmatic tool in toxicology. *Toxicology.* ^{https://doi.org/10.1016/j.tox.2013.08.011}.
- Vrijheid, M., Fossati, S., Maitre, L., et al., 2020. Early-life environmental exposures and childhood obesity: an exposome-wide approach. *Environ. Health Perspect.* (Submitted for publication).
- Vrijheid, M., Slama, R., Robinson, O., Chatzi, L., Coen, M., van den Hazel, P., Thomsen, C., Wright, J., Athersuch, T.J., Avellana, N., Basagana, X., Brochot, C., Buccini, L., Bustamante, M., Carracedo, A., Casas, M., Estivill, X., Fairley, L., van Gent, D., Gonzalez, J.R., Granum, B., Gražulevičienė, R., Gutzkow, K.B., Julvez, J., Keun, H.C., Kogevinas, M., McEachan, R.R.C., Meltzer, H.M., Sabidó, E., Schwarze, P.E., Siroux, V., Sunyer, J., Wang, E.J., Zeman, F., Nieuwenhuijsen, M.J., 2014. The human early-life exposome (HELIX): project rationale and design. *Environ. Health Perspect.* 122, 535-544. ^{https://doi.org/10.1289/ehp.1307204}.
- Wahl, A., Kasela, S., Carnero-Montoro, E., van Iterson, M., Štambuk, J., Sharma, S., van den Akker, E., Klaric, L., Benedetti, E., Razdorov, G., Trbojević-Akmazić, I., Vučković, F., Ugrina, I., Beekman, M., Deelen, J., van Heest, D., Heijmans, B.T., B.I.O.S. Consortium, Wuhrer, M., Plomp, R., Keser, T., Šimurina, M., Pavić, T., Gudelj, I., Krištić, J., Grallert, H., Kunz, S., Peters, A., Bell, J.T., Spector, T.D., Milani, L., Slagboom, P.E., Lau, G., Gieger, C., 2018. IgG glycosylation and DNA methylation are interconnected with smoking. *Biochim. Biophys. Acta - Gen. Subj.* 1862, 637-648. 10.1016/j.bbagen.2017.10.012.
- Watkins, D.J., Welenius, G.A., Butler, R.A., Bartell, S.M., Fletcher, T., Kelsey, K.T., 2014. Associations between serum perfluorooctyl acids and LINE-1 DNA methylation. *Environ. Int.* 63, 71-76. ^{https://doi.org/10.1016/j.envint.2013.10.018}.
- Wild, C.P., 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Prev. Biomarkers* 14.
- Wright, J., Small, N., Raynor, P., Tuffnell, D., Bhopal, R., Cameron, N., Fairley, L., Lawlor, D.A., Parslow, R., Petherick, E.S., Pickett, K.E., Wablinger, D., West, J., 2013. Cohort Profile: the born in bradford multi-ethnic family cohort study. *Int. J. Epidemiol.* 42, 978-991. ^{https://doi.org/10.1093/ije/dys112}.
- Yakinci, G., Paç, A., Küçükbay, F.Z., Tayfun, M., Gül, A., 1997. Serum zinc, copper, and magnesium levels in obese children. *Pediatr. Int.* 39, 339-341. ^{https://doi.org/10.1111/j.1442-200X.1997.tb03748.x}.
- Yu, Q., Li, B., 2017. mma: an r package for mediation analysis with multiple mediators. *J. Open Res. Softw.* 5. ^{https://doi.org/10.5334/jors.160}.
- Zeyda, M., Stulig, T.M., 2009. Obesity, inflammation, and insulin resistance - a mini-review. *Gerontology* 55, 379-386. ^{https://doi.org/10.1159/000212758}.
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301-320. ^{https://doi.org/10.1111/j.1467-9868.2005.00503.x}.

II. 3. Supplementary materials

Supplementary Material II.1: Exposure levels in 1,173 mother-child pairs from the HELIX cohort.

Exposure name	Exposure window	Group	Unit - transformation	Modality	Mean +/- SE	n (%)	Assessment	Detailed exposure name
Arsenic	Postnatal	Metals	µg/L – Log2		0.6 ± 0		By biomarkers	
Arsenic	Pregnancy	Metals	µg/L – Log2		0.5 ± 0		By biomarkers	
Cadmium	Postnatal	Metals	µg/L – Log2		-3.4 ± 0		By biomarkers	
Cadmium	Pregnancy	Metals	µg/L – Log2		-1.3 ± 0		By biomarkers	
Cobalt	Postnatal	Metals	µg/L – Log2		-1.5 ± 0		By biomarkers	
Cobalt	Pregnancy	Metals	µg/L – Log2		-1.4 ± 0		By biomarkers	
Caesium	Postnatal	Metals	µg/L – Log2		1.3 ± 0		By biomarkers	
Caesium	Pregnancy	Metals	µg/L – Log2		1.5 ± 0		By biomarkers	
Copper	Postnatal	Metals	µg/L – Log2		10.8 ± 0		By biomarkers	
Copper	Pregnancy	Metals	µg/L – Log2		11.4 ± 0		By biomarkers	
Mercury	Postnatal	Metals	µg/L – Log2		0.4 ± 0		By biomarkers	
Mercury	Pregnancy	Metals	µg/L – Log2		1.6 ± 0		By biomarkers	
Manganese	Postnatal	Metals	µg/L – Log2		4 ± 0		By biomarkers	
Manganese	Pregnancy	Metals	µg/L – Log2		4.3 ± 0		By biomarkers	

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Molybdenum	Postnatal	Metals	$\mu\text{g/L} - \text{Log2}$	0.5 ± 0	By biomarkers	
Molybdenum	Pregnancy	Metals	$\mu\text{g/L} - \text{Log2}$	0.2 ± 0	By biomarkers	
Lead	Postnatal	Metals	$\mu\text{g/L} - \text{Log2}$	3.9 ± 0	By biomarkers	
Lead	Pregnancy	Metals	$\mu\text{g/L} - \text{Log2}$	4.1 ± 0	By biomarkers	
Thallium	Postnatal	Metals	Undetected	1087 (0.93)	By biomarkers	
Thallium	Postnatal	Metals	Detected	86 (0.07)	By biomarkers	
Thallium	Pregnancy	Metals	Undetected	1134 (0.97)	By biomarkers	
Thallium	Pregnancy	Metals	Detected	39 (0.03)	By biomarkers	
DDE	Postnatal	OCs	$\text{ng/g lipids} - \text{Log2}$	5.4 ± 0	By biomarkers	Dichlorodiphenyldichloroethylene
DDE	Pregnancy	OCs	$\text{ng/g lipids} - \text{Log2}$	6.6 ± 0	By biomarkers	Dichlorodiphenyldichloroethylene
DDT	Postnatal	OCs	$\text{ng/g lipids} - \text{Log2}$	0.1 ± 0	By biomarkers	Dichlorodiphenyltrichloroethane
DDT	Pregnancy	OCs	$\text{ng/g lipids} - \text{Log2}$	1.5 ± 0	By biomarkers	Dichlorodiphenyltrichloroethane
HCB	Postnatal	OCs	$\text{ng/g lipids} - \text{Log2}$	4 ± 0	By biomarkers	Hexachlorobenzene
HCB	Pregnancy	OCs	$\text{ng/g lipids} - \text{Log2}$	3.9 ± 0	By biomarkers	Hexachlorobenzene
PCB 118	Postnatal	OCs	$\text{ng/g lipids} - \text{Log2}$	1.9 ± 0	By biomarkers	Polybrominated diphenyl ether 47,Dichlorodiphenyldichloroethylene ether-118
PCB 118	Pregnancy	OCs	$\text{ng/g lipids} - \text{Log2}$	2.1 ± 0	By biomarkers	Polybrominated diphenyl ether 47,Dichlorodiphenyldichloroethylene ether-119

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

PCB 138	Postnatal	OCs	ng/g lipids – Log2	3.2 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether-138
PCB 138	Pregnancy	OCs	ng/g lipids – Log2	4 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether-138
PCB 153	Postnatal	OCs	ng/g lipids – Log2	4.3 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether- 153
PCB 153	Pregnancy	OCs	ng/g lipids – Log2	4.8 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether- 154
PCB 170	Postnatal	OCs	ng/g lipids – Log2	0.9 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether-170
PCB 170	Pregnancy	OCs	ng/g lipids – Log2	2.7 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether-171
PCB 180	Postnatal	OCs	ng/g lipids – Log2	2.5 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether- 180
PCB 180	Pregnancy	OCs	ng/g lipids – Log2	4 ± 0	By biomarkers	Polybrominated diphenyl ether-153,Dichlorodiphenyldichloroethylene ether- 181
PCBs (sum)	Postnatal	OCs	ng/g lipids – Log2	6.4 ± 0	By biomarkers	
PCBs (sum)	Pregnancy	OCs	ng/g lipids – Log2	7.3 ± 0	By biomarkers	
DEP	Postnatal	OP Pesticides	µg/L – Log2	1.2 ± 0.1	By biomarkers	Diethyl phosphate (DEP)
DEP	Pregnancy	OP Pesticides	µg/L – Log2	2.5 ± 0	By biomarkers	Diethyl phosphate (DEP)
DETP	Postnatal	OP Pesticides	µg/L – Log2	-0.7 ± 0.1	By biomarkers	Diethyl thiophosphate (DETP)
DETP	Pregnancy	OP Pesticides	µg/L – Log2	-0.1 ± 0.1	By biomarkers	Diethyl thiophosphate (DETP)

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

DMDTP	Postnatal	OP Pesticides	Undetected	960 (0.82)	By biomarkers	Dimethyldithiophosphate
DMDTP	Postnatal	OP Pesticides	Detected	213 (0.18)	By biomarkers	Dimethyldithiophosphate
DMP	Postnatal	OP Pesticides	µg/L – Log2	0.7 ± 0.1	By biomarkers	Dimethyl phosphate
DMP	Pregnancy	OP Pesticides	µg/L – Log2	3.7 ± 0	By biomarkers	Dimethyl phosphate
DMTP	Postnatal	OP Pesticides	µg/L – Log2	2.2 ± 0	By biomarkers	Dimethyl thiophosphate
DMTP	Pregnancy	OP Pesticides	µg/L – Log2	3 ± 0	By biomarkers	Dimethyl thiophosphate
PBDE 153	Postnatal	PBDEs	µg/L – Log2	-2.6 ± 0	By biomarkers	Polybrominated diphenyl ether 153
PBDE 153	Pregnancy	PBDEs	µg/L – Log2	-0.8 ± 0.1	By biomarkers	Polybrominated diphenyl ether 153
PBDE 47	Postnatal	PBDEs	µg/L – Log2	-1.3 ± 0	By biomarkers	Polybrominated diphenyl ether 47
PBDE 47	Pregnancy	PBDEs	µg/L – Log2	0.1 ± 0	By biomarkers	Polybrominated diphenyl ether 47
PFHXS	Postnatal	PFASs	µg/L – Log2	-1.6 ± 0	By biomarkers	Perfluorohexane sulfonate
PFHXS	Pregnancy	PFASs	µg/L – Log2	-0.2 ± 0	By biomarkers	Perfluorohexane sulfonate
PFNA	Postnatal	PFASs	µg/L – Log2	-1.1 ± 0	By biomarkers	Perfluorononanoate
PFNA	Pregnancy	PFASs	µg/L – Log2	0.2 ± 0	By biomarkers	Perfluorononanoate
PFOA	Postnatal	PFASs	µg/L – Log2	0.6 ± 0	By biomarkers	Perfluorooctanoate
PFOA	Pregnancy	PFASs	µg/L – Log2	1.9 ± 0	By biomarkers	Perfluorooctanoate
PFOS	Postnatal	PFASs	µg/L – Log2	1.1 ± 0	By biomarkers	Perfluorooctanoate
PFOS	Pregnancy	PFASs	µg/L – Log2	3.4 ± 0	By biomarkers	Perfluorooctanoate

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

PFUNDA	Postnatal	PFASs	µg/L – Log2	-4.6 ± 0	By biomarkers	Perfluoroundecanoate
PFUNDA	Pregnancy	PFASs	µg/L – Log2	-1.8 ± 0	By biomarkers	Perfluoroundecanoate
BPA	Postnatal	Phenols	µg/g – Log2	2.9 ± 0	By biomarkers	Bisphenol A
BPA	Pregnancy	Phenols	µg/g – Log2	2.4 ± 0	By biomarkers	Bisphenol A
BUPA	Postnatal	Phenols	µg/g – Log2	-2.7 ± 0	By biomarkers	N-Butyl paraben
BUPA	Pregnancy	Phenols	µg/g – Log2	1.7 ± 0.1	By biomarkers	N-Butyl paraben
ETPA	Postnatal	Phenols	µg/g – Log2	0.5 ± 0	By biomarkers	Ethyl paraben
ETPA	Pregnancy	Phenols	µg/g – Log2	3.3 ± 0.1	By biomarkers	Ethyl paraben
MEPA	Postnatal	Phenols	µg/g – Log2	4.1 ± 0.1	By biomarkers	Methyl paraben
MEPA	Pregnancy	Phenols	µg/g – Log2	7.8 ± 0.1	By biomarkers	Methyl paraben
OXBE	Postnatal	Phenols	µg/g – Log2	1.5 ± 0.1	By biomarkers	Oxybenzone
OXBE	Pregnancy	Phenols	µg/g – Log2	3.4 ± 0.1	By biomarkers	Oxybenzone
PRPA	Postnatal	Phenols	µg/g – Log2	-1.3 ± 0.1	By biomarkers	Propyl paraben
PRPA	Pregnancy	Phenols	µg/g – Log2	5.9 ± 0.1	By biomarkers	Propyl paraben
TRCS	Postnatal	Phenols	µg/g – Log2	-0.3 ± 0.1	By biomarkers	Triclosan
TRCS	Pregnancy	Phenols	µg/g – Log2	3.8 ± 0.1	By biomarkers	Triclosan
MBzP	Postnatal	Phthalates	µg/g – Log2	3.2 ± 0	By biomarkers	Mono benzyl phthalate
MBzP	Pregnancy	Phthalates	µg/g – Log2	3.8 ± 0	By biomarkers	Mono benzyl phthalate

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

MECPP	Postnatal	Phthalates	µg/g – Log2	6 ± 0	By biomarkers	Mono-2-ethyl 5-carboxypentyl phthalate
MECPP	Pregnancy	Phthalates	µg/g – Log2	6 ± 0	By biomarkers	Mono-2-ethyl 5-carboxypentyl phthalate
MEHHP	Postnatal	Phthalates	µg/g – Log2	5.2 ± 0	By biomarkers	Mono-2-ethyl-5-hydroxyhexyl phthalate
MEHHP	Pregnancy	Phthalates	µg/g – Log2	5 ± 0	By biomarkers	Mono-2-ethyl-5-hydroxyhexyl phthalate
MEHP	Postnatal	Phthalates	µg/g – Log2	2.4 ± 0	By biomarkers	Mono-2-ethylhexyl phthalate
MEHP	Pregnancy	Phthalates	µg/g – Log2	3.7 ± 0	By biomarkers	Mono-2-ethylhexyl phthalate
MEOHP	Postnatal	Phthalates	µg/g – Log2	4.5 ± 0	By biomarkers	Mono-2-ethyl-5-oxohexyl phthalate
MEOHP	Pregnancy	Phthalates	µg/g – Log2	4.6 ± 0	By biomarkers	Mono-2-ethyl-5-oxohexyl phthalate
MEP	Postnatal	Phthalates	µg/g – Log2	5.3 ± 0	By biomarkers	Mono-2-ethyl-5-oxohexyl phthalate
MEP	Pregnancy	Phthalates	µg/g – Log2	8.2 ± 0	By biomarkers	Mono-2-ethyl-5-oxohexyl phthalate
MiBP	Postnatal	Phthalates	µg/g – Log2	5.4 ± 0	By biomarkers	Mono-iso-butyl phthalate
MiBP	Pregnancy	Phthalates	µg/g – Log2	6.1 ± 0	By biomarkers	Mono-iso-butyl phthalate
MnBP	Postnatal	Phthalates	µg/g – Log2	4.6 ± 0	By biomarkers	Mono-n-butyl phthalate
MnBP	Pregnancy	Phthalates	µg/g – Log2	5.7 ± 0	By biomarkers	Mono-n-butyl phthalate
OH-MiNP	Postnatal	Phthalates	µg/g – Log2	2.6 ± 0	By biomarkers	Mono-4-methyl-7-hydroxyoctyl phthalate
OH-MiNP	Pregnancy	Phthalates	µg/g – Log2	0.7 ± 0	By biomarkers	Mono-4-methyl-7-hydroxyoctyl phthalate
oxo-MiNP	Postnatal	Phthalates	µg/g – Log2	1.7 ± 0	By biomarkers	Mono-4-methyl-7-oxooctyl phthalate
oxo-MiNP	Pregnancy	Phthalates	µg/g – Log2	0.9 ± 0	By biomarkers	Mono-4-methyl-7-oxooctyl phthalate

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

DEHP (sum of metabolites)	Postnatal	Phthalates	$\mu\text{g/g} - \text{Log2}$	7 ± 0	By biomarkers	Sum of DEHP metabolites
DEHP (sum of metabolites)	Pregnancy	Phthalates	$\mu\text{g/g} - \text{Log2}$	7.8 ± 0	By biomarkers	Sum of DEHP metabolites
Cotinine	Postnatal	Tobacco Smoke	Undetected	957 (0.82)	By biomarkers	
Cotinine	Postnatal	Tobacco Smoke	Detected	216 (0.18)	By biomarkers	
Cotinine	Pregnancy	Tobacco Smoke	$\mu\text{g/L} - \text{None}$	1.5 ± 0	By biomarkers	
PMabsorbance (preg)	Pregnancy	Air Pollution	$10^{-5} \text{ m}^{-1} - \text{Log}$	1 ± 0	By environmental models or questionnaires	
NO2 (preg)	Pregnancy	Air Pollution	$\mu\text{g/m}^3 - \text{Log}$	3.5 ± 0	By environmental models or questionnaires	
PM10 (preg)	Pregnancy	Air Pollution	$\mu\text{g/m}^3 - \text{None}$	23.9 ± 0.2	By environmental models or questionnaires	
PM2.5 (preg)	Pregnancy	Air Pollution	$\mu\text{g/m}^3 - \text{None}$	15.1 ± 0.1	By environmental models or questionnaires	
NO2 (year)	Postnatal	Air Pollution	$\mu\text{g/m}^3 - \text{Log}$	3.5 ± 0	By environmental models or questionnaires	
PM10 (year)	Postnatal	Air Pollution	$\mu\text{g/m}^3 - \text{None}$	25.9 ± 0.2	By environmental models or questionnaires	
PM2.5 (year)	Postnatal	Air Pollution	$\mu\text{g/m}^3 - \text{None}$	13.7 ± 0.1	By environmental models or questionnaires	
PMabsorbance (year)	Postnatal	Air Pollution	$10^{-5} \text{ m}^{-1} - \text{Log}$	0.8 ± 0	By environmental models or questionnaires	
Accessibility (bus lines 300m)	Pregnancy	Built Environment	$\text{m} / \text{km}^2 - \text{Dic}$	0.7 ± 0	By environmental models or questionnaires	
Accessibility (bus stops 300m)	Pregnancy	Built Environment	$\text{m} / \text{km}^2 - \text{Dic}$	2.5 ± 0	By environmental models or questionnaires	
Built density (300m)	Pregnancy	Built Environment	$\text{m}^2 \text{ built} / \text{km}^2$	418.9 ± 4.3	By environmental models or questionnaires	
Connectivity (300m)	Pregnancy	Built Environment	number of intersections / km^2	12.2 ± 0.1	By environmental models or questionnaires	

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Facility richness (300m)	Pregnancy	Built Environment	-	0.1 ± 0	By environmental models or questionnaires
Land use (300m)	Pregnancy	Built Environment	-	0.4 ± 0	By environmental models or questionnaires
Population density	Pregnancy	Built Environment	people / km ²	77.2 ± 1.3	By environmental models or questionnaires
Walkability	Pregnancy	Built Environment	-	0.3 ± 0	By environmental models or questionnaires
Accessibility (bus lines 300m)	Postnatal	Built Environment	m / km ² - Dic	0.7 ± 0	By environmental models or questionnaires
Accessibility (bus lines 300m - school)	Postnatal	Built Environment	m / km ² - Dic	0.7 ± 0	By environmental models or questionnaires
Accessibility (bus stops 300m - school)	Postnatal	Built Environment	m / km ² - Log	2.5 ± 0	By environmental models or questionnaires
Built density (300m)	Postnatal	Built Environment	m ² built / km ² – Square root	390.3 ± 4.6	By environmental models or questionnaires
Built density (300m - school)	Postnatal	Built Environment	m ² built / km ² – Square root	409.8 ± 4.3	By environmental models or questionnaires
Connectivity density (300 m)	Postnatal	Built Environment	number of intersections / km ² - Log	5.3 ± 0	By environmental models or questionnaires
Connectivity density (300m - school)	Postnatal	Built Environment	number of intersections / km ² - Log	5.4 ± 0	By environmental models or questionnaires
Facility density (300m - school)	Postnatal	Built Environment	n / km ² - Log	3.4 ± 0	By environmental models or questionnaires
Facility richness	Postnatal	Built Environment	-- None	0.1 ± 0	By environmental models or questionnaires
Facility richness (300m - school)	Postnatal	Built Environment	-- None	0.1 ± 0	By environmental models or questionnaires
Land use (300m)	Postnatal	Built Environment	-- None	0.4 ± 0	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Land use (300m - school)	Postnatal	Built Environment	-- None	0.4 ± 0	By environmental models or questionnaires
Population density	Postnatal	Built Environment	people / km ² – Square root	69.4 ± 1.3	By environmental models or questionnaires
Population density (school)	Postnatal	Built Environment	people / km ² – Square root	69.8 ± 1.1	By environmental models or questionnaires
Walkability index	Postnatal	Built Environment	-- None	0.3 ± 0	By environmental models or questionnaires
Facility density (300m)	Pregnancy	Built Environment	n / km ² – Log	3.4 ± 0	By environmental models or questionnaires
Facility density (300m)	Postnatal	Built Environment	n / km ² – Log	3.3 ± 0	By environmental models or questionnaires
Accessibility (bus stops 300m)	Postnatal	Built Environment	m / km ² – Log	2.7 ± 0	By environmental models or questionnaires
Indoor PMabsorbance	Postnatal	Indoor air	10 ⁻⁵ m ⁻¹ – Log	0.4 ± 0	By environmental models or questionnaires
Indoor benzene	Postnatal	Indoor air	µg/m ³ – Log	1.2 ± 0	By environmental models or questionnaires
Indoor NO2	Postnatal	Indoor air	µg/m ³ – Log	4.3 ± 0	By environmental models or questionnaires
Indoor PM2.5	Postnatal	Indoor air	µg/m ³ – Log	3 ± 0	By environmental models or questionnaires
Indoor BTEX	Postnatal	Indoor air	µg/m ³ – Log	3.6 ± 0	By environmental models or questionnaires
Alcohol intake	Pregnancy	Lifestyle		0 811 (0.69)	By environmental models or questionnaires
Alcohol intake	Pregnancy	Lifestyle		1 362 (0.31)	By environmental models or questionnaires
Cereals intake	Pregnancy	Lifestyle		1 459 (0.39)	By environmental models or questionnaires
Cereals intake	Pregnancy	Lifestyle		2 393 (0.34)	By environmental models or questionnaires
Cereals intake	Pregnancy	Lifestyle		3 321 (0.27)	By environmental models or questionnaires
Dairy intake	Pregnancy	Lifestyle		1 399 (0.34)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Dairy intake	Pregnancy	Lifestyle	2	404 (0.34)	By environmental models or questionnaires
Dairy intake	Pregnancy	Lifestyle	3	370 (0.32)	By environmental models or questionnaires
Fast-food intake	Pregnancy	Lifestyle	1	257 (0.22)	By environmental models or questionnaires
Fast-food intake	Pregnancy	Lifestyle	2	127 (0.11)	By environmental models or questionnaires
Fast-food intake	Pregnancy	Lifestyle	3	789 (0.67)	By environmental models or questionnaires
Fish and seafood intake	Pregnancy	Lifestyle	1	482 (0.41)	By environmental models or questionnaires
Fish and seafood intake	Pregnancy	Lifestyle	2	358 (0.31)	By environmental models or questionnaires
Fish and seafood intake	Pregnancy	Lifestyle	3	333 (0.28)	By environmental models or questionnaires
Folic acid supplementation	Pregnancy	Lifestyle	0	524 (0.45)	By environmental models or questionnaires
Folic acid supplementation	Pregnancy	Lifestyle	1	649 (0.55)	By environmental models or questionnaires
Fruits intake	Pregnancy	Lifestyle	1	447 (0.38)	By environmental models or questionnaires
Fruits intake	Pregnancy	Lifestyle	2	357 (0.3)	By environmental models or questionnaires
Fruits intake	Pregnancy	Lifestyle	3	369 (0.31)	By environmental models or questionnaires
Legumes intake	Pregnancy	Lifestyle	1	277 (0.24)	By environmental models or questionnaires
Legumes intake	Pregnancy	Lifestyle	2	509 (0.43)	By environmental models or questionnaires
Legumes intake	Pregnancy	Lifestyle	3	387 (0.33)	By environmental models or questionnaires
Meat intake	Pregnancy	Lifestyle	1	381 (0.32)	By environmental models or questionnaires
Meat intake	Pregnancy	Lifestyle	2	384 (0.33)	By environmental models or questionnaires
Meat intake	Pregnancy	Lifestyle	3	408 (0.35)	By environmental models or questionnaires
Moderate physical activity (t3)	Pregnancy	Lifestyle	None or sometimes	565 (0.48)	By environmental models or questionnaires
Moderate physical activity (t3)	Pregnancy	Lifestyle	Often	283 (0.24)	By environmental models or questionnaires
Moderate physical activity (t3)	Pregnancy	Lifestyle	Very Often	325 (0.28)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Vigorous physical activity (t3)	Pregnancy	Lifestyle	Low	514 (0.44)	By environmental models or questionnaires
Vigorous physical activity (t3)	Pregnancy	Lifestyle	Medium-High	659 (0.56)	By environmental models or questionnaires
Vegetables intake	Pregnancy	Lifestyle	1	375 (0.32)	By environmental models or questionnaires
Vegetables intake	Pregnancy	Lifestyle	2	411 (0.35)	By environmental models or questionnaires
Vegetables intake	Pregnancy	Lifestyle	3	387 (0.33)	By environmental models or questionnaires
Bakery products intake	Postnatal	Lifestyle	1	409 (0.35)	By environmental models or questionnaires
Bakery products intake	Postnatal	Lifestyle	2	451 (0.38)	By environmental models or questionnaires
Bakery products intake	Postnatal	Lifestyle	3	313 (0.27)	By environmental models or questionnaires
Soda intake	Postnatal	Lifestyle	1	460 (0.39)	By environmental models or questionnaires
Soda intake	Postnatal	Lifestyle	2	418 (0.36)	By environmental models or questionnaires
Soda intake	Postnatal	Lifestyle	3	295 (0.25)	By environmental models or questionnaires
Breakfast cereals intake	Postnatal	Lifestyle	1	408 (0.35)	By environmental models or questionnaires
Breakfast cereals intake	Postnatal	Lifestyle	2	384 (0.33)	By environmental models or questionnaires
Breakfast cereals intake	Postnatal	Lifestyle	3	381 (0.32)	By environmental models or questionnaires
Caffeinated drinks	Postnatal	Lifestyle	1	746 (0.64)	By environmental models or questionnaires
Caffeinated drinks	Postnatal	Lifestyle	2	163 (0.14)	By environmental models or questionnaires
Caffeinated drinks	Postnatal	Lifestyle	3	264 (0.23)	By environmental models or questionnaires
Dairy products intake	Postnatal	Lifestyle	1	396 (0.34)	By environmental models or questionnaires
Dairy products intake	Postnatal	Lifestyle	2	402 (0.34)	By environmental models or questionnaires
Dairy products intake	Postnatal	Lifestyle	3	375 (0.32)	By environmental models or questionnaires
Sleep duration	Postnatal	Lifestyle	10.3 ± 0		By environmental models or questionnaires
Fast-food intake	Postnatal	Lifestyle	1	651 (0.55)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Fast-food intake	Postnatal	Lifestyle	2	382 (0.33)	By environmental models or questionnaires
Fast-food intake	Postnatal	Lifestyle	3	140 (0.12)	By environmental models or questionnaires
KIDMED score	Postnatal	Lifestyle		2.8 ± 0.1	By environmental models or questionnaires
Moderate and vigorous PA	Postnatal	Lifestyle		40.3 ± 0.7	By environmental models or questionnaires
Organic food intake	Postnatal	Lifestyle	1	409 (0.35)	By environmental models or questionnaires
Organic food intake	Postnatal	Lifestyle	2	468 (0.4)	By environmental models or questionnaires
Organic food intake	Postnatal	Lifestyle	3	296 (0.25)	By environmental models or questionnaires
Cat at home	Postnatal	Lifestyle	0	967 (0.82)	By environmental models or questionnaires
Cat at home	Postnatal	Lifestyle	1	206 (0.18)	By environmental models or questionnaires
Dog at home	Postnatal	Lifestyle	0	1006 (0.86)	By environmental models or questionnaires
Dog at home	Postnatal	Lifestyle	1	167 (0.14)	By environmental models or questionnaires
Other pets at home	Postnatal	Lifestyle	No	728 (0.62)	By environmental models or questionnaires
Other pets at home	Postnatal	Lifestyle	Yes	445 (0.38)	By environmental models or questionnaires
Processed meat intake	Postnatal	Lifestyle	1	427 (0.36)	By environmental models or questionnaires
Processed meat intake	Postnatal	Lifestyle	2	484 (0.41)	By environmental models or questionnaires
Processed meat intake	Postnatal	Lifestyle	3	262 (0.22)	By environmental models or questionnaires
Readymade food intake	Postnatal	Lifestyle	1	581 (0.5)	By environmental models or questionnaires
Readymade food intake	Postnatal	Lifestyle	2	293 (0.25)	By environmental models or questionnaires
Readymade food intake	Postnatal	Lifestyle	3	299 (0.25)	By environmental models or questionnaires
Sedentary behaviour	Postnatal	Lifestyle		239.6 ± 3.8	By environmental models or questionnaires
Bread intake	Postnatal	Lifestyle	1	531 (0.45)	By environmental models or questionnaires
Bread intake	Postnatal	Lifestyle	2	360 (0.31)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Bread intake	Postnatal	Lifestyle	3	282 (0.24)	By environmental models or questionnaires
Cereals intake	Postnatal	Lifestyle	1	395 (0.34)	By environmental models or questionnaires
Cereals intake	Postnatal	Lifestyle	2	394 (0.34)	By environmental models or questionnaires
Cereals intake	Postnatal	Lifestyle	3	384 (0.33)	By environmental models or questionnaires
Fish and seafood intake	Postnatal	Lifestyle	1	468 (0.4)	By environmental models or questionnaires
Fish and seafood intake	Postnatal	Lifestyle	2	356 (0.3)	By environmental models or questionnaires
Fish and seafood intake	Postnatal	Lifestyle	3	349 (0.3)	By environmental models or questionnaires
Fruits intake	Postnatal	Lifestyle	1	386 (0.33)	By environmental models or questionnaires
Fruits intake	Postnatal	Lifestyle	2	396 (0.34)	By environmental models or questionnaires
Fruits intake	Postnatal	Lifestyle	3	391 (0.33)	By environmental models or questionnaires
Total fat intake	Postnatal	Lifestyle	1	467 (0.4)	By environmental models or questionnaires
Total fat intake	Postnatal	Lifestyle	2	323 (0.28)	By environmental models or questionnaires
Total fat intake	Postnatal	Lifestyle	3	383 (0.33)	By environmental models or questionnaires
Meat intake	Postnatal	Lifestyle	1	444 (0.38)	By environmental models or questionnaires
Meat intake	Postnatal	Lifestyle	2	329 (0.28)	By environmental models or questionnaires
Meat intake	Postnatal	Lifestyle	3	400 (0.34)	By environmental models or questionnaires
Potatoes intake	Postnatal	Lifestyle	1	417 (0.36)	By environmental models or questionnaires
Potatoes intake	Postnatal	Lifestyle	2	506 (0.43)	By environmental models or questionnaires
Potatoes intake	Postnatal	Lifestyle	3	250 (0.21)	By environmental models or questionnaires
Sweets intake	Postnatal	Lifestyle	1	382 (0.33)	By environmental models or questionnaires
Sweets intake	Postnatal	Lifestyle	2	394 (0.34)	By environmental models or questionnaires
Sweets intake	Postnatal	Lifestyle	3	397 (0.34)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Vegetables intake	Postnatal	Lifestyle		1	594 (0.51)	By environmental models or questionnaires
Vegetables intake	Postnatal	Lifestyle		2	224 (0.19)	By environmental models or questionnaires
Vegetables intake	Postnatal	Lifestyle		3	355 (0.3)	By environmental models or questionnaires
Yogurt intake	Postnatal	Lifestyle		1	525 (0.45)	By environmental models or questionnaires
Yogurt intake	Postnatal	Lifestyle		2	264 (0.23)	By environmental models or questionnaires
Yogurt intake	Postnatal	Lifestyle		3	384 (0.33)	By environmental models or questionnaires
Humidity (preg)	Pregnancy	Meteorological	% - None		76 ± 0.3	By environmental models or questionnaires
Humidity (t1)	Pregnancy	Meteorological	% - None		76 ± 0.3	By environmental models or questionnaires
Pressure (preg)	Pregnancy	Meteorological	Bar - None		995.1 ± 0.4	By environmental models or questionnaires
Pressure (t1)	Pregnancy	Meteorological	Bar - None		994.7 ± 0.4	By environmental models or questionnaires
Temperature (preg)	Pregnancy	Meteorological	°C - None		11.4 ± 0.1	By environmental models or questionnaires
Temperature (t1)	Pregnancy	Meteorological	°C - None		10.8 ± 0.2	By environmental models or questionnaires
Blue spaces (300 m)	Pregnancy	Natural Spaces	m - Log	0	1085 (0.92)	By environmental models or questionnaires
Blue spaces (300 m)	Pregnancy	Natural Spaces	m - Log	1	88 (0.08)	By environmental models or questionnaires
Green spaces (300 m)	Pregnancy	Natural Spaces	m - Log	0	292 (0.25)	By environmental models or questionnaires
Green spaces (300 m)	Pregnancy	Natural Spaces	m - Log	1	881 (0.75)	By environmental models or questionnaires
NDVI (100 m)	Pregnancy	Natural Spaces	NDVI - None		0.4 ± 0	By environmental models or questionnaires
Blue spaces (300 m)	Postnatal	Natural Spaces	m - Log	0	1077 (0.92)	By environmental models or questionnaires
Blue spaces (300 m)	Postnatal	Natural Spaces	m - Log	1	96 (0.08)	By environmental models or questionnaires
Blue spaces (300m - school)	Postnatal	Natural Spaces	m - Log	0	1092 (0.93)	By environmental models or questionnaires
Blue spaces (300m - school)	Postnatal	Natural Spaces	m - Log	1	81 (0.07)	By environmental models or questionnaires
Green spaces (300 m)	Postnatal	Natural Spaces	m - Log	0	245 (0.21)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Green spaces (300 m)	Postnatal	Natural Spaces	m - Log	1	928 (0.79)	By environmental models or questionnaires
Green spaces (300m - school)	Postnatal	Natural Spaces	m - Log	0	263 (0.22)	By environmental models or questionnaires
Green spaces (300m - school)	Postnatal	Natural Spaces	m - Log	1	910 (0.78)	By environmental models or questionnaires
NDVI (100 m)	Postnatal	Natural Spaces	NDVI - None	0.4 ± 0		By environmental models or questionnaires
NDVI (100 m - school)	Postnatal	Natural Spaces	NDVI - None	0.4 ± 0		By environmental models or questionnaires
Traffic noise (24h)	Pregnancy	Noise		1	551 (0.47)	By environmental models or questionnaires
Traffic noise (24h)	Pregnancy	Noise		2	207 (0.18)	By environmental models or questionnaires
Traffic noise (24h)	Pregnancy	Noise		3	259 (0.22)	By environmental models or questionnaires
Traffic noise (24h)	Pregnancy	Noise		4	156 (0.13)	By environmental models or questionnaires
Traffic noise (night)	Pregnancy	Noise		1	911 (0.78)	By environmental models or questionnaires
Traffic noise (night)	Pregnancy	Noise		2	152 (0.13)	By environmental models or questionnaires
Traffic noise (night)	Pregnancy	Noise		3	72 (0.06)	By environmental models or questionnaires
Traffic noise (night)	Pregnancy	Noise		4	24 (0.02)	By environmental models or questionnaires
Traffic noise (night)	Pregnancy	Noise		5	14 (0.01)	By environmental models or questionnaires
Traffic noise (24h)	Postnatal	Noise		1	616 (0.53)	By environmental models or questionnaires
Traffic noise (24h)	Postnatal	Noise		2	190 (0.16)	By environmental models or questionnaires
Traffic noise (24h)	Postnatal	Noise		3	250 (0.21)	By environmental models or questionnaires
Traffic noise (24h)	Postnatal	Noise		4	117 (0.1)	By environmental models or questionnaires
Traffic noise (24h - school)	Postnatal	Noise	dB - None	54.1 ± 0.2		By environmental models or questionnaires
Traffic noise (night)	Postnatal	Noise		1	934 (0.8)	By environmental models or questionnaires
Traffic noise (night)	Postnatal	Noise		2	94 (0.08)	By environmental models or questionnaires
Traffic noise (night)	Postnatal	Noise		3	86 (0.07)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Traffic noise (night)	Postnatal	Noise		4	59 (0.05)	By environmental models or questionnaires
Family affluence score	Postnatal	Socio-eco capital	Low		134 (0.11)	By environmental models or questionnaires
Family affluence score	Postnatal	Socio-eco capital	Middle		469 (0.4)	By environmental models or questionnaires
Family affluence score	Postnatal	Socio-eco capital	High		570 (0.49)	By environmental models or questionnaires
Contact with family and friends	Postnatal	Socio-eco capital	Less than once a week		50 (0.04)	By environmental models or questionnaires
Contact with family and friends	Postnatal	Socio-eco capital	Once a week		338 (0.29)	By environmental models or questionnaires
Contact with family and friends	Postnatal	Socio-eco capital	(almost) Daily		785 (0.67)	By environmental models or questionnaires
House crowding	Postnatal	Socio-eco capital	- N o n e	4.3 ± 0		By environmental models or questionnaires
Social participation	Postnatal	Socio-eco capital	None		667 (0.57)	By environmental models or questionnaires
Social participation	Postnatal	Socio-eco capital	1 organisation		332 (0.28)	By environmental models or questionnaires
Social participation	Postnatal	Socio-eco capital	2 or more organisations		174 (0.15)	By environmental models or questionnaires
Cigarette	Pregnancy	Tobacco Smoke	- N o n e	0.6 ± 0.1		By environmental models or questionnaires
Active smoking (preg)	Pregnancy	Tobacco Smoke	no		998 (0.85)	By environmental models or questionnaires
Active smoking (preg)	Pregnancy	Tobacco Smoke	yes		175 (0.15)	By environmental models or questionnaires
ETS	Postnatal	Tobacco Smoke	no exposure		745 (0.64)	By environmental models or questionnaires
ETS	Postnatal	Tobacco Smoke	exposure		428 (0.36)	By environmental models or questionnaires
Parental smoking	Postnatal	Tobacco Smoke	neither		722 (0.62)	By environmental models or questionnaires
Parental smoking	Postnatal	Tobacco Smoke	one		325 (0.28)	By environmental models or questionnaires
Parental smoking	Postnatal	Tobacco Smoke	both		126 (0.11)	By environmental models or questionnaires
Maternal smoking (active and ETS)	Pregnancy	Tobacco Smoke	no exposure		624 (0.53)	By environmental models or questionnaires

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Maternal smoking (active and ETS)	Pregnancy	Tobacco Smoke	only passive exposure smoker	374 (0.32)	By environmental models or questionnaires
Maternal smoking (active and ETS)	Pregnancy	Tobacco Smoke	smoker	175 (0.15)	By environmental models or questionnaires
Inverse distance to nearest road	Pregnancy	Traffic	m^{-1} - None	-2.5 ± 0	By environmental models or questionnaires
Road traffic load (100 m)	Pregnancy	Traffic	Vehm./da y.m -cube root	71.9 ± 1.9	By environmental models or questionnaires
Traffic density on nearest road	Pregnancy	Traffic	veh/day.m -cube root	12.5 ± 0.2	By environmental models or questionnaires
Inverse distance to nearest road	Postnatal	Traffic	m^{-1} - None	-3.5 ± 0	By environmental models or questionnaires
Road traffic load (100 m)	Postnatal	Traffic	veh/day.m -cube root	71.7 ± 2	By environmental models or questionnaires
Inverse distance to nearest road (school)	Pregnancy	Traffic	m^{-1} - None	16.2 ± 0.3	By environmental models or questionnaires
Traffic load of major roads (100 m)	Pregnancy	Traffic		0	915 (0.78)
Traffic load of major roads (100 m)	Pregnancy	Traffic		1	258 (0.22)
Traffic load of major roads (100 m - school)	Pregnancy	Traffic		0	884 (0.75)
Traffic load of major roads (100 m - school)	Pregnancy	Traffic		1	289 (0.25)
Water Brominated THMs	Pregnancy	Water disinfection by-products	ng/L - None	1.6 ± 0.1	By environmental models or questionnaires
Water Chloroform	Pregnancy	Water disinfection by-products	ng/L - None	1.5 ± 0	By environmental models or questionnaires
Water THMs	Pregnancy	Water disinfection by-products	ng/L - None	3 ± 0	By environmental models or questionnaires

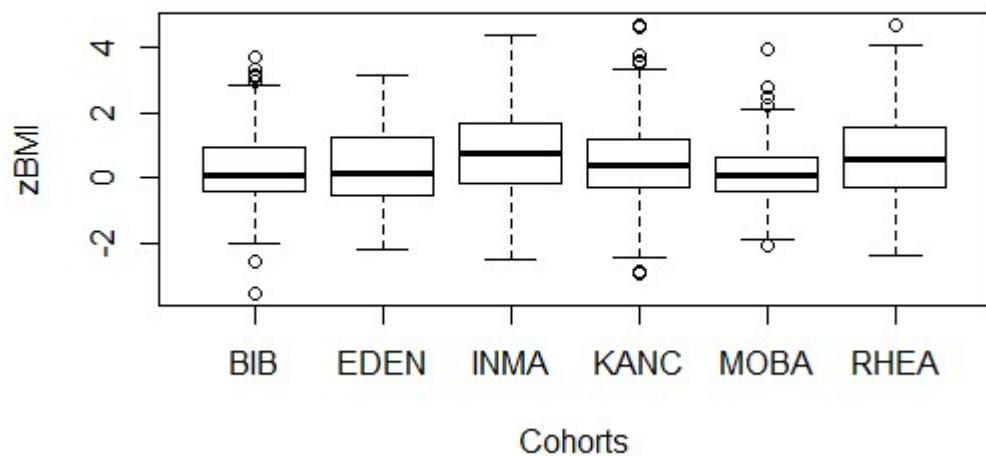
OCs : Organochlorine compounds ; Brominated ; PFAS: Perfluorinated alkylated substances; PBDEs: Brominated compounds Ops: Organophosphate pesticide metabolites

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach
using preselected methylation marks

Supplementary Material II.2: Pathways identified as relevant for zBMI relying on KEGG database, and corresponding numbers of genes and enhancer CpGs

Pathway	Number of genes	Number of enhancer CpGs
Fat digestion and absorption	41	84
Fatty acid elongation	30	93
Fatty acid degradation	44	79
Biosynthesis of unsaturated fatty acids	23	66
Vitamin digestion and absorption	24	80
Bile secretion	71	304
PPAR signalling pathway	74	202
Insulin resistance	107	478
Regulation of lipolysis in adipocytes	54	299
Adipocytokine signalling pathway	69	194
Type II diabetes mellitus	46	269
Ribosome biogenesis in eukaryotes	105	52
Thyroid hormone signalling pathway	116	524
Fanconi anemia pathway	54	51
GnRH signalling pathway	93	518
Prolactin signalling pathway	70	271

Supplementary Material II.3: Boxplot of child zBMI in HELIX data, by cohorts.



CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

Supplementary Material II.4: Population characteristics by cohort

Cohorts (number of individuals)	BIB		EDEN		INMA		KANC		MOBA		RHEA	
Characteristic	Mean (SD)	n (%)										
Child BMI	16.0 (2.0)		17.9 (2.7)		18.1 (3.1)		16.4 (2.3)		16.3 (1.9)		16.8 (2.6)	
Child sex												
Female	91 (45)		63 (43)		98 (46)		91 (46)		98 (46)		88 (44)	
Male	112 (55)		83 (57)		117 (54)		107 (54)		114 (54)		111 (56)	
Child age (year)	6.6 (0.2)		10.7 (0.5)		8.82 (0.6)		6.5 (0.5)		8.5 (0.5)		6.5 (0.3)	
Maternal education												
low	95 (47)		10 (7)		50 (23)		12 (6)		0 (0)		9 (5)	
middle	36 (18)		53 (36)		91 (42)		69 (35)		43 (20)		110 (55)	
high	72 (35)		83 (57)		74 (34)		117 (59)		169 (80)		80 (40)	
Maternal pre- pregnancy BMI (kg/m ²)	28.2 (5.3)		23.3 (4.2)		24 (4.7)		27.6 (5)		22.7 (3.3)		24.1 (4.3)	
Parity before index pregnancy												
0	86 (42)		70 (48)		116 (54)		85 (43)		97 (46)		76 (38)	
1	57 (28)		51 (35)		90 (42)		57 (29)		88 (42)		87 (44)	
2 or more	60 (30)		25 (17)		9 (4)		56 (28)		27 (13)		36 (18)	
Trimester of conception												
January- March	73 (36)		51 (35)		46 (21)		63 (32)		75 (35)		60 (30)	
April-June	20 (10)		30 (21)		48 (22)		28 (14)		43 (20)		65 (33)	
July- September	34 (17)		24 (16)		60 (28)		64 (32)		35 (17)		43 (22)	
October- December	76 (37)		41 (28)		61 (28)		43 (22)		59 (28)		31 (16)	
Maternal smoking status during pregnancy												
no exposure	137 (67)		99 (68)		78 (36)		102 (52)		190 (90)		18 (9)	

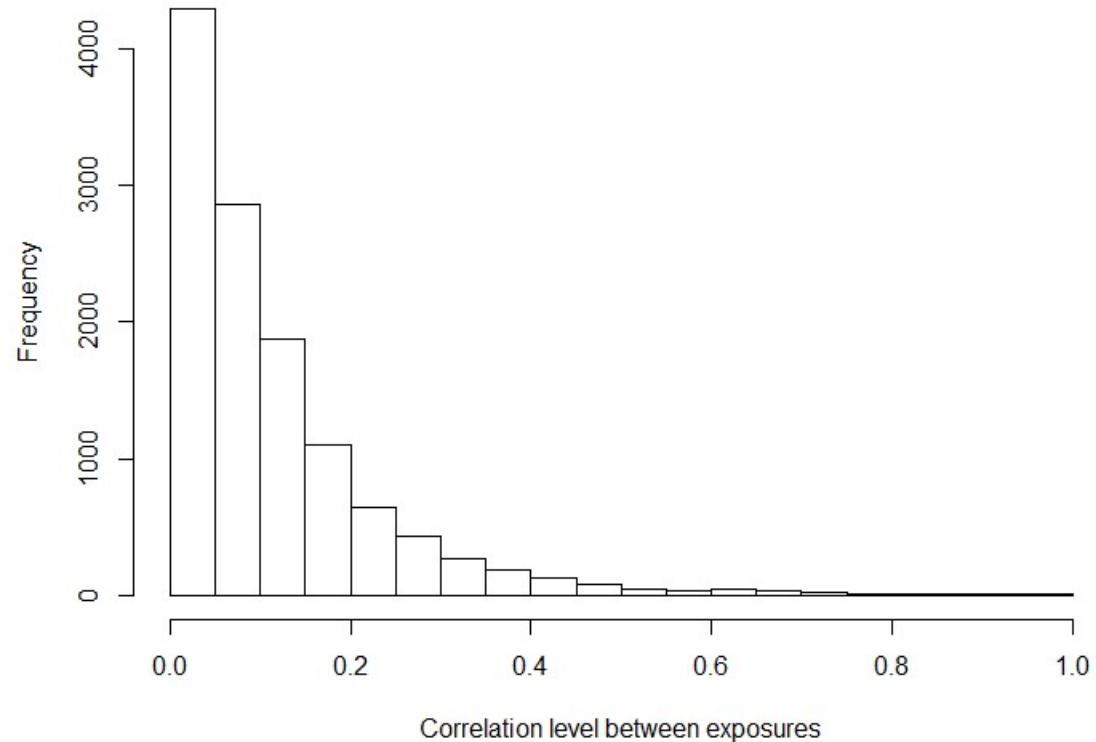
CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach using preselected methylation marks

only passive exposure	39 (19)	16 (11)	84 (39)	84 (42)	12 (6)	139 (70)
smoker	27 (13)	31 (21)	53 (25)	12 (6)	10 (5)	42 (21)
Postnatal tobacco smoke exposure						
<i>not exposed</i>	145 (71)	108 (74)	146 (68)	115 (58)	171 (81)	60 (30)
<i>exposed</i>	58 (29)	38 (26)	69 (32)	83 (42)	41 (19)	139 (70)
Maternal age (years)	28.7 (5.8)	30.8 (4.9)	31.9 (4)	29.1 (4.9)	32.7 (3.7)	30.9 (4.8)
Birthweight						
< 2500g	13 (6)	5 (3)	5 (2)	5 (3)	6 (3)	6 (3)
2500 to 3500g	114 (56)	92 (63)	141 (66)	75 (38)	97 (46)	143 (72)
3500 to 4000g	56 (28)	36 (25)	62 (29)	84 (42)	76 (36)	43 (22)
≥ 4000g	20 (10)	13 (9)	7 (3)	34 (17)	33 (16)	7 (4)
Breastfeeding duration						
< 10.8 weeks	68 (33)	81 (55)	60 (28)	32 (16)	32 (15)	88 (44)
10.8 to 34.9 weeks	87 (43)	52 (36)	98 (46)	61 (31)	42 (20)	79 (40)
> 34.9 weeks	48 (24)	13 (9)	57 (27)	105 (53)	138 (65)	32 (16)
Parents born in the country of inclusion						
None	81 (40)	0 (0)	1 (0)	8 (4)	42 (20)	2 (1)
Only one	31 (15)	10 (7)	10 (5)	0 (0)	1 (0)	6 (3)
Both	91 (45)	136 (93)	204 (95)	190 (96)	169 (80)	191 (96)
Ethnicity						
African	7 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Asian	13 (6)	0 (0)	0 (0)	0 (0)	6 (3)	0 (0)
European ancestry	87 (43)	146 (100)	215 (100)	198 (100)	203 (96)	199 (100)
Native American	0 (0)	0 (0)	0 (0)	0 (0)	2 (1)	0 (0)

CHAPTER II: Early-life Exposures and child Body Mass Index: a Meet-in-the-Middle approach
using preselected methylation marks

Pakistani	79 (39)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	17 (8)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)

Supplementary Material II.5: Distribution of pairwise coefficients of correlation within quantitative variables of the full exposome assessed in 1,173 mother-child pairs from HELIX cohort



Supplementary Material II.6: Adjusted associations between the reduced methylome (2284 CpGs) and zBMI in 1,173 children from the HELIX cohort (ExWAS model, step b) of the Meet-in-the-Middle approach. Results are presented only for CpGs with a (FDR-corrected for multiple hypothesis testing) p-value below 0.05 in ExWAS.

CpG site	Gene	Effect estimate	95% CI	Uncorrected p-Value	FDR-corrected p-Value
cg23098018	PIK3CD	3.64	2.12 5.16	2.80x10 ⁻⁶	3.20x10 ⁻³
cg01943221	PIK3CD	4.16	2.47 5.85	1.51x10 ⁻⁶	3.20x10 ⁻³
cg06695691	SPATA5	-3.99	-5.70 -2.27	5.65x10 ⁻⁶	4.30x10 ⁻³
cg03781224	MGLL	-5.79	-8.38 -3.20	1.29x10 ⁻⁵	5.96x10 ⁻³
cg25905215	NFKB1	5.23	2.88 7.57	1.31x10 ⁻⁵	5.96x10 ⁻³
cg09706586	PIK3CD	3.37	1.81 4.93	2.46x10 ⁻⁵	8.19x10 ⁻³
cg11947782	PIK3CD	3.05	1.63 4.48	2.80x10 ⁻⁵	8.19x10 ⁻³
cg18288462	ELOVL3	-3.70	-5.43 -1.96	3.14x10 ⁻⁵	8.19x10 ⁻³
cg12228229	DLG4	-9.94	-14.63 -5.24	3.58x10 ⁻⁵	8.19x10 ⁻³
cg27110374	GRB2	2.68	1.41 3.94	3.48x10 ⁻⁵	8.19x10 ⁻³
cg15526535	TNFRSF1B	2.31	1.18 3.44	6.48x10 ⁻⁵	0.01
cg09900893	RPS6KA1	-4.54	-6.78 -2.31	7.10x10 ⁻⁵	0.01
cg05805445	SPATA5	2.64	1.29 4.00	1.31x10 ⁻⁴	0.02
cg09035699	ACSL6	2.55	1.23 3.86	1.56x10 ⁻⁴	0.02
cg14841483	ACSL6	2.80	1.37 4.24	1.35x10 ⁻⁴	0.02
cg14003265	TRAF2	2.74	1.33 4.15	1.51x10 ⁻⁴	0.02
cg02423534	ADCY6	-2.69	-4.08 -1.31	1.37x10 ⁻⁴	0.02
cg00810292	TBC1D4	1.66	0.81 2.50	1.28x10 ⁻⁴	0.02
cg12085119	IRS2	3.31	1.63 4.99	1.12x10 ⁻⁴	0.02
cg16702014	ABCC1	-2.13	-3.23 -1.04	1.38x10 ⁻⁴	0.02
cg01312837	CREBBP	-4.14	-6.27 -2.01	1.44x10 ⁻⁴	0.02
cg22435313	PRKCA	2.00	0.97 3.04	1.52x10 ⁻⁴	0.02
cg22284398	PRKCE	-2.98	-4.53 -1.43	1.73x10 ⁻⁴	0.02
cg02105211	ITPR1	-1.57	-2.39 -0.75	1.74x10 ⁻⁴	0.02
cg02099877	NCEH1	-3.07	-4.68 -1.46	1.89x10 ⁻⁴	0.02
cg11010552	ACSL6	-2.44	-3.72 -1.15	2.08x10 ⁻⁴	0.02

CpG site	Gene	Effect estimate	95% CI	Uncorrected p-Value	FDR-corrected p-Value
cg05004855	CAMK2A	-3.84	-5.86 -1.82	2.03x10 ⁻⁴	0.02
cg09365147	REV3L	-1.18	-1.80 -0.56	2.09x10 ⁻⁴	0.02
cg19791262	HK3	-3.22	-4.94 -1.49	2.68x10 ⁻⁴	0.02
cg18681426	ELOVL5	-4.29	-6.61 -1.98	2.78x10 ⁻⁴	0.02
cg13941235	RXRA	-2.13	-3.27 -0.98	2.72x10 ⁻⁴	0.02
cg04056757	GRB2	3.12	1.45 4.78	2.53x10 ⁻⁴	0.02
cg07298473	NR1H3	-2.34	-3.61 -1.06	3.28x10 ⁻⁴	0.02
cg09877009	PRKCQ	-3.72	-5.76 -1.69	3.44x10 ⁻⁴	0.02
cg13832670	CREB3L2	-1.98	-3.08 -0.88	4.18x10 ⁻⁴	0.03
cg23875758	SREBF1	-5.80	-9.02 -2.59	4.14x10 ⁻⁴	0.03
cg00793946	PRKCE	-9.05	-14.07 -4.02	4.35x10 ⁻⁴	0.03
cg04730825	ABCC1	-2.47	-3.85 -1.10	4.41x10 ⁻⁴	0.03
cg07217499	CACNA1C	-3.87	-6.07 -1.67	5.87x10 ⁻⁴	0.03
cg19542445	CACNA1C	-2.98	-4.68 -1.28	5.96x10 ⁻⁴	0.03
cg07382687	CREB3L2	-2.32	-3.65 -0.99	6.30x10 ⁻⁴	0.03
cg12978800	PRKAG2	-2.49	-3.93 -1.06	6.75x10 ⁻⁴	0.03
cg13487983	RXRA	-4.39	-6.91 -1.86	6.75x10 ⁻⁴	0.03
cg16401207	PRKG1	-3.91	-6.15 -1.66	6.67x10 ⁻⁴	0.03
cg25338454	ITPR2	-1.43	-2.26 -0.61	6.45x10 ⁻⁴	0.03
cg27340723	ADCY9	-1.91	-3.02 -0.81	7.21x10 ⁻⁴	0.04
cg09265397	NOTCH3	-2.55	-4.02 -1.07	7.54x10 ⁻⁴	0.04
cg23257225	ADORA1	-3.82	-6.04 -1.60	7.72x10 ⁻⁴	0.04
cg19782686	SPATA5	-1.43	-2.26 -0.59	8.33x10 ⁻⁴	0.04
cg18912768	ABCB11	-2.14	-3.41 -0.87	9.64x10 ⁻⁴	0.04
cg18390025	ELOVL3	-1.57	-2.50 -0.64	9.75x10 ⁻⁴	0.04
cg00536939	NR1H3	-1.62	-2.59 -0.66	9.59x10 ⁻⁴	0.04
cg16942632	CAMKK2	-1.19	-1.89 -0.49	8.96x10 ⁻⁴	0.04
cg19982668	MAP3K3	-2.32	-3.70 -0.95	9.56x10 ⁻⁴	0.04
cg24961795	PLCG1	1.67	0.68 2.66	9.29x10 ⁻⁴	0.04
cg13092108	RPS6KA1	-2.97	-4.74 -1.20	1.03x10 ⁻³	0.04

CpG site	Gene	Effect estimate	95% CI	Uncorrected p-Value	FDR-corrected p-Value
cg18537222	PPARG	-1.08	-1.73 -0.43	1.08x10 ⁻³	0.04
cg05379597	CAMK2G	-2.04	-3.26 -0.81	1.12x10 ⁻³	0.04
cg09473249	ABCC1	-6.15	-9.84 -2.46	1.11x10 ⁻³	0.04
cg12005026	PTEN	-2.37	-3.80 -0.94	1.22x10 ⁻³	0.05
cg04885396	ABCC1	-1.71	-2.76 -0.67	1.28x10 ⁻³	0.05
cg23369234	CACNA1C	-5.80	-9.34 -2.26	1.35x10 ⁻³	0.05

Supplementary Material II.7: Adjusted associations between exposures and CpGs associated with childhood zBMI in 1,173 children from HELIX cohort (ExWAS model, step c) of the Meet-in-the-Middle approach. Results are presented only for CpGs with a (FDR-corrected for multiple hypothesis testing) p-value below 0.05 in ExWAS.

CpG sites	Gene	Exposure	Effect	CI 95%	Uncorrected p-Value	FDR corrected
			estimate			p -Value
cg09035699	ACSL6	Copper - Postnatal	0.01	8.77x10 ⁻³	0.02	4.23x10 ⁻¹¹
cg12085119	IRS2	Copper - Postnatal	9.30x10 ⁻³	6.42x10 ⁻³	0.01	3.40x10 ⁻¹⁰
cg22435313	PRKCA	Copper - Postnatal	0.01	0.01	0.02	4.19x10 ⁻¹⁰
cg07217499	CACNA1C	Copper - Postnatal	-6.85x10 ⁻³	-9.05x10 ⁻³	-4.65x10 ⁻³	1.39x10 ⁻⁹
cg09365147	REV3L	Copper - Postnatal	-0.02	-0.03	-0.02	1.19x10 ⁻⁸
cg04056757	GRB2	Copper - Postnatal	8.37x10 ⁻³	5.46x10 ⁻³	0.01	2.14x10 ⁻⁸
cg05004855	CAMK2A	Copper - Postnatal	-6.79x10 ⁻³	-9.19x10 ⁻³	-4.39x10 ⁻³	3.58x10 ⁻⁸
cg24961795	PLCG1	Copper - Postnatal	0.01	9.02x10 ⁻³	0.02	3.42x10 ⁻⁸
cg05805445	SPATA5	Copper - Postnatal	0.01	6.50x10 ⁻³	0.01	4.23x10 ⁻⁸
cg18537222	PPARG	Copper - Postnatal	-0.02	-0.03	-0.01	5.71x10 ⁻⁸
cg07298473	NR1H3	Copper - Postnatal	-0.01	-0.01	-6.81x10 ⁻³	5.70x10 ⁻⁸
cg23098018	PIK3CD	Copper - Postnatal	8.54x10 ⁻³	5.36x10 ⁻³	0.01	1.59x10 ⁻⁷
cg18681426	ELOVL5	Copper - Postnatal	-5.66x10 ⁻³	-7.76x10 ⁻³	-3.56x10 ⁻³	1.53x10 ⁻⁷
cg19782686	SPATA5	Copper - Postnatal	-0.02	-0.02	-9.68x10 ⁻³	2.05x10 ⁻⁷
cg12005026	PTEN	Copper - Postnatal	-8.98x10 ⁻³	-0.01	-5.59x10 ⁻³	2.51x10 ⁻⁷
cg02105211	ITPR1	Copper - Postnatal	-0.02	-0.02	-9.68x10 ⁻³	2.79x10 ⁻⁷
cg11010552	ACSL6	Copper - Postnatal	-9.91x10 ⁻³	-0.01	-6.13x10 ⁻³	3.17x10 ⁻⁷
cg27110374	GRB2	Copper - Postnatal	1.00x10 ⁻²	6.16x10 ⁻³	0.01	3.60x10 ⁻⁷
cg04730825	ABCC1	Copper - Postnatal	-9.05x10 ⁻³	-0.01	-5.51x10 ⁻³	5.99x10 ⁻⁷
cg18390025	ELOVL3	Copper - Postnatal	-0.01	-0.02	-8.08x10 ⁻³	6.80x10 ⁻⁷
cg00536939	NR1H3	Copper - Postnatal	-0.01	-0.02	-7.71x10 ⁻³	8.67x10 ⁻⁷
cg01943221	PIK3CD	Copper - Postnatal	7.16x10 ⁻³	4.30x10 ⁻³	0.01	1.00x10 ⁻⁶
cg15526535	TNFRSF1B	BPA - Postnatal	-8.70x10 ⁻³	-0.01	-5.23x10 ⁻³	1.01x10 ⁻⁶
cg05379597	CAMK2G	Copper - Postnatal	-9.99x10 ⁻³	-0.01	-6.01x10 ⁻³	9.92x10 ⁻⁷
cg14003265	TRAF2	BPA - Postnatal	-6.92x10 ⁻³	-9.70x10 ⁻³	-4.14x10 ⁻³	1.16x10 ⁻⁶
						6.22x10 ⁻⁴

CpG sites	Gene	Exposure	Effect	CI 95%	Uncorrected p-Value	FDR corrected	
			estimate			p-Value	
cg14841483	ACSL6	Copper - Postnatal	8.41x10 ⁻³	5.02x10 ⁻³	0.01	1.28x10 ⁻⁶	6.60x10 ⁻⁴
cg22284398	PRKCE	Copper - Postnatal	-7.72x10 ⁻³	-0.01	-4.59x10 ⁻³	1.53x10 ⁻⁶	7.61x10 ⁻⁴
cg25338454	ITPR2	Copper - Postnatal	-0.01	-0.02	-8.58x10 ⁻³	1.77x10 ⁻⁶	8.18x10 ⁻⁴
cg04885396	ABCC1	Copper - Postnatal	-0.01	-0.02	-6.79x10 ⁻³	1.72x10 ⁻⁶	8.18x10 ⁻⁴
cg16942632	CAMKK2	Copper - Postnatal	-0.02	-0.02	-0.01	1.92x10 ⁻⁶	8.56x10 ⁻⁴
cg19982668	MAP3K3	Copper - Postnatal	-8.60x10 ⁻³	-0.01	-5.06x10 ⁻³	2.17x10 ⁻⁶	9.36x10 ⁻⁴
cg01312837	CREBBP	Copper - Postnatal	-5.52x10 ⁻³	-7.80x10 ⁻³	-3.23x10 ⁻³	2.45x10 ⁻⁶	1.03x10 ⁻³
cg11947782	PIK3CD	Copper - Postnatal	8.19x10 ⁻³	4.78x10 ⁻³	0.01	2.68x10 ⁻⁶	1.09x10 ⁻³
cg09706586	PIK3CD	Copper - Postnatal	7.41x10 ⁻³	4.30x10 ⁻³	0.01	3.30x10 ⁻⁶	1.30x10 ⁻³
cg02423534	ADCY6	Copper - Postnatal	-8.34x10 ⁻³	-0.01	-4.82x10 ⁻³	3.79x10 ⁻⁶	1.45x10 ⁻³
cg16401207	PRKG1	Copper - Postnatal	-5.13x10 ⁻³	-7.30x10 ⁻³	-2.96x10 ⁻³	3.92x10 ⁻⁶	1.46x10 ⁻³
cg27340723	ADCY9	Copper - Postnatal	-0.01	-0.01	-5.78x10 ⁻³	6.42x10 ⁻⁶	2.33x10 ⁻³
cg22284398	PRKCE	BPA - Postnatal	5.79x10 ⁻³	3.26x10 ⁻³	8.33x10 ⁻³	8.19x10 ⁻⁶	2.89x10 ⁻³
cg02099877	NCEH1	Copper - Postnatal	-6.81x10 ⁻³	-9.84x10 ⁻³	-3.78x10 ⁻³	1.14x10 ⁻⁵	3.93x10 ⁻³
cg15526535	TNFRSF1B	Copper - Postnatal	9.64x10 ⁻³	5.33x10 ⁻³	0.01	1.23x10 ⁻⁵	4.12x10 ⁻³
cg18912768	ABCB11	Copper - Postnatal	-8.58x10 ⁻³	-0.01	-4.72x10 ⁻³	1.38x10 ⁻⁵	4.50x10 ⁻³
cg18390025	ELOVL3	BPA - Postnatal	9.34x10 ⁻³	5.11x10 ⁻³	0.01	1.63x10 ⁻⁵	5.07x10 ⁻³
cg16702014	ABCC1	Copper - Postnatal	-9.82x10 ⁻³	-0.01	-5.37x10 ⁻³	1.63x10 ⁻⁵	5.07x10 ⁻³
cg19982668	MAP3K3	BPA - Postnatal	6.30x10 ⁻³	3.43x10 ⁻³	9.17x10 ⁻³	1.78x10 ⁻⁵	5.42x10 ⁻³
cg13832670	CREB3L2	Copper - Postnatal	-9.75x10 ⁻³	-0.01	-5.30x10 ⁻³	1.83x10 ⁻⁵	5.43x10 ⁻³
cg07298473	NR1H3	BPA - Postnatal	6.77x10 ⁻³	3.67x10 ⁻³	9.87x10 ⁻³	1.94x10 ⁻⁵	5.65x10 ⁻³
cg13092108	RPS6KA1	Copper - Postnatal	-6.00x10 ⁻³	-8.76x10 ⁻³	-3.24x10 ⁻³	2.19x10 ⁻⁵	6.12x10 ⁻³
cg09035699	ACSL6	BPA - Postnatal	-6.50x10 ⁻³	-9.49x10 ⁻³	-3.51x10 ⁻³	2.19x10 ⁻⁵	6.12x10 ⁻³
cg00810292	TBC1D4	Copper - Postnatal	0.01	6.54x10 ⁻³	0.02	3.01x10 ⁻⁵	8.01x10 ⁻³
cg04056757	GRB2	BPA - Postnatal	-5.05x10 ⁻³	-7.41x10 ⁻³	-2.68x10 ⁻³	3.05x10 ⁻⁵	8.01x10 ⁻³
cg24961795	PLCG1	BPA - Postnatal	-8.54x10 ⁻³	-0.01	-4.54x10 ⁻³	3.04x10 ⁻⁵	8.01x10 ⁻³
cg23369234	CACNA1C	Copper - Postnatal	-2.94x10 ⁻³	-4.33x10 ⁻³	-1.56x10 ⁻³	3.14x10 ⁻⁵	8.08x10 ⁻³
cg03781224	MGLL	Copper - Postnatal	-3.93x10 ⁻³	-5.81x10 ⁻³	-2.06x10 ⁻³	4.04x10 ⁻⁵	9.89x10 ⁻³
cg12085119	IRS2	BPA - Postnatal	-4.94x10 ⁻³	-7.29x10 ⁻³	-2.59x10 ⁻³	3.92x10 ⁻⁵	9.89x10 ⁻³

CpG sites	Gene	Exposure	Effect	CI 95%	Uncorrected p-Value	FDR corrected	
			estimate			p-Value	p-Value
cg09265397	NOTCH3	Copper - Postnatal	-6.95x10 ⁻³	-0.01	-3.64x10 ⁻³	4.06x10 ⁻⁵	9.89x10 ⁻³
cg07382687	CREB3L2	PFOS - Postnatal	8.85x10 ⁻³	4.62x10 ⁻³	0.01	4.38x10 ⁻⁵	0.01
cg09900893	RPS6KA1	Copper - Postnatal	-4.54x10 ⁻³	-6.72x10 ⁻³	-2.36x10 ⁻³	4.74x10 ⁻⁵	0.01
cg18537222	PPARG	BPA - Postnatal	0.01	6.61x10 ⁻³	0.02	4.66x10 ⁻⁵	0.01
cg22435313	PRKCA	PFOS - Postnatal	-0.01	-0.02	-5.86x10 ⁻³	4.79x10 ⁻⁵	0.01
cg04730825	ABCC1	PFOS - Postnatal	8.37x10 ⁻³	4.28x10 ⁻³	0.01	6.34x10 ⁻⁵	0.01
cg14841483	ACSL6	BPA - Postnatal	-5.61x10 ⁻³	-8.36x10 ⁻³	-2.86x10 ⁻³	6.52x10 ⁻⁵	0.01
cg13092108	RPS6KA1	PFOS - Postnatal	6.43x10 ⁻³	3.24x10 ⁻³	9.61x10 ⁻³	7.88x10 ⁻⁵	0.02
cg05004855	CAMK2A	PFOS - Postnatal	5.59x10 ⁻³	2.81x10 ⁻³	8.37x10 ⁻³	8.41x10 ⁻⁵	0.02
cg06695691	SPATA5	Copper - Postnatal	-5.67x10 ⁻³	-8.50x10 ⁻³	-2.84x10 ⁻³	8.89x10 ⁻⁵	0.02
cg07382687	CREB3L2	Copper - Postnatal	-7.35x10 ⁻³	-0.01	-3.67x10 ⁻³	9.41x10 ⁻⁵	0.02
cg05379597	CAMK2G	BPA - Postnatal	6.41x10 ⁻³	3.18x10 ⁻³	9.64x10 ⁻³	1.05x10 ⁻⁴	0.02
cg19791262	HK3	Copper - Postnatal	-5.59x10 ⁻³	-8.42x10 ⁻³	-2.76x10 ⁻³	1.14x10 ⁻⁴	0.02
cg19791262	HK3	PFOS - Postnatal	6.35x10 ⁻³	3.09x10 ⁻³	9.61x10 ⁻³	1.38x10 ⁻⁴	0.03
cg23875758	SREBF1	Copper - Postnatal	-2.98x10 ⁻³	-4.50x10 ⁻³	-1.45x10 ⁻³	1.37x10 ⁻⁴	0.03
cg22284398	PRKCE	PFOS - Postnatal	7.05x10 ⁻³	3.43x10 ⁻³	0.01	1.43x10 ⁻⁴	0.03
cg01312837	CREBBP	PFOS - Postnatal	5.11x10 ⁻³	2.47x10 ⁻³	7.75x10 ⁻³	1.53x10 ⁻⁴	0.03
cg13832670	CREB3L2	BPA - Postnatal	6.90x10 ⁻³	3.30x10 ⁻³	0.01	1.77x10 ⁻⁴	0.03
cg19782686	SPATA5	PFOS - Postnatal	0.01	6.16x10 ⁻³	0.02	1.81x10 ⁻⁴	0.03
cg19542445	CACNA1C	Copper - Postnatal	-5.49x10 ⁻³	-8.37x10 ⁻³	-2.60x10 ⁻³	2.01x10 ⁻⁴	0.04
cg27340723	ADCY9	BPA - Postnatal	6.80x10 ⁻³	3.22x10 ⁻³	0.01	2.01x10 ⁻⁴	0.04
cg25905215	NFKB1	Copper - Postnatal	3.94x10 ⁻³	1.86x10 ⁻³	6.01x10 ⁻³	2.08x10 ⁻⁴	0.04
cg23098018	PIK3CD	Humidity (preg) - Pregnancy	0.03	0.01	0.04	2.16x10 ⁻⁴	0.04
cg24961795	PLCG1	PFOS - Postnatal	-0.01	-0.02	-4.99x10 ⁻³	2.46x10 ⁻⁴	0.04
cg03781224	MGLL	PFOS - Postnatal	4.00x10 ⁻³	1.84x10 ⁻³	6.16x10 ⁻³	2.93x10 ⁻⁴	0.05
cg19982668	MAP3K3	PFOS - Postnatal	7.59x10 ⁻³	3.50x10 ⁻³	0.01	2.89x10 ⁻⁴	0.05

Supplementary Material II.8: Sensitivity Analysis I: adjusted associations between the whole exposome and zBMI in 1,173 children from the HELIX cohort (2 multivariate agnostic approaches, one prenatal, one postnatal, ignoring the methylome). Results are presented only for exposures with a (FDR-corrected for multiple hypothesis testing) p-value lower than 0.05.

Exposure group	Exposure variable	Unit	Transformation	Effect estimate*	95%CI	Uncorrected p -value	FDR-corrected p -value
Metals	Copper - Postnatal	µg/L	Log2	0.18	0.10 - 0.27	2.48x10 ⁻⁵	3.16x10 ⁻³
Organochlorines	HCB - Postnatal	ng/g lipids	Log2	-0.44	-0.57 - -0.31	5.65x10 ⁻¹¹	1.44x10 ⁻⁸

* Adjusted change in mean zBMI for each increase by 1 in transformed exposure level. Models were adjusted for all (respectively prenatal and postnatal exposome variables) and maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for the postnatal model for birth weight, passive smoking during childhood, breastfeeding duration.

Supplementary Material II.9: Characteristics of the CpG selected by a methylome wide analysis on the whole methylome (row percentages).

CpGs	Number of CpGs	Number (%) of CpGs selected by MWAS on the whole methylome - Benjamini Hochberg correction
All	386,518	1788 (0.46%)
Belonging to a pathway AND enhancer	2,284	28 (1.22%)
Neither belonging to a pathway nor being an enhancer	384,234	1760 (0.46%)

Supplementary Material II.10: Sensitivity analysis III - adjusted association between the whole methylome and zBMI in 1,173 children from the HELIX cohort (ExWAS model, step b of the Meet-in-the-Middle approach applied to the whole methylome). Results are presented only for CpGs with a (FDR - corrected for multiple hypothesis testing) p-value below 0.05 in ExWAS.

This supplementary is provided in the appendix III due to its large size.

Supplementary Material II.11: Sensitivity analysis III, Meet-in-the-Middle without CpGs preselection: adjusted associations between the exposome and CpGs associated with zBMI in 1,173 children from the HELIX cohort (ExWAS model adjusted on zBMI, step c of the Meet-in-the-Middle approach applied on the whole methylome). Results are presented only for exposures associated with a (stringently corrected for multiple hypothesis testing) p-value of less than 0.05 in exposure-CpGs ExWAS, with CpGs being previously selected in a CpGs-zBMI ExWAS.

Exposure	Number of CpGs associated with the exposure and with zBMI	Number of corresponding genes
Copper - postnatal	1110	677
Bisphenol A (BPA) -postnatal	449	293
Perfluorooctanesulfonic acid (PFOS) - postnatal	180	133
Humidity - Pregnancy	47	38
PCBs (sum) - Pregnancy	10	10
PCB 170 - Pregnancy	9	7
MEHP - Postnatal	8	7
MiBP - Postnatal	7	7
DETP - Pregnancy	3	3
DDT - Postnatal	2	1
PBDE 153 - Postnatal	2	2
PCB 138 - Pregnancy	2	2
PFHXS - Postnatal	2	2
Thallium - Postnatal	2	2

Traffic noise (24h) - Postnatal	2	2
Caesium - Postnatal	1	1
Fast-food intake - Pregnancy	1	1
Green spaces (300 m) - Pregnancy	1	1
HCB - Postnatal	1	1
MBzP - Postnatal	1	1
OH-MiNP - Pregnancy	1	1
PM2.5 - Pregnancy	1	1
Population density - Postnatal	1	1
PRPA - Pregnancy	1	1
Social participation - Postnatal	1	1
Soda intake - Postnatal	1	1
Vegetables intake - Pregnancy	1	1
Yogurt intake - Postnatal	1	1

Supplementary Material II.12: Sensitivity analysis IV: Meet-in-the-Middle approach considering the cell-types as the intermediate layer. Adjusted associations at steps b),c) and d).

A.

Cell types	Effect estimate*	95%CI		Unadjusted p-Value	FDR adjusted p -Value
NK-cells	0.87	-0.82	2.56	0.31	0.38
B-cells	-0.33	-2.19	1.53	0.72	0.72
CD4+ T-cells	-2.22	-3.31	-1.13	7.00x10⁻⁵	4.20x10⁻⁴
CD8+ T-cells	-1.91	-3.35	-0.47	9.20x10⁻³	0.02
Granulocytes	0.87	0.23	1.51	7.86x10⁻³	0.02
Monocytes	3.11	0.70	5.53	0.01	0.02

* Adjusted change in mean zBMI for each increase by 1 in cell-type level. Models were adjusted for maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for postnatal exposures birth weight, passive smoking during childhood, breastfeeding duration.

B.

Cell types	Exposure variable	Effect estimate*	95%CI	Unadjusted p-Value	FDR adjusted p -Value
CD4+ T-cells	Copper - Postnatal	-0.01	-0.02 -8.76x10 ⁻³	6.56x10 ⁻⁹	5.67x10 ⁻⁶
Granulocytes	Copper - Postnatal	0.02	0.01 0.03	8.50x10 ⁻⁸	3.67x10 ⁻⁵
CD8+ T-cells	Copper - Postnatal	-7.68x10 ⁻³	-0.01 -4.27x10 ⁻³	1.09x10 ⁻⁵	3.13x10 ⁻³
Granulocytes	BPA - Postnatal	-0.01	-0.02 -6.96x10 ⁻³	3.36x10 ⁻⁵	7.26x10 ⁻³
CD4+ T-cells	Humidity - Pregnancy	-0.04	-0.06 -0.02	2.36x10 ⁻⁴	0.04

* Adjusted change in transformed exposure for each increase by 1 in cell-type level. Models were adjusted for maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for postnatal exposures birth weight, passive smoking during childhood, breastfeeding duration.

C.

Exposure group	Exposure variable	Unit	Transformation	Effect estimate*	95%CI	Unadjusted p-Value	FDR adjusted p - Value
Metals	Copper - Postnatal	µg/L	Log2	0.22	0.14 0.30	3.57x10 ⁻⁷	1.07x10 ⁻⁶
Phenols	BPA - Postnatal	µg/g	Log2	-0.07	-0.14 2.83x10 ⁻⁴	0.05	0.08
Meteorological	Humidity - Pregnancy	%	None	0.05	-0.34 0.44	0.81	0.81

* Adjusted change in mean zBMI for each increase by 1 in transformed exposure level. Models were adjusted for maternal BMI, maternal education, maternal smoking during pregnancy, parental country, cohort, parity, trimester of conception, ethnicity, child age and child sex and additionally only for postnatal exposures birth weight, passive smoking during childhood, breastfeeding duration.

CHAPTER III: Early-life Exposures and child lung function: a modified Meet-in-the-Middle approach using preselected methylation marks

The work presented in this small chapter is a study performed on Helix data similar to the one presented in Chapter II, but considering another child health outcome: the child lung function, measured with FEV₁ (see below). We used the same oMITM design, but with one major modification: we followed through its logical next step the strategy initiated in our first study to preselect a part of the methylome according to possible pathways from the exposome to the outcome; for the first step of our oMITM design, a drastic preselection of the methylome was performed according to MWAS studies about lung function available in the literature, instead of a test of association between the methylome and the outcome.

This work was the topic of an oral communication presented at the annual congress of the International Society of Environmental Epidemiology in Utrecht in 2019:

Cadiou, S., Agier, L., Bustamante, M., Maitre, L., Basagana, X., Vrijheid, M., Siroux, V., Slama, R., “Using DNA methylation to characterize more efficiently associations between the exposome and child lung function”, ISEE 2019 congress, Utrecht, Netherlands

III. 1. Abstracts

III.1.1. English abstract

Background: Early environmental exposures may influence lung function. Most pollutants have small effect sizes and some are correlated, potentially limiting the statistical power of agnostic exposome-wide association study (ExWAS). DNA methylation, which may act as a mediator for some exposures, could be used in exposome-health studies to increase power by reducing the exposome to exposures with biologically plausible mechanisms.

Aim: To assess the relations between the exposome and child lung function (FEV₁, forced expiratory volume in one second) with a method consisting in reducing the exposome dimension using DNA methylation.

Methods: Among 919 mother-child pairs from Helix cohorts, exposures to 216 environmental factors were assessed during pregnancy or at age 6-10 years. Genome-wide DNA methylation levels in peripheral blood at 6–10 years were measured using HumanMethylation450 BeadChip, filtered and corrected for batch effects. An oriented Meet-in-the-Middle, consisting in a 3-step statistical approach, was applied: (i) selecting *a priori* relevant DNA CpG sites for FEV₁ according to a review of the literature; (ii) selecting exposures significantly associated with at least one of these CpGs, using an ExWAS approach adjusted for FEV₁ and confounders; (iii) identifying by linear regression the exposures from this reduced set associated with FEV₁.

Results: 314 CpGs enhancers from 23 candidate genes were selected at step 1. Step 2 identified a single exposure, postnatal blood copper level, which was associated with one CpG site located on *ARMC2* gene. In step 3, copper was found significantly associated with lower FEV₁. A classical ExWAS analysis on FEV₁ corrected for multiple comparisons did not identify statistically significant association; copper was among the 6 exposures associated with FEV₁ when no multiple testing correction was applied.

Conclusion: Our 3-steps approach identified one exposure associated with lower FEV₁, postnatal blood copper level, while an agnostic ExWAS reported no significant association. Further research is needed to quantify the efficiency of this approach.

III.1.2. French abstract

Contexte : Les expositions environnementales précoces peuvent influencer la fonction pulmonaire. La plupart des polluants ont des effets de faible ampleur et certains sont corrélés, ce qui peut limiter la puissance statistique des études d'association à l'échelle de l'exposome (ExWAS). La méthylation de l'ADN, qui peut agir comme médiateur pour certaines expositions, pourrait être utilisée dans les études exposome-santé pour augmenter la puissance en restreignant l'exposome à un ensemble d'expositions avec des mécanismes d'effets biologiquement plausibles.

Objectif : tester une méthode consistant à réduire la dimension de l'exposome en utilisant la méthylation de l'ADN dans le cadre d'une étude d'association exposome-VEMS (Volume Expiratoire Maximal par seconde).

Méthodes : L'exposition à 216 facteurs environnementaux a été évaluée pour 919 paires mère-enfant des cohortes du projet Helix, pendant la grossesse ou à l'âge de 6-10 ans. Les niveaux de méthylation de l'ADN à l'échelle du génome dans le sang périphérique à 6-10 ans ont été mesurés à l'aide de la puce HumanMethylation450 BeadChip, filtrés et corrigés des effets batch. Une approche statistique « Meet-in-the-Middle orientée », consistant en 3 étapes, a été appliquée : (i) sélectionner a priori les sites CpG pertinents pour le VEMS à partir d'une revue de la littérature existante ; (ii) sélectionner les expositions significativement associées à au moins un de ces CpG, en utilisant une approche ExWAS ajustée sur le VEMS et les facteurs de confusion pertinents; (iii) identifier par régression linéaire les expositions associées au VEMS au sein de cet exposome réduit.

Résultats : 314 CpGs enhancers provenant de 23 gènes candidats ont été sélectionnés à l'étape 1. L'étape 2 a identifié une seule exposition, le taux de cuivre dans le sang postnatal, qui a été associée à un site CpG situé sur le gène *ARMC2*. À l'étape 3, on a constaté que le cuivre était associé de manière significative à un VEMS plus faible. Une analyse classique ExWAS sur le VEMS corrigé pour les comparaisons multiples n'a pas identifié d'association statistiquement significative ; le cuivre figurait parmi les 6 expositions associées au VEMS lorsqu'aucune correction pour les comparaisons multiples n'était appliquée.

Conclusion : Notre approche en 3 étapes a identifié une exposition associée à un VEMS plus faible, le niveau de cuivre dans le sang postnatal, alors qu'une analyse ExWAS agnostique n'a révélé aucune association significative. Des travaux supplémentaires sont nécessaires pour quantifier l'efficacité de l'approche oMITM.

III. 2. Background

Early environmental exposures may influence lung function (Vernet et al., 2017; Vrijheid et al., 2016). Effects of smoking and air pollution exposures are well known: maternal prenatal as well as postnatal smoking is associated with child asthma (Hofhuis et al., 2003; Wang and Pinkerton, 2008), and prenatal exposure to air pollution is associated with both decreased lung function and asthma (Latzin et al., 2009). There is also moderate evidence of the effect of persistent organic compounds on lung function (Gascon et al., 2014; Hansen et al., 2014), as well as emerging concerns for other man-made substances (Qin et al., 2017; Vernet et al., 2017). However, exploring the effects of the exposome on lung function involves the challenges of limited power and specificity that we discussed earlier: most pollutants are expected to have small effect sizes and some are correlated. A previous ExWAS study on HELIX data did not find any association between the exposome and child lung function (Agier et al., 2019). As differential DNA methylation in blood cells can be associated with decreased lung function, it seems relevant to use our oMITM design on Helix data of exposome, blood methylome and lung function, to study causal environmental predictors of child respiratory health. In this short report, we aimed to apply an oriented Meet-in-the-Middle approach to exposome and lung function, strongly relying on a priori knowledge about the link between methylome and lung function.

III. 3. Methods

Among 1031 mother-child pairs from Helix cohorts, exposures to 216 environmental factors were assessed during pregnancy or at age 6-10 years, as previously done and detailed in Cadiou et al. (2020) (see Chapter II). The forced expiratory volume in one second, expressed in % (FEV₁) was measured with standardized spirometry tests during a clinical examination at age 6-10 years. FEV₁ predicted values were computed after filtering and according to reference equations from the Global Lung Initiative (Agier et al., 2019; Quanjer et al., 2012). Genome-wide DNA methylation

levels in peripheral blood sampled at the time of the clinical examination were measured using HumanMethylation450 BeadChip, filtered and corrected for batch effects (see Chapter II). They were available for 919 children. Relevant potential confounders for the exposome-FEV₁ association were *a priori* selected: child sex, child age, child height, parental country of birth, breast feeding duration, season of conception, presence of older siblings, parental education level, maternal age, maternal pre-pregnancy body-mass index (BMI), postnatal passive smoking status, prenatal maternal active and passive smoking status and ethnicity. An oriented MITM adapted from (Cadiou et al., 2020) was applied. It consists in a 3-step statistical approach: a) selecting *a priori* enhancers CpG sites on genes relevant for FEV₁ according to the literature (Li et al., 2013) (but without direct test of association between them and the outcome, differently from (Cadiou et al., 2020)) ; b) selecting exposures significantly associated with at least one of these CpGs, using an ExWAS approach adjusted for FEV₁ and confounders; b) identifying by linear regression the exposures from this reduced set associated with FEV₁. All the ExWAS-type analyses were corrected for multiple testing using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). An agnostic ExWAS as well as a simplified MITM without step a) were performed as sensitivity analyses.

Table III.1: Preselected genes related to FEV1 according to (Li et al., 2013) and corresponding number of enhancers CpGs available in Helix data.

Genes	Number of corresponding enhancer CpGs
ADAM19	6
ARMC2	8
C10orf11	37
CCDC38	1
CDC123	2
CFDP1	8
DAAM2	11
FAM13A	9
GPR126	7
HDAC4	67
HHIP	2
HTR4	2
LRP1	10
MECOM	31
MFAP	2
MMP15	9
PID1	1
PTCH1	5
RARB	16
SPATA9	1
TGFB2	4
THSD4	57
TNS1	18

III. 4. Results

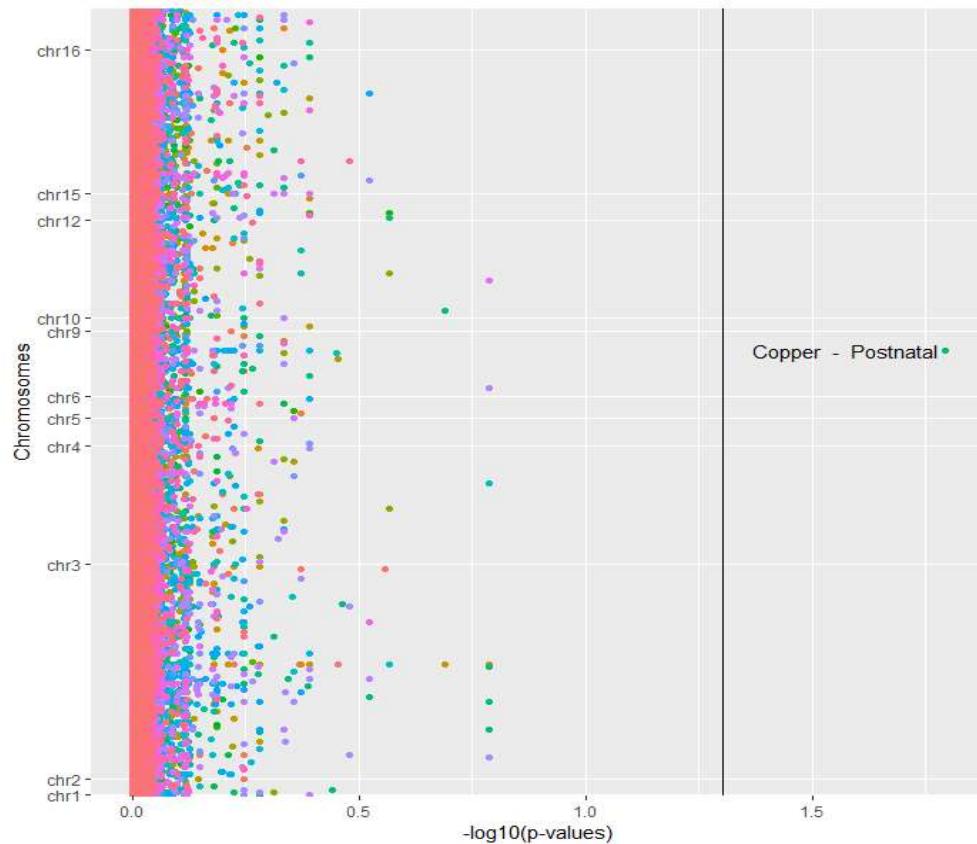
314 CpGs enhancers from 23 candidate genes were selected at step a) (see Table III.1). Step b) identified a single exposure, postnatal blood copper level, which was associated with one CpG site (cpg04642300, coefficient -0.301 (CI: -0.414; -0.187), adjusted p-value: 0.016) located on gene *ARMC2* (Figure III.1). *ARMC2* (Armadillo Repeat Containing 2) is a protein encoding gene. Some variants have been identified as significantly associated with low FEVs (in non-asthmatic adults (Yamada et al., 2016)); in 7-year-old children with asthma (Kreiner-Møller et al., 2014). Some CpGs on this gene have been identified as significantly associated with bronchopneumopathy in smoking adults (Busch et al., 2016). In step c) copper was found significantly associated with lower FEV₁ (coefficient of -5.72, p-value, 0.0412). No CpG were associated with the outcome in the sensitivity analysis considering the whole methylome (minimum corrected p-value: 0.342).

A classical ExWAS analysis on FEV₁ corrected for multiple comparisons did not identify any statistically significant association. Copper was among the 6 exposures associated with FEV₁ when no correction for multiple testing was applied (see Table III.2).

Table III.2: Agnostic ExWAS corrected for relevant potential confounders and corrected for multiple testing relating the exposome and child FEV₁.

Exposures	Effect estimate	95% Confidence Interval		Uncorrected p-Value	FDR corrected p-Value
Alcohol intake - Pregnancy	1.96	0.44	3.48	0.0118	1
ETPA - Postnatal	-0.62	-1.11	-0.14	0.0122	1
PFOA - Pregnancy	-1.39	-2.68	-0.11	0.0336	1
House crowding - Postnatal	-0.78	-1.51	-0.06	0.0342	1
Copper - Postnatal	-5.72	-11.22	-0.23	0.0412	1
Vegetables intake - Postnatal	0.81	0.01	1.61	0.0475	1

Figure III.1: Manhattan plot of the step b) of the oMITM approach: adjusted-values of adjusted association tests between preselected CpG and exposome. Each color represents a different exposure. The vertical black line is the significant threshold at 0.05.



III. 5. Discussion

Our 3-step oMITM approach identified one exposure associated with lower FEV₁, postnatal blood copper level, while an agnostic ExWAS reported no significant association. Copper exposure has not been found previously associated with respiratory outcomes in the literature besides Helix (Agier et al., 2019) but is known for its role in inflammatory diseases, as we discussed in the previous chapter (Cadiou et al., 2020). The fact that copper was a hit in both oMITM approaches on BMI and FEV₁ on Helix data may mean that it is linked to a general inflammation process which can be observed at the methylome. With our 3-step approach, we preselected the pathways whereby the exposures could act on lung function: this was a way to add external knowledge without performing a priori selection on the exposome itself, which would not be relevant in a

discovery study. Interestingly, in this case, assuming that copper is a true predictor of FEV₁ (which as we discussed is plausible), the oMITM approach was more sensitive than the ExWAS agnostic approach: indeed, while no exposure was associated with FEV₁ when corrected with multiple testing, one exposure was associated with one CpG relevant for FEV₁, allowing to select a non-empty reduced exposome. This highlights the gain in power allowed by the dimension reduction of the exposome: the association with copper was not significant due to the correction for multiple comparisons in the ExWAS but was significant in the oMITM approach due to the decreased size of the reduced exposome. Thus, we can hypothesize that a relevant dimension reduction using the oMITM approach can increase sensitivity compared to its agnostic counterpart.

Note: In an ongoing study based on Sepages cohort, we applied a modified oMITM implementation to relate the prenatal exposome to birth weight, taking advantage of the maternal methylome. The preliminary results are presented in Appendix II.

CHAPTER IV: Performance of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health: a simulation study under various causal structures

In the two previous chapter and in Appendix II, we presented three studies on real data relating the exposome to child outcomes, relying on an innovative oMITM design. Considering the results and their comparison with the results of the agnostic counterparts of our oMITM, we made the hypotheses that: 1. our oMITM may be more specific, and possibly even more sensitive, than an agnostic design if the methylome lies on the pathway from some exposures to the outcome, at a cost in sensitivity in cases of exposures effects not involving the methylome; 2. that the adjustment on the outcome that we proposed in the second step of our MITM may help to get rid of some reverse causal associations. We chose to perform a simulation study in a realistic setting similar to our first study (chapter II) and under various causal structures to compare the performance of our MITM design to other methods involving or not the use of the methylome, which is presented in this chapter.

Cadiou, S., Basagana, X., González, JR., Lepeule, J., Siroux, V., Vrijheid, M., Slama, R., “Performance of approaches relying on intermediate high-dimensional data to decipher causal relationships between the exposome and health: a simulation study under various causal structures”, submitted

IV. 1. Abstracts

IV.1.1. English abstract

Challenges in the assessment of the health effects of the exposome, defined as encompassing all environmental exposures from the prenatal period onwards, include a possibly high rate of false positive signals. It might be overcome using data dimension reduction techniques. Data on biological layers lying between the exposome and its possible health consequences, such as the methylome, may help reducing exposome dimension. We aimed to quantify the performances of approaches relying on the incorporation of an intermediary biological layer to relate the exposome and health, and compare them with agnostic approaches ignoring the intermediary layer. We performed a Monte-Carlo simulation, in which we generated realistic exposome and intermediary layer data by sampling with replacement real data from the Helix exposome project. We generated a Gaussian outcome assuming linear relationships between the three data layers, in 2381 scenarios under five different causal structures, including mediation and reverse causality. We tested 3 agnostic methods considering only the exposome and the health outcome: ExWAS (for Exposome-Wide Association study), DSA, LASSO; and 3 methods relying on an intermediary layer: two implementations of our new oriented Meet-in-the-Middle (oMITM) design, using ExWAS and DSA, and a mediation analysis using ExWAS. Methods' performances were assessed through their sensitivity and FDP (False-Discovery Proportion). The oMITM-based methods generally had lower FDP than the other approaches, possibly at a cost in terms of sensitivity; FDP was in particular lower under a structure of reverse causality and in some mediation scenarios. The oMITM–DSA implementation showed better performances than oMITM–ExWAS, especially in terms of FDP. Among the agnostic approaches, DSA showed the highest performance. Integrating information from intermediary biological layers can help lowering FDP in studies of the exposome health effects; in particular, oriented-MITM seems less sensitive to reverse causality than agnostic exposome-health association studies.

IV.1.2. French abstract

Les nombreux défis dans l'évaluation des effets sur la santé de l'exposome, défini comme l'ensemble des expositions environnementales subies à partir de la période prénatale, incluent un taux éventuellement élevé de faux positifs. Ce défi pourrait être surmonté en utilisant des techniques de réduction de la dimension. Les données sur les couches biologiques situées entre l'exposome et ses éventuelles conséquences sur la santé, telles que le méthylome, peuvent aider à réaliser une telle réduction de la dimension de l'exposome. Nous avons cherché à quantifier les performances des approches reposant sur l'incorporation d'une couche biologique intermédiaire pour mettre en relation l'exposome et la santé, et à les comparer avec des approches agnostiques ignorant la couche intermédiaire. Nous avons réalisé une simulation de Monte-Carlo, dans laquelle nous avons générée des données réalistes d'exposome et de couche intermédiaire en échantillonnant des données réelles du projet HELIX. Nous avons générée un outcome gaussien en postulant des relations linéaires entre les trois couches de données, dans 2381 scénarios sous cinq structures causales différentes, y compris la médiation et la causalité inverse. Nous avons testé 3 méthodes agnostiques ne considérant que l'exposome et l'outcome de santé : ExWAS (étude d'association à l'échelle de l'exposome), DSA, LASSO ; et 3 méthodes reposant sur des données intermédiaires : deux implémentations de notre nouveau design « Meet-in-the-Middle orienté » (oMITM), utilisant ExWAS et DSA, et une analyse de médiation utilisant ExWAS. Nous avons évalué la sensibilité des méthodes et le taux de faux positifs (FDP). Les méthodes oMITM avaient généralement un FDP plus faible que les autres approches ; c'était notamment le cas dans une structure de causalité inverse et dans certains scénarios de médiation (parfois à un coût en termes de sensibilité). L'implémentation oMITM-DSA a montré de meilleures performances qu'oMITM-ExWAS. Parmi les approches agnostiques, DSA a montré les meilleures performances. L'utilisation d'informations provenant de couches biologiques intermédiaires peut contribuer à réduire le FDP dans les études des effets de l'exposome sur la santé ; en particulier, oMITM semble moins sensible à la causalité inverse que les études agnostiques d'association exposome-santé.

IV. 2. Introduction

The exposome concept acknowledges that individuals are exposed simultaneously to a multitude of environmental factors from conceptions onwards (Wild, 2005). The exposome, understood as the totality of the individual environmental (i.e. non-genetic exogenous) factors, may explain an important part of the variability in chronic diseases risk (Manrai et al., 2017; Sandin et al., 2014; Visscher et al., 2012). During the last decade, environmental epidemiology started embracing the exposome concept (see e.g. (Agier et al., 2019; Lenters et al., 2016; Patel et al., 2010)). Such studies typically face an issue encountered in many fields (Runge et al., 2019), that of efficiently identifying the causal predictors of an outcome among a set of possibly correlated variables of intermediate to high dimension (currently, a few hundred to a few thousand variables). The correlation within the exposome (Tamayo-Uria et al., 2019) was shown to entail a possibly high rate of false positive findings, in particular when using ExWAS (exposome-wide association study), i.e. parallel univariate models with correction for multiple testing (Agier et al., 2016). Recent studies, typically conducted among a few hundred or thousand subjects, are also expected to have limited power (Chung et al., 2019; Siroux et al., 2016; Slama and Vrijheid, 2015; Vermeulen et al., 2020). In addition, they can suffer from reverse causality: if exposures are measured by biomarkers at the same time as the outcome, this opens the possibility of the health outcome influencing some components of the exposome. For example, the serum concentration of persistent compounds can be influenced by the amount of body fat, which is related to health outcomes such as obesity or cardiovascular disorders (Cadiou et al., 2020). The potential for reverse causality is even stronger if biomarkers of effect (e.g. biomarkers of oxidative stress or inflammation) are considered to be part of the exposome, as sometimes advocated (Rappaport, 2012; Vermeulen et al., 2020). Indeed, these may also be consequences of the considered health outcome.

Benchmark studies and reviews tried to identify which statistical methods could help to face some of these issues (Agier et al., 2016; Barrera-Gómez et al., 2017; Lazarevic et al., 2019; Lenters et al., 2018). Dimension reduction tools were a relevant option to consider (Chadeau-Hyam et al., 2013).

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Dimension reduction can be achieved by purely statistical approaches, or rely on external (e.g., biological) information. Past simulation studies focused on statistical dimension reduction techniques and generally assumed a simple causal structure and that the variability of the outcome explained by the exposome was higher than 5% (Agier et al., 2016; Barrera-Gómez et al., 2017; Lengers et al., 2018): within this framework, they showed that dimension reduction techniques such as regression-based variable selection methods simultaneously considering multiple variables were more efficient than the ExWAS to control the false positive rate (Agier et al., 2016). When it comes to non-purely statistical dimension reduction approaches, it may be relevant to try relying on biological parameters, including ‘omic (methylome, transcriptome, metabolome...), inflammatory or immunologic markers, possibly acting as intermediary factors between the exposome and health. This logic is embodied in the Meet-in-the-Middle (MITM) design (Chadeau-Hyam et al., 2011; Jeong et al., 2018), which detects “intermediary” biomarkers associated with both exposures and the health outcome. We recently applied a tailored MITM design (Cadiou et al., 2020), named hereafter “oriented Meet-in-the-Middle” (oMITM), with a dimension reduction aim, and using methylation data to reduce exposome dimension.

Here, we make the hypothesis that oMITM could 1) allow lowering the high 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 (Cadiou et al., 2020). Specifically, we aimed to test if methods relying on intermediary multidimensional biological data allow to more efficiently identify the causal predictors of a health outcome among a large number of environmental factors. We both considered methods making use of information on potential mediators of the health effects of exposures and agnostic methods ignoring the intermediate layer, and compared their sensitivity and false positive rate. Data were generated assuming five different possible causal models, including reverse causality, for realistically low values of the share of the outcome variability explained by the exposome. After comparing the

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health methods using simulated data, in a second section, we use causal inference theory to discuss which designs may be most adapted under each possible causal structure.

IV. 3. Materials and methods

IV.3.1. Overview of the simulation

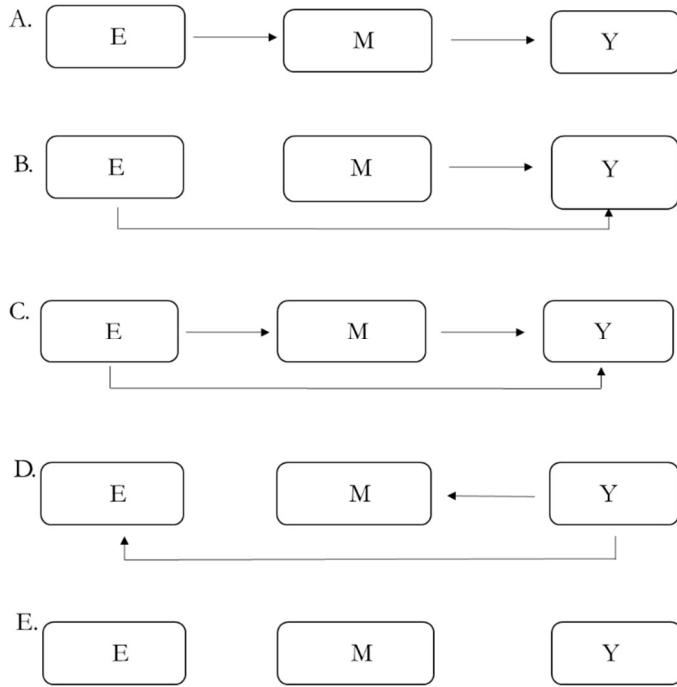
We relied on a Monte-Carlo simulation to compare the efficiency of methods aiming at identifying which components of the exposome influenced a health outcome under various causal models and hypotheses (altogether defining a total of 2381 scenarios). Exposome, intermediary layer and outcome data were generated under these various scenarios. For each scenario, 100 datasets were simulated (see below). The 6 methods compared, as well as two control methods (see below), were applied to each dataset and their performances were assessed, and synthesized over all datasets related to a given scenario.

IV.3.2. Causal structures considered

Five different causal structures were considered (see Figure IV.1): in structures A, B, C the exposome (E) affects the outcome (Y) directly or indirectly. In A, there is no direct effect from E to Y, all the effect being mediated by the intermediary layer (i.e. “indirect effect” in the mediation analysis terminology (Vanderweele and Vansteelandt, 2009)). B assumed a causal link from M to Y and a direct effect from E to Y, without mediation through M. C assumed both a direct and an indirect effect of E on Y. Structure D is a situation with reverse causal links from Y to M and from Y to E. Structure E assumed total independence between the three layers.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.1: Causal structures considered in the simulation study of the efficiency of studies relating a layer of predictors E (e.g., the exposome) to a layer of possibly intermediary parameters (e.g. biological parameters such as DNA methylation) M and a health outcome or parameter Y.



IV.3.3. Generation of independent realistic exposome, methylome and outcome data, and addition of causal relations within them

To build datasets according to these causal structures, we first generated independent variables corresponding to a set of exposures (our exposome) and a biological layer (e.g., corresponding to metabolomic signals or methylation levels at various sites on the DNA) by independently sampling with replacement real data on the exposome and DNA methylome from 1173 individuals of HELIX project (Cadiou et al., 2020; Maitre et al., 2018; Vrijheid et al., 2014). For the exposome, 173 quantitative variables corresponding to the exposures were obtained from the real prenatal and postnatal child exposome data of Helix, selecting only the quantitative exposures and covariates. Variables were then standardized and bounded (each standardized value greater than 3 in absolute value were replaced by a value lower than 3 in absolute value randomly drawn in the distribution). For the intermediary layer, 2284 quantitative variables corresponding to the CpGs were obtained

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health from the real methylome data of Helix and the a priori selection of CpGs performed in Cadiou et al. (2020) by selecting only enhancers CpGs belonging to selected pathways. These variables were standardized.

From this sampled dataset, in which the exposure and the methylome were, by construction, independent, we used linear models to possibly add an hypothesized effect of some exposures on the methylome, and to generate a health outcome possibly related to E and/or M according to the above-mentioned causal structures (Figure IV.1): in causal structures A, B and C, assuming a causal effect of the exposome or the methylome on the outcome, the outcome (Y) was drawn from a normal distribution to which potential effects of E and M were added. The variance of this distribution was set to ensure that the total variability explained by E and M was that defined by the desired scenario. To simulate a reverse causal link (structure D, Figure IV.1) and a situation without causal link between the three layers (structure E), we generated the outcome by bootstrapping the real child BMI data of HELIX cohorts; a linear effect of the outcome was added to the exposome and to the methylome for causal structure D. BMI was standardized according to WHO guidelines (Cadiou et al., 2020; de Onis et al., 2007).

For each structure, different scenarios varying the intensity of the hypothesized associations and the number of predictors from the different layers were generated: in particular, for the structures displaying an effect of E on Y, the total variability of Y explained by E and M, fixed within a scenario, varied between 0.01 and 0.4 and the number of true predictors of Y within E varied between 1 and 25; the number of elements of M with an effect on Y varied between 10 and 100 in the causal structures assuming such an effect. The parameters of the different scenarios are detailed in Supplementary Table V.1. For each scenario considered, 100 datasets were simulated.

The simulation (detailed in Supplementary Material V.1) additionally made the following assumptions:

- All direct effects of a variable on another were assumed to be linear.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

- The magnitude (i.e., slope) of all effects from the predictor variables of a given layer (e.g. E) on the predicted variables of another layer were identical within a given scenario.
- A variable from M could not be affected by more than one exposure. In consequence, when multiple exposures were assumed to affect the intermediary layer, the number of variables affected was a multiple of the number of exposures.

Table IV.1: Details of the methods compared in the simulation study.

Name	Description	References	Name used in figure
<i>Agnostic methods</i>			
ExWAS with Benjamini-Hochberg correction	Independent linear regressions corrected for multiple testing using Benjamini-Hochberg correction. The output corresponds to exposures associated with the outcome.	(Benjamini and Hochberg, 1995)	ExWAS
Lasso	Penalized regression model relying on a generalized linear framework developed by Tibshirani (Tibshirani, 1996). The LASSO penalty (L^1) added to the loss function promotes sparsity and performs variable selection through shrinkage: the lowest regression coefficients, corresponding to the least informative predictors, are attributed a zero value, according to a penalty parameter λ . As advised by Tibshirani (Tibshirani, 1996) and implemented in the <i>glmnet</i> package (Friedman et al., 2010), λ is determined by minimizing the prediction root mean squared error (RMSE) using 10-fold cross-validation. λ sequences tested in the cross-validation process is a sequence of 100 values deterministically determined from the data (Friedman et al., 2010). Exposures with non-zero coefficients in the final model using optimal lambda are the output of this selection method.	(Tibshirani, 1996) (Friedman et al., 2010).	LASSO
DSA (Deletion Substitution Addition) algorithm	DSA is an iterative linear regression model search algorithm (Sinisi and van der Laan 2004) following three constraints: maximum order of interaction amongst predictors, the maximum power for a given predictor, and the maximum model size. At each iteration, the following three steps are allowed: a) removing a term, b) replacing one term with another, and c) adding a term to the current model. The search for the best model starts with the intercept model and identifies an optimal model for each model size. The final model is selected by minimizing the value of the RMSE using 5-fold cross-validated data. We allowed no polynomial or interaction terms, and made no restriction on the number of predictors. Exposures selected by DSA are the output of this selection method.	(Sinisi and van der Laan 2004)	DSA

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Methods incorporating information from an intermediary layer

Oriented Meet in the Middle - ExWAS	Design of the Meet in the Middle approach from (Cadiou et al., 2020), using ExWAS-type corrected for multiple testing using Benjamini-Hochberg correction for all three steps. 3 steps: a. tests of association between the intermediary layer and the outcome with an ExWAS type approach corrected for multiple comparisons using Benjamini and Hochberg procedure; b. tests of association between each exposure and the intermediary variables found associated with the outcome in step, adjusted on the outcome, corrected for multiple testing using the Benjamini-Hochberg procedure. Correction for multiple testing takes into account all the tests performed at this step (i.e. number of exposures x number of intermediary variables found associated with the outcome in step a); c. Test of the associations between exposures found associated with at least one intermediary variable at step b and the outcome, using an ExWAS design corrected for multiple comparisons. Correction for multiple testing takes into account all the tests performed at this step (i.e. number of exposures found associated with at least one CpG at step b.). Exposures found associated with the outcome in step c. are the output of this selection method.	(Cadiou et al., 2020) (Benjamini and Hochberg, 1995)	oMTM-ExWAS
Oriented Meet in the Middle - DSA	Design of the oriented Meet in the Middle approach from (Cadiou et al., 2020), using ExWAS-type corrected for multiple testing using Benjamini-Hochberg correction for the two first steps and DSA for the last steps. 3 steps: a. tests of association between the intermediary layer and the outcome with an ExWAS type approach corrected for multiple comparisons using Benjamini and Hochberg procedure; b. tests of association between each exposure and the intermediary variables found associated with the outcome in step, adjusted on the outcome, corrected for multiple testing using the Benjamini-Hochberg procedure. Correction for multiple testing takes into account all the tests performed at this step (i.e. number of exposures x number of intermediary variables found associated with the outcome in step a); c. DSA algorithm (implemented as described above) is applied to select exposures associated with the outcome among the exposures found associated with at least one intermediary variable at step b. Exposures found associated with the outcome in step c. are the output of this selection method.	(Cadiou et al., 2020) (Benjamini and Hochberg, 1995) (Sinisi and van der Laan 2004)	oMTM-DSA
Mediation	Mediation analysis in 3 causal steps: a. ExWAS using Benjamini-Hochberg correction; b. Tests of the associations between the exposures selected in step a. and each intermediary variable, corrected for multiple comparisons using Benjamini-Hochberg correction; c. tests of the association of each intermediary variable with the outcome adjusted on each exposure found associated with the outcome at step a., corrected for multiple testing using Benjamini Hochberg procedure. Exposures for which corrected p-values are significant for at least one intermediary variable site in both step b and c are the output of this selection method.	(MacKinnon et al., 2002; Vanderweele and Vansteelandt, 2009)	Mediation

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Control methods

Steps 1 and 2 of the oriented Meet-in-the-Middle	2 first steps of the design of the oriented Meet in the Middle approach using ExWAS-type, corrected for multiple testing with Benjamini-Hochberg correction (Cadiou et al., 2020). a. tests of association between the intermediary layer and the outcome with an ExWAS type approach corrected for multiple comparisons using Benjamini and Hochberg procedure; b. tests of association between each exposure and the intermediary variables found associated with the outcome in step a, adjusted on the outcome, corrected for multiple testing using Benjamini-Hochberg procedure. Correction for multiple testing takes into account all the tests performed at step b (i.e. number of exposures x number of intermediary variables found associated with the outcome in step a). Exposures found associated with at least one intermediary variable in step b. are the output of this selection method.	(Cadiou et al., 2020)	oMITM – steps 1 and 2
ExWAS on a random subsample	ExWAS with FDR correction on a set of n_R random exposures, where n_R is the number of exposures in the reduced exposome when applying oMITM -ExWAS on the same dataset. Exposures found associated with the outcome are the output of this selection method.	(Benjamini and Hochberg, 1995)	ExWAS on subsample

IV.3.4. Methods to relate the exposome and health compared

For each generated dataset, we applied 8 different statistical methods, detailed in Table IV.1:

- three “agnostic” methods ignoring the intermediary layer: ExWAS with Benjamini-Hochberg correction (Benjamini and Hochberg, 1995), Lasso (Friedman et al., 2019; Tibshirani, 1996), Deletion Substitution Addition (DSA) algorithm (Sinisi and van der Laan, 2004);
- three methods using the intermediary layer to reduce the dimension of the exposome: two implementations of our oMITM-design (Cadiou et al., 2020) and a mediation analysis using parallel simple linear regressions (Küpers et al., 2015; MacKinnon et al., 2002);
- two “control” methods: “ExWAS steps 1 and 2” and “ExWAS on subsample”, meant to inform the comparison between the results of the previous methods (see below and Table IV.1).

The oMITM design, detailed in Table IV.1 and implemented by Cadiou et al. (2020), consists in three series of association tests: *a*) between the intermediary layer M and the outcome Y, allowing to identify components of M associated with Y; *b*) between the components of the intermediary layer selected at step a) and the exposures E, with an adjustment on the outcome; *c*) between the

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health exposures selected at step b) and the outcome (see (Cadiou et al., 2020) for details). Various statistical methods can be used at steps a, b, c). We tested two different implementations of the oMITM design: the first one (oMITM-ExWAS) used ExWAS-type methods at all steps, i.e. a series of parallel linear regression models (one per tested predictor) corrected for multiple testing using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995); the second oMITM implementation used an ExWAS-type approach at steps *a*) and *b*) and DSA algorithm at step *c*). DSA (Sinisi and van der Laan, 2004) is an iterative linear regression model search algorithm, which has been shown to provide the best performance (assessed as the compromise between sensitivity and FDP) in studies relating the exposome to health, compared to other common methods including ExWAS (Agier et al., 2016). It performs data-adaptive estimation through loss-based cross-validated estimator selection. DSA was not considered for step a) and b) as, as a wrapper method, it is not computationally feasible to use it on a set of covariates of dimension higher than a few hundred. The third “agnostic” method used, LASSO, is a regularized linear regression, adding a penalty term (L^1) to the loss function (Tibshirani, 1996). For the mediation design, using the 3 causal steps defined by the seminal article of Baron and Kenny (1986), we implemented ExWAS-type analysis at each step, in order to allow comparison with MITM-ExWAS.

IV.3.5. Assessing scenarios’ characteristics and methods’ performances

To assess the characteristics of each scenario, variabilities of Y explained by the true predictors of E, by the true predictors of M and by both were measured and their mean and standard deviation were computed over the 100 runs. For causal structures A and C, the variability explained by E for each variable of M affected by E was also measured and averaged. Then mean and standard deviation of this averaged variability were computed over the 100 runs. For structures D and E, the variability explained by Y was measured for each variable of M or exposure predicted and means and standard deviations were computed across the exposome and the intermediary layer.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

To compare methods, for each scenario of causal structures A, B and C, false discovery proportion (FDP) and sensitivity to identify true predictors within the exposome were measured and mean and standard deviation were computed. FDP was defined as the proportion of exposures that were not causal predictors among the exposures selected. When no exposure was selected, FDP was set to 0. Sensitivity was defined as the proportion of exposures selected among the true causal predictors. For scenario from structures D and E, for which there were no true predictors of Y, the mean and standard deviations of the number of predictors found were computed over the 100 runs. The “sensitivity” to detect exposures affected by Y was also computed. In causal structures A, B and C, methods’ performances were compared in term of FDP, sensitivity and accuracy (defined as the sum of sensitivity and $1 - FDP$).

The script, developed in R, is provided in Supplementary Material V.2.

IV.3.6. Comparisons between oMITM, mediation and direct association test using structural causal modelling theory in a three-variable scheme

We used the theory of structural causal modelling (Pearl, 2009, 1995) to identify in which causal situations a causal association could be expected to be identified using the-Middle design in the simpler situation of three unidimensional variables (i.e. one exposure, one CpG, one outcome, ignoring the higher dimension of E and M in our simulation). Twenty-five Directed Acyclic Graphs (DAG) were assessed, corresponding to the 27 theoretical possibilities combining 3 variables with 3 modalities (no causal link, causal link, reverse causal link) without the two diagrams corresponding to cyclic graphs ($E \rightarrow M \rightarrow Y \rightarrow E$ and $Y \rightarrow M \rightarrow E \rightarrow Y$). For each causal structure, potential bias were identified for each association test through the existence of spurious association between two variables because of a backdoor path not controlled for or because of adjustment for a collider (Pearl, 2009, 1995). This allowed to determine if oMITM would be able to show an association, assuming that statistical power was sufficient. We determined for each causal structure if the design was expected to provide a false-positive, false-negative, true-positive or true negative

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health finding, according to the theoretical output (exposure selected or not) and the presence of a direct causal link from the exposure to the outcome in the causal structure considered. Similar analyses were done for the mediation design (see Table IV.1), for a design similar to the oMITM but without adjustment on the outcome in the second step *b*) (which corresponds to the MITM design most commonly implemented in the literature (Chadeau-Hyam et al., 2011)), and for the basic association test between E and Y ignoring M.

IV. 4. Results

IV.4.1. Causal structures assuming an effect of the exposome on health

The characteristics of the scenarios under causal structures assuming an effect of the exposome on health (structures A, B and C) are summarized in Supplementary Table V.2. On average over these three structures, DSA and oMITM-DSA provided the highest accuracy; FDP was lower for oMITM-DSA and sensitivity higher for DSA (Table IV.2). When we considered the three causal structures separately, the method most accurate differed between causal structures.

When we assumed that the totality of the effect of E on Y was mediated by M (structure A), the variability of Y explained by E was necessarily lower than under the other causal structures with direct E-Y relation (Supplementary Table V.2). Overall, the method maximizing the accuracy was oMITM-DSA (Table IV.2). It was immediately followed by the oMITM-ExWAS and then the mediation analysis. Average sensitivity was higher than 0.095 for all the agnostic and non-agnostic methods and it increased with the variability of E explained by Y. The method displaying the lowest FDP was oMITM-DSA (average FDP across scenarios, 0.038), which also showed one of the lowest sensitivities on average (0.095); however, as soon as the variability explained by the exposome was above 0.1, its sensitivity was above 0.70 while its FDP remained below 0.20 (see Figure V.2). In a few scenarios (when the variability explained by the exposome was between 0.05 and 0.1, see Figure 2 and Supplementary Figure 1), oMITM-DSA even showed a better sensitivity than its agnostic counterpart, DSA, with a similar FDP. When the variability explained by the exposome was low (below 0.01), oMITM-DSA did not select any predictor, contrarily to DSA, which showed a non-null FDP in this range of variabilities. oMITM-ExWAS and mediation had an average FDP and an average sensitivity that were both of 0.1. Overall, the reduced exposome selected by the two oMITM designs (after steps 1 and 2 of oMITM) contained more true predictors than a random set of exposures of the same dimension; this can be seen by comparing the sensitivity of oMITM-ExWAS to the sensitivity of *ExWAS on subsample* (Figure IV.3A), which was lower in all scenarios. Interestingly, the FDP of oMITM-ExWAS and ExWAS on subsample were

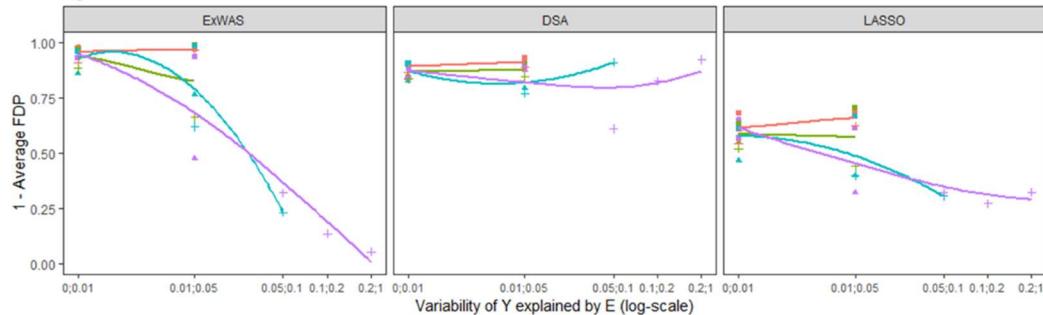
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health similar and lower than the FDP of ExWAS. This shows the influence of the dimension on the FDP for ExWAS-based methods and illustrates the benefit of the dimension reduction steps provided by oMITM.

Coming to the agnostic methods, DSA and ExWAS displayed similar global performances (Table IV.2), but DSA showed better (lower) FDP in the few scenarios for which the variability explained by E was higher than 0.1 (Figure IV.2A and B). LASSO displayed the highest FDP (average, 0.41) and had a high FDP even when the variability explained by the predictors was low (Figure IV.2B), as, contrarily to the other methods, it most often selected a non-null number of variables in these situations (Supplementary Figure V.1C).

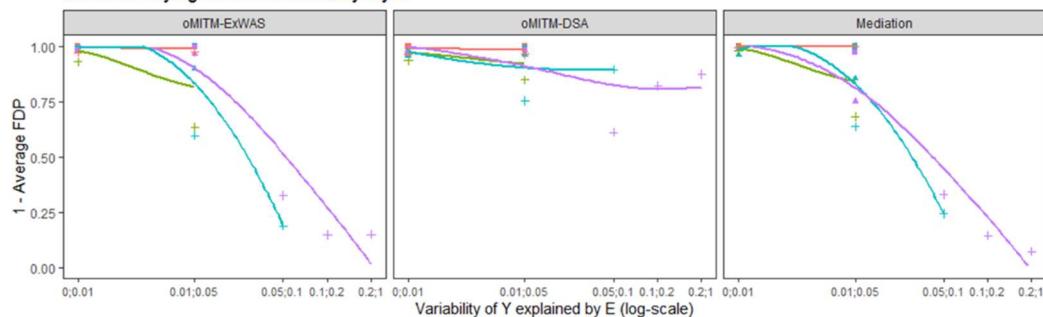
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.2: **A.** 1- FDP and **B.** Sensitivity under causal structure A (see Figure IV.1) for all compared methods; performances were averaged across scenarios according to categories of variabilities of Y explained by E (x-axis) and by M (color) and categories of mean variability of a covariate from M affected by E by E. Values were smoothed to give the trend according to averaged categories of variabilities of Y explained by E and by M. (in color)

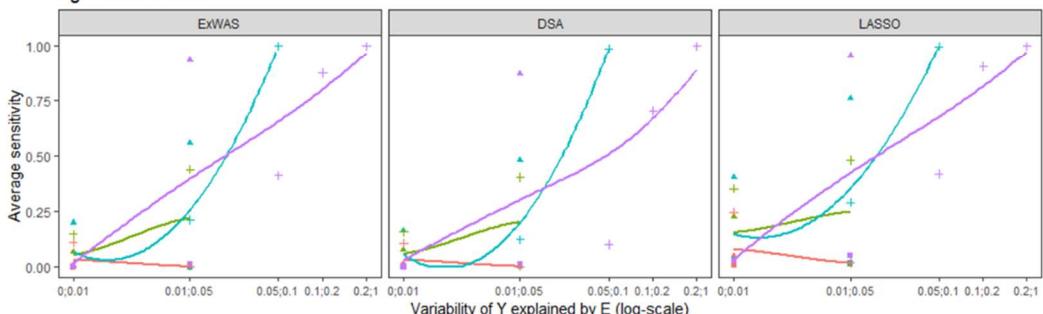
A. Agnostic methods



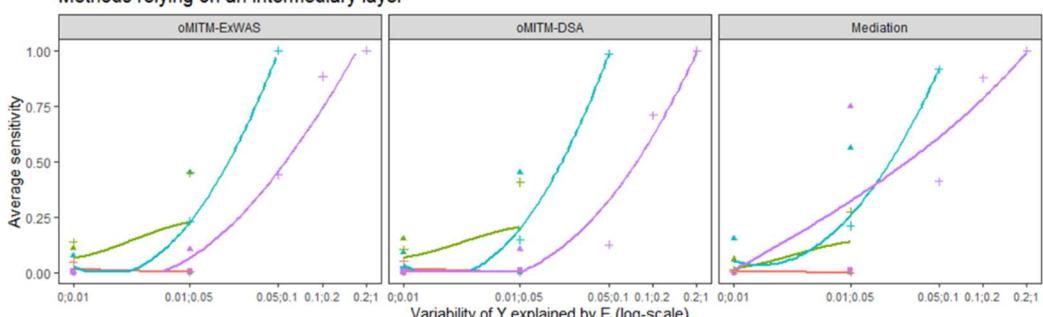
Methods relying on an intermediary layer



B. Agnostic methods



Methods relying on an intermediary layer



Mean variability of a mediator explained by E

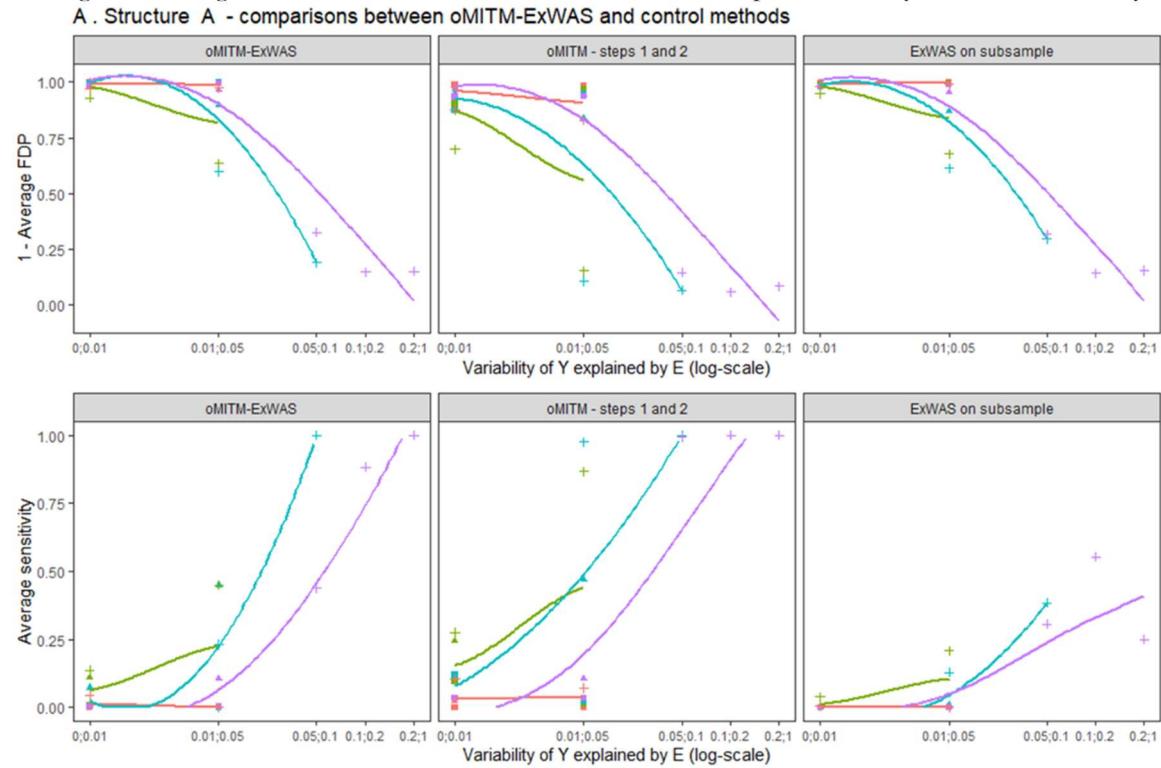
• 0:0.005 ▲ 0.005:0.01 ■ 0.01:0.15 + 0.15:0.2

Variability of Y explained by M

— 0:0.05 — 0.05:0.1 — 0.1:0.4 — 0.4:1

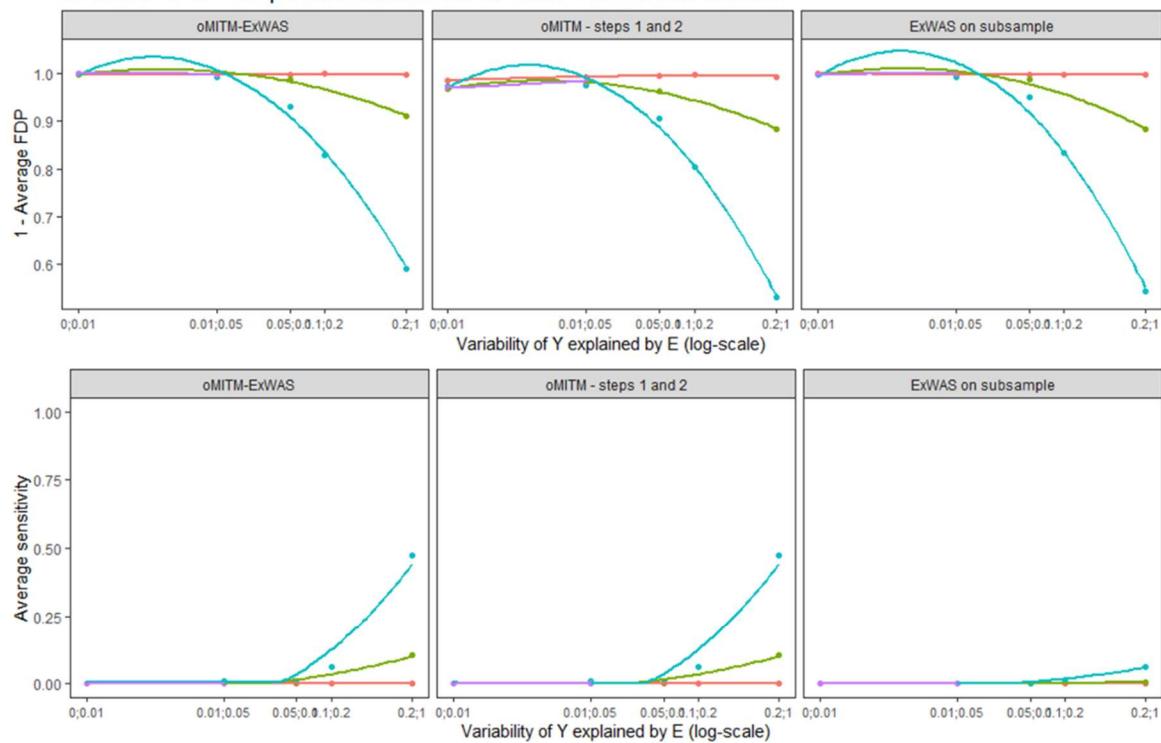
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.3: Comparisons between oMITM-ExWAS and control methods (*oMITM-steps 1 and 2* and *ExWAS on subsample*) performance (1-Average FDP and sensitivity) for causal structures A, B and C. Performances were averaged across scenarios according to categories of variabilities of Y explained by E (x-axis) and by M (color). Values were smoothed to give the trend according to averaged categories of variabilities of Y explained by E and by M.

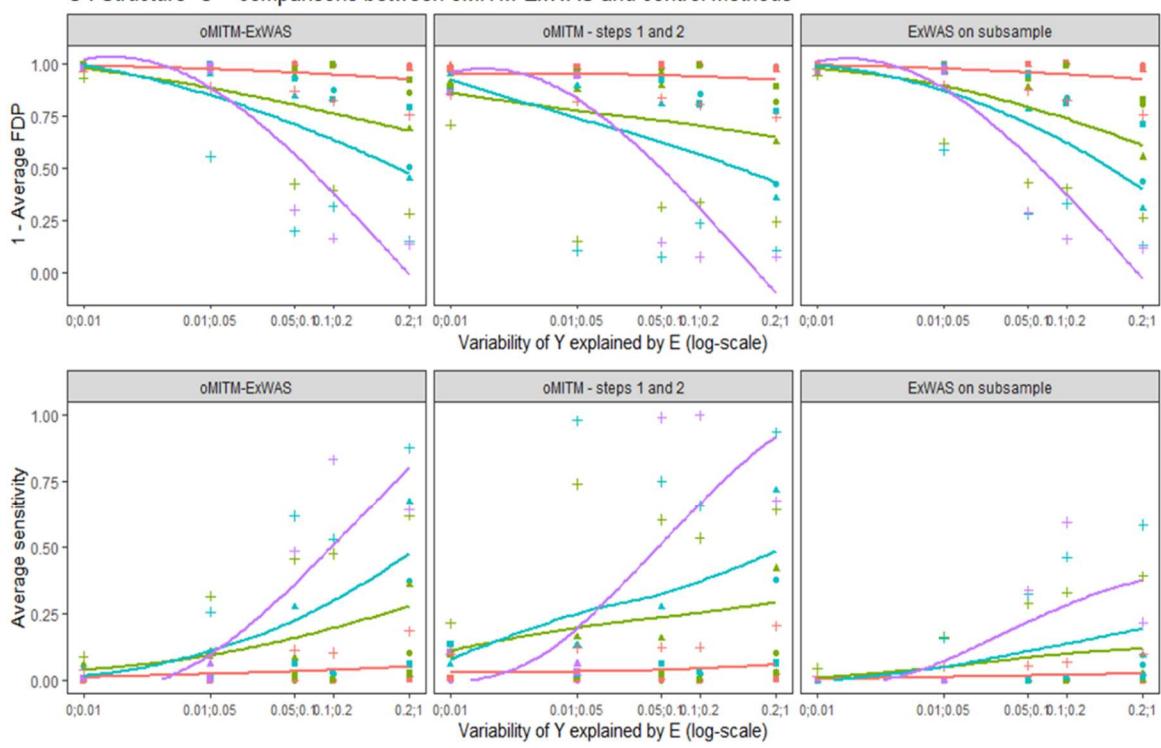


CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

B . Structure B - comparisons between oMITM-ExWAS and control methods



C . Structure C - comparisons between oMITM-ExWAS and control methods



Mean variability of a mediator explained by E Variability of Y explained by M

- 0;0.005 ▲ 0.005;0.01 ■ 0.01;0.15 + 0.15;0.2
- 0;0.05 — 0.05;0.1 — 0.1;0.4 — 0.4;1

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

When we assumed that the exposome directly influenced health (without mediation by the intermediate layer, structure B), all methods relying on information from the intermediary layer unsurprisingly showed very low sensitivity (lower than 0.010); they also had very low FDP (lower than 0.013, Table IV.2), as they did not select any exposure in most scenarios (see Supplementary Figure V.2C). Coming to the agnostic methods, their sensitivity increased with the variability of Y explained by E (Figure IV.4B). Among both types of methods, the one maximizing accuracy was DSA, which performed far better than the other methods (Table IV.2). oMITM-DSA ranked second in terms of accuracy: there were some scenarios (when both variabilities explained by E and M were higher than 0.1) in which oMITM methods selected some exposures that were true predictors (Figure IV.4B and Supplementary Figure V.2A). In these scenarios, oMITM-DSA showed good sensitivity (average, 50%) and very good FDP (lower than 15%). Indeed, counter-intuitively, for these scenarios, the reduced exposome selected by oMITM design was non-empty and contained more true predictors than would be selected by chance (this can be seen in Figure IV.3B by comparing the sensitivity of oMITM-ExWAS to the sensitivity of *ExWAS on subsample*, which was always lower). On the contrary, mediation provided a null sensitivity, always failing to detect true predictors (Figure IV.4B). This relatively good behavior of oMITM under causal structure B can be explained by the selection bias (Hernán et al., 2004) induced in step *b*) of the oMITM design when adjusting on Y: a spurious link between E and Y is created, leading to add some causal predictors of Y in the reduced exposome.

For structure C, the situation with both direct and indirect effects of the exposome on health, performances ranged between those observed in scenarios A and B; oMITM-DSA and DSA were, again, the methods with the highest accuracy (Figure IV.5).

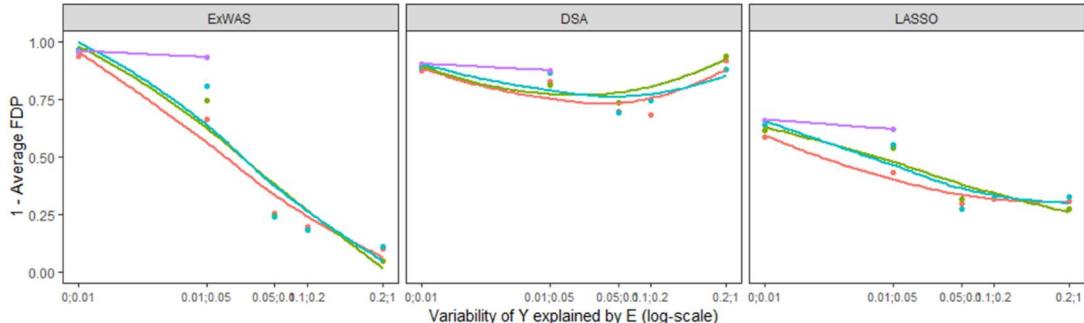
Table IV.2: Performance for every method under each causal structure. For structures A, B and C, FDP (average mean and standard error across scenarios), sensitivity (average mean and standard error across scenarios) and accuracy, defined as 1 - FDP + sensitivity (average mean across scenarios). For structure D, number of hits (average mean and standard error across scenarios) and sensitivity to find the exposures predicted by Y (average mean and standard error across scenarios). For structure E, number of hits (average mean and standard error across scenarios). For each performance indicator and for each structure, an * indicates the method with the best performance for a given causal structure. ^a: Proportion of exposures influenced by Y identified by the approach.

Methods	Causal structure A			Causal structure B			Causal structure C			Causal structure D		Causal structure E
	FDP (SD)	Sensitivity (SD)	Accuracy	FDP (SD)	Sensitivity (SD)	Accuracy	FDP (SD)	Sensitivity (SD)	Accuracy	Number of hits (SD)	Sensitivity to predicted exposures (SD) ^a	Number of hits (SD)
<i>Agnostic methods</i>												
ExWAS	0.132 (0.199)	0.126 (0.048)	0.994	0.388 (0.188)	0.363 (0.077)	0.975	0.361 (0.222)	0.288 (0.105)	0.927	6.622 (1.242)	0.554 (0.052)	0.32 (0.909)
DSA	0.123 (0.308)	0.113 (0.054)	0.99	0.169 (0.284)	0.279 (0.082)	1.110*	0.172 (0.306)	0.216 (0.087)	1.044*	5.935 (2.625)	0.182 (0.022)	0.13 (0.661)
LASSO	0.413 (0.430)	0.158 (0.098)*	0.745	0.528 (0.317)	0.395 (0.106)*	0.8671	0.540 (0.341)	0.320 (0.127)*	0.780	41.4 (17.483)	0.463 (0.124)	2.56 (5.472)
<i>Methods incorporating information from an intermediary layer</i>												
oMITM-ExWAS	0.094 (0.065)	0.105 (0.043)	1.011	0.012 (0.032)	0.010 (0.010)	0.998	0.109 (0.085)	0.088 (0.051)	0.979	0.014 (0.128)	0.001 (0.008)	0 (0)*
oMITM-DSA	0.038 (0.102) *	0.095 (0.049)	1.057*	0.009 (0.046)	0.010 (0.010)	1.001	0.043 (0.108) *	0.073 (0.045)	1.03	0.003 (0.022)*	2x10 ⁻⁴ (0.002)	0 (0)*
Mediation	0.097 (0.081)	0.105 (0.055)	1.008	0.003 (0.020)*	0.000 (0.003)	0.997	0.098 (0.083)	0.068 (0.044)	0.970	1.214 (0.4)	0.13 (0.034)	0 (0)*
<i>Control methods</i>												
ExWAS on subsample	0.091 (0.091)	0.041 (0.047)	0.959	0.014 (0.039)	0.001 (0.005)	0.987	0.110 (0.100)	0.043 (0.040)	0.932	0.002 (0.015)	8x10 ⁻⁵ (0.001)	0 (0)
oMITM steps 1 and 2	0.177 (0.158)	0.176 (0.058)	0.999	0.028 (0.110)	0.010 (0.011)	0.982	0.164 (0.151)	0.132 (0.062)	0.968	0.026 (0.219)	0.001 (0.013)	0 (0)

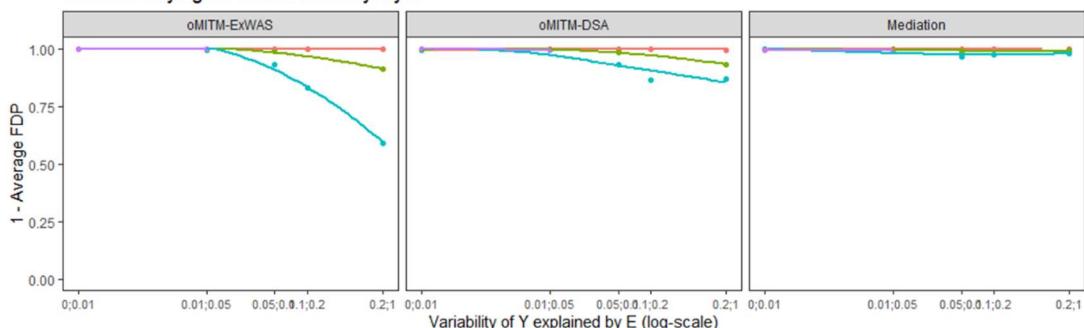
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.4: **A.** 1- FDP; **B.** sensitivity under causal structure B (see Figure IV.1). Performances were averaged across scenarios according to categories of variabilities of Y explained by E (x-axis) and by M (color). Values were smoothed to give the trend according to averaged categories of variabilities of Y explained by E and by M.

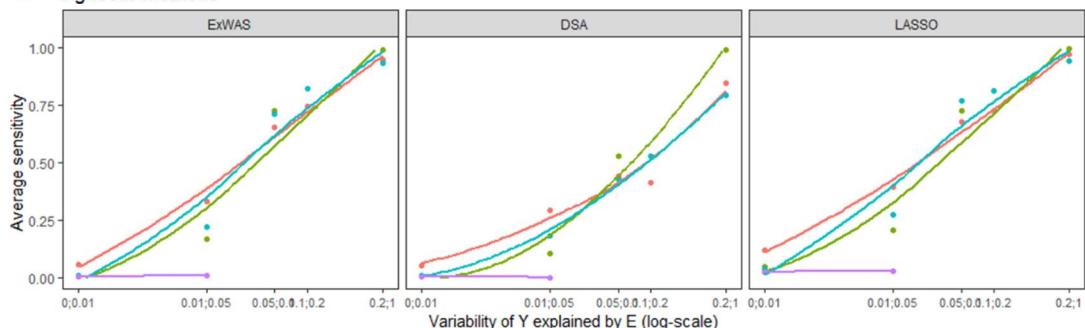
A. Agnostic methods



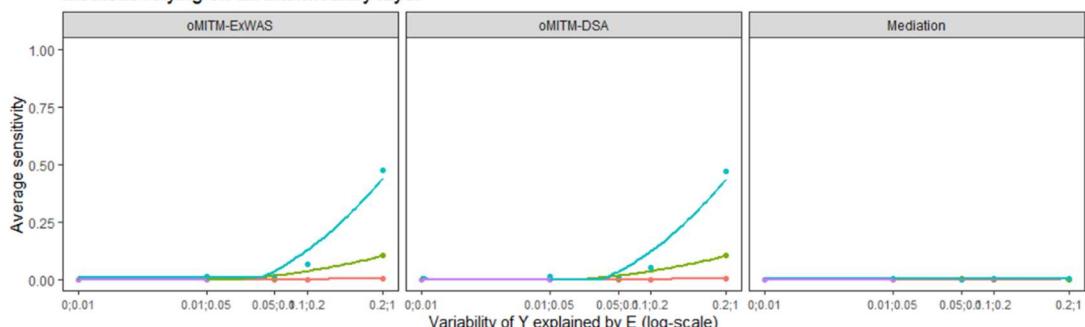
Methods relying on an intermediary layer



B. Agnostic methods



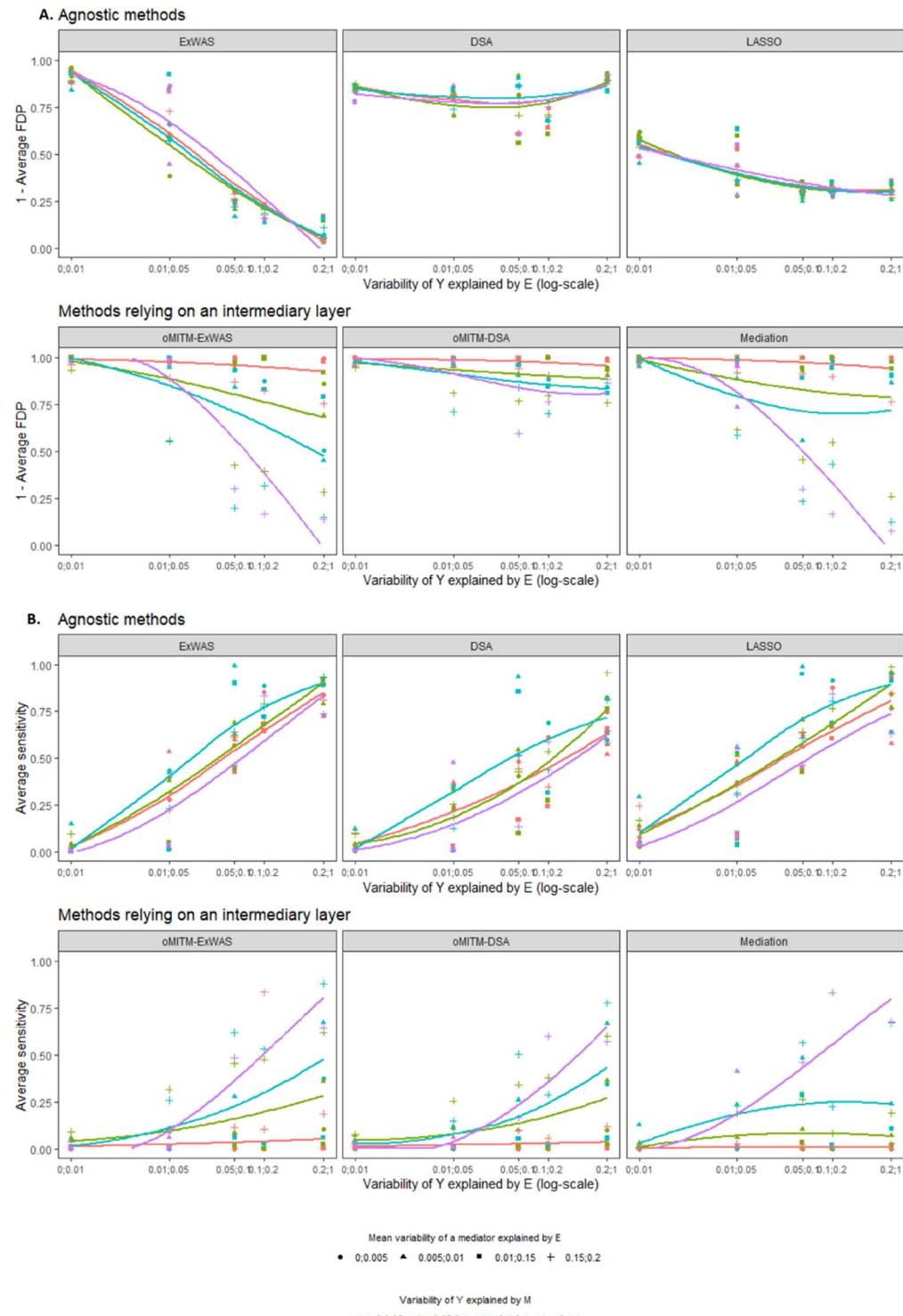
Methods relying on an intermediary layer



Variability of Y explained by M
— 0.05 — 0.05;0.1 — 0.1;0.4 — 0.4;1

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.5: **A.** 1- FDP; **B.** Sensitivity under causal structure C. Performances were averaged across scenarios according to categories of variabilities of Y explained by E (x-axis) and by M (color) and of mean variability of an element of M affected by E by E. Values were smoothed to give the trend according to performance averaged categories of variabilities of Y explained by E and by M.



IV.4.2. Causal structures without effect of the exposome on health

In a situation with a causal link from Y to E (corresponding to reverse causality, structure D, scenarios described in Supplementary Table V.3), all agnostic methods displayed a non-null number of hits, with the number of hits increasing when the variability of E explained by Y increased (Figure IV.6B and Supplementary Figure V.4A). This is consistent with the fact that these methods cannot distinguish an influence of E on Y from an influence of Y on E: as shown in Figure IV.6, as the variability of exposures explained by Y increased, exposures were more often selected as hits. This proportion of hits had values similar to the sensitivity displayed by these agnostic methods in structures A, B and C.

Table IV.3: Number of hits (average mean and standard error across scenarios), sensitivity to find the exposures predicted by Y (average mean and standard error across scenarios) under causal structures D and E. For each performance indicator and for each causal structure, an * indicates the method minimizing the indicator.

Structure	Causal structure D		Causal structure E	
Methods	Number of hits (SD)	Sensitivity to predicted exposures (SD)	Number of hits (SD)	Sensitivity to predicted exposures (SD)
<i>Agnostic methods</i>				
ExWAS	6.622 (1.242)	0.554 (0.052)	0.32 (0.909)	-
DSA	5.935 (2.625)	0.182 (0.022)	0.13 (0.661)	-
LASSO	41.4 (17.483)	0.463 (0.124)	2.56 (5.472)	-
<i>Methods incorporating information from the intermediary layer</i>				
oMITM-ExWAS	0.014 (0.128)	0.001 (0.008)	0 (0)*	-
oMITM-DSA	0.003 (0.022)	2x10 ⁻⁴ (0.002)	0 (0)*	-
Mediation	1.214 (0.4)	0.13 (0.034)	0 (0)*	-
<i>Control methods</i>				
ExWAS on subsample	0.002 (0.015)*	8x10 ⁻⁵ (0.001)*	0 (0)*	-
oMITM steps 1 and 2	0.026 (0.219)	0.001 (0.013)	0 (0)*	-

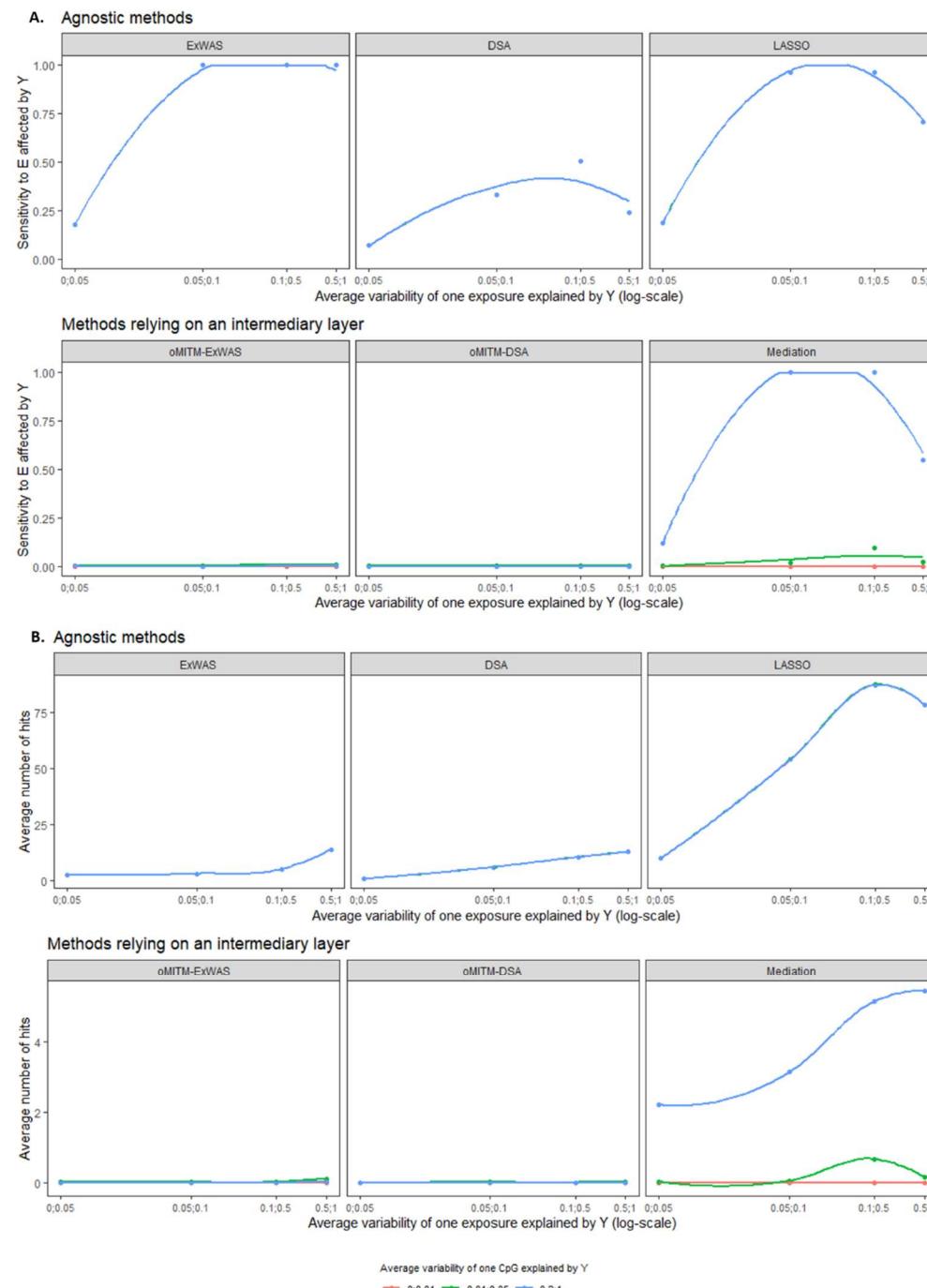
Both oMITM methods selected no exposure most of the time (Figure IV.6 and Table IV.2Table IV.3). On the contrary, the mediation analysis showed a non-null number of hits as soon as the mean variability of E explained by Y was higher than 0.05 and the mean variability of M explained by Y was higher than 0.3.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

The structure without any causal link (structure E) can be seen as the limit of all four precedent structures when the strength of all associations approaches zero. All methods using methylome information selected no exposure, while agnostic methods erroneously selected some exposures, with LASSO showing the highest error rate (Table IV.2 and Figure IV.7).

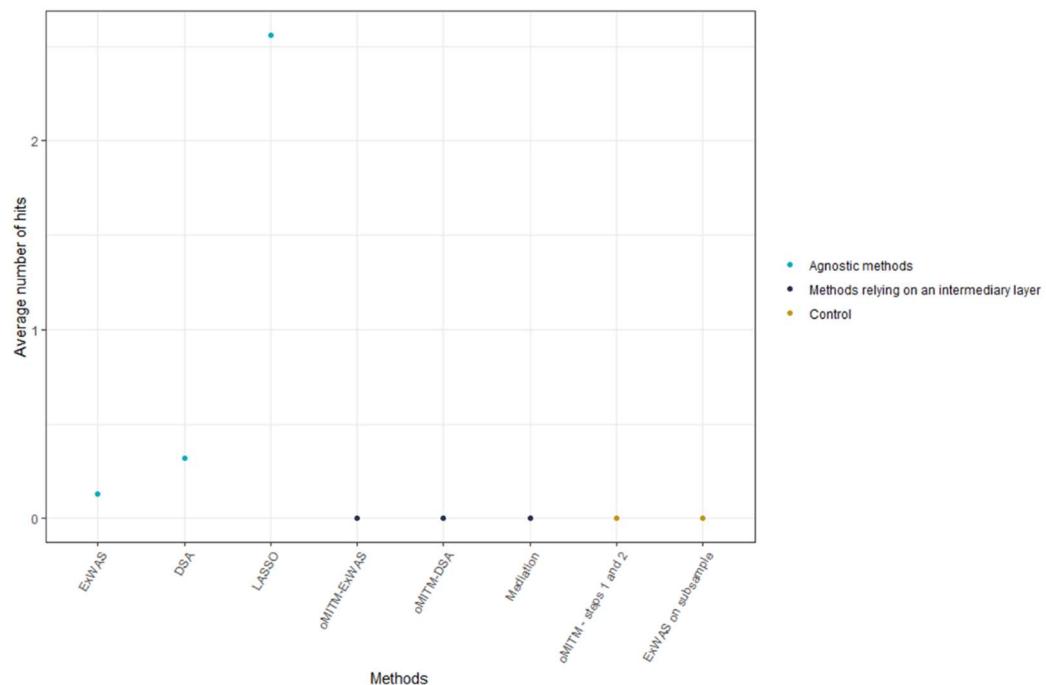
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.6: **A.** Proportion of exposures influenced by Y wrongly identified, and **B.** number of hits under causal structure D. Values were averaged across scenarios according to categories of variabilities of one exposure explained by Y (x-axis) and one element of M explained by Y (color).



CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Figure IV.7: Average number of covariates selected per method under causal structure E.



IV.4.3. Comparisons between methods using causal inference theory

Applying causal inference theory, we compared the number of possible causal structures under which various analytical strategies would be able to identify a true effect of an exposure on health in ideal situations of large sample size. The results are synthesized in Table IV.4, while details of results for each causal situation are displayed in Supplementary Table 4. In Supplementary Table 5, the step-by-step results for oMITM are detailed.

A test of association between E and Y ignoring M was expected to properly identify all situations in which E influenced Y (0 false negative, 9 true positive results, Table IV.4), but also identified associations corresponding to reverse causality (10 false positive results, Table IV.4). Among the methods using the intermediary variable M, oMITM and MITM without adjustment on Y both displayed 2 false negatives (structures J and K, Supplementary Table V.4). The mediation test showed 2 additional false negatives (Table IV.4): in particular, contrarily to oMITM, it was not able to detect the structure A in which E affects Y indirectly through M (structure A, Supplementary Table V.4, Figure IV.1). Coming to false positives, oMITM was the design minimizing the false positive findings (6 versus at least 8 for any other design). MITM method led to false positives in two situations of reverse causality to which oMITM was not sensitive (structures D and Q, Supplementary Table 4). The mediation method displayed similarly to MITM 8 false positives.

Overall, oMITM was the design giving true results (true positive or true negative) in the highest number of causal structures (17, versus 15 for tests of association ignoring M and for MITM not adjusted for Y, and 13 for mediation, Table IV.4).

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Table IV.4: Number of true causal links detected, false causal links detected, true causal links non-detected, false causal links non-detected by different designs among the 25 causal structures considering all possible links between 3 unidimensional layers. The analysis has been made using causal inference theory and the results for each of all 25 causal structures are detailed in Supplementary Table V.4 The columns giving true results (i.e. true positive or true negative) are displayed in bold. *: a design similar to our oMITM design but without adjusting on Y at step b). This design corresponds to Meet-in-the-Middle commonly implemented in the literature.

Methods	True causal link		No causal link		Total		
	Association detected (true positive)	No association detected (false negative)	Association detected (false positive)	No association detected (true negative)	True results (true negative and true positive)	False results (false negative and false positive)	All
Test of association	9 (36%)	0 (0%)	10 (40%)	6 (24%)	15 (60%)	10 (40%)	25 (100%)
oMITM	7 (28%)	2 (8%)	6 (24%)	10 (40%)	17 (68%)	8 (32%)	25 (100%)
Mediation	5 (20%)	4 (16%)	8 (32%)	8 (32%)	13 (52%)	12 (48%)	25 (100%)
MITM without adjusting on Y*	7 (28%)	2 (8%)	8 (32%)	8 (32%)	13 (52%)	12 (48%)	25 (100%)

IV. 5. Discussion

Our simulations highlighted that the oMITM design has high accuracy under various causal structures. In particular, it allows to avoid false-positive associations in some structures corresponding to reverse causality more efficiently than all other tested designs which detected the spurious association, in particular those not making use of the intermediary layer. Moreover, in the causal structures with a direct effect of the exposome on the outcome for which other methods sometimes suffer from a low specificity, it allows increasing specificity while conserving a good accuracy compared to other methods.

IV.5.1. Strengths and limitations

We implemented a simulation considering five different causal structures to identify in which contexts specific methods making use of information from an intermediary biological layer could be more efficient than specific agnostic algorithms to identify components of the exposome influencing health. Former simulations about the performance of statistical methods to assess exposome-health associations generally considered simpler causal structures, without any intermediate layer nor reverse causality (Agier et al., 2016; Barrera-Gómez et al., 2017; Lenters et al., 2018). Other simulations considered multi-layered data, but often with an aim distinct from ours, such as the quantification of the share of the effect of an exposure on an outcome mediated by a high dimension intermediate layer (Barfield et al., 2017; Tobi et al., 2018).

We only studied experimentally 5 of the 25 possible causal structures theoretically possible, deferring the discussion about the remaining causal structures to the qualitative assessment of the simplified DAGs (which did not assume that either E or M had a dimension larger than one). We selected the 5 structures that we thoroughly tested so as to cover what we considered to be the most realistic situations in an exposome setting; the reader interested in another specific structure may modify our code to study it more deeply. We considered separately these causal structures, while in reality, with multidimensional exposures and intermediary layers, several causal structures are expected to co-exist: for example, an exposure may only act directly on Y while another exposure could act directly and via an indirect effect mediated by M. Models performances estimated for different causal structures should not be compared one with another as the weight of scenarios with high or low variability explained by predictors were not the same across different causal structures. Within-structure comparisons/reasonings are more relevant.

In some of the considered situations, the variability of Y explained by E was very low, which seemed realistic to us. This corresponds to a situation of “rare and weak” event (Donoho and Jin, 2008), which may be more plausible than higher values of explained outcome variability assumed

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health in previous simulations (Agier et al., 2016; Barrera-Gómez et al., 2017). Thus, we chose to include these scenarios even if most methods showed very low power, as it may represent the performance encountered in real studies. This led to point to major difference in terms of specificity between methods. Situations in which E explained a large share of the variability of Y (above 20%) were hard to reach in the causal model corresponding to mediation (structure A), which should be seen as a realistic feature of our simulation rather than a limitation thereof. This was a consequence of our choice not to simulate scenarios with strong effects of E on M (maximum average share of variability in M explained by E, 20%).

We assumed that the dimension of our intermediary layer was 2284; this value corresponded to the dimension of a set of variables representing DNA methylation sites selected on the basis of their a priori relevance for the considered outcome (Cadiou et al., 2020); this is also a realistic size for biological information of other nature, such as metabolomic or immunological markers. The dimension of the intermediary layer in which the information is diluted is expected to impact the efficiency of approaches relying on this layer.

Coming to our causal inference analysis, the main limitation is that we analyzed only low-dimensional DAGs (with three variables), whereas the analyzed designs are meant to be used in higher dimension.

IV.5.2. Summary of methods' performances

Our oMITM is an innovative design, used here in two flavors (oMITM-ExWAS and oMITM-DSA). It shows similarities with a mediation design and especially with the Meet-in-the-Middle framework described in the literature (Chadeau-Hyam et al., 2011; Jeong et al., 2018; Vineis et al., 2013). It is notably distinguished from the classical Meet-in-the-Middle by that: 1) it does not aim to discover intermediary biomarkers but to reduce exposome dimension in the context of an exposome-outcome association; this explains the order chosen for the different steps; 2) we added an adjustment on the outcome in the test of association between the exposure and the potential mediators. Overall, our oMITM design showed good performance compared to agnostic methods. Due to our adjustment on the outcome (leading to what corresponds to a “selection bias”, as defined by Hernán et al. (2004)), oMITM can identify some true predictors even in structures under which there is no indirect effect of E on Y through M (causal structure B). We explained why this can happen in the theoretical part of our work (see paragraph 3.3). In situations of reverse causation without link between E and M, the additional adjustment on Y of our oMITM design also allowed to avoid false positives due to reverse causality. In situations of mediation without any direct effect of the exposures, the reduced exposome was relevant; under this causal structure, oMITM allowed

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

to decrease FDP in most scenarios, and in some scenarios to increase sensitivity. The replacement of ExWAS by DSA in the last step of the oMITM design increases performance, in particular in term of FDP when the effect of the exposures on the outcome was high. oMITM could be further enhanced by replacing the ExWAS-type methods used at step *b*) and *c*) by selection methods more adapted to the high dimension (see for example the reviews of (Fan and Lv, 2010; Lazarevic et al., 2019)).

We used an ExWAS-based implementation of mediation analysis (Küpers et al., 2015) to allow comparisons with the oMITM design (through the oMITM-ExWAS). However, alternative mediation implementation, more adapted to multidimensional mediators, have been proposed (Barfield et al., 2017; Blum et al., 2020; Chén et al., 2018).

Moving now to the agnostic methods, Deletion-Addition-Substitution algorithm was the best agnostic method in situations involving a causal effect of the exposome on the health outcome. As shown before by Agier et al. (2016), DSA provided a better compromise between sensitivity and specificity than ExWAS. However, it is prone to suffer from reverse causality, like all other agnostic methods. Our results on ExWAS are consistent with those from Agier et al. (2016) when R^2 was higher than 0.1. When R^2 was lower than 0.01, ExWAS often selected no exposures and thus exhibited a FDP of 0 whereas the two other agnostic methods (DSA and LASSO) showed non-null FDP and null or very low sensitivity. LASSO was the worst performing agnostic method; in particular, it displayed a very high FDP. In a case of correlation between a true predictor and other variables, LASSO is known to select one variable among a set of correlated variables (Leng et al., 2006). The high rate of false positive findings that we observed may be explained by our choice of a penalty parameter (the parameter which minimizes the error of prediction during the cross-validation process (Tibshirani, 1996)) optimized for prediction. Elastic-Net (Friedman et al., 2010), which was designed to improve the performance of Lasso when predictors are correlated, could have been tested here. However, Agier et al. (Agier et al., 2016) already showed that DSA provided better performance than Elastic-Net in the context of a realistic exposome.

IV.5.3. Consistency between our structural causal modelling analysis and experimental simulation-based

Although simplified in its design, our analysis based on DAGs yield results consistent with the more elaborate simulation study, which considered an exposome of dimension 173 and an intermediate layer of dimension 2284. In particular, in the causal structure of reverse causality (Y influencing E and M) without link between E and M (structure D), the oMITM method provided no hit (Figure IV.6), as predicted by the analyses of DAGs (Supplementary Table IV.4). Similarly, in structure B, we observed a non-null sensitivity of oMITM due to selection bias when the variabilities in Y explained by both E and M were above a certain level.

Moreover, the behavior of oMITM in a structure of reverse causality is also consistent with the results of a previous study using oMITM-ExWAS to relate the exposome and child BMI in Helix data using methylome (Cadiou et al., 2020). Indeed, as detailed in Cadiou et al (2020), an agnostic ExWAS applied on the same data resulted in 20 significant associations, with the majority likely to be due to reverse causality: most of these hits corresponded to lipophilic substances (such as polychlorobiphenyls (PCB)), measured in blood at the same time as the outcome. They were negatively associated with BMI, whereas toxicological studies based on a prospective design suggested obesogenic effect of such components (Heindel & vom Saal, 2009; Thayer et al., 2012). As they are stored in fat, a plausible explanation is that these associations are due to increased fat levels in obese subjects, entailing a higher amount of PCBs stored in fat and, conversely, a lowering of circulating PCB levels in blood (Cadiou et al., 2020). The reduced exposome obtained with oMITM, which consisted of 4 exposures, did not contain any of these hits of the agnostic analysis suspected to be due to reverse causality, except PFOS level. Thus, we can hypothesize that for these exposures, this situation corresponded to one of the cases of reverse causality situations discussed above, in which the oMITM design is not expected to identify exposures influenced by the outcome. This is consistent with the simulation results and highlights that the benefit of oMITM may come from the dimension reduction performed in the two first steps. The fact that blood postnatal level of PFOS, another compound suspected of reverse causality, was selected by the oMITM-ExWAS approach may be a consequence of the fact that oMITM is not expected to avoid all situations of reverse causality (as shown by our causal discovery analysis (Supplementary Table IV.4)).

IV.5.4. The need to rely on causal knowledge

We illustrated under which causal structures the results from previous exposome-health simulations (Agier et al., 2016) are expected to be true and that methods always imply underlying causal assumptions which are difficult to verify in an exposome setting. We showed that the use of additional information through the use of methylome layer can help to deal with reverse causality and thus decrease the false positive findings. This illustrates the affirmation of Hernan (2019) that “causal analyses typically require not only good data and algorithms, but also domain expert knowledge.” In our case, the use of an intermediate layer and our design, which itself relies on the assumption of three distinctive biological layers, added some *a priori* information. However, oMITM is still expected to lead to false positive findings in several causal structures corresponding to reverse causality. Further knowledge, for example on the causal link between the exposome and the intermediate layer, could help discarding these non-causal associations. Our work also illustrates that classical designs, such as mediation and classical Meet-in-the-Middle procedure, are not robust to violations of the strong assumptions they make about the underlying causal structure. Especially, a significant mediation or classical Meet-in-the-Middle result should not be interpreted as a causal clue supplementing the association between a factor or an outcome, unless strong knowledge about the intermediary variables *a priori* makes their mediating role very likely: as we demonstrated, in the causal structure D, which featured (reverse) causal links from the outcome to the potential mediators and to the exposure, both mediation test and basic association test can result in significant associations. Similarly (see theoretical results for structure D), a classic Meet-in-the-Middle framework without adjustment on the outcome at the second step would also lead to significant associations. Interestingly, in such a situation, even a longitudinal design may not be sufficient to get rid of reverse causality (see the DAG provided in Supplementary Figure 5 for an example). Thus, the statement about the Meet-in-the-Middle procedure that “*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.* » (Vineis et al., 2020) must be interpreted cautiously: associations rising from an epidemiological study should be supplemented by toxicological and biological knowledge. Overall, our work confirms that the uncertainty about the causal framework deserves to be taken in consideration when applying statistical methods to exposome and health data: first, it is of course crucial to understand the underlying causal assumptions behind the existing model, and to take them into account when interpreting epidemiologic results; secondly, multilayer approaches such as our oMITM design can be more robust than agnostic approaches when the causal model is uncertain.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

From a practical point of view, in an exposome health study where intermediary data are available, if strong prior knowledge about the outcome or the nature of the intermediary layer makes one specific causal structure very likely, one may choose the method(s) with a design adapted to this causal structure according to a comparative causal analysis such as the one we performed. The oMITM should in particular be preferred if there are reasons to expect associations due to reverse causality (e.g. in the case of a cross-sectional design). A multilayer design should be preferred to an agnostic one if both are adapted to the hypothesized underlying structure as the first one could help increase the specificity. Once the design is chosen, the statistical methods (e.g. DSA, ExWAS) for the implementation of this design should be chosen according to the dimensions of the considered layer(s), relying on simulations studies. For example, in an exposome settings and with an intermediary layer of intermediate dimension, our own simulations showed that respectively DSA and ExWAS may be adapted for the implementation of the different steps of an oMITM design.

If little a priori knowledge is available about the underlying causal structure, one could use either an agnostic approach (if one tends to favor sensitivity over specificity, e.g. in a rather exploratory study) or oMITM, which proved to be robust, if one tends to favor specificity.

IV. 6. Supplementary materials

Supplementary Table IV.1: Meaning and ranges of parameters used in each causal structure to simulate the link between layers. *When a scenario did not fulfil the ‘multiplicity constraint’ (see Supplementary Material V.1), it was ignored.

Structure	Parameters	Meaning of parameters	Range of parameters	Theoretical number of scenarios	Effective number of scenarios tested*
A	n_E_{E->M->Y}	Numbers of exposures having an effect on at least one intermediary variable which has an effect on Y	1, 3, 10, 25	384	384
	n_E_{E->Y}	Number of exposures having a direct effect on Y	0		
	n_E_{E->M}	Numbers of exposures having an effect on at least one intermediary variable having no effect on Y	0		
	n_M_{E->M}	Number of intermediary variables affected by an exposure having no effect on Y	0		
	n_M_{E->M->Y}	Number of intermediary variables affected by an exposure and having an effect on Y	10, 18, 25 ,100		
	n_M_{M->Y}	Number of intermediary variables non-affected by an exposure having an effect on Y	0		
	R2	Total variability of Y explained by E and M	0.01, 0.05, 0.1, 0.4		
	β	Coefficient of the effect of an exposure on an intermediary variable	0.0001, 0.001, 0.01, 0.1, 0.5		
	β'	Coefficient of the effect of an intermediary variable on Y	0.01		
	β''	Coefficient of the effect of an exposure on Y	0		
B	n_E_{E->M->Y}	Numbers of exposures having an effect on at least one intermediary variable which has an effect on Y	0	320	180
	n_E_{E->Y}	Number of exposures having a direct effect on Y	1, 3, 10, 25		
	n_E_{E->M}	Numbers of exposures having an effect on at least one intermediary variable having no effect on Y	0		
	n_M_{E->M}	Number of intermediary variables affected by an exposure having no effect on Y	0		

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

		Number of intermediary variables affected by an exposure and having an effect on Y	0
	$n_{M_{E \rightarrow M \rightarrow Y}}$	Number of intermediary variables non-affected by an exposure having an effect on Y	10, 18, 25, 100
	R^2	Total variability of Y explained by E and M	0.01, 0.05, 0.1, 0.4
	β	Coefficient of the effect of an exposure on an intermediary variable	0.0001, 0.001, 0.01, 0.1, 0.5
	β'	Coefficient of the effect of an intermediary variable on Y	0.01
	β''	Coefficient of the effect of an exposure on Y	0
C	$n_{E_{E \rightarrow M \rightarrow Y}}$	Numbers of exposures having an effect on at least one intermediary variable which has an effect on Y	1, 3, 10, 25
	$n_{E_{E \rightarrow Y}}$	Number of exposures having a direct effect on Y	1, 3, 10, 25
	$n_{E_{E \rightarrow Y} U n_{E_{E \rightarrow M \rightarrow Y}}}$	Number of exposures having both a direct effect on Y and an effect on at least one intermediary variable which has an effect on Y	$n_{E_{E \rightarrow Y}}$
	$n_{E_{E \rightarrow M}}$	Numbers of exposures having an effect on at least one intermediary variable having no effect on Y	0
	$n_{M_{E \rightarrow M}}$	Number of intermediary variables affected by an exposure having no effect on Y	0
	$n_{M_{E \rightarrow M \rightarrow Y}}$	Number of intermediary variables affected by an exposure and having an effect on Y	10, 18, 25, 100
	$n_{M_{M \rightarrow Y}}$	Number of intermediary variables non-affected by an exposure having an effect on Y	0
	R^2	Total variability of Y explained by E and M	0.01, 0.05, 0.1, 0.4
	$n_{M_{Y \rightarrow M}}$	Number of intermediary variables affected by Y	10, 18, 25, 100
	β	Coefficient of the effect of an exposure on an intermediary variable	0.0001, 0.001, 0.01, 0.1, 0.5
	β'	Coefficient of the effect of an intermediary variable on Y	0.01
	β''	Coefficient of the effect of an exposure on Y	0.0001, 0.001, 0.01, 0.1, 0.5

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

D	$n_{M_{M \rightarrow Y}}$	Number of intermediary variables having an effect on Y	0	576	576
	$n_{E_{Y \rightarrow E}}$	Number of exposures affected by Y	1, 3, 10, 25		
	β'	Coefficient of the effect of an intermediary variable on the outcome	0		
	γ	Coefficient of a non-zero effect of the outcome on an intermediary variable	0.0001, 0.001, 0.01, 0.1, 0.5, 2		
	γ'	Coefficient of the effect of the outcome on an exposure	0.0001, 0.001, 0.01, 0.1, 0.5, 2		
E	None	N.A.	N.A.	1	1

Supplementary Table IV.2: Characteristics of scenarios for structure A, B and C.

Causal structure	Causal structure A						Causal structure B						Causal structure C					
	Descriptive statistics*	Min.	25 th centile	Median	Mean	75 th centile	Max.	Min.	25 th centile	Median	Mean	75 th centile	Max.	Min.	25 th centile	Median	Mean	75 th centile
Total variability of Y explained by E	3.46 x10 ⁻⁴	6.49 x10 ⁻⁴	0.004	0.016	0.010	0.289	4.34 x10 ⁻⁴	0.004	0.015	0.073	0.069	0.409	7.570 x10 ⁻⁴	0.004	0.013	0.062	0.059	0.406
Total variability of Y explained by M	0.014	0.054	0.098	0.157	0.206	0.428	0.004	0.011	0.043	0.087	0.100	0.428	0.005	0.020	0.057	0.112	0.111	0.427
Mean variability of one intermediary variable affected by E explained by E	3.87 x10 ⁻⁴	5.07 x10 ⁻⁴	0.009	0.041	0.013	0.189	0.000	0.000	0.000	0.000	0.000	0.000	3.869 x10 ⁻⁴	5.073 x10 ⁻⁴	0.009	0.041	0.013	0.189

*For each structure, various scenarios were simulated using the range of parameters detailed in Supplementary Table 1. Variabilities of Y explained by respectively E and M and mean variability of one intermediary variable affected by E explained by E were measured for each scenario, and descriptive statistics for these measures were computed across structure.

Supplementary Table IV.3: Characteristics of the scenarios simulated for structure D

	Min.	25thcentile	Median	Mean	75th.centile	Max.
Average variability of an intermediary variable affected by Y explained by Y*	4.18Ex10 ⁻⁴	4.48 x10 ⁻⁴	0.006	0.170	0.204	0.804
Average variability of an exposure affected by Y explained by Y*	3.72 x10 ⁻⁴	4.94 x10 ⁻⁴	0.019	0.295	0.628	0.988

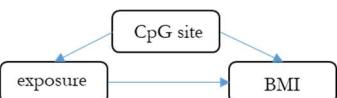
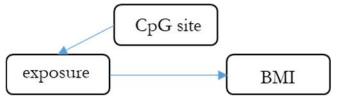
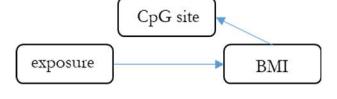
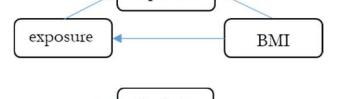
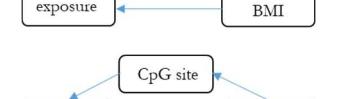
*Various scenarios were simulated under structure D using the range of parameters detailed in Supplementary Table 1. Average variability of an intermediary variable affected by Y explained by Y and Average variability of an exposure affected by Y explained by Y were measured for each scenario, and descriptive statistics for these measures were computed across structure.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

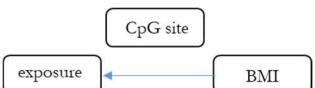
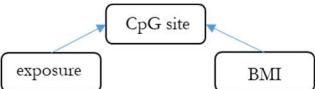
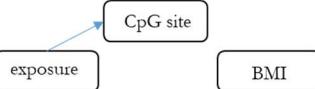
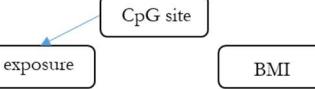
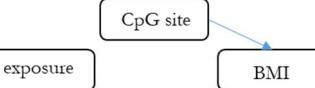
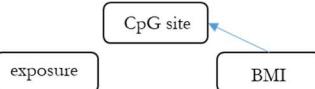
Supplementary Table IV.4: DAG analysis for different designs when considering all possible links between 3 unidimensional layers (e.g. an exposure, a CpG site, and BMI) according to causal inference theory.

	DAG	Causal link from E to Y	Mediation from E to Y	Association between E and Y	Detected by oMITM	Detected by MITM (without correcting on Y)	Detected by mediation analysis	Wanted to detect (i.e. direct or indirect causal link between E and Y)
A	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	Yes	No	Yes	Yes	Yes	Yes	Yes
B	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	No	Yes	Yes	Yes	No	No	Yes
C	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	Yes	Yes	Yes	Yes	Yes	Yes	Yes
D	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	No	No	Yes	No	Yes	Yes	No
E	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	No	No	No	No	No	No	No
F	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	Yes	No	Yes	Yes	Yes	Yes	Yes
G	<pre> graph TD exposure[exposure] --> CpGsite[CpG site] exposure --> BMI[BMI] CpGsite --> BMI </pre>	Yes	No	Yes	Yes	Yes	No	Yes

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

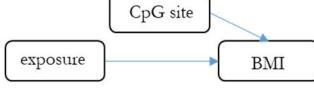
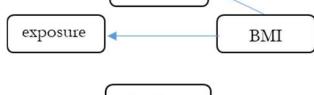
H		Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
I		Yes	No	Yes	Yes	Yes	Yes	No	Yes
J		Yes	No	Yes	No	Yes	Yes	Yes	Yes
K		Yes	No	Yes	No	No	No	No	Yes
L		No	No	Yes	Yes	Yes	Yes	Yes	No
M		No	No	Yes	Yes	Yes	No	No	No
N		No	No	Yes	Yes	Yes	Yes	Yes	No
O		No	No	Yes	Yes	Yes	Yes	Yes	No
P		No	No	Yes	No	No	Yes	No	No
Q		No	No	Yes	No	Yes	Yes	No	No

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

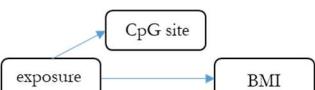
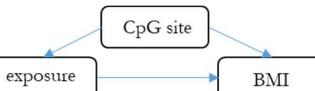
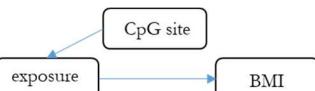
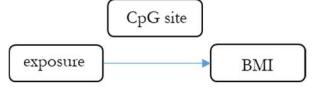
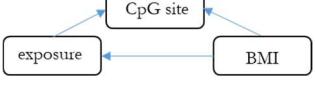
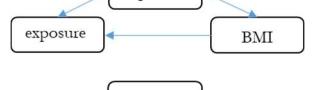
R		No	No	Yes	No	No	No	No
S		No	No	No	No	No	No	No
T		No	No	No	No	No	No	No
U		No	No	Yes	Yes	Yes	Yes	No
V		No	No	Yes	Yes	Yes	Yes	No
W		No	No	No	No	No	No	No
X		No	No	No	No	No	No	No
Y		No	No	No	No	No	No	No

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

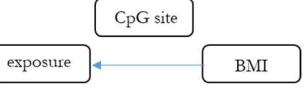
Supplementary Table IV.5: Details of causal inference analysis for the oMITM design applied to 3 variables (e.g. an exposure, a CpG site, and BMI) according to causal inference theory in all possible causal structures.

	DAG	Causal link from E to Y	Mediation from E to Y	Bias in oMITM step 1	Bias in oMITM step 2	Selected as to be tested in oMITM (assuming perfect power)	Detected in oMITM (assuming perfect power)	Wanted to detect (i.e. direct or indirect causal link between E and Y)
A		Yes	No	No	No	Yes	Yes	Yes
B		No	Yes	No	Yes (selection bias)	Yes (selection bias)	Yes	Yes
C		Yes	Yes	Yes (uncorrected for confounder)	No	Yes	Yes	Yes
D		No	No	No	No	No	No	No
E		No	No	No	No	No	No	No
F		Yes	No	Yes (uncorrected for confounder)	No	Yes	Yes	Yes

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

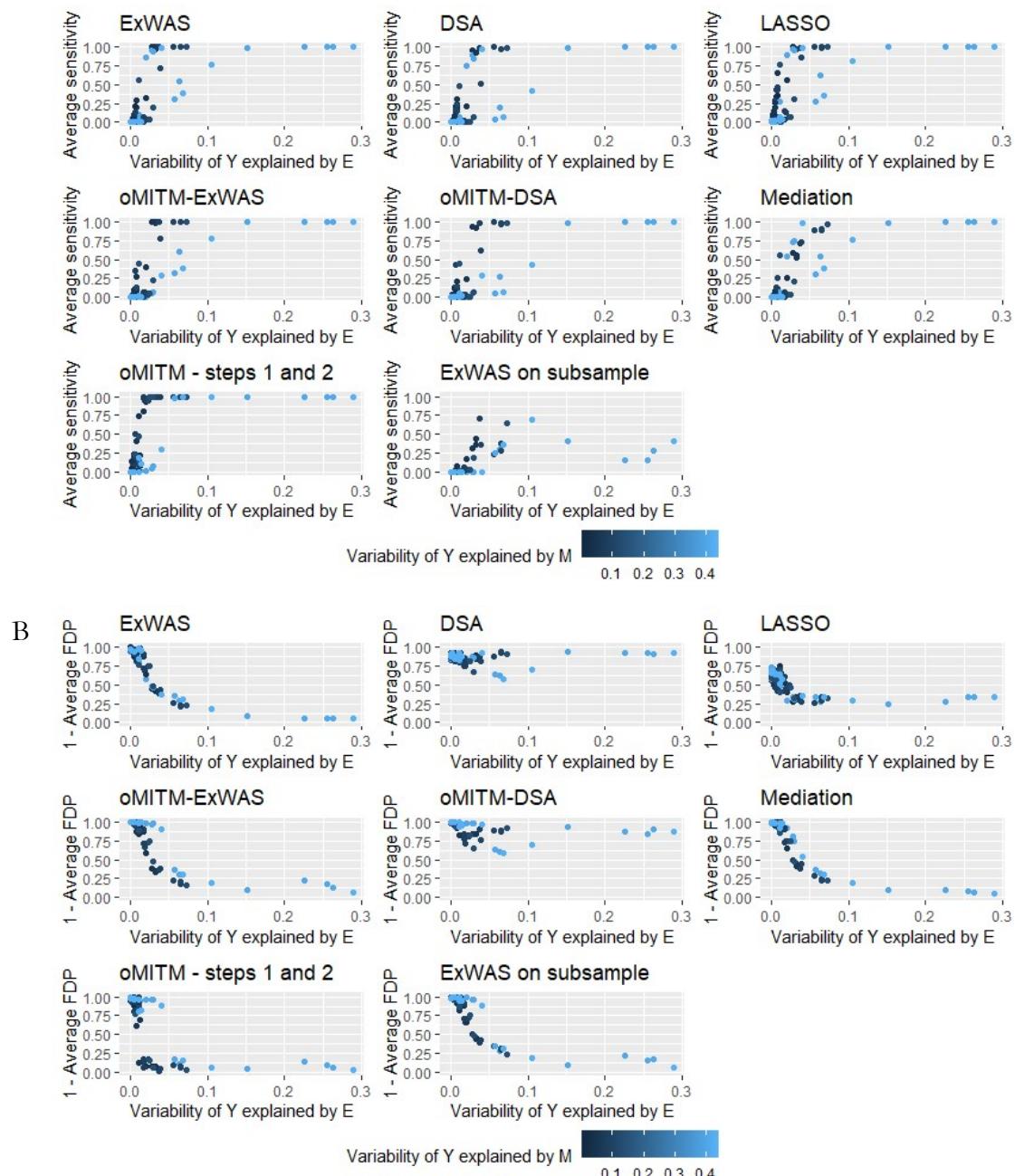
G		Yes	No	Yes (uncorrected for confounder)	No	Yes	Yes	Yes
H		Yes	No	No	Yes (selection bias)	Yes	Yes	Yes
I		Yes	No	No	No	No	Yes	Yes
J		Yes	No	No	No	No	No	Yes
K		Yes	No	No	No	No	No	Yes
L		No	No	No	No	Yes	Yes	No
M		No	No	No	No	Yes	Yes	No
N		No	No	No	No	Yes	Yes	No
O		No	No	No	No	Yes	Yes	No
P		No	No	No	No	No	No	No

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Q		No	No	No	No	No	No	No
R		No	No	No	No	No	No	No
S		No	No	No	No	Yes	No	No
T		No	No	No	No	No	No	No
U		No	No	No	No	Yes	Yes	No
V		No	No	No	No	Yes	Yes	No
W		No	No	No	No	No	No	No
X		No	No	No	No	No	No	No
Y		No	No	No	No	No	No	No

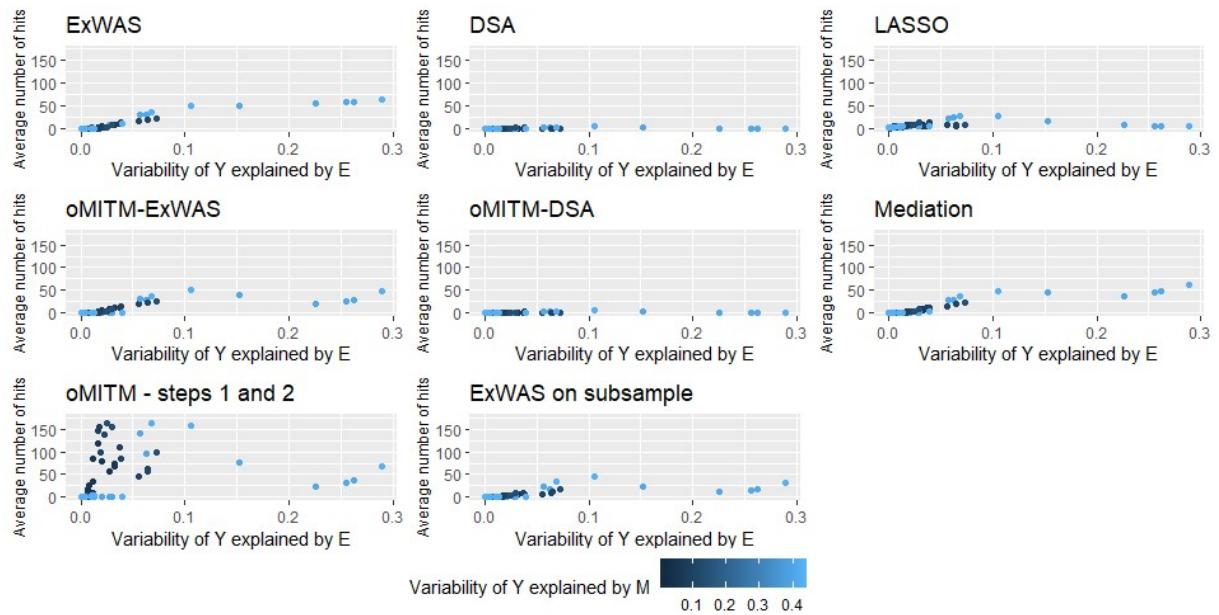
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

- A Supplementary Figure IV.1: Causal structure A, average sensitivity (A), 1- FDP (B) and number of hits (C) according to the variability of Y explained by E, method by method. Color scale gives the magnitude of variability of Y explained by M.



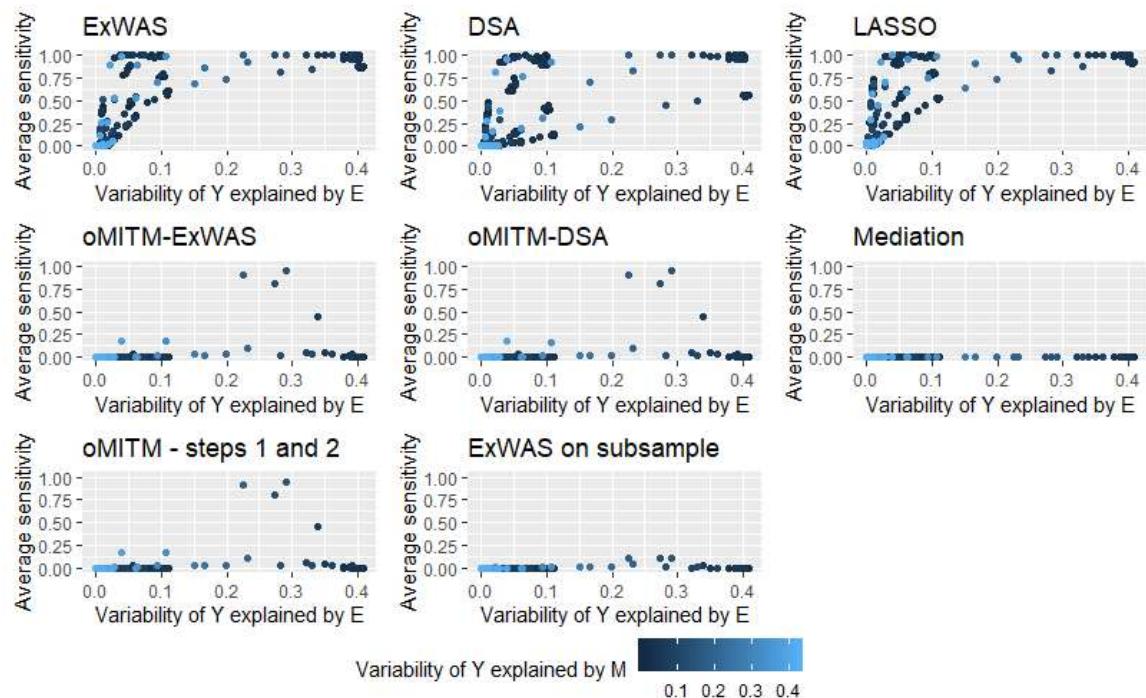
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

C.



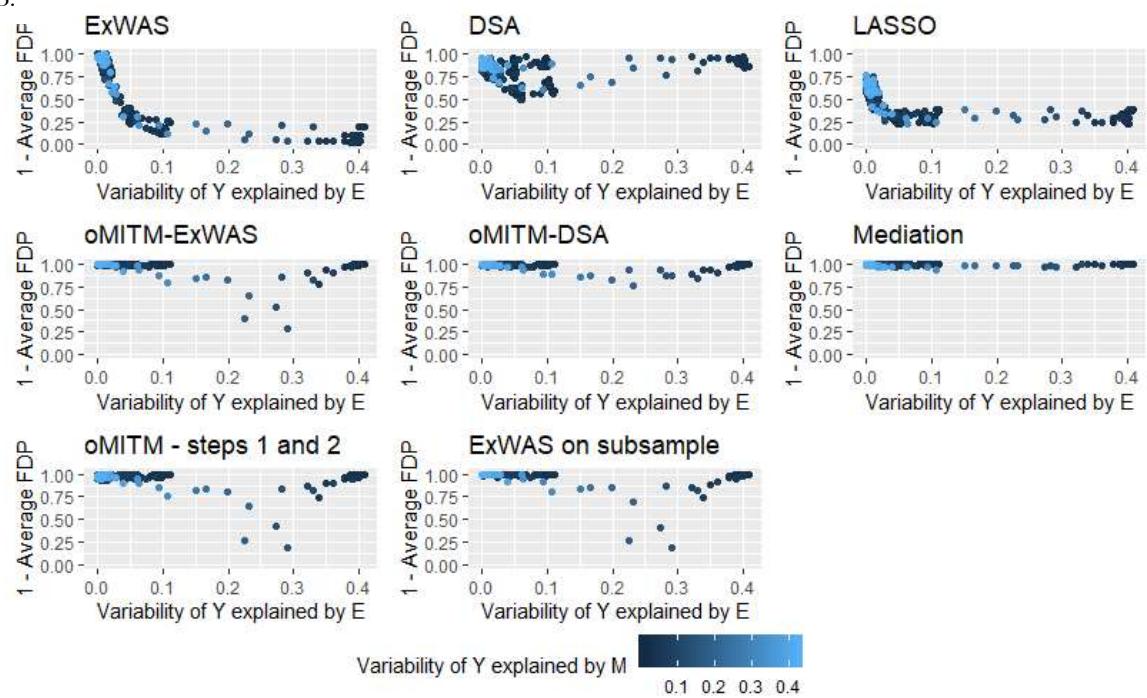
Supplementary Figure IV.2: Causal structure B, average sensitivity (A), 1- FDP (B) and number of hits (C) according to the variability of Y explained by E methods by methods. Color scale gives the magnitude of variability of Y explained by M.

A.

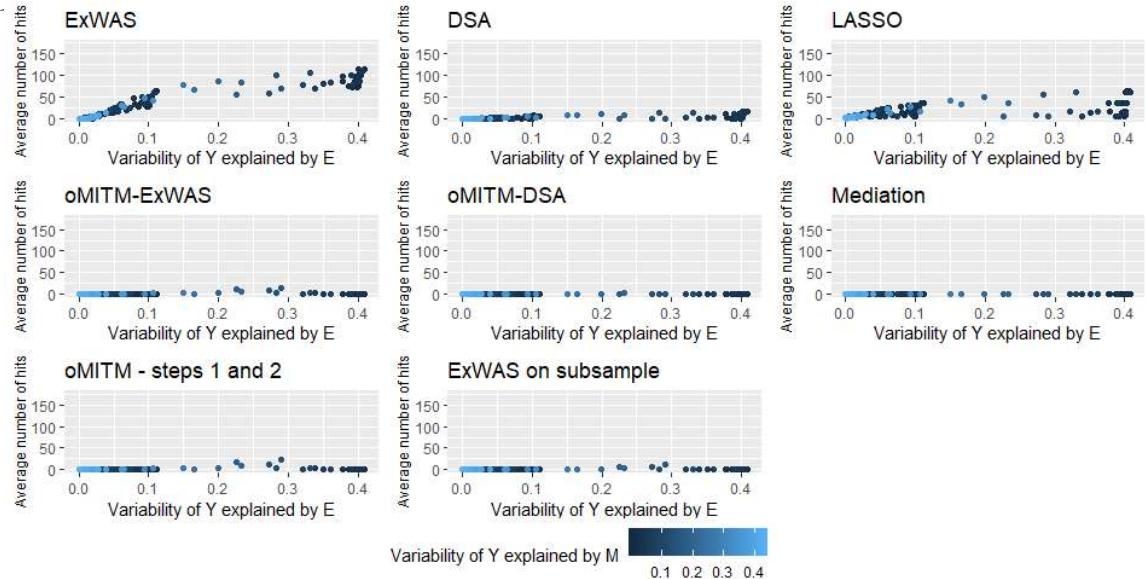


CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

B.

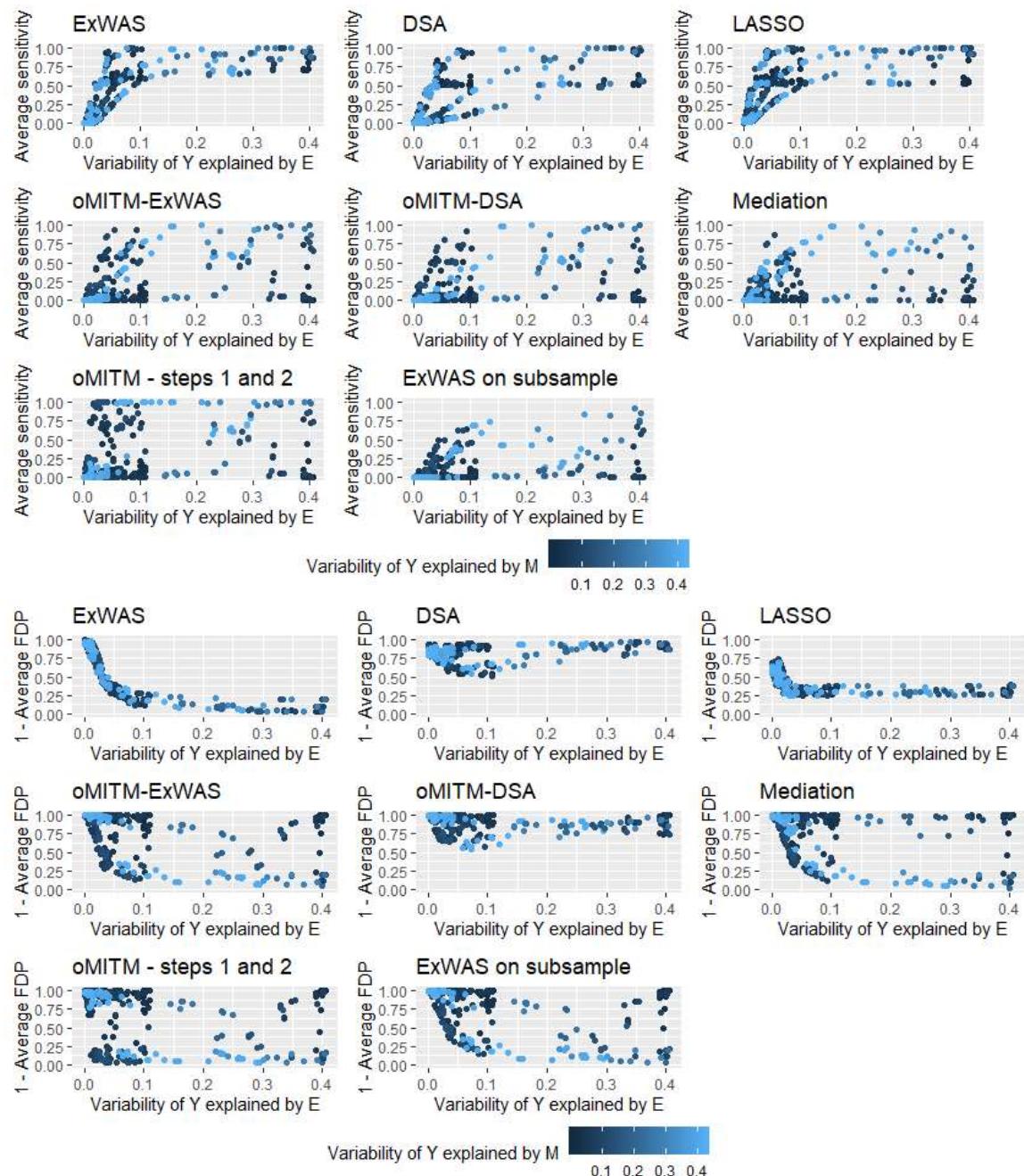


C



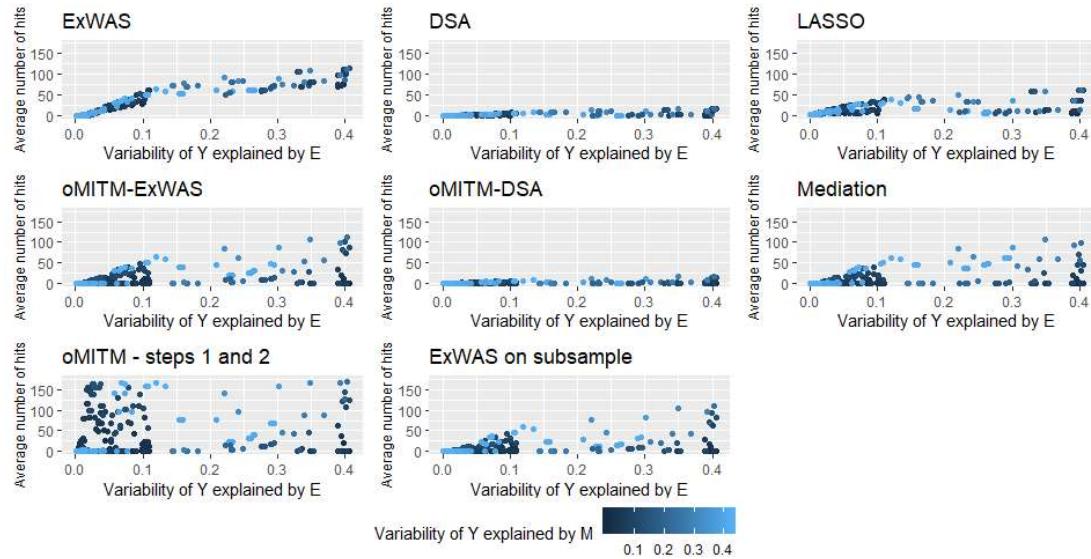
CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Supplementary Figure IV.3: Causal structure C, average sensitivity (A), 1- FDP (B) and number of hits (C) according to the variability of Y explained by E, method by method. Color scale gives the magnitude of variability of Y explained by M.

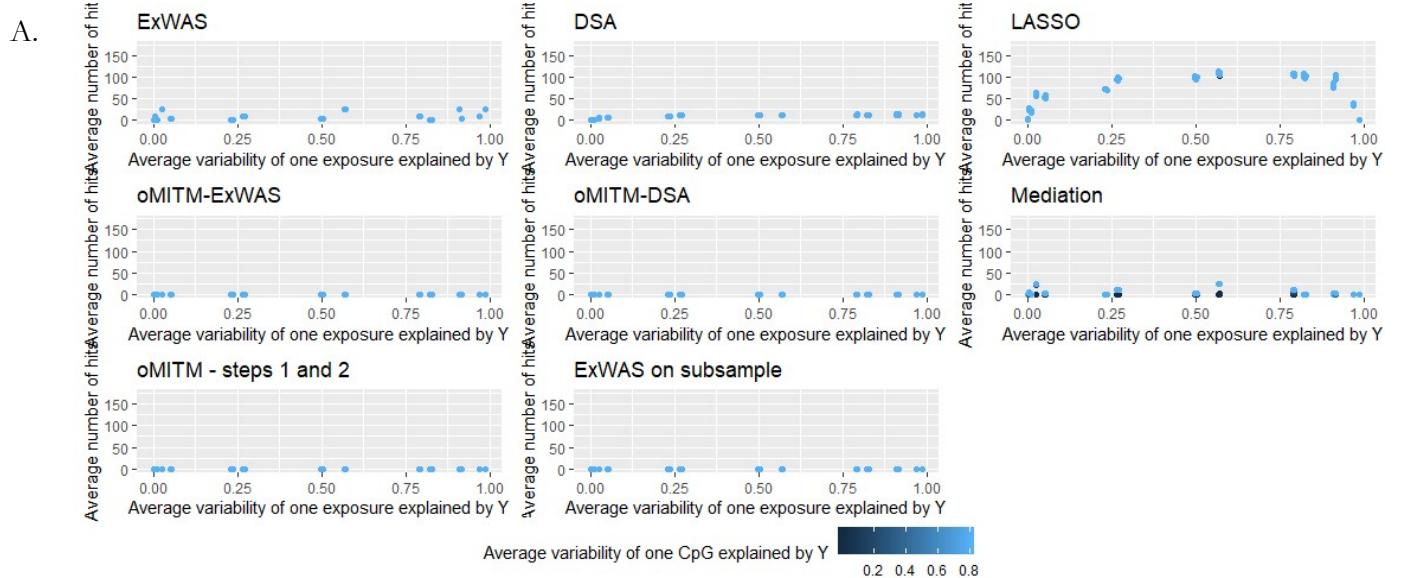


CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

C.

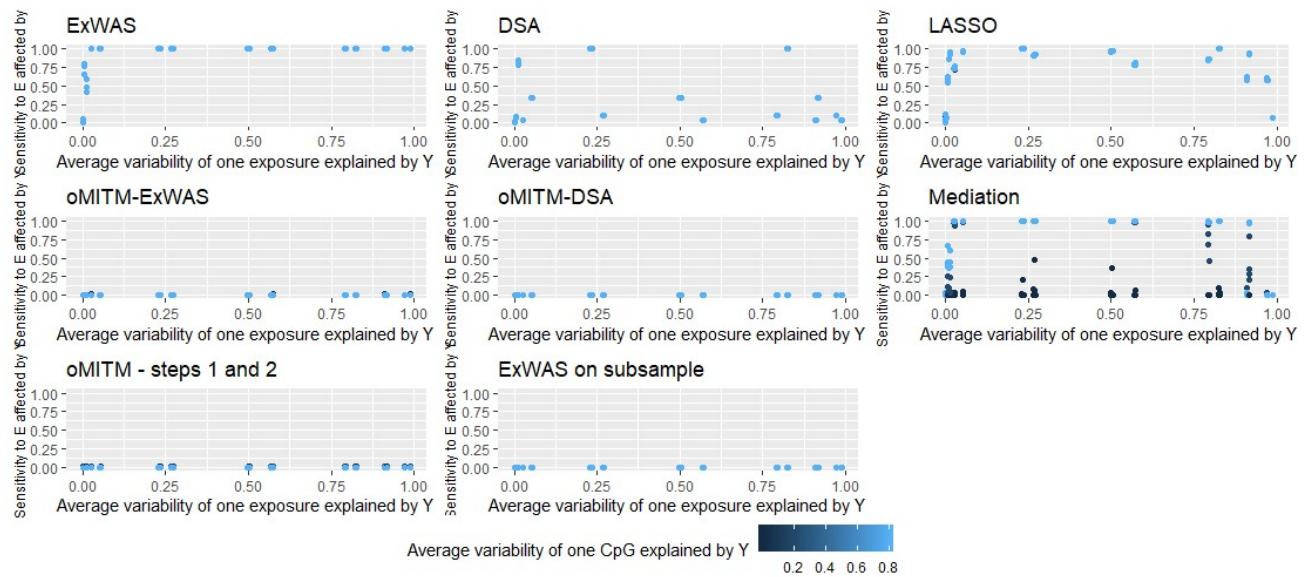


Supplementary Figure IV.4: causal structure D, average number of hits (A), and sensitivity to detect exposures affected by Y (B) according to the variability of one exposure affected by Y explained by Y, methods by methods, in causal structure D (reverse causality). Color scale gives the magnitude.

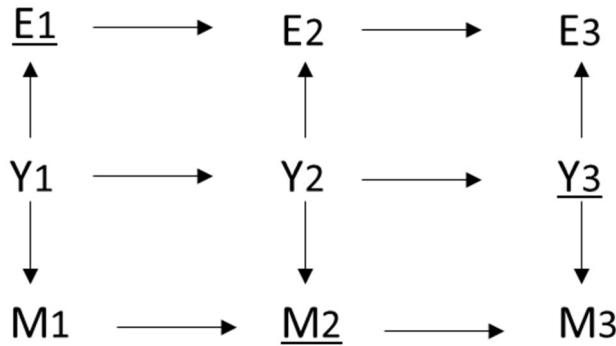


CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

B.



Supplementary Figure IV.5: Example of a causal situation in which a classical Meet-in-the-Middle framework without our additional adjustment would conclude to causal associations between E_1 and Y_3 whereas there is no causal influence of E_1 on Y_3 through M . This causal situation corresponds to a longitudinal design, with 3 different measurement times for exposure (E), a potential mediator (M) and outcome (Y). The underlined variables are those included in the analysis. A classical MITM design would find, assuming perfect power, an association between E_1 and M_2 (due to Y_1 confounding), an association between M_2 and Y_3 (due to Y_2 confounding) and an association between E_1 and Y_3 (due to Y_1 confounding).



CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Supplementary Material IV.1: Detailed simulation methods

In these paragraphs, a variable of the intermediary layer M is called a CpG.

2 similar implementations were used to simulate in one hand causal structures A, B and C, and in the other hand structures D and D.

1. Causal structures A, B and C:

Generation of links between E and M

An effect of the exposures on the methylome was generated according to

$$\text{For each CpG } i \text{ affected by E, } m_i = m_{\text{boot } i} + \sum_{k=1}^{k=173} \beta_{ki} E_k \quad (1)$$

where $m_{\text{boot } i}$ is the vector containing all values of methylation at the CpG i in the matrix M bootstrapped from the real data, m_i is the vector containing all new simulated values for this CpG site after the addition of an effect of the exposome, and E_k is the vector containing all values for the predictor k of the exposome. Regressions coefficients were all set to zeros except for the exposures randomly exposures variables for which we created a causal link. The number of CpGs affected by E, the number of exposures affecting each CpG and the values of the non-zero regressions coefficients were fixed as described in the “definition of each scenario” section.

Generation of the health outcome

The health outcome Y was generated as a function of the exposome and the methylome according to: $Y = \sum_{j=1}^{j=2284} \beta'_j m_j + \sum_{k=1}^{k=173} \beta''_k E_k + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (2)$

where Y is the vector of the generated outcome, m_j is the vector containing all values of the predictor j from the methylome, E_k is the vector containing all values for the predictor k of the exposome, and ε represents the residuals of the regression models, computed from a residual variance σ^2 . The number and the values of non-zero regression coefficients were defined as described in the “definition of each scenario” section below.

The value of σ^2 is computed to ensure that the total variability of Y explained by E and M is equal to the parameter R^2 which varies between scenarios.

Implementation of each causal structure:

In order to easily define different causal situations, we define in the methylome matrix M 4 matrices, with no intersection: $M_0, M_{M \rightarrow Y}, M_{E \rightarrow M \rightarrow Y}, M_{E \rightarrow M}$.

$M_{E \rightarrow M \rightarrow Y}$ contains all the methylome variable which are both affected by E and affecting Y, i.e. for a methylation variable j belonging to $M_{E \rightarrow M \rightarrow Y}$, at least one $\beta_{i=j,k}$ is non-zero in the equation (1) and β'_j is non-zero in equation (2).

$M_{M \rightarrow Y}$ contains all the methylome variables which are not affected by E but affecting Y, i.e. for a methylation variable j belonging to $M_{M \rightarrow Y}$, all $\beta_{i=j,k}$ are zero in the equation (1) and at least β'_j is non-zero in equation (2).

$M_{E \rightarrow M}$ contains all the methylome variables which are affected by E but not affecting Y, i.e. for a methylation variable j belonging to $M_{E \rightarrow M}$, at least one $\beta_{i=j,k}$ is non-zero in the equation (1) and β'_j is zero in equation (2).

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

M_0 contains all the methylome variables which are neither affected by E nor affecting Y , i.e. for a methylation variable j belonging to M_0 , all $\beta_{i=j,k}$ are zero in the equation (1) and $\beta'_{j=i}$ is zero in equation (2).

M can therefore be seen as the concatenation of $M_{E \rightarrow M \rightarrow Y}, M_{M \rightarrow Y}, M_{E \rightarrow M}$ and M_0 , which we will now called our M submatrices.

We also defined 3 submatrices in E with possible intersections: $E_{E \rightarrow M \rightarrow Y}$, $E_{E \rightarrow M}$ and $E_{E \rightarrow Y}$.

$E_{E \rightarrow M \rightarrow Y}$ contains all the exposome variables which affect at least one CpG site belonging to $M_{E \rightarrow M \rightarrow Y}$, i.e. for an exposure variable k belonging to $E_{E \rightarrow M \rightarrow Y}$, there is at least one methylation variable i for which $\beta_{k,i}$ is non-zero in the equation (1) and $\beta'_{j=i}$ is non-zero in equation (2).

$E_{E \rightarrow M}$ contains all the exposome variables which affect at least one CpG site belonging to $M_{E \rightarrow M}$, i.e. for an exposure variable k belonging to $E_{E \rightarrow M}$, there is at least one methylation variable i for which $\beta_{k,i}$ is non-zero in the equation (1) and $\beta'_{j=i}$ is zero in equation (2).

Therefore, $E_{E \rightarrow M} \cup E_{E \rightarrow M \rightarrow Y}$ contains all the exposome variables affecting at least one CpG Site.

$E_{E \rightarrow Y}$ contains all the exposome variables which affect directly M i.e., for an exposure variable k belonging to $E_{E \rightarrow Y}$, β''_{k} is non-zero in equation (2).

To model the influence of exposures via some pathways, we constrained the size of $M_{E \rightarrow M}$ and $M_{E \rightarrow M \rightarrow Y}$ (i.e. the two sets of CpGs affected by E) to be a multiple of the size of respectively $E_{E \rightarrow M}$ and $E_{E \rightarrow M \rightarrow Y}$. Each exposure of $E_{E \rightarrow M}$ has a non-zero effect only on $nM_{E \rightarrow M}/nE_{E \rightarrow M}$ CpGs (i.e. the number of CpGs belonging to $M_{E \rightarrow M}$ divided by the number of exposures belonging to $YE_{E \rightarrow M}$). Similarly, each exposure of $E_{E \rightarrow M \rightarrow Y}$ has a non-zero effect only on $nMEY/nE_{E \rightarrow M \rightarrow Y}$ CpGs (i.e. the number of CpGs belonging to $M_{E \rightarrow M \rightarrow Y}$ divided by the number of exposures belonging to $YE_{E \rightarrow M \rightarrow Y}$). We call this further this constraint the “multiplicity constraint”.

Moreover, to simplify the simulations, we set that in each M submatrices all the effects from the methylome are identical, i.e. for two methylome variables i and j belonging to the same submatrix, $\beta'_{i}=\beta'_{j}$ in (2). We also set that the effect of one exposure of $E_{E \rightarrow M}$ (respectively $E_{E \rightarrow M \rightarrow Y}$) is identical for all the CpGs affected by a non-zero effect of $E_{E \rightarrow M}$ (respectively $E_{E \rightarrow M \rightarrow Y}$), i.e. $\forall k \text{ tq } \beta_{j,k} \neq 0$ and $\beta_{i,k} \neq 0$, $\beta_{j,k}=\beta_{i,k}$ in (1).

Last, we can control the recovering between the different set of predictors from E , i.e. recovering between $E_{E \rightarrow M \rightarrow Y}$, $E_{E \rightarrow M}$ and $E_{E \rightarrow Y}$.

Therefore the 3 causal structure A, B, and C can be defined by:

- Situation A: $M_{M \rightarrow Y}=M_{E \rightarrow M}=E_{E \rightarrow Y}=E_{E \rightarrow M \rightarrow Y}=\emptyset$ and $M_{E \rightarrow M \rightarrow Y} \neq \emptyset$ and $E_{E \rightarrow M \rightarrow Y} \neq \emptyset$.
- Situation B: $M_{M \rightarrow Y} \neq \emptyset$, $E_{E \rightarrow Y} \neq \emptyset$ and $M_{E \rightarrow M}=E_{E \rightarrow M}=E_{E \rightarrow M \rightarrow Y}=M_{E \rightarrow M \rightarrow Y}=\emptyset$.
- Situation C: $M_{M \rightarrow Y}=M_{E \rightarrow M}=\emptyset$ and $M_{E \rightarrow M \rightarrow Y} \neq \emptyset$, $E_{E \rightarrow Y} \neq \emptyset$ and $E_{E \rightarrow M \rightarrow Y} \neq \emptyset$, and $E_{E \rightarrow Y} \cup E_{E \rightarrow M \rightarrow Y}=E_{E \rightarrow Y}=E_{E \rightarrow M \rightarrow Y}$

The values of parameters used can be found in table 1.

2. Causal structures D and E:

Reverse causality links:

A linear causal effect from Y was added to E and M :

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

For each CpG i affected by Y , $m_i = m + \gamma_i Y$ (3)

where m is the vector containing all values of methylation at the CpG i in the matrix M bootstrapped from the real data, m_i is the vector containing all new simulated values for this CpG site after the addition of an effect of the outcome, and Y is the vector of the simulated outcome. Regressions coefficients were all set to zeros except for the CpGs sites selected for which we created a causal link. The number of CpGs affected by Y , and the values of the non-zero regressions coefficients were fixed as described in the “definition of each scenario” section.

Similarly, for each exposure E_k affected by Y , $E_k = E_{kboot} + \gamma'_k Y$ (4)

where E_{kboot} is the vector containing all exposure k values in the matrix M bootstrapped from the real data, E_k is the vector containing all new simulated values for this exposure after the addition of an effect of the outcome, and Y is the vector of the simulated outcome. Regressions coefficients were all set to zeros except for the exposures randomly selected for which we created a causal link. The number of exposures affected by Y , and the values of the non-zero regressions coefficients were fixed as described in the “definition of each scenario” section.

Definition of each scenario

In order to easily define different causal situations, we define in the methylome matrix M 3 matrices, with no intersection: M_0 , $M_{M \rightarrow Y}$ and $M_{Y \rightarrow M}$.

$M_{Y \rightarrow M}$ is the matrix containing all CpGs affected by Y , i.e. γ_i is non-zero in (3).

Similarly, we defined in E a subset $E_{Y \rightarrow E}$, which is a matrix containing all exposures affected by Y , i.e. γ'_k is non-zero in (4).

Thus, causal structures are defined:

- Causal structure D : $E_{Y \rightarrow E} = \emptyset$, $M_{Y \rightarrow M} = \emptyset$.
- Causal structure E: $E_{Y \rightarrow E} \neq \emptyset$, $M_{Y \rightarrow M} \neq \emptyset$

The values of parameters used can be found in Supplementary Table IV.1.

CHAPTER IV: A simulation study of approaches relying on intermediate high-dimension data to decipher causal relationships between the exposome and health

Supplementary Material IV.2: Commented simulation script

Due to its size, Supplementary Material IV.2 is provided in Appendix III.

This script will also be available on github (<https://github.com/SoCadiou>) once the corresponding draft will be published.

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

In this last chapter, we focused on a practical question encountered when we tried to use some machine learning algorithms: the instability of some selected subset. In the simulation study presented in this chapter, we studied the performance and the stability of some algorithms commonly used to relate the exposome with an outcome and took the example of LASSO to show that applying a stabilization step can modify performance.

This work is currently under review in Epidemiology journal:

Cadiou S., Slama R., “Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology”, under review

V. 1. Abstracts

V.1.1. English abstract

Background: Machine-learning algorithms are increasingly used in epidemiology to identify true predictors of a health outcome when many potential predictors are measured. However, these algorithms can provide different outputs when repeatedly applied on the same dataset. Such instability can compromise research reproducibility. We aimed to illustrate that commonly-used algorithms are unstable and, with the example of LASSO, that the stabilization method choice is crucial.

Methods: In a simulation study, we tested the stability and performance of widely-used machine-learning algorithms (LASSO, Elastic-Net and DSA). We then assessed the effectiveness of six methods to stabilize LASSO, and their impact on performance. We assumed that a linear combination of factors drawn from a simulated set of 173 quantitative variables assessed in 1301 subjects influenced to varying extents a continuous health outcome. Model stability, sensitivity and False-Discovery-Proportion (FDP) were assessed.

Results: All tested algorithms were unstable. For LASSO, stabilization methods improved stability without ensuring perfect stability, a finding confirmed by an application to an exposome study. Stabilization methods also affected performance. Specifically, stabilization based on hyperparameter optimization, frequently implemented in epidemiology, increased dramatically the FDP when predictors explained a low share of outcome variability. In contrast, stabilization based on stability selection procedure often decreased the FDP, while sometimes simultaneously lowering sensitivity.

Discussion: Epidemiologists wishing to rely on machine-learning methods for variable selection should care about instability. Stabilizing a model can impact its performance. For LASSO, addressing estimation stability rather than prediction stability should be preferred when one aims to identify true predictors.

V.1.2. French abstract

Contexte : Les algorithmes d'apprentissage automatique sont de plus en plus utilisés en épidémiologie pour identifier les prédicteurs causaux d'un outcome de santé lorsque de nombreux prédicteurs potentiels sont disponibles. Toutefois, ces algorithmes peuvent fournir des résultats différents lorsqu'ils sont appliqués de manière répétée sur le jeu de données. Une telle instabilité compromet la reproductibilité de la recherche. Nous avons voulu illustrer que les algorithmes couramment utilisés sont instables et, avec l'exemple de LASSO, que le choix d'une méthode de stabilisation est crucial.

Méthodes : Nous avons réalisé une étude de simulation pour tester la stabilité et les performances d'algorithmes de sélection de variables largement utilisés (LASSO, Elastic-Net et DSA). Nous avons évalué l'efficacité de six méthodes de stabilisation et leur impact sur les performances de LASSO. Nous avons supposé qu'une combinaison linéaire de facteurs tirés d'un exposome simulé de 173 variables quantitatives évaluées chez 1301 sujets influençait à des degrés divers un outcome de santé continu. La stabilité, la sensibilité et la proportion de faux positifs (FDP) du modèle ont été évaluées.

Résultats : Tous les algorithmes testés étaient effectivement instables. Pour LASSO, les méthodes de stabilisation ont amélioré la stabilité sans assurer une stabilité parfaite, un résultat confirmé par une application à une étude exposome réelle. Elles ont également affecté les performances. En particulier, la stabilisation basée sur l'optimisation des hyperparamètres, fréquemment mise en œuvre en épidémiologie, a augmenté de façon spectaculaire le FDP lorsque la variabilité de l'outcome expliquée par les prédicteurs était faible. En revanche, la stabilisation basée sur la procédure de 'stability selection' a souvent réduit le FDP, tout en diminuant parfois simultanément la sensibilité.

Discussion : Les épidémiologistes qui souhaitent s'appuyer sur des méthodes d'apprentissage automatique pour la sélection des variables doivent se préoccuper de la stabilité des modèles. La stabilisation d'un modèle peut avoir un impact sur ses performances. Pour LASSO, les méthodes traitant de la stabilité de l'estimation plutôt que de la stabilité de la prédiction doivent être préférées lorsque l'objectif est l'identification de prédicteurs causaux.

V. 2. Introduction

Thanks to the development of high throughput sensitive biochemical assays and the wider availability of environmental models, exposome studies now allow considering several hundred or thousand exposures in a given study population. Such exposome studies raise many issues, in terms of exposure assessment, handling of measurement error and missing data and consideration of possible mixture effects (Agier et al., 2020b; Siroux et al., 2016; Slama and Vrijheid, 2015; Vermeulen et al., 2020). They also raise more statistical challenges, encountered in other areas relying on ‘omics data, such as genomic, epigenomic or metabolomic studies. Specifically, as the ratio of the number of potential predictors of a health outcome to the number of observations increases, the efficiency of multiple regression models to identify true predictors decreases (Courvoisier et al., 2011; Fan et al., 2019): for example, the classical maximum-likelihood estimator may be biased when the ratio of the number of variables to the number of individuals is typically of 0.2 or more (Sur and Candès, 2019). More complex machine learning algorithms, such as LASSO (Least Absolute Shrinkage and Selection Operator), which performs variable selection through shrinkage, according to a penalty parameter λ (Tibshirani, 1996), ElasticNet, a penalized regression algorithm relying on both the LASSO and ridge penalties (Zou and Hastie, 2005), or Deletion-Substitution Addition algorithm (DSA) (Sinisi and van der Laan, 2004) may be more adapted for variable selection in this setting of intermediate to high dimension, as underlined by recent simulations (Agier et al., 2016; Tibshirani, 1996; Zou and Hastie, 2005). Although they can be used for purely predictive approaches (consisting in predicting the outcome probability or expected value without identifying its true predictors), they are increasingly used in epidemiology in multi-exposures (or exposome) studies (Agier et al., 2019; Forns et al., 2016; Gängler et al., 2019; Huang et al., 2019; Lenters et al., 2016; Mustieles et al., 2017; Nieuwenhuijsen et al., 2019; Philippat et al., 2019; Vrijheid et al., 2020) and in studies relying on omics data (Benton et al., 2017; Cho et al., 2010; Zhou and Lo, 2018), in order to select predictors whose associations with an outcome are often interpreted as in favor of an underlying causal relationship.

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

A feature limiting the use of these methods relates to their possible lack of *stability*. In machine learning, *stability* is the property that a small perturbation in the input does not change the learned model, and thus the model prediction (Bousquet and Elisseeff, 2002; Poggio et al., 2004). Here, we will focus on the instability corresponding to a variation in the model output in the absence of modification in the observations. An example is when applying a model to a dataset would select covariates A, E and G as associated with the outcome and applying again the same model to exactly the same data would this time select covariates A and B. This type of instability relates to the fact that some machine-learning algorithms have a random component (Elisseeff, 2005), for example if they use bootstrap or cross-validation. LASSO and ElasticNet are unstable when their hyperparameter(s) are determined by a cross-validation approach minimizing the prediction error (Bach, 2008; Lazarevic et al., 2019; Lenters et al., 2016; Meinshausen and Bühlmann, 2010), as done in their default implementation (Friedman et al., 2019) (see Table V.1). DSA has been reported to be unstable in real exposome studies (Agier et al., 2019; Nieuwenhuijsen et al., 2019; Warembourg et al., 2019), but its instability has only been mentioned in one simulation study (Agier et al., 2020b).

Instability limits results generalizability and research reproducibility (Bousquet and Elisseeff, 2002; Lee et al., 2013; Nogueira et al., 2017; Poggio et al., 2004). It might be perceived as a fatal drawback of machine-learning methods and hinder their diffusion among epidemiologists used to the stability of classical regression models and concerned with the possibility of researchers cherry-picking the most “convincing” results if models are unstable.

Instability has little been studied in epidemiology. Lazarevic et al. (2019) expressed concerns about the reproducibility of results due to the instability of ElasticNet (Lazarevic et al., 2019). In machine learning research, different strategies were developed to address instability. Some strategies address the *estimation* stability, which is the stability of the model estimates (selected variables and the associated parameters) (Lim and Yu, 2016), while other strategies address *prediction* stability, i.e. they focus on the stability of the predicted value of the dependent variable, for example the disease risk.

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Both approaches can lead to different results, because the predicted disease risk or outcome expected value can remain quite stable in the presence of changes in the selected variables, e.g. by replacing a variable by another one strongly correlated to it or by adding a variable not associated with the outcome. Methods relying on prediction stability are often based on prediction error minimization by cross-validation. Lazarevic et al. (2019) advised to consider methods relying on estimation stability, a strategy supported by theoretical work in the statistical field (Lim and Yu, 2016). However, to our knowledge, in the cases of Elastic-Net and LASSO, most multi-exposures studies aiming at selecting relevant explanatory variables (supposed to possibly causally influence the outcome) relied on hyperparameters optimization by repeated cross-validation to stabilize results, which corresponds to prediction stability approaches (Forns et al., 2016; Gängler et al., 2019; Huang et al., 2019; Lenters et al., 2016; Mustieles et al., 2017; Philippat et al., 2019).

We aimed to highlight the existence of instability in algorithms commonly-used for variable selection (LASSO, DSA and Elastic-Net) and contrast it with the stability of the traditional linear model. We also aimed to compare several stabilization methods in the case of LASSO; we focused on LASSO as an algorithm with a wide array of proposed stabilization methods. With this example, we point that model stability cannot be considered independently of model performance dimensions such as sensitivity and false discovery proportion (FDP, the proportion of selected variables not genuinely related to the outcome) and thus that the implementation of a stabilization method is crucial and cannot be considered as a free add-on by epidemiologists who intend to use an originally unstable algorithm.

V. 3. Methods

V.3.1. Simulation study of LASSO, DSA and ExWAS under various correlation structures

We first performed a Monte-Carlo simulation to assess the stability and performance of LASSO, DSA, Elastic-NET and ExWAS. ExWAS (Exposome-Wide Associations Study) corresponds to parallel univariate regressions corrected for multiple testing (Nieuwenhuijsen et al., 2019). The implementation of these variable selection algorithms is detailed in Table V.1.

We simulated an exposome of 173 Gaussian quantitative variables among 1301 subjects (as in the Helix exposome project) (Haug et al., 2018; Tamayo-Uria et al., 2019) from three different correlation matrices representing the correlation structure within the exposome: no correlation between the covariates; a realistic correlation structure, computed from Helix project data (Tamayo-Uria et al., 2019), with a median coefficient of correlation between any pair of exposures of 0.12; and an identical correlation of 0.5 between all covariates pairs. We generated an outcome according to a multivariate linear regression model, with the number of true predictors fixed to 10. Scenarios considered three different values of R^2 , the variance explained by the true predictors of the exposome (0.001, 0.1, 0.4), as well as the three different correlation structures.

LASSO was implemented using *default LASSO*, in which the hyperparameter λ is chosen so as to minimize the root mean squared error (RMSE) of prediction derived from 10-fold cross-validation; after this cross-validation step, the model's result are obtained by fitting a single LASSO model with this optimal value of the hyperparameter (Tibshirani, 1996). Similarly, we used the default implementation of Elastic-Net, corresponding to an RMSE-based cross-validation to choose hyperparameters followed by a single model run using these hyperparameters (see Table V.1 for details) (Friedman et al., 2019). A stabilized version of Elastic-Net, using repeated cross-validation, was also implemented (see Supplementary Figure V.1). We quantified stability, sensitivity and FDP of each method (see below).

Table V.1: Implementation details for ExWAS, Elastic-Net and DSA.

Method name	Description of the method	References
ExWAS	Univariate regressions corrected for multiple testing, known as ExWAS (Exposome-Wide Association Study). Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was used to correct multiple testing. Variable selection is performed by selecting all the variable for which the association test -p-value was below the significance threshold of 5%, after correction for multiple testing, as usually done (Agier et al., 2019, 2016; Vrijheid et al., 2020).	(Agier et al., 2016; Nieuwenhuijsen et al., 2019)
ElasticNet	Penalized regression model using a weighted mixture of LASSO (Tibshirani, 1996) and ridge (Hoerl and Kennard, 1970) penalties. The ridge penalty accommodates correlated variables and ensures numerical stability, but does not shrink coefficients exactly to zero, and thus cannot perform variable selection. Penalty is calibrated with a parameter λ , and an other tuning parameter α controlled the mixing proportion of the two penalties. As advised by Zou & Hastie, 2005 (Zou and Hastie, 2005), the two hyperparameters are determined by two-dimensional cross-validation, implemented using the <i>glmnet</i> package (Friedman et al., 2010).	Elastic_Net method (Zou and Hastie, 2005) <i>glmnet</i> package (Friedman et al., 2010) Two-dimensional cross-validation implementation was similar to the one used in a simulation study (Agier et al., 2016).
DSA	DSA (Deletion Substitution Addition) algorithm is an iterative linear regression model search algorithm (Sinisi and Van Der Laan, 2004) following three constraints: maximum order of interaction amongst predictors, the maximum power for a given predictor, and the maximum model size. At each iteration, the following three steps are allowed: a) removing a term, b) replacing one term with another, and c) adding a term to the current model. The search for the best model starts with the intercept model and identifies an optimal model for each model size. The final model is selected by minimizing the value of the RMSE using 5-fold cross-validated data. We allowed no polynomial or interaction terms, and made no restriction on the number of predictors.	(Sinisi and van der Laan, 2004)

V.3.2. Simulation study of stabilizations methods for LASSO in a realistic exposome setting

To compare some stabilizations methods for LASSO, we performed a second Monte-Carlo simulation: we expanded the number of simulation scenarios, considering between 1 and 25 true predictors and letting R^2 vary from 0.0001 to 0.8. In this second simulation, the exposome was realistically simulated by sampling with replacement quantitative variables of Helix exposome dataset (Haug et al., 2018; Tamayo-Uria et al., 2019), which had been beforehand normalized, standardized and bounded (i.e., a value greater than 3 in absolute value was replaced by a value lower than 3 in absolute value randomly drawn in the distribution). In addition to default LASSO (Tibshirani, 1996), we implemented six variants with different stabilization methods: LASSO-1 SE, which is similar to default LASSO, but uses the largest λ located within one RMSE of the λ value which minimizes the RMSE; this strategy is known to be more parsimonious, as increasing λ values more strongly penalize the model and tend to select fewer variables (Krstajic et al., 2014) ; two methods (CV_1 and CV_2) optimizing the hyperparameter λ by repeating the cross-validation procedure 100 times and averaging the results using two different procedures (Table V.2); two implementations of the *stability selection* proposed by Meinshausen and Bühlman ($Meinshausen_1$ and $Meinshausen_2$), which repeatedly ran the algorithm on subsamples of the observations while varying λ over large ranges and finally provided as outputs the covariates most frequently selected across all these runs (Meinshausen and Bühlmann, 2010). A conceptual difference is that, in contrast to CV_1 and CV_2 , these two methods do not rely on cross-validation and do not fix an optimal hyperparameter but follow a logic of *model averaging* (Claeskens and Hjort, 2008). We finally tested an approach (*Mix*) that we developed as a mixture of the principles of the two previous stabilization procedures (Table V.2): empirical selection probability was derived for each variable from repeated runs on random subsamples as in stability selection, but using the optimal λ parameter determined by cross-validation instead of varying λ on a large range of values. We quantified stability, sensitivity and FDP of each method (see below).

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Table V.2: Details of the implemented LASSO methods.

LASSO method	Description of the method and of the stabilization process	References
Default LASSO	A penalized regression model relying on a generalized linear framework (Tibshirani, 1996). The LASSO penalty promotes sparsity and performs variable selection through shrinkage: the lowest regression coefficients, corresponding to the least informative predictors, are attributed a zero value, according to a penalty parameter λ . As advised by Zou et al. (Tibshirani, 1996) and implemented in <i>glmnet</i> package (Friedman et al., 2019), λ was determined by minimizing the prediction root mean squared error (RMSE) using 10-fold cross-validation (i.e., the data were partitioned into 10 subsets; for each subset, models were trained on the other 9 partitions and fitted on the left-out subset, over which the RMSE was estimated). λ sequences tested in the cross-validation process were sequences of 100 values deterministically determined from the data (Friedman et al., 2019, 2010).	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019)
LASSO_1SE	Similar to default LASSO, with the difference that the penalty parameter λ chosen after 10-fold cross-validation was the <i>largest</i> among the λ values giving an error within 1 standard error of the minimum RMSE, (Friedman et al., 2019) instead of the value minimizing the RMSE.	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019)
CV₁	LASSO with 10-fold cross-validation was repeated 100 times on the same dataset. Penalty parameter minimizing the RMSE averaged across the 100 runs was used to fit the final LASSO model. The principle is similar to the one of bootstrap averaging (“bagging”) (Breiman, 2004), but considers always the same dataset with a different seed for cross-validation instead of bootstrapped samples. A similar stabilization method was used with ElasticNet (Huang et al., 2019; Linters et al., 2016; Philippat et al., 2019).	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019)
CV₂	LASSO with 10-fold cross-validation was repeated 100 times on the same dataset. RMSE curves as a function of λ were averaged over the 100 runs. The averaged curve allowed to determine the optimal λ optimizing the RMSE, which was used to fit the final LASSO model. The principle is similar to bootstrap averaging (“bagging”) (Breiman, 2004), but considers always the same dataset with a different seed for cross-validation, instead of bootstrapped samples.	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019)
Meinshausen₁	Implementation of the stability selection on LASSO (Meinshausen and Bühlmann, 2010): LASSO was run on 100 subsamples of half the size of the initial dataset on a range of 100 different values of the penalty parameter. A probability to be selected was derived empirically for each variable. Variables having an empirical probability greater than a selection threshold T ($T=0.85$) were retained in the final model. For all subsamples, the range of λ values used was the one deterministically computed by <i>glmnet</i> package on the complete dataset.	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019) Stability selection (Meinshausen and Bühlmann, 2010)
Meinshausen₂	Alternative implementation of the stability selection (Meinshausen and Bühlmann, 2010). Similar to Meinshausen ₁ above, with two differences: $T=0.95$ and the range of λ used was different for each subsample and deterministically computed by <i>glmnet</i> package on each subsample.	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019) Stability selection (Meinshausen and Bühlmann, 2010)
Mix	Implementation of the stability selection (Meinshausen and Bühlmann, 2010) on the cross-validated LASSO. In Meinshausen et al. (Meinshausen and Bühlmann, 2010), empirical probabilities of selection for each covariate were derived from results of the algorithm run on a range of penalty parameters on different subsamples. Here, we derived empirical probabilities from runs on a range of subsamples but only from the model fitted with the optimal	LASSO method (Tibshirani, 1996) <i>glmnet</i> package (Friedman et al., 2019) Stability selection (Meinshausen and Bühlmann, 2010)

penalty parameter obtained by cross-validation. More precisely, LASSO with 10-fold cross-validation was run on 100 random subsamples each having half of the observations of the initial dataset. Empirical probabilities to be selected in the model optimizing RMSE in a subsample were then derived for each variable. Variables having an empirical probability of selection greater than $T = 0.5$ were retained in the final model. This original implementation is in principle similar to stability selection (Meinshausen and Bühlmann, 2010), but considers always the same dataset with a different seed for cross-validation instead of subsamples.

The code is provided in Supplementary Material V.1.

V.3.3. Indicators of stability and performance

In both simulations, for each scenario, we generated 30 datasets (Efron and Tibshirani, 1993) upon which each performance indicator was assessed. In order to assess stability, defined as the presence of variations in the model output in the absence of modification in the observations, each method was run 15 times on each of the 30 datasets generated for each scenario. The stability of the set of variables selected as predictors in each of the 15 runs of each dataset was quantified using averaged Sorensen index. Sorensen index is one of the most commonly-used measure of similarity (Magurran, 2004). For two runs based on datasets with similar covariates, Sorensen index is defined as twice the number of selected covariates common to both runs divided by the sum of the number of covariates selected for each run (Boulesteix and Slawski, 2009). The index was averaged over all pairs of runs done with a given dataset. Averaged Sorensen index has a value of 0 when there is no intersection between the sets of selected variables in all runs based on the same dataset (total instability) and of 1 when the selected variables are the same in all runs based on the same dataset, or when no covariate is selected in any of the runs (Boulesteix and Slawski, 2009). As an alternative measure of stability, we also counted the number of variables selected in at least 20% and 60% of the runs on a same dataset.

We assessed two dimensions of model's performance: FDP (the proportion of false-positive among the predictors selected by the algorithms) and sensitivity (the proportion of true predictors selected by the algorithm among the true predictors). Averaged FDP and sensitivity were computed

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

for each dataset by averaging respectively FDP and sensitivity across repeated runs, allowing to estimate FDP and sensitivity for each scenario.

All simulations were performed with R software using *glmnet* (Friedman et al., 2019, 2010) and DSA (Sinisi and van der Laan, 2004) packages.

V.3.4. Application: using LASSO to relate the exposome to child body mass index

We illustrated instability in the context of an exposome study. Previous studies (Cadiou et al., 2020; Vrijheid et al., 2020) tried to identify components of the prenatal and postnatal exposomes associated with child body mass index (BMI, the mass in kilograms divided by the squared height in meters). Within Helix project, this has been done using ExWAS and DSA (Cadiou et al., 2020; Vrijheid et al., 2014). We repeated this analysis by using default LASSO and the different stabilized LASSO presented in Table V.2. BMI, the outcome considered, was measured between 6 and 10 year of life in 1301 children from the 6 European cohorts (Chatzi et al., 2017; Grazuleviciene et al., 2015; Guxens et al., 2012; Heude et al., 2016; Magnus et al., 2016; Wright et al., 2013) involved in the Helix project. 216 prenatal (measured during mother pregnancy) and postnatal exposures (measured at the time of the child clinical examination) were considered, measured by biomarkers in urine or blood or by environmental models. Exposures belonged to 15 families: metals, organochlorines, organophosphate pesticides, polybrominated diphenyl ethers (PBDE), perfluorinated alkylated substances (PFAS), phenolic compounds, phthalates, built environment exposures, indoor air exposures, lifestyle factors, meteorological data, natural spaces quantification, noise, traffic, socio-economic capital and concentrations of disinfection by-products in drinking water. Among them, we retained the 173 variables corresponding to quantitative exposures. Details of exposome and covariates assessment (Tamayo et al., 2018), as well as relevant adjustment factors selection (Cadiou et al., 2020) have been published elsewhere. In statistical analysis, an age-and-sex-standardized z-score (de Onis et al., 2007), named hereafter zBMI, was used to take into account the age-related shift in BMI in childhood. Adjustment factors were taken into account by

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

preliminarily computing residuals of zBMI in a multivariate model considering all relevant covariates. We applied default LASSO as well as all studied stabilized LASSO algorithms to relate the full (i.e. prenatal and postnatal) exposome to zBMI. We then applied the methods to relate only the prenatal exposome to zBMI, as we expected the magnitude of link between prenatal exposome and zBMI to be lower than that between the postnatal exposome and zBMI (Vrijheid et al., 2020). In both cases, we computed averaged Sorenson index by repeating 15 times each method and counted the average number of variables selected. As default LASSO and stabilized LASSO are variable selection algorithms that do not estimate the model's coefficients, we ran in a second step a multivariate linear model including for each run all selected exposures as well as the relevant covariates to assess the direction of associations.

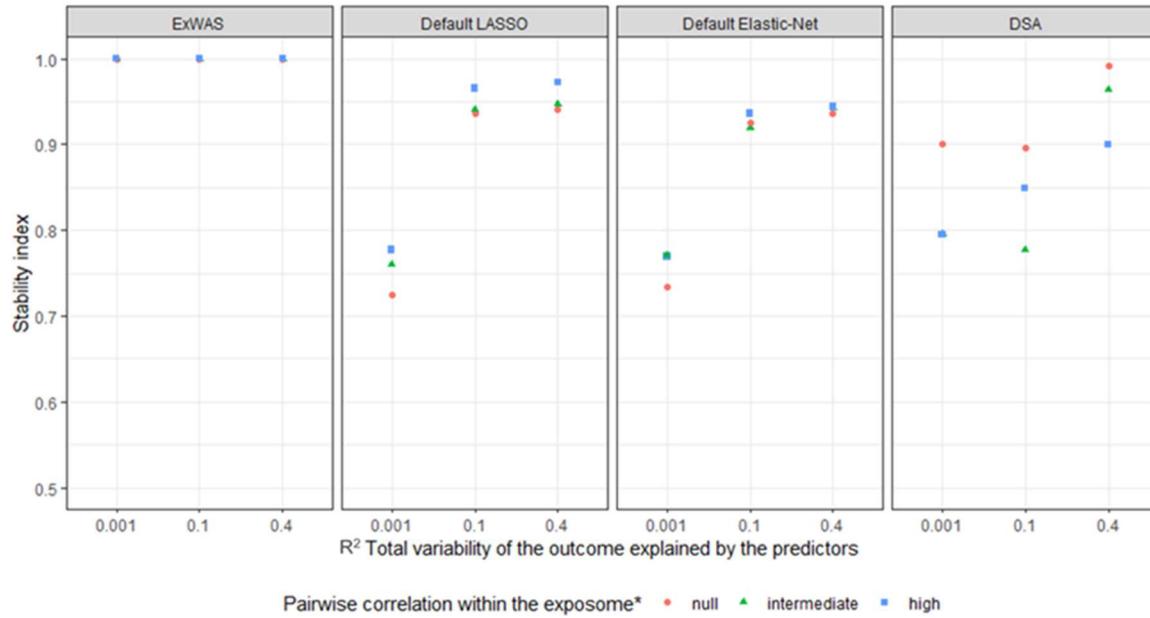
V. 4. Results

V.4.1. Stability and performances of the models' default implementations

The simulation highlighted ExWAS as the only stable model (Sorensen index, 1; Figure V.1). Elastic-Net, DSA and LASSO default implementations were unstable (Figure V.1, Supplementary Figure V.1): default LASSO had an average Sorensen index between 0.77 and 0.95 when the correlation among predictors was the realistic structure from Helix exposome project, with a mean correlation of 0.12 (Figure V.1). For all three algorithms, instability was stronger when the variability explained was 0.001 than when it was 0.1 or 0.4. Elastic-Net was slightly less stable than default LASSO, in particular when the correlation was high. DSA was more stable than default LASSO when the variability of the outcome explained by the true predictors (R^2) was below 0.1, but less stable than default LASSO when R^2 was above 0.1. The correlation among potential predictors influenced stability (Figure V.1): for LASSO and Elastic-Net, a higher correlation was associated with higher stability in most cases. For DSA, Sorensen stability index was highest in the absence of correlation among covariates.

Coming to sensitivity and FDP (Supplementary Figure V.1A and B), DSA was the method with the lowest FDP, with FDP levels always lower than 70%; FDP was lower than 35% when correlation was not high (i.e., not 0.5); DSA did not select any predictor when R^2 was 0.001 but, when R^2 was higher, DSA had a non-null sensitivity (higher than 15%). Elastic-Net and LASSO showed considerably higher FDP than DSA (FDP higher than 40% in all scenarios for LASSO and higher than 68% for Elastic-Net). Elastic-Net showed higher FDP and sensitivity than LASSO, as theoretically expected (Zou and Hastie, 2005). Elastic-Net had a FDP of more than 90% when R^2 was 0.001. ExWAS showed high sensitivity (higher than 60% when R^2 was higher than 0.001) but was the algorithm with the highest FDP when the correlation was non-zero, reaching a FDP of 95% when the correlation was high (Supplementary Figure V.1A).

Figure V.1: Stability index (mean Sorensen index) of ExWAS, default LASSO, Elastic-Net and DSA for 3 different structures of pairwise correlations between the predictors. Simulations assumed the existence of 10 true predictors in a set of 173 tested predictors. Stability is reported as a function of R^2 , the share of the outcome variability explained by the true predictors.



*: *null* corresponds to an absence of correlation between true predictors; *intermediate* corresponds to the realistic correlation observed within real data from Helix project exposome (Vrijheid et al., 2014); *high* corresponds to an identical pairwise correlation of 0.5 among all pairs of covariates.

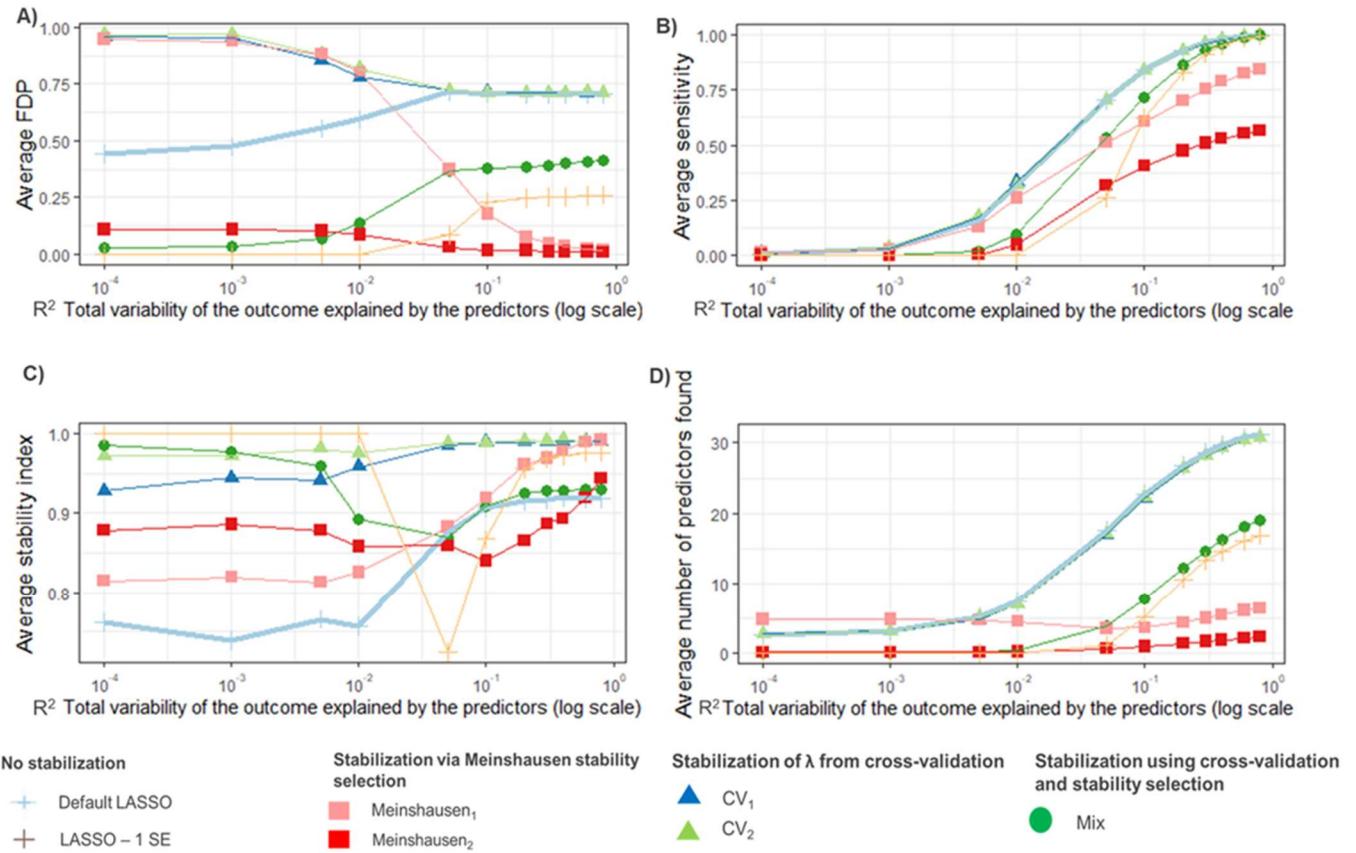
V.4.2. Effectiveness of stabilization methods

The application of stabilization methods generally increased the stability of LASSO, compared to default LASSO, but without allowing perfect stability (Figure V.2C). Comparing stabilization methods, mean Sorensen index was lowest for Meinshausen₂ and highest for CV₂ (Supplementary Table V.1, Supplementary Figure V.3C).

For methods using cross-validation (CV₁ and CV₂) and for Meinshausen₁ method, stability increased with the outcome variability explained by the predictors (Figure V.2C, Supplementary Figure V.2C). For LASSO-1SE and Mix, stability index followed a U-shaped curve: when R^2 was very small, both methods tended not to select any variable, leading to very good stability, after which, as methods began to select some predictors, stability decreased when R^2 increased; stability increased again for higher R^2 .

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Figure V.2: Performance and stability of various stabilization methods of LASSO. Values are averages of various scenarios with 1 to 25 true predictors of the health outcome. Performance is reported as a function of the total variability explained by the predictors (log scale). **A.** False discovery proportion. **B.** Sensitivity. **C.** Sorensen stability index. **D.** Number of hits (covariates selected by the model).



FDP: False Detection Proportion. See Table V.2 for explanations regarding the compared methods.

Regarding Elastic-Net, the comparison between the performance of default Elastic-Net and of an Elastic-Net with repeated cross-validation showed that this stabilization method effectively allowed to stabilize the model with a pattern similar to that observed for LASSO-CV₁: repeating the cross-validation-process allowed to increase stability (Supplementary Figure V.1).

V.4.3. Relation between stabilization and model performance

Stabilization generally influenced LASSO model's performance. For R^2 lower than 0.1, stability obtained from repeated cross-validation came at a cost of a strongly increased FDP (mean FDP, 0.83 and 0.85 for CV₁ and CV₂, respectively, versus 0.58 for default LASSO; Supplementary Table V.1). When R^2 was greater than 0.1, CV₁ and CV₂ provided a clear stability gain with very small

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

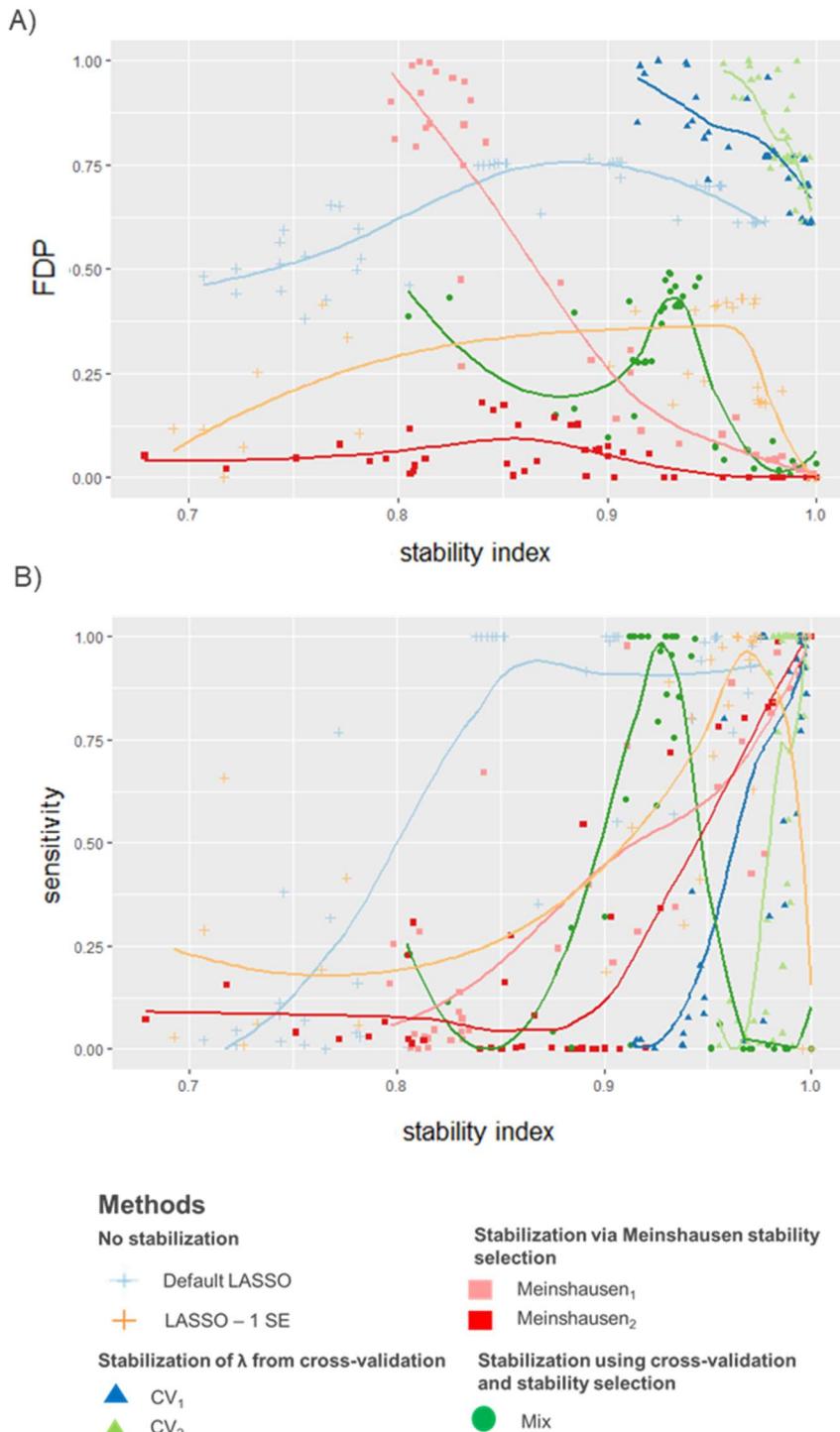
impacts on sensitivity and FDP compared to default LASSO (Figure V.2A and B). However, in this R^2 range, all three methods had a very high FDP (mean of 0.70-0.71).

Regarding stabilization methods based on stability selection principle, Meinshausen₁ and Meinshausen₂ showed excellent performances when R^2 was greater than 0.1 (FDP of 0.04 and 0.01, respectively, sensitivity of 0.78 and 0.53, respectively) but mixed performances when R^2 was lower. Specifically, Meinshausen₂ showed lower FDP than Meinshausen₁ and default LASSO, at a cost of a lower sensitivity (Figure V.2A and B). The *Mix* method showed the best performances when R^2 was lower than 0.1 (average FDP, 0.17; average sensitivity, 0.23; average Sorensen index, 0.93, see Supplementary Table V.1), followed by LASSO-1SE, which had slightly lower sensitivity and FDP than the *Mix* method. The very good stability and FDP of LASSO-1SE when R^2 was below 0.01 were linked to a lack of sensitivity; when sensitivity was non-null and R^2 was lower than 0.1, LASSO-1SE showed limited stability.

An interesting pattern of the two stability selection methods Meinshausen₁ and Meinshausen₂ was that when the method was stable (Sorensen index above 0.95), performance was excellent both in terms of sensitivity and FDP (Figure V.3). For the *Mix* method and LASSO-1SE, a high stability always corresponded to a low FDP and either to a null (no exposures selected) or a very high sensitivity. Such a relationship between stability and FDP was not observed for default LASSO and for methods based on RMSE minimization by repeated cross-validation, for which high levels of the stability index were observed for low values of sensitivity and high FDP values.

Regarding Elastic-Net, stabilization by repeated cross-validation (a logic similar to LASSO-CV₁) showed again similarities with what was observed for LASSO: it modified the average number of predictors selected compared to the default Elastic-Net when R^2 was low (Supplementary Figure V.1D). When R^2 was lower than 0.1, FDP, which was already extremely high (higher than 90%) for default Elastic-Net, was also slightly increased (reaching 98% when correlation was high, Supplementary Figure V.1A) after stabilization by repeated cross-validation.

Figure V.3: Variation of performance according to stability. **A.** False discovery proportion as a function of model stability. **B.** Sensitivity as a function of model stability (mean Sorensen index). Performance values were smoothed using LOESS method. Values are averages over 15 model runs.



V.4.4. Application: using LASSO to relate exposome to child body mass index in Helix data

When relying on Helix real data, default LASSO identified in average 58.3 exposures (out of 173 candidate covariates) as related to zBMI and was unstable (Table V.3, Supplementary Table V.3 and Supplementary Figure V.4). CV₁ and CV₂ displayed very similar results, with a slightly lower number of variables selected than by default LASSO and a higher stability (CV₂ even ensuring perfect stability). This similar behavior between default LASSO, CV₁ and CV₂, as well as the relative stability of default LASSO make it plausible, from the simulation results (Figure V.1 and Supplementary Figure V.1), that the share of outcome variability explained by the true predictors be larger than 0.08. In this situation, one would expect (Figure V.1 and Supplementary Figure V.1) a sensitivity between 0.7 and 1 for the three methods, but also a large FDP (higher than 0.75), which may explain the high number of selected variables. The three methods based on stability selection selected much fewer variables: the Mix method selected on average 21.5 variables. The lower number of variables selected by the Mix method is consistent with what is expected from our simulation (Figure V.1): for this range of scenarios (R^2 above 0.08) the Mix method has a sensitivity similar to default LASSO but a lower FDP. Meinshausen₂ selected on average less than 1 predictor and was far less stable. Last, Meinshausen₁ was perfectly stable and selected on average 5 predictors. For the expected range of scenarios, Meinshausen₂ is expected to show an almost null FDP (a situation generally observed in our simulation in all scenarios in which Meinshausen₂ was almost perfectly stable, Figure V.2) and a non-null sensitivity. We could thus hypothesize that all the 5 exposures selected by Meinshausen₂ are truly associated with zBMI. This hypothesis is coherent with toxicological and epidemiological literature, as discussed in previous studies on the same data (Cadiou et al., 2020; Vrijheid et al., 2020). Indeed, the highlighted positive association of blood post-natal copper level with higher BMI (see Supplementary Table V.2) is a plausible association, as copper toxicity and ability to induce oxidative stress is well-known in human (Brewer, 2010; Uriu-Adams and Keen, 2005) and from animal models (Galhardi et al., 2004; Pereira

CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

et al., 2016). The link between zBMI and the four other variables selected by Meinshausen₂ (postnatal blood levels of DDE, PBDE, HCB and PCB170, all negatively associated with zBMI, see Supplementary Table V.2) may also correspond to true structural associations. As discussed previously, they may be caused by reverse causality: these lipophilic compounds are stored in fat and therefore a low blood circulating level can be caused by a high BMI causing higher fat storage, even if these compounds are generally expected to have harmful effects (Cadiou et al., 2020), which would be identified in prospective studies possibly relying on other biological matrices to assess PCBs. Additional results for the analysis on the pregnancy exposome are detailed in Supplementary Material V.1, Supplementary Table V.3 and Supplementary Figure V.4.

Table V.3: Results of the application of default LASSO and various LASSO stabilization methods to relate an exposome of 173 prenatal and postnatal quantitative exposures to zBMI in 1301 mother-child pairs of the Helix cohorts.

Stabilization method	Sorense n index	Number of selected exposures*	Computation time (in seconds)*	Number of exposures selected at least once
None (default LASSO)	0.957	58.3	1.16	68
LASSO_1SE	0.752	12.3	1.32	20
CV ₁	0.998	58.1	194.34	59
CV ₂	1.000	58.0	114.64	58
Meinshausen ₁	1.000	5.0	14.32	5
Meinshausen ₂	0.492	1.0	14.58	2
Mix	0.910	21.5	172.21	28

* computed after repetition of each method 15 times.

V. 5. Discussion

Our simulation study brings practical insights on issues related to the stability of some algorithms sometimes referred to as belonging to the field of machine learning used in epidemiology. First, we confirmed that the default implementation of LASSO, ElasticNet and DSA were not stable (Agier et al., 2019; Meinshausen and Bühlmann, 2010; Nieuwenhuijsen et al., 2019; Warembourg et al., 2019). On the contrary, EXWAS, which relies on parallel simple (i.e., considering each covariate one at a time) regression models estimated by the least squares approach, was as expected stable, but had a large proportion of false positive signals as soon as there was some correlation between the potential predictors, as previously reported (Agier et al., 2016). With the data structures that we explored, LASSO and DSA, in spite of their instability, showed relatively low FDP, which makes them attractive for epidemiologists aiming to identify true predictors of an outcome. In contrast, Elastic-Net had a higher FDP, especially when the influence of the explanatory variable on the outcome was low, making this algorithm not adapted to this true predictors selection problem – in many situations, FDP was above 50%, meaning that less than half of the selected variables were true predictors. Second, although all stabilization methods did improve LASSO stability, none of them, including stabilizing hyperparameter based on repeated cross-validation as usually done (Huang et al., 2019; Lenters et al., 2016; Philippat et al., 2019), allowed ensuring perfect stability of the set of selected variables. This conclusion was illustrated with our application based on real Helix data, in which some stabilization methods even showed less stable results than for the default LASSO. Third, stabilizing LASSO affected its specificity and sensitivity, showing that the choice of a stabilization method bears strong consequences on performance. To our knowledge, this important feature has not been clearly described in the literature. Thus, this example of LASSO suggests that selecting an algorithm on the basis of its expected performance should be done considering simultaneously its stabilization method, as different stabilization methods can be expected to differently alter performances.

V.5.1. Strengths and limitations

Some limitations need to be acknowledged. We considered a narrow definition of stability, which is more generally understood as the robustness to small perturbations in the observations or other input parameters (Bousquet and Elisseeff, 2002; Poggio et al., 2004). Here we only considered instability due to the random process in algorithms, thus without changes in the dataset. We considered this to be the form of instability most worrying and least familiar to the epidemiologists, who are used to seeing results change when data change, even slightly, but not when re-running a model on the same data. We focused on sensitivity and FDP as indicators of models' performances; bias in the effect estimates is another relevant indicator, but we considered the issue of bias in a parameter affecting an outcome to be secondary in the context of many models showing high FDP or low sensitivity, that is, being unable to identify the true predictors. Although we considered a large number of scenarios with ample variations in model predictive ability and number of true predictors, we focused on a continuous outcome (whereas methods that we considered could also be used for example with binary outcomes) (Lenters et al., 2018) and did not consider possible non-linear relations or interactions between exposures. We chose an “intermediate” dimension corresponding to current exposome studies, which were our motivating example. Last, we did not test all possible stabilization methods suggested for LASSO: other existing approaches, possibly relevant for epidemiologists, also rely on cross-validation but consider other metrics in addition to error prediction in the cross-validation process, in a logic similar to LASSO-1SE (Lim and Yu, 2016; Roberts and Nowak, 2014), or adapt the LASSO algorithm itself by adding another penalty term (Ternès et al., 2016; Zou, 2006). The methods we chose are the most commonly-used for variable selection in epidemiology and give an insight on the two main categories of stabilization methods: methods based on the optimization and the stabilization of the hyperparameter and methods relying on a logic of *model averaging*. For DSA, there is to our knowledge no stabilization method validated in the literature, which would be an interesting development given the relatively good performances of the non-stabilized version of model.

V.5.2. Stability selection, a relevant approach when selecting true predictors is the aim

Regarding the modification of the performance coming with the stabilization process, our results offer new insights as to which stabilization methods are the most adapted. Stability selection methods, which pick up the variables most often selected over a large number of model runs with different values of the hyperparameter, can be seen as a model averaging approach, while default LASSO relies on repeated cross-validation to define the hyperparameter optimal for prediction, followed by a single model run with the selected “optimal” hyperparameter value. Stability selection methods often provided increased performance compared to the default LASSO (and to the two LASSO stabilized with repeated cross-validations), with a considerably lowered FDP; this result was consistent with a study on survival models (Khan et al., 2016). In particular, when the variability explained by the true predictor was higher than 0.2, the Mix method was able to show similar sensitivity than default LASSO with a considerably lower FDP (and yet a non-null sensitivity); Meinshausen₂ stabilization method provided extremely low FDP (lower than 5%, compared to about 70% for default LASSO), with a sensitivity above 50%. When R^2 was low, both methods also provided strong improvement for FDP almost without loss in sensitivity. Overall, on a realistic range of low R^2 values, the Mix method that we developed offered the best compromise between sensitivity and FDP. Similarly, LASSO-1SE allowed to improve performance compared to default LASSO. In our real data example, stability-selection based stabilization methods also seemed to perform better than default LASSO. Thus, some stabilized versions of LASSO also have an added value in terms of improvement (decrease) of the false detection rate. This pattern can be understood by recalling that machine learning typically focuses on prediction accuracy, the ability to correctly predict (disease) risk given a subject’s characteristics, which is what cross-validation-based approaches rely on. In contrast, epidemiologists wishing to identify causal predictors of health are rather interested in *feature* selection (Hernán et al., 2019), the ability to select causal predictors of the outcome. However, *prediction* performance and *selection* performance are not equivalent, in particular in a high-dimension setting (Hernán et al., 2019; Leng et al., 2006). In high

dimension, a stable and accurate prediction (e.g., of disease risk) can be obtained by different sets of predictors that may contain false positives (for example, variables correlated with a true predictor, which are to some extent “exchangeable” from a prediction perspective). A model switching between these different predictors in successive runs or adding unnecessary variables with little influence on the predicted risk may have a good predictive power while being unstable and of limited value when it comes to identifying true (causal) predictors of the outcome. Thus, models aiming at optimizing predictive ability, like default LASSO, are not always the most effective in terms of feature selection: (Hernán et al., 2019; Zou, 2006) for default LASSO, hyperparameter optimization of RMSE with cross-validation (which focuses on the accuracy of risk prediction), even if it allows high predictive performance, leads to selecting models *including* the true model rather than selecting the true model itself (Leng et al., 2006; Zou, 2006). These models also have tendency to select predictors even when there is no signal in the data (Belloni et al., 2011), i.e. in our case when no covariate is associated with the outcome. In our realistic exposome settings, we additionally showed that stabilization by repeating the cross-validation process (LASSO-CV₁ and LASSO CV₂) suffered from strongly inflated FDP even compared to default LASSO. This is of practical importance as, in environmental epidemiology, so far, most publications using LASSO or Elastic-Net relied on repeated cross-validation for stabilization. Overall, the stabilization methods which take into account other criteria than the prediction accuracy are likely to be appropriate to the “true predictor selection” problem faced by epidemiologists. Moreover, a relationship between FDP and stability was observed for these methods: for stability selection methods (including the Mix method) as well as for LASSO-1SE, when stability was good, FDP was low. This relation between FDP and stability (Figure V.3B) makes it tempting to consider stability as an indicator of a low FDP; if general, this feature is interesting since, while FDP cannot be estimated on one’s real data, stability can. For example, it allowed us to choose when to trust the results provided by Meinshausen₂ method in our application to real data.

V.5.3. Importance of the calibration of model averaging approaches

These remarks are in favor of the use of stability selection-type methods, which follow a logic of model averaging, rather than methods stabilizing the hyperparameters, when the aim is variable selection. The rationale behind Meinshausen stability selection, i.e. picking the predictors most often selected among various runs in a logic of model averaging, involves the choice of a threshold (the proportion of runs containing a variable for this variable to be selected by the final model). In practice, the performance of Meinshausen LASSO strongly depended on the implementation chosen: in particular, Meinshausen₁ showed extremely high FDP when the variability explained by the true predictors was low, while the more stringent threshold chosen for Meinshausen₂ provided better performance. The strong impact of the choice of the threshold advocates in favor of performing simulations taking into account the stabilization step to fine-tune it. More generally, when running several times an algorithm in order to select the variables selected in a proportion T of the runs, the choice of the threshold T may have major consequences on results (see Supplementary Figure V.3).

V.5.4. Practical consequences and conclusion

The practical consequences of this work can be summarized in five points: 1) Many commonly-used algorithms used for high- or intermediary-dimension data are unstable, a finding previously reported that we illustrated in a realistic exposome setting. 2) As we illustrated with LASSO, not all stabilization methods provide effective stabilization; epidemiologists should therefore assess the results' stability after having used a stabilization method, e.g. by simply re-running the model 10 times. 3) Adding a stabilization step to an existing variable-selection algorithm is likely to change its performance, and not all stabilization methods allow to simultaneously increase stability and model performance. A practical consequence is that if one had chosen a method based on its expected performance according to a simulation study conducted ignoring stability, this expected performance is likely to change if a stabilization step is added. 4) In particular, for LASSO, stabilization methods based on the averaging of optimal hyperparameters obtained by cross-

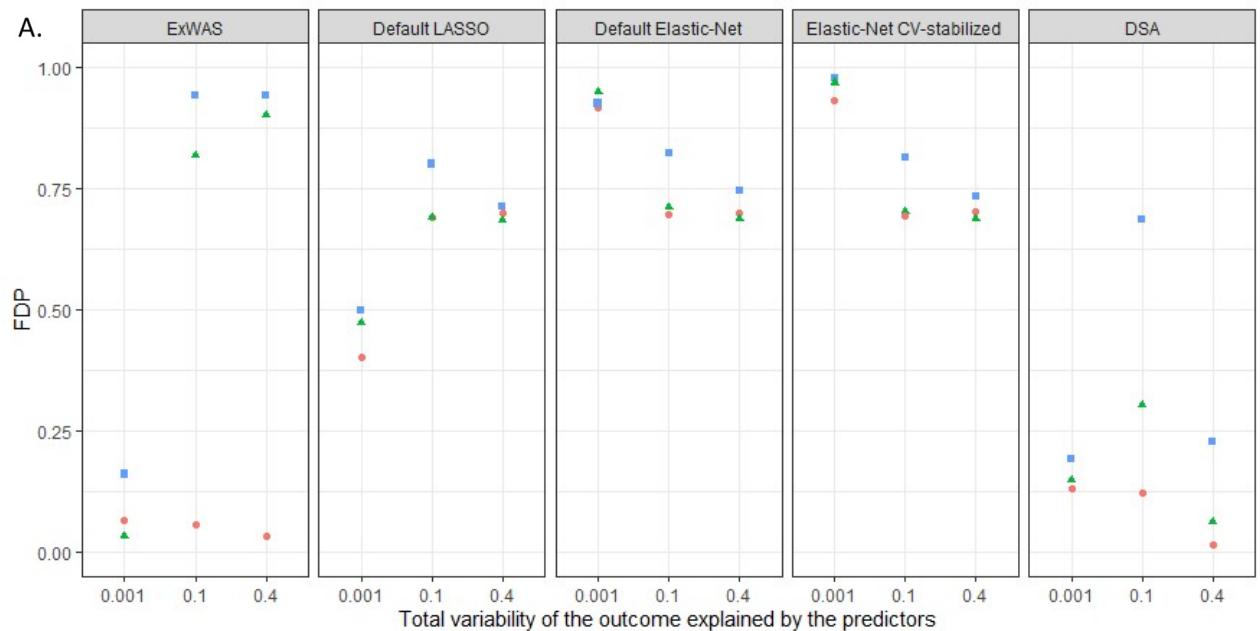
CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

validation (CV_1 and CV_2) dramatically increased the false-discovery rate when the variability of the outcome explained by the predictors was low, which may correspond to many epidemiological studies. With these common stabilization methods, scientists are left with the poor alternative between a stable result likely to include many false positives, and results with a lower false positive rate but selecting different variables in successive runs. 5) When searching for true predictors of an outcome, implementation of stability selection (Meinshausen and Bühlmann, 2010) or of similar methods that do not use the prediction accuracy as the only criterion may be more appropriate than hyperparameters optimization.

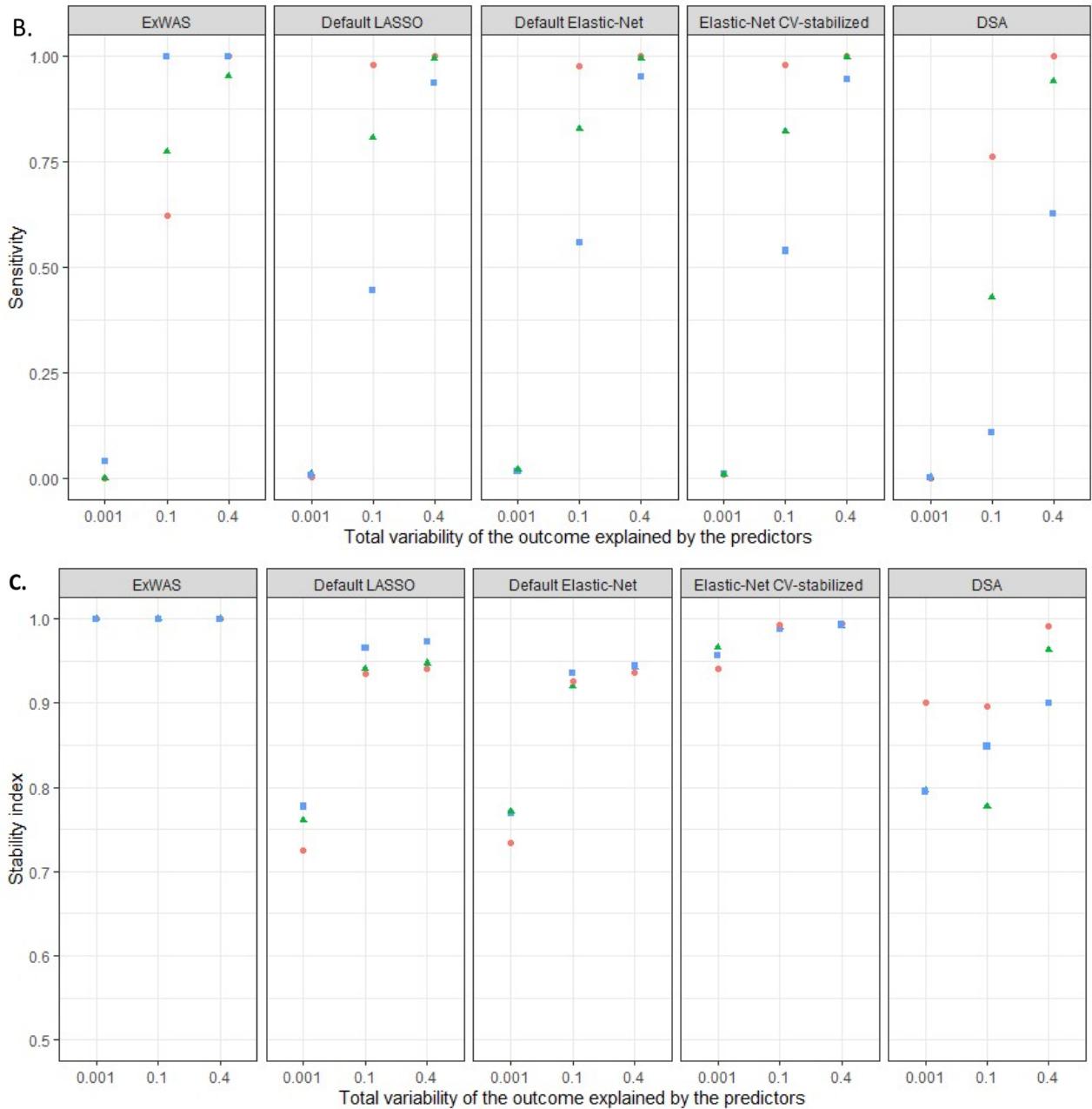
In conclusion, stabilization methods are worth applying and may make some complex machine learning algorithms more attractive to epidemiologists, but should be seen as something that inherently modifies the model considered and, in particular, its performance, rather than a small add-on that comes for free.

V. 7. Supplementary Materials

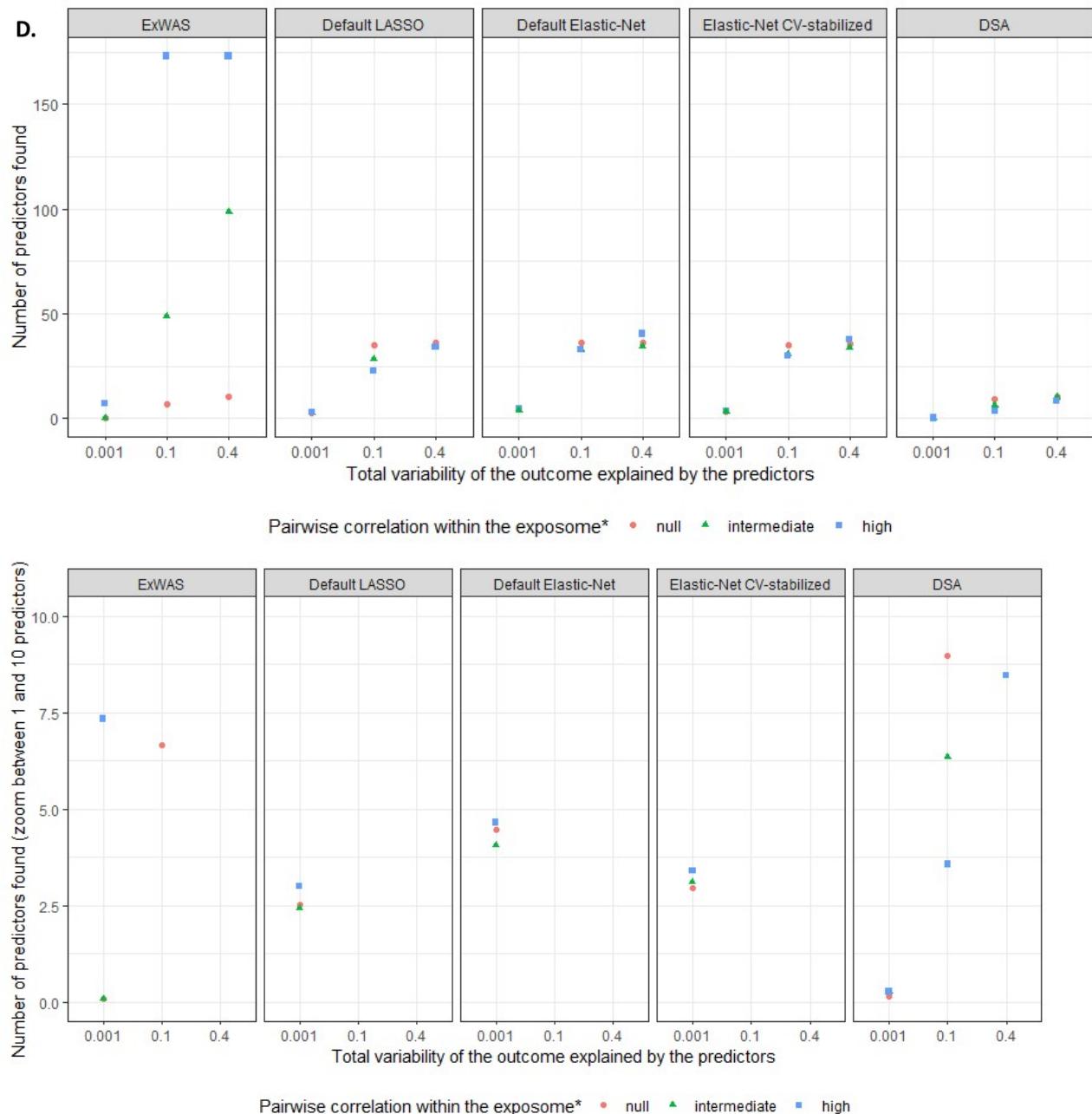
Supplementary Figure V.1: Performance and stability of some common algorithms as a function of the total outcome variability explained by the predictors in situations in which 10 predictors explained the outcome. Colors show the level of correlation within the simulated exposome. **A.** Average false discovery proportion. **B.** Average sensitivity. **C.** Average (Sorenson) stability index. **D.** Average number of hits. **E.** Average number of hits, with a zoom between 0 and 10 *: null corresponds to an absence of correlation between true predictors; intermediate corresponds to the realistic correlation observed within real Helix exposome; high corresponds to a pairwise correlation of 0.5. Elastic-Net CV-stabilized corresponds to Elastic-Net stabilized with a methodology similar as LASSO-CV1: a cross-validation process selecting the value of the parameter α minimizing the RMSE is repeating 100 times. The average value obtained for α is then used to repeat 100 times the cross-validation process selecting the optimal λ parameter minimizing RMSE: the average value of λ is used to fit the final model (this second step is exactly similar to LASSO-CV₁).



CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

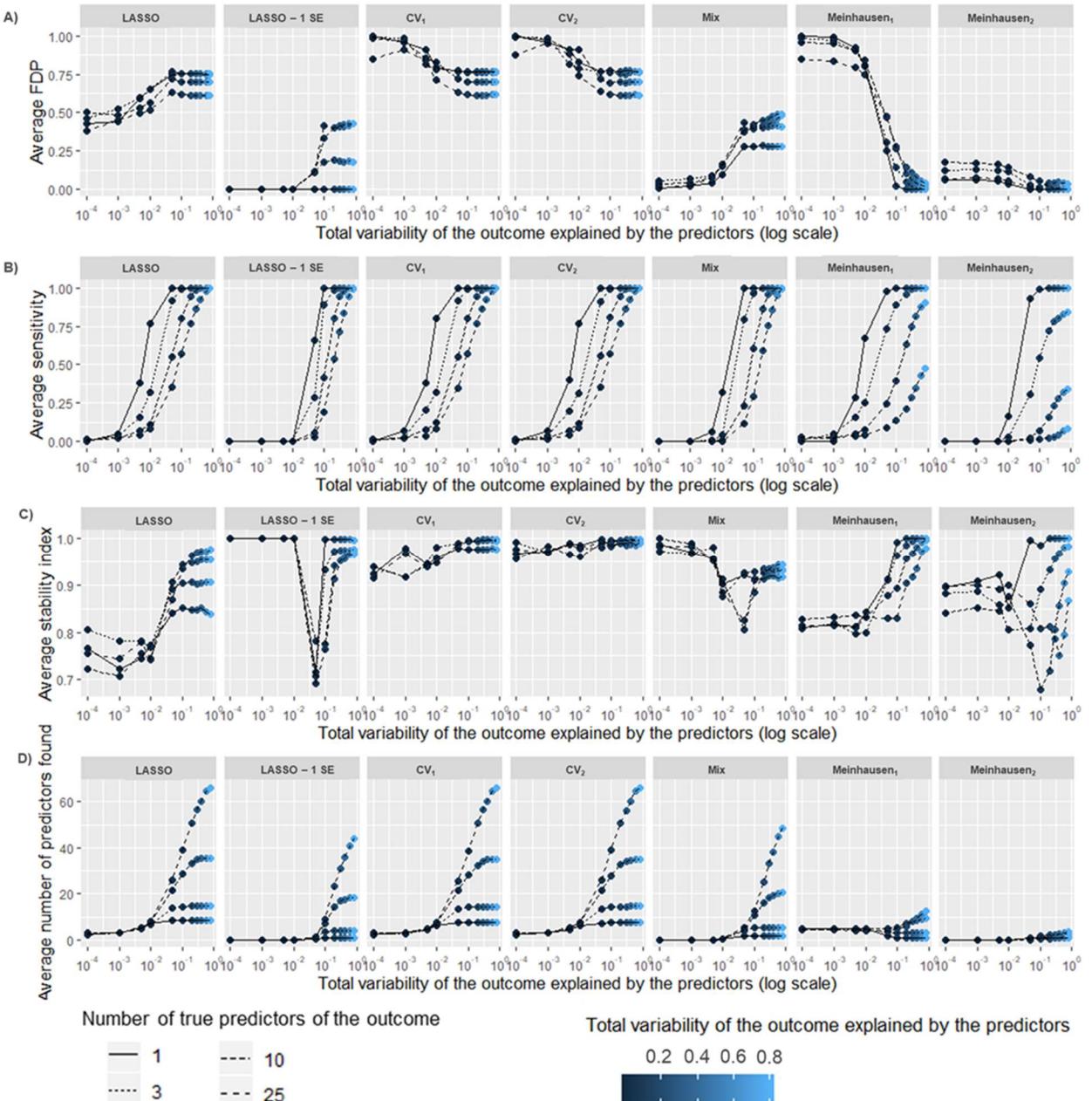


CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology



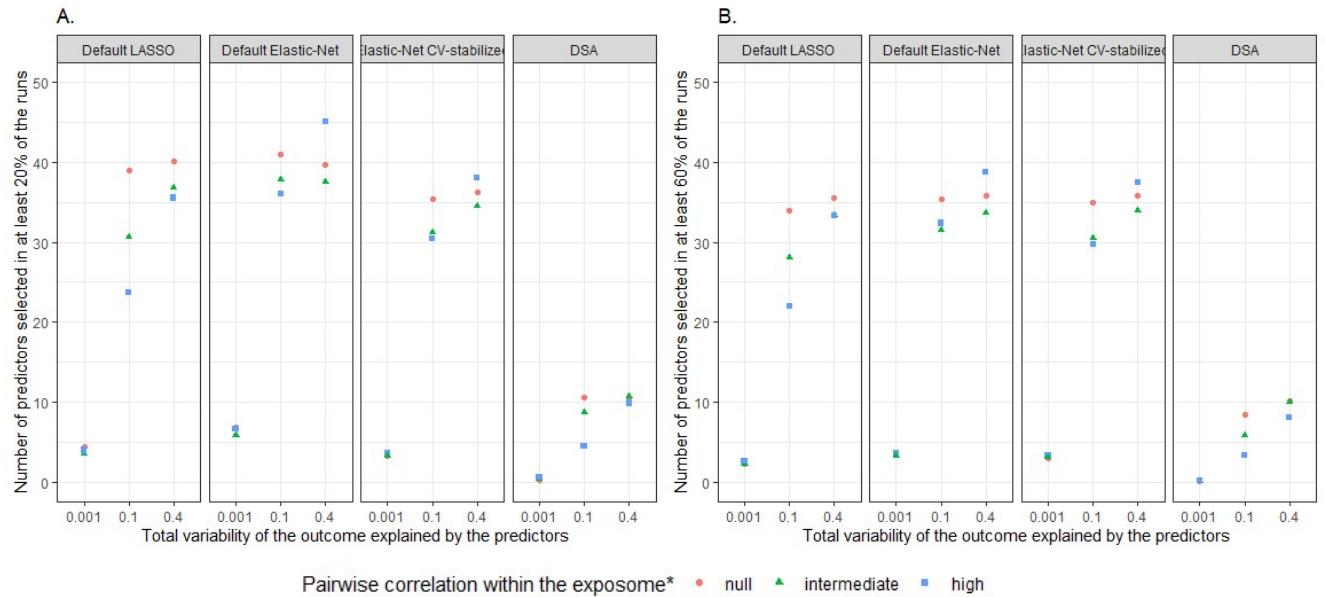
CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Supplementary Figure V.2: Performance and stability of the LASSO stabilization methods as a function of the total outcome variability explained by the predictors (log scale). Colors also show the total variability explained by the predictors. Each line corresponds to a different number of true predictors, in contrast with Figure V.1, in which results from scenarios with different numbers of true predictors were averaged. **A.** Average false discovery proportion. **B.** Average sensitivity. **C.** Average stability index. **D.** Average number of hits.



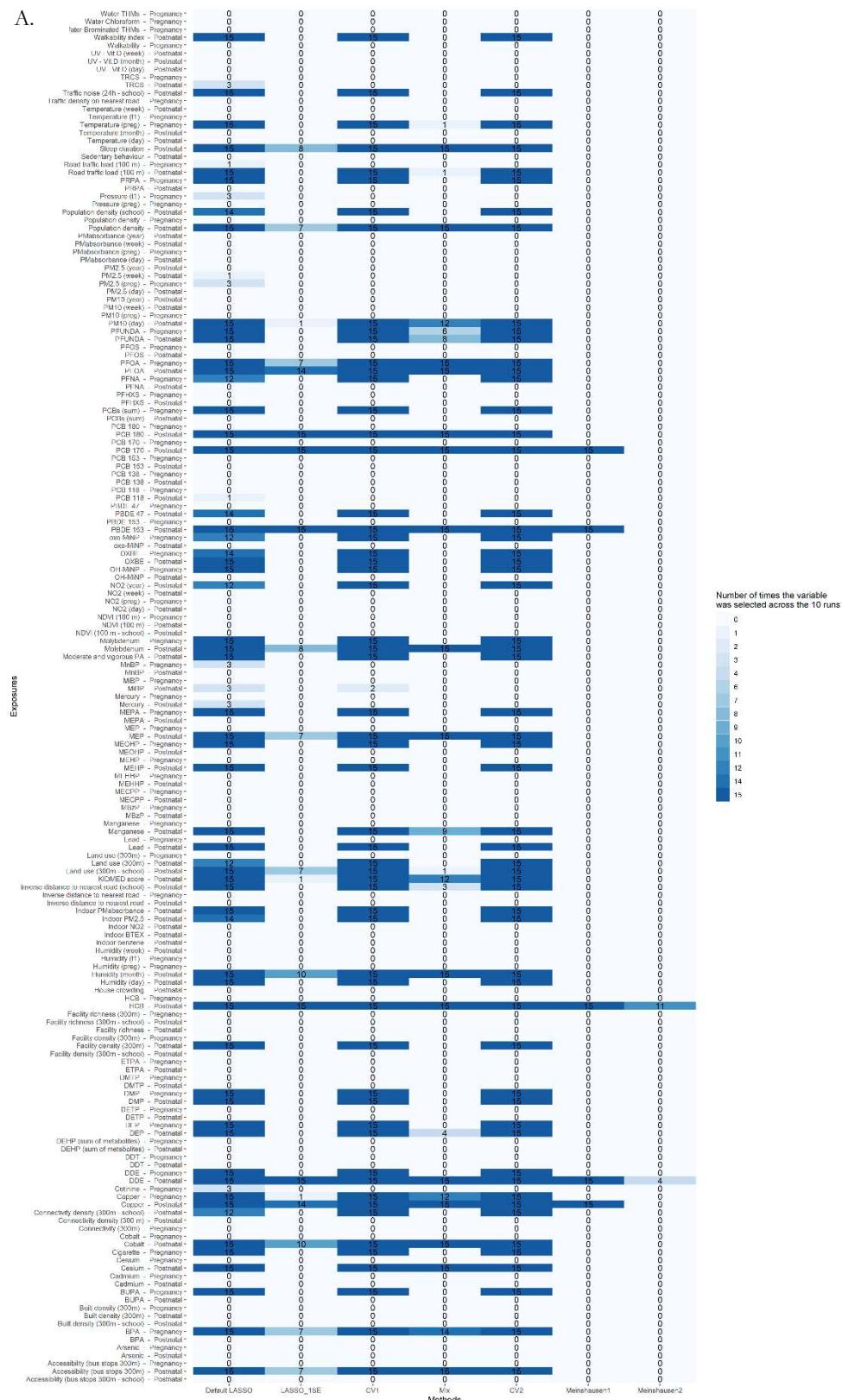
CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Supplementary Figure V.3: Average number of predictors selected in 20% (A.) and 60% (B.) of the runs on a same dataset for unstable methods, as a function of the total variability of the outcome explained by the true predictors. Colors show the level of correlation within the simulated exposome. *: null corresponds to an absence of correlation between true predictors; intermediate corresponds to the realistic correlation observed within real Helix exposome; high corresponds to a pairwise correlation of 0.5.

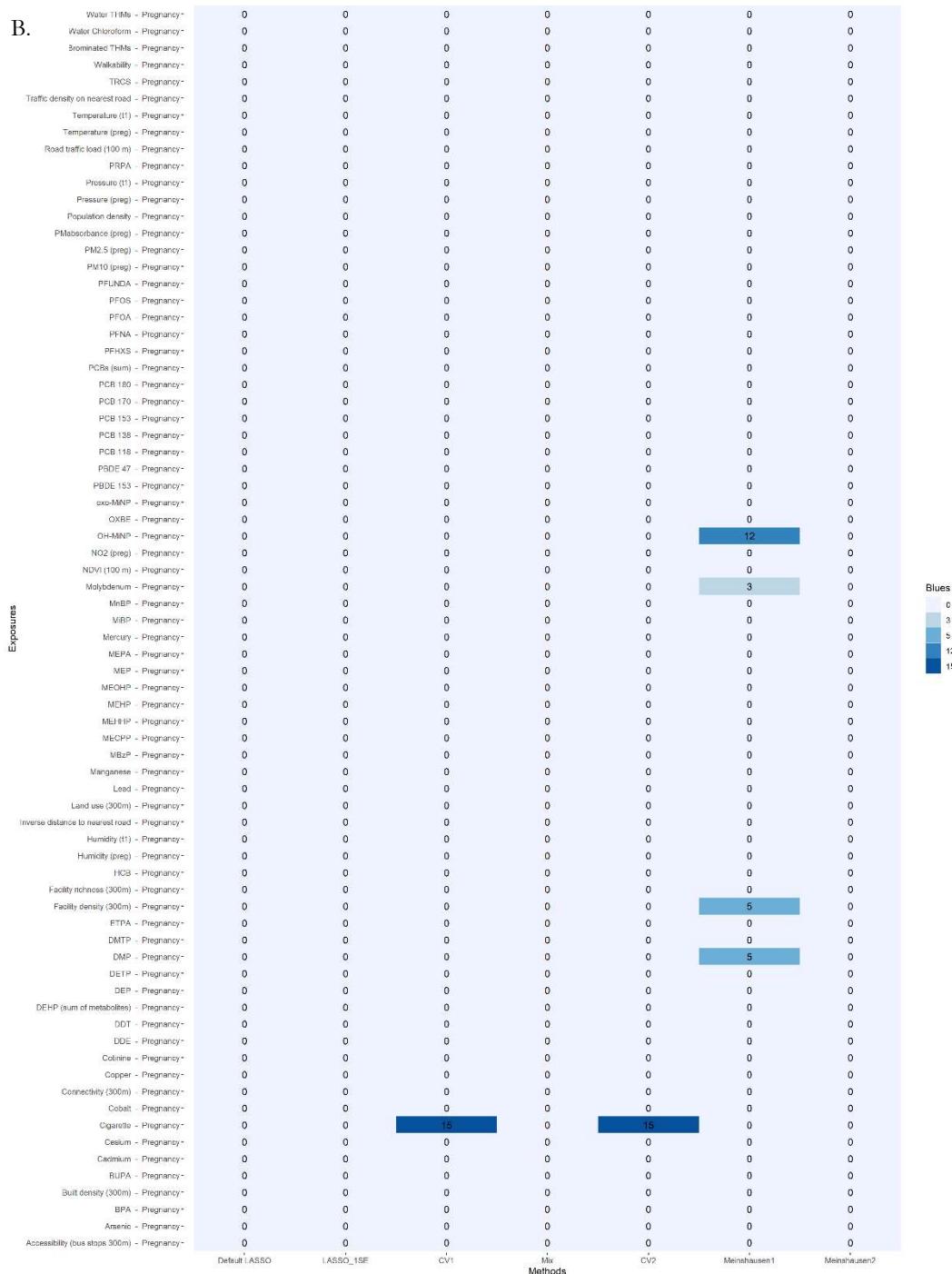


CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Supplementary Figure V.4: Occurrence of selection for each exposure when applying 10 times default LASSO and all tested stabilized LASSO to relate an exposome of 173 prenatal and postnatal quantitative exposures (A) or only the smaller exposome of 74 prenatal quantitative variables (B), to zBMI in 1301 mother-child pairs of the Helix cohorts.



CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology



CHAPTER V: Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology

Supplementary Table V.1: Distribution of stability index, sensitivity and False Discovery Proportion (FDP), across all scenarios and categorized according to the total variability explained by the true predictors (>1 and ≤ 1) for different stabilization methods of LASSO.

	All scenarios												Scenarios with $R^2 \leq 0.1$												Scenarios with $R^2 > 0.1$											
	FDP				Sensitivity				Stability index				FDP				Sensitivity				Stability index				FDP				Sensitivity				Stability index			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max				
Default LASSO	0.64	0.11	0.38	0.76	0.63	0.42	0.00	1.00	0.85	0.09	0.71	0.98	0.58	0.12	0.38	0.76	0.34	0.38	0.00	1.00	0.80	0.07	0.71	0.94	0.70	0.06	0.61	0.76	0.97	0.06	0.77	1.00	0.92	0.05	0.84	0.98
LASSO - 1 SE	0.14	0.18	0.00	0.43	0.51	0.46	0.00	1.00	0.95	0.09	0.69	1.00	0.05	0.11	0.00	0.42	0.15	0.30	0.00	1.00	0.93	0.12	0.69	1.00	0.25	0.18	0.00	0.43	0.94	0.12	0.53	1.00	0.97	0.02	0.91	1.00
CV ₁	0.78	0.11	0.61	1.00	0.63	0.42	0.00	1.00	0.97	0.03	0.91	1.00	0.83	0.11	0.62	1.00	0.35	0.38	0.00	1.00	0.96	0.03	0.91	0.99	0.71	0.06	0.61	0.77	0.97	0.06	0.77	1.00	0.99	0.01	0.98	1.00
CV ₂	0.78	0.12	0.61	1.00	0.63	0.42	0.00	1.00	0.99	0.01	0.96	1.00	0.85	0.12	0.62	1.00	0.35	0.38	0.00	1.00	0.98	0.01	0.96	1.00	0.71	0.06	0.61	0.77	0.97	0.06	0.77	1.00	0.99	0.00	0.98	1.00
Mix	0.27	0.17	0.01	0.49	0.55	0.45	0.00	1.00	0.93	0.04	0.80	1.00	0.17	0.16	0.01	0.43	0.23	0.36	0.00	1.00	0.93	0.05	0.80	1.00	0.40	0.08	0.28	0.49	0.95	0.11	0.59	1.00	0.93	0.01	0.91	0.94
Meinshausen ₁	0.39	0.40	0.00	1.00	0.50	0.41	0.00	1.00	0.91	0.08	0.80	1.00	0.69	0.32	0.02	1.00	0.26	0.33	0.00	1.00	0.85	0.05	0.80	0.99	0.04	0.04	0.00	0.14	0.78	0.28	0.21	1.00	0.98	0.03	0.90	1.00
Meinshausen ₂	0.05	0.05	0.00	0.18	0.31	0.39	0.00	1.00	0.88	0.08	0.68	1.00	0.08	0.06	0.00	0.18	0.13	0.29	0.00	0.99	0.87	0.06	0.68	1.00	0.01	0.02	0.00	0.05	0.53	0.40	0.02	1.00	0.90	0.09	0.72	1.00

Supplementary Table V.2: Results of the application of default LASSO and LASSO stabilization methods to relate an exposome of 74 prenatal quantitative exposures to zBMI in 1301 mother-child pairs of the Helix cohorts.

Stabilization method	Sorensen index	Number of selected exposures*	Computation time (in seconds)*	Number of exposures selected at least one time
None (default LASSO)	1.000	0.00	0.17	0
LASSO_1SE	1.000	0.00	0.17	0
CV ₁	1.000	1.00	13.92	1
CV ₂	1.000	1.00	15.40	1
Meinshausen ₁	0.467	1.67	0.98	4
Meinshausen ₂	1.000	0.00	0.96	0
Mix	1.000	0.00	11.86	0

* computed after repetition of each method 15 times.

Supplementary Table V.3: List of variables selected by default LASSO and all tested stabilized LASSO for each of the 10 runs applied to relate an exposome of 173 prenatal and postnatal quantitative exposures (**A**) or only the smaller exposome of 74 prenatal quantitative variables (**B**), to zBMI in 1301 mother-child pairs of the Helix cohorts. For each run, the direction of association with zBMI in a multivariate model including all exposures selected in the run and adjusted for relevant covariates is also given.

Due to its size, this supplementary is provided as a separated file in Appendix III.

Supplementary Material V.1: Commented simulation script

Due to its size, this supplementary is provided as a separated file in Appendix III.

This script will also be available on github (<https://github.com/SoCadiou>) once the corresponding draft will be published.

Supplementary Material V.2: Application: using LASSO to relate pregnancy exposome to child body mass index in Helix data

In the analysis relating only the prenatal exposures to zBMI, default LASSO did not select any exposure and was stable (Supplementary Table V.2). CV₁ and CV₂ displayed different results: they both selected one exposure (not the same for the two methods) and were stable (Supplementary Table V.3 and Supplementary Figure V.4). This difference of behavior between default LASSO and the cross-validation based methods made it plausible that the true predictors from the prenatal exposome explained less than 8% of zBMI (see Figure V.1), which is consistent with the fact that the pregnancy maternal exposome may show links of lower magnitude with child zBMI than the exposome assessed during the year before the child examination. For this range of scenarios, CV₁ and CV₂ are expected to show a very high FDP and should not be trusted, despite their stability. This is also the case for Meinshausen₂, which selected on average more than one exposure, and which is also expected to show high FDP (high FDP is more largely expected for Meinshausen₂, as soon as it is unstable, according to Figure V.3). The two other stability-selection-based methods did not select any exposure, which is consistent with an expected very low FDP in this range of scenarios.

CHAPTER VII: Discussion

In this PhD report we tried to provide new insights on the way to relevantly reduce the dimension of the exposome in the context of exposome-health studies; we focused on a strategy consisting in using information from intermediate layers between exposome and health to help performing this dimension reduction. We applied this strategy to the study of the effects of the early-life exposome on child health (growth and lung function).

In this discussion, after briefly summarizing our main results, we will question such a strategy and our choices regarding its implementation. We will first discuss the central question of dimension reduction, in exposome studies and within our particular context involving the use of a high-dimensional intermediate layer: how to relevantly choose among the possible approaches to reduce the dimension of a layer, according to the study objective? Then we will discuss the possible designs combining multiple layers in the context of studies aiming at selecting causal predictors of a health outcome. Last, we will replace our work within the larger challenges encountered today in environmental epidemiology and exposome research.

VII. 1. Overview of results

We performed three different studies on real data, using different implementation of our oMITM design (chapters II, III and Appendix II). All three studies share the main characteristics of our oMITM: they used an intermediary layer (the methylome), whose dimension was reduced borrowing information from the outcome of interest; then this reduced methylome was related to the exposome conditionally to the outcome, in order to build a reduced exposome. In a last step, a selection of exposures related to the outcome was performed within this reduced exposome. The methods chosen for the implementation of oMITM differed between the three studies. In our two studies on Helix data (chapters II and III), a priori information on the methylome was used by selecting CpG related to the outcome according to external databases. The two studies identified copper as related with the outcomes (respectively BMI and FEV₁ at 6-10 years), a finding which

may be relevant giving the role of copper in the regulation of inflammation and which is extensively discussed in chapters II and III. Compared to their agnostic counterparts ignoring the methylome, the oMITM design implementations possibly allowed to discard associations likely to be due to reverse causality, such as lipophilic compounds assessed at the same time at the outcome (chapter II) and to increase sensitivity by identifying copper, which an agnostic approach failed to do (chapter III). In the study on SEPAGES cohort (Appendix II), no a priori information about the methylome was used: as the relevance of pathways databases may be discussed (Mubeen et al., 2019), we chose a supervised dimension reduction without a priori, the PLS method, in order to try to relevantly summarize the needed information from the methylome layer. The whole oMITM did not identify any significant association between the (smaller) exposome and birth weight in this study of 438 mother-child pairs.

In chapters IV and V, we performed simulations to strengthen our choice of design and implementations. In a first simulation, we tested the relevance of our oMITM design compared to other designs, either “agnostic” or using an intermediary layer. We showed that our oMITM design allowed improved specificity compared to its agnostic counterpart (at a cost in sensitivity, when the methylome is not involved on the path from exposures to the outcome) and is less prone to suffer from reverse causality bias (chapter IV). In our second simulation (chapter V), we studied the stability algorithms commonly-used to select exposures related to health, and showed with the example of LASSO that methods taking into account the stability of selection seems to be more adapted in the context of a study aiming at minimizing the false discoveries than methods aiming at optimizing prediction, which are currently the most used.

VII. 2. Dimension reduction approaches in the context of exposome studies

A motivation of this PhD work was related to the challenges of limited specificity and power of exposome studies previously identified by our team (Agier et al., 2016; Slama and Vrijheid, 2015).

As we detailed in the introduction, relevant dimension reduction of the exposome may be a way to address them (see I.2.1). Dimension reductions techniques can also be necessary to visualize high dimensional data and thus better understand the model; or for computational reasons, to reduce algorithmic costs (Van Der Maaten et al., 2009). All these objectives are encountered in environmental epidemiology at different levels according to the aim of the study and the layers considered. Different methods can be used depending on the objectives: in particular, we illustrated that the dimension reduction can be done using a priori knowledge on the structuration of the data or without it (for example with agnostic variable selection algorithms, as described in chapter V). When it is data-driven, it can be supervised (using another layer, which is a more subtle way to incorporate a priori information) or unsupervised (Guyon and Elisseeff, 2003).

With our oMITM design, we chose to rely on the information coming from an intermediate layer of high dimension to perform this dimension reduction, and thus we also had to perform dimension reduction on this intermediary layer. What are the possible dimension reduction techniques that can be applied to these layers and in which situations are they relevant?

VII.2.1. Which dimension reduction approaches when selection of variable(s) of interest is the aim?

In exposome studies, as selection of variable is a key aim, extraction techniques are a priori not relevant. However, not all selection methods are adapted to our problem. Indeed, epidemiologists aim at selecting variables with causal biological meaning, with an ultimate public health aim to be able to operate on these variables to modify the outcome. Thus, a supervised selection method must be applied. But beforehand, another dimension reduction method can be applied to the exposome to “simplify the problem”: this is what we did with the primary implementation of our oriented Meet-in-the-Middle (chapters II and IV) design which allowed to build a *reduced exposome*, on which a basic selection method, ExWAS, was then applied. What are the other dimension reductions techniques that we could have considered? A summary is provided in Table VII.1, but

it is important to notice that in an exposome study, using directly a priori knowledge to focus on one (or a few more) a priori chosen exposure(s) is an option which does not take directly into account the potential of the exposome in the sense that the interest of the exposome approach lies in the ability to consider 1) potential predictors simultaneously; 2) potential predictors whose effect on health is currently not well known. Thus, focusing on specific compounds whose harmfulness is known would not be a strategy in an exploratory exposome study, whereas on the contrary it would be relevant in specific confirmatory studies or studies aiming at precisely quantifying known effects. However, some a priori knowledge on the exposome can be used for example to discard exposures with no possibility of effects (for example irrelevant exposure windows).

Various data-driven selection techniques are available. They can be divided in three categories. The first one are *filtering methods*: they select features on the basis of their scores in various statistical tests (for example, their correlation with the outcome variable), usually as a pre-processing step before applying a learning algorithm. Multiple linear regression (MLR) as well as ExWAS-type methods can be considered as data-driven filtering methods, and we used them to perform a selection on the exposome by relating it with the outcome of interest on our three real-data studies (chapters II, III and Appendix II). Indeed, in the context of exposome studies, such methods can be used as the main analysis if they have sufficient specificity, and rather as a pre-processing step if there are sensitive but with low specificity. Thus, for example, ExWAS without correction for multiple comparison could be used as a pre-processing filtering method, whereas ExWAS with correction can be used for the main analysis (as soon as the correlation is not too high, which would have a strong negative impact on the specificity, as we showed in chapter V). With our MITM design, we used also such methods as a pre-processing step involving an intermediate layer: indeed, at the second step of our original MITM design (chapter II, IV and Appendix II) or in our lung function study at the first step (chapter III), we ran univariate tests involving another layer allowing to rank and to select some variables to build our reduced exposome, on which a more stringent selection method was then applied.

A second category of selection techniques are the *wrapper methods*, which consist of iterative searches of a subset of variables: at each iteration, some variables are added or removed to try to strengthen the inference (Guyon and Elisseeff, 2003). These methods are usually very computationally demanding, but may be adapted to the intermediate dimension of the exposome. DSA is an example of such a method: it has been used in exposome studies (Agier et al., 2019; Nieuwenhuijsen et al., 2019) in spite of its lack of stability, which we studied in chapter V.

Last, the *embedded methods* are methods which have inbuilt feature-selection methods, like for example Elastic-Net and LASSO, which we have studied in chapter V.

The three categories could be used to tackle exposome studies challenges. However, embedded methods may be not specific enough, as most of them were designed in the machine learning field to optimize the predictive ability of models, which may imply the selection of too many variables, as we showed with LASSO and discussed in chapter V. Indeed, as demonstrated by Guyon and Elisseeff (2003), for learning tasks, “noise and reduction and consequently better class separation may be obtained by adding variables that are presumably redundant.” This may explain why, in our simulation study of the stabilization methods of LASSO (chapter V), the stabilization method aiming at optimizing the prediction showed lower performance in term of specificity than stability selection (Meinshausen and Bühlmann, 2010), a method focusing on selection of variables of interest in a logic of model averaging. Overall, the wrapper methods may be the most adapted for the main analysis in exposome studies, whereas filtering methods should be potentially used as a preprocessing step.

VII.2.2. Which dimension reduction approaches when information concentration is the aim?

When the aim is not selection of relevant variables but concentration of a diluted information, other methods can be envisaged (see Table VII.1). In a strategy aiming at using a layer to inform the structure of another layer, extraction methods may be well suited, in particular when the layer is of high dimension, which was the case of the methylome. More generally, all methods reducing

dimension can be used as soon as they are not too specific, i.e. if they do not restrict too much the quantity of information. In other words, in this case, the compromise between sensitivity and specificity (here in the sense of detection of available information) is in favor of sensitivity, as it is a pre-processing step, whereas in the case of selection that we discussed in the previous paragraph, specificity may be favored. As soon as the dimension is reduced enough to make the information usable, the presence of redundant variables is not a problem. This was one the motivations of the use of ExWAS-type method (MWAS) to reduce the dimension of the methylome (chapters II and IV). Lack of biological interpretability is not a problem either: for example, in our MITM design, selecting a CpG correlated with a causal CpG on the pathway from exposures to health is not an issue as soon as it still enables to select the relevant exposures. This also explains why extraction methods may be adapted in this context. Thus, in the Sepages study (Appendix II), we chose to modify our MITM design by using PLS in the first step, which enabled to perform a huge supervised dimension reduction by building one summary new variable. In the same logic, interestingly, in chapter II, when we performed a sensitivity analysis of our MITM analysis by using 6 variables describing cell-types instead of our methylome of 2284 CpG as our intermediary variables, we found the same results: the cell types can be considered as a smaller number of variables carrying the same information for our problem than our high-dimensional methylome, which made us hypothesize that the information that we obtained from the methylome was majorly due to an inflammation process also observable from the cell-types counts. This also illustrated that this step of information concentration can be performed using *a priori* external knowledge. This was also what we did when we chose to focus on enhancers CpGs belonging to relevant biological pathways (chapters II and III). Following the same logic of extraction dimension reduction and reliance on external knowledge, we could also have considered a strategy aiming at summarizing the methylation information pathway by pathway.

Overall, multiple methods exist to deal with high dimensional layer when interpretability is not necessary. They could be used to optimize prediction but also to concentrate information in order to perform supervised dimension reduction on another layer.

Table VII.1: Possible strategies of dimension reduction for the exposome and methylome layers

	Exposome layer	Methylome (or another intermediate layer)
Dimension	Intermediate ($\sim 10^2$)	High ($\sim 10^6$)
Average correlation	Intermediate (~ 0.1)	Intermediate (>0.1)
Objective within our strategy	Understand its causal link with an outcome	Provide information to help dimension reduction of the exposome
Why reducing dimension?	To select relevant variables of interest	To make information usable
Direct use of a priori knowledge in dimension reduction?	No direct a priori selection in a discovery exposome approach (but useful to discard totally irrelevant variables, e.g. variables in an irrelevant exposure window, before the analysis)	Yes, if domain knowledge is available
Use of selection methods for dimension reduction?	Yes	Yes
Filtering methods	With a priori knowledge on the layer: no Without a priori: yes, for example ExWAS (<i>chap. II and III</i>), MLR (<i>appendix II</i>)	With a priori on the layer: yes, for example preselection of relevant features according to domain knowledge (<i>chap. II, chap. III</i>) Without a priori: yes, for example MWAS (<i>chap II, chap IV</i>)
Wrapper methods	Yes, for example DSA (<i>chap. IV and V</i>)	No, they are often computationally not feasible in high dimension
Embedded methods	Yes, for example LASSO or Elastic-Net (<i>chap. V</i>)	Yes, for example high dimensional LASSO or random forest
Use of extraction methods for dimension reduction	No, selection is the aim	Yes
Unsupervised	/	Without a priori: yes, for example ACP With a priori: yes, for example summary variables by biological pathway
Supervised	/	Yes, for example PLS (<i>appendix II</i>)
Favouring sensitivity or specificity?	Specificity (with non-null sensitivity)	Sensitivity (as well information is usable)

VII. 3. Multilayer designs to identify causal predictors

In the previous paragraphs, we discussed which methods could be adapted to perform dimension reduction on the layers that we considered in this PhD project. Assuming the question of the dimension reduction solved, the design of the statistical analysis would still be a major question: how to combine the information from the two layers? As we showed in our simulation study under various causal structures (chapter IV), the choice of the design is essential if the aim is to select *causal* predictors.

VII.3.1. Infer causality in exposome studies

In epidemiology, the first condition to select causal variables is to choose a relevant design of data collection (for example a longitudinal design). Then, knowledge-based adjustment on relevant confounders is necessary. However, it is still impossible to distinguish between correlated variables on which no previous information is available or be sure to discard reverse causality (as we show in chapter IV (see Supplementary Figure IV.5), even with a longitudinal design, reverse causality can still lead to spurious association). Some selection methods among those we discussed in paragraph VI.2.2. have been built reckoning the causal inference theory and consider estimators derived from counterfactuals inference in order to perform selection of causal predictors: for example, in DSA, a derivative-based importance measure aiming at measuring the ‘causal’ effect of a variable can be implemented (Sinisi and van der Laan, 2004). The ‘targeted learning’ proposed by Van der Laan (van der Laan et al., 2007; van der Laan and Starmans, 2014) also considers inference theory to choose the parameters to estimate. These methods, which can be considered “data-driven causal modelling”, like the Bayesian networks that we discussed in the introduction (paragraph I.2.3), may be adapted to our problem, as selection methods aiming at inference rather than prediction should be preferred in the context of discovery exposome studies.

VII.3.2. Multiple layers design as a clue to overcome the challenges of data-driven causal modelling

However, the possibility of inferring causality solely from data has been contested in the literature on causal discovery, in the line of the idea of ‘no causes in, no causes out’ first introduced by Cartwright: it states that “old causal knowledge must be supplied for new causal knowledge to be had” when trying to decipher causality of an observed phenomenon (Cartwright and Nancy, 1994). Big data (defined as available datasets characterized by their high volume, high velocity of acquisition and high variety (Canali, 2016)) had once been considered a solution (Canali, 2016): the abundance of meaningful correlations was supposed to allow to get rid of the need of prior causal theory or knowledge (Mayer-Schönberger and Cukier, 2013). But these claims have later been contradicted (Leonelli, 2014; Titiunik, 2014) and the indispensable use of hypotheses in research based on Big Data highlighted (Ratti, 2015). Hernan underlined that a real causal assignation always involves expert knowledge: this is what makes the difference between prediction and counterfactual prediction (Hernán et al., 2019), i.e. causal selection. This can be linked to the fact that theoretical conditions needed to make causal inference possible using only data (e.g. when using Bayesian networks) seem impossible to reach in reality: for example, it is only under the faithfulness condition (i.e. that all probabilistic dependencies and independencies characterizing the variables studied should be considered) that a learned Bayesian network can be used for inference (Ghiara, 2019). More generally, all the causal theory used in DAG and underlying the causal learning methods involves the absence of external confounders. In practice, this seems unreachable in observational studies, unless the ‘whole system’ is considered, which would make the complexity infinite. In particular, we can hypothesize that in most biological systems studied, the faithfulness condition involves to consider a number of features which makes the curse of dimensionality a problem. Thus, as we discussed before, dimension reduction techniques would be needed; but they would be incompatible with the concept of purely data-driven causal modeling: indeed, selection of variables involves *a priori* knowledge (even supervised selection techniques involve previous

knowledge as a link with an outcome is postulated) and variables extraction prevents the causal interpretation, as it provides new variables without biological meaning. Thus, even when using big data, data-driven causal modelling cannot be enough to infer causality, expert knowledge is additionally needed and we suggest here that this could be explained using the faithfulness condition.

More precisely, the Russo-Williamson thesis suggests that to infer causality, a mechanism must be supplied in addition to an observed association (i.e. a probabilistic dependency) (Russo and Williamson, 2007). Such a mechanism generally means an external explanation of *how*, and corresponds to the expert knowledge advocated by Hernan (2019). However, we must recall here that all human knowledge derives solely from experience (Hume, 1740): in particular, what is considered as external expert knowledge used for causality always derives from previous observation(s). For example, it is interesting to note that information from biological databases (e.g. pathways or genetic database) is discussed by Canali as a supplementary knowledge enabling to infer causality in exposome studies (Canali, 2016), whereas Leonelli discussed it as part of data-driven sciences (Leonelli, 2014). Thus, the notion of « mechanism » proposed by Russo & Williamson must rather be interpreted as probabilistic dependencies observed at a different scale: for example, toxicological observations at the level of the cell to support an adverse observed effect at the population level. Thus, working with multiple biological layers as we did in this PhD work may be a step on the path to causality: this is what Canali suggested when he considered the Meet-in-the-Middle design proposed by Chadeau-Hyam and Vineiss (similar to our oMITM but without our additional adjustment on the outcome) as a way to investigate disease causation by searching for intermediate biomarkers (Canali, 2016; Chadeau-Hyam et al., 2011). Considering the structuration of our data in different layers from different scales is a way to add expert knowledge. However, we showed both with theoretical and simulated works (chapter IV) that the Meet-in-the-Middle design advocated by Chadeau-Hyam et al. (2011) is not sufficient to affirm causality. In

particular, it is prone to reverse causality, similarly to the mediation design. On the contrary, our oMITM design seems to be more robust to the variety of underlying causal structures, in the sense that we developed in chapter IV. But, even considering an implementation with perfect power, it is still not sufficient to get rid of all situations of reverse causality and it is not able to detect all true causal associations. Combining results of different designs (for example running a mediation test and a oMITM test and compare results according as in Table IV.4 and Supplementary Table IV.4) could increase the probability of deciphering the true causal structures, but overall, causality could not be affirmed but only discussed with more or less likeliness. Overall, data-driven causal modelling is not enough and adding knowledge is always delicate, but working with structured layers and adequate robust designs may help.

VII. 4. Perspectives and conclusion

Thus, our results show that to perform a causal discovery exposome study, one may rely on intermediate layer and choose adapted dimension reduction method(s), design(s) and implementation(s), according to the structure of data and the aim of study: in particular choosing whether high sensitivity or specificity should be favored is needed. In this chapter, we discussed the dimension reduction methods (VI.2.) as well as the most adequate analytical designs (VI.3); the possible implementations of our proposed designs were extensively discussed in chapters IV and V. One question remaining is the choice of the intermediate layer. We chose to focus on the methylome layer, but other ‘omics’ (or non ‘omics’) layers could have been considered and may be helpful to inform the exposome health relations. For example, transcriptome may also contain biomarkers of exposures and diseases, making mediation or oMITM analyses relevant (Winckelmans et al., 2017a, 2017b). Microbiome, whose causal relationship with health outcome is complex, would also be a potentially relevant intermediate layer (Sohn and Li, 2019), as well as inflammatory or immunological markers.

Overall, this PhD work can give us insights of a possible future for exposome studies. Our work supports the assessment of omics intermediate layers, as done by most of recent exposome studies (see Table I.1): indeed, we demonstrated that these layers are not only useful to better understand mechanisms of effect of exposures on health as often suggested, but also offer possibilities to improve the detection of causal predictors of health among the exposome compared to “agnostic exposome studies”. However, our work also illustrated the curse of dimensionality, which constitutes a threat for exposome studies, as we underlined the systematic gain in term of specificity obtained by dimension reduction for intermediate or high dimension data (see in particular in Chapter IV the evolution of FDP according to the size of the reduced exposome, Figure IV.3). Thus, the development of omics assessment in exposome studies should go together the enhancement of dimension reduction tools. This is becoming crucial as novel analytical techniques make it easier to measure omics by millions in one sample at low-cost. As we discussed previously (see VI.I.1.), when the aim is to use intermediate layer to inform the exposome-health relation, many statistical methods can be used, including the powerful extraction methods. But, we also showed (see Chapter II and III) the relevance of dimension reduction relying on a priori knowledge, especially from the toxicological field: if for the moment, information from toxicological databases are not always reliable (Mubeen et al., 2019), consolidated and usable pathways database would be promising tools for the exposome research. A novel application, in the line of our work, could be to use such tools to build summary variables by pathways and use them to relevantly reduce the exposome dimension.

On the contrary, less dimension reduction strategies are relevant for the exposome than for the intermediary layer, as we discussed above (see VI.1.2), whereas current sample size of exposome projects limit the specificity and power of exposome studies to detect causal predictors of health (see chapters IV and V and Appendix II). As we showed, informed dimension reduction and relevant statistical variable selection methods can help lower FDP and even in some cases increase sensitivity, if an adequate analytical design is chosen. However, both of these strategies would still

be limited by a too high dimension: an increase in the components of the exposome assessed should come together with an increase in sample size to make strategies that we pointed applicable and to provide sufficient power. Indeed, as we developed in appendix II, the oMITM cannot for example be more sensitive than the corresponding agnostic approach without correction for multiple testing. This is of particular importance as the assessment of exposures by biomarkers also makes a wider assessment of exposome easier. Such an increase in the factors assessed would of course be valuable to better describe the personal environment, but it could lead to assessing of thousands to billions of factors in a few individuals, which will result in datasets unexploitable for causal inference of environmental effects on health.

To conclude, we provided new insights on how the use of intermediate layers may help to inform exposome health study and in particular to help to tackle the challenge of reverse causality and low specificity, with the aim of detecting the causal predictors of a health outcome. To replace our problematic among the challenges of the environmental epidemiology field, if omics are assessed conjointly with a minimized measurement error on the exposome, the use of methods such as those we proposed to identify with increased specificity causal predictors of health may help to focus on relevant compounds and to build adequate analytical designs to estimate measures of association and then perform health impact assessment, with an ultimate aim to help public policies.

References

- Agay-Shay, K., Martinez, D., Valvi, D., Garcia-Estebar, R., Basagaña, X., Robinson, O., Casas, M., Sunyer, J., Vrijheid, M., 2015. Exposure to endocrine-disrupting chemicals during pregnancy and weight at 7 years of age: A multi-pollutant approach. *Environ. Health Perspect.* 123, 1030–1037. <https://doi.org/10.1289/ehp.1409049>
- Agier, L., Basagaña, X., Hernandez-Ferrer, C., Maitre, L., Tamayo Uria, I., Urquiza, J., Andrusaityte, S., Casas, M., de Castro, M., Cequier, E.M., Chatzi, L., Donaire, D., Giorgis-Allemand, L., Ramon Gonzalez, J., Grazuleviciene, R., Gützkow, K.B., Haug, L.S., Sakhi, A.K., McEachan, R.R.C., Meltzer, H.M., Nieuwenhuijsen, M., Robinson, O., Roumeliotaki, T., Sunyer, J., Thomsen, C., Vafeiadi, M., Valentin, A., West, J., Wright, J., Siroux, V., Vrijheid, M., Slama, R., 2020a. Association Between the Pregnancy Exposome and Fetal Growth [WWW Document]. *Int. J. Epidemiol.* <https://doi.org/10.2139/ssrn.3258668>
- Agier, L., Basagaña, X., Maitre, L., Granum, B., Bird, P.K., Casas, M., Oftedal, B., Wright, J., Andrusaityte, S., de Castro, M., Cequier, E., Chatzi, L., Donaire-Gonzalez, D., Grazuleviciene, R., Haug, L.S., Sakhi, A.K., Leventakou, V., McEachan, R., Nieuwenhuijsen, M., Petracienciene, I., Robinson, O., Roumeliotaki, T., Sunyer, J., Tamayo-Uria, I., Thomsen, C., Urquiza, J., Valentin, A., Slama, R., Vrijheid, M., Siroux, V., 2019. Early-life exposome and lung function in children in Europe: an analysis of data from the longitudinal, population-based HELIX cohort. *Lancet Planet. Heal.* 3, e81–e92. [https://doi.org/10.1016/S2542-5196\(19\)30010-5](https://doi.org/10.1016/S2542-5196(19)30010-5)
- Agier, L., Portengen, L., Chadeau-Hyam, M., Basagaña, X., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J.R., Nieuwenhuijsen, M.J., Vineis, P., Vrijheid, M., Slama, R., Vermeulen, R., Hyam, M.C., Basagaña, X., Allemand, L.G., Siroux, V., Robinson, O., Vlaanderen, J., González, J.R., Nieuwenhuijsen, M.J., Vineis, P., Vrijheid, M., Slama, R., Vermeulen, R., 2016. A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations. *Environ. Health Perspect.* 124, 1848–1856. <https://doi.org/10.1289/EHP172>
- Agier, L., Slama, R., Basagaña, X., 2020b. Relying on repeated biospecimens to reduce the effects of classical-type exposure measurement error in studies linking the exposome to health. *Environ. Res.* 186, 109492. <https://doi.org/10.1016/j.envres.2020.109492>
- Armstrong, B.G., 1998. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup. Environ. Med.* 55, 651–656. <https://doi.org/10.1136/oem.55.10.651>
- Baccarelli, A., Wright, R.O., Bollati, V., Tarantini, L., Litonjua, A.A., Suh, H.H., Zanobetti, A., Sparrow, D., Vokonas, P.S., Schwartz, J., 2009. Rapid DNA Methylation Changes after Exposure to Traffic Particles. *Am. J. Respir. Crit. Care Med.* 179, 572–578. <https://doi.org/10.1164/rccm.200807-1097OC>
- Bach, F., 2008. Bolasso: model consistent Lasso estimation through the bootstrap. <https://doi.org/10.1145/1390156.1390161>
- Barfield, R., Shen, J., Just, A.C., Vokonas, P.S., Schwartz, J., Baccarelli, A.A., VanderWeele, T.J., Lin, X., 2017. Testing for the indirect effect under the null for genome-wide mediation analyses. *Genet. Epidemiol.* 41, 824–833. <https://doi.org/10.1002/gepi.22084>
- Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173–82.

- Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J.R., Nieuwenhuijsen, M., Vineis, P., Vrijheid, M., Vermeulen, R., Slama, R., Basagaña, X., 2017. A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ. Heal. A Glob. Access Sci. Source* 16, 74. <https://doi.org/10.1186/s12940-017-0277-6>
- Belbasis, L., Savvidou, M.D., Kanu, C., Evangelou, E., Tzoulaki, I., 2016. Birth weight in relation to health and disease in later life: An umbrella review of systematic reviews and meta-analyses. *BMC Med.* 14, 147. <https://doi.org/10.1186/s12916-016-0692-5>
- Bell, M.L., Belanger, K., Ebisu, K., Gent, J.F., Lee, H.J., Koutrakis, P., Leaderer, B.P., 2010. Prenatal exposure to fine particulate matter and birth weight: Variations by particulate constituents and sources. *Epidemiology* 21, 884–891. <https://doi.org/10.1097/EDE.0b013e3181f2f405>
- Bellman, R.E., 1961. Adaptive Control Processes: A Guided Tour [WWW Document]. Princeton, NJ. URL https://books.google.fr/books?hl=fr&lr=&id=iwbWCgAAQBAJ&oi=fnd&pg=PR9&ots=bDK5XqD25i&sig=KBl_vhNG67fA26OW4b0WF2JLSRU&redir_esc=y#v=onepage&q&f=false (accessed 5.28.20).
- Belloni, A., Chernozhukov, V., Hansen, C.B., 2011. Inference for high-dimensional sparse econometric models. *Adv. Econ. Econom.* Tenth World Congr. Vol. 3, Econom. 245–295. <https://doi.org/10.1017/CBO9781139060035.008>
- Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B.* <https://doi.org/10.2307/2346101>
- Benjamini, Y., Yekutieli, D., 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* <https://doi.org/10.2307/2674075>
- Benton, M.C., Sutherland, H.G., Macartney-Coxson, D., Haupt, L.M., Lea, R.A., Griffiths, L.R., 2017. Methylome-wide association study of whole blood DNA in the Norfolk Island isolate identifies robust loci associated with age. *Aging (Albany. NY)*. 9, 753–768. <https://doi.org/10.18632/aging.101187>
- Blum, M.G.B., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., Slama, R., 2020. Challenges Raised by Mediation Analysis in a High-Dimension Setting. *Environ. Health Perspect.* 128, 055001. <https://doi.org/10.1289/EHP6240>
- Bogdanos, D.P., Smyk, D.S., Rigopoulou, E.I., Sakkas, L.I., Shoenfeld, Y., 2015. Infectomics and autoinfectomics: a tool to study infectious-induced autoimmunity. *Lupus* 24, 364–73. <https://doi.org/10.1177/0961203314559088>
- Bonvallot, N., Tremblay-Franco, M., Chevrier, C., Canlet, C., Warembourg, C., Cravedi, J.-P., Cordier, S., 2013. Metabolomics Tools for Describing Complex Pesticide Exposure in Pregnant Women in Brittany (France). *PLoS One* 8, e64433. <https://doi.org/10.1371/journal.pone.0064433>
- Boulesteix, A.-L., Durif, G., Lambert-lacroix, S., Peyre, J., Strimmer, K., 2018. Package ‘*plsgenomics*’.
- Boulesteix, A.-L., Slawski, M., 2009. Stability and aggregation of ranked gene lists. *Brief. Bioinform.* 10, 556–568. <https://doi.org/10.1093/bib/bbp034>
- Bousquet, O., Elisseeff, A., 2002. Stability and Generalization. *J. Mach. Learn. Res.* 2, 499–526.

- Breiman, L., 2004. Bagging predictors. *Mach. Learn.* 24, 123–140.
<https://doi.org/10.1007/bf00058655>
- Brewer, G.J., 2010. Copper toxicity in the general population. *Clin. Neurophysiol.* 121, 459–460.
<https://doi.org/10.1016/j.clinph.2009.12.015>
- Brulle, R.J., Pellow, D.N., 2006. ENVIRONMENTAL JUSTICE: Human Health and Environmental Inequalities. *Annu. Rev. Public Health* 27, 103–124.
<https://doi.org/10.1146/annurev.publhealth.27.021405.102124>
- Bunin, A., Sisirak, V., Ghosh, H.S., Grajkowska, L.T., Hou, Z.E., Miron, M., Yang, C., Ceribelli, M., Uetani, N., Chaperot, L., Plumas, J., Hendriks, W., Tremblay, M.L., Häcker, H., Staudt, L.M., Green, P.H., Bhagat, G., Reizis, B., 2015. Protein Tyrosine Phosphatase PTPRS Is an Inhibitory Receptor on Human and Murine Plasmacytoid Dendritic Cells. *Immunity* 43, 277–288. <https://doi.org/10.1016/j.jimmuni.2015.07.009>
- Burling, T.A., Bigelow, G.E., Robinson, J.C., Mead, A.M., 1991. Smoking during pregnancy: Reduction via objective assessment and directive advice. *Behav. Ther.* 22, 31–40.
[https://doi.org/10.1016/S0005-7894\(05\)80241-2](https://doi.org/10.1016/S0005-7894(05)80241-2)
- Busch, R., Qiu, W., Lasky-Su, J., Morrow, J., Criner, G., DeMeo, D., 2016. Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations. *Respir. Res.* 17, 143. <https://doi.org/10.1186/s12931-016-0459-8>
- Buschdorf, J.P., Ong, M.L., Ong, S.X., MacIsaac, J.L., Chng, K., Kobor, M.S., Meaney, M.J., Holbrook, J.D., 2016. Low birth weight associates with hippocampal gene expression. *Neuroscience* 318, 190–205. <https://doi.org/10.1016/j.neuroscience.2016.01.013>
- Buuren, S. van, Groothuis-Oudshoorn, K., 2011. **mice** : Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Cadiou, S., Bustamante, M., Agier, L., Andrusaityte, S., Basagaña, X., Carracedo, A., Chatzi, L., Grazuleviciene, R., Gonzalez, J.R., Gutzkow, K.B., Maitre, L., Mason, D., Millot, F., Nieuwenhuijsen, M., Papadopoulou, E., Santorelli, G., Saulnier, P.J., Vives, M., Wright, J., Vrijheid, M., Slama, R., 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>
- Canali, S., 2016. Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data Soc.* 3, 205395171666953.
<https://doi.org/10.1177/2053951716669530>
- Carrico, C., Gennings, C., 2013. Characterization of a Weighted Quantile Score Approach for Highly Correlated Data in Risk Analysis Scenarios. *Dep. Biostat. Ph.D.*, 150.
<https://doi.org/10.1007/s13253-014-0180-3.Characterization>
- Cartwright, Nancy, 1994. Nature's Capacities and Their Measurement. OUP Cat.
- Casas, M., Basagaña, X., Sakhi, A.K., Haug, L.S., Philippat, C., Granum, B., Manzano-Salgado, C.B., Brochot, C., Zeman, F., de Bont, J., Andrusaityte, S., Chatzi, L., Donaire-Gonzalez, D., Giorgis-Allemand, L., Gonzalez, J.R., Gracia-Lavedan, E., Grazuleviciene, R., Kampouri, M., Lyon-Caen, S., Pañella, P., Petraviciene, I., Robinson, O., Urquiza, J., Vafeiadi, M., Vernet, C., Waiblinger, D., Wright, J., Thomsen, C., Slama, R., Vrijheid, M., 2018. Variability of urinary concentrations of non-persistent chemicals in pregnant women and school-aged children. *Environ. Int.* 121, 561–573.
<https://doi.org/10.1016/J.ENVINT.2018.09.046>

- Cervantes, S., Fontcuberta-Pisunyer, M., Servitja, J.M., Fernandez-Ruiz, R., García, A., Sanchez, L., Lee, Y.S., Gomis, R., Gasa, R., 2017. Late-stage differentiation of embryonic pancreatic β -cells requires Jarid2. *Sci. Rep.* 7, 1–14. <https://doi.org/10.1038/s41598-017-11691-2>
- Chadeau-Hyam, M., Athersuch, T.J., Keun, H.C., De Iorio, M., Ebbels, T.M.D., Jenab, M., Sacerdote, C., Bruce, S.J., Holmes, E., Vineis, P., 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>
- Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B., Vermeulen, R.C.H., 2013. Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers. *Environ. Mol. Mutagen.* 54, 542–557. <https://doi.org/10.1002/em.21797>
- Chamberlain, P.P., Qian, X., Stiles, A.R., Cho, J., Jones, D.H., Lesley, S.A., Grabau, E.A., Shears, S.B., Spraggan, G., 2007. Integration of inositol phosphate signaling pathways via human ITPK1. *J. Biol. Chem.* 282, 28117–28125. <https://doi.org/10.1074/jbc.M703121200>
- Chatzi, L., Leventakou, V., Vafeiadi, M., Koutra, K., Roumeliotaki, T., Chalkiadaki, G., Karachaliou, M., Daraki, V., Kyriklaki, A., Kampouri, M., Fthenou, E., Sarri, K., Vassilaki, M., Fasoulaki, M., Bitsios, P., Koutis, A., Stephanou, E.G., Kogevinas, M., 2017. Cohort Profile: The Mother-Child Cohort in Crete, Greece (Rhea Study). *Int. J. Epidemiol.* 46, 1392–1393k. <https://doi.org/10.1093/ije/dyx084>
- Chén, O.Y., Crainiceanu, C., Ogburn, E.L., Caffo, B.S., Wager, T.D., Lindquist, M.A., 2018. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* 19, 121–136. <https://doi.org/10.1093/biostatistics/kxx027>
- Chersich, M.F., Urban, M., Olivier, L., Davies, L.A., Chetty, C., Viljoen, D., 2012. Universal prevention is associated with lower prevalence of fetal alcohol spectrum disorders in Northern cape, South Africa: A multicentre before-after study. *Alcohol Alcohol.* 47, 67–74. <https://doi.org/10.1093/alc/alc145>
- Cho, S., Kim, K., Kim, Y.J., Lee, J.-K., Cho, Y.S., Lee, J.-Y., Han, B.-G., Kim, H., Ott, J., Park, T., 2010. Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. *Ann. Hum. Genet.* 74, 416–428. <https://doi.org/10.1111/j.1469-1809.2010.00597.x>
- Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72, 3–25. <https://doi.org/10.1111/j.1467-9868.2009.00723.x>
- Chung, M.K., Buck Louis, G.M., Kannan, K., Patel, C.J., 2019. Exposome-wide association study of semen quality: Systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes. *Environ. Int.* 125, 505–514. <https://doi.org/10.1016/j.envint.2018.11.037>
- Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge Books.
- Courvoisier, D.S., Combescure, C., Agoritsas, T., Gayet-Ageron, A., Perneger, T. V., 2011. Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *J. Clin. Epidemiol.* 64, 993–1000. <https://doi.org/10.1016/j.jclinepi.2010.11.012>
- Crews, D., Gore, A.C., 2011. Life imprints: living in a contaminated world. *Environ. Health Perspect.* 119, 1208–10. <https://doi.org/10.1289/ehp.1103451>

- Dadvand, P., Ostro, B., Figueras, F., Foraster, M., Basagaña, X., Valentín, A., Martínez, D., Beelen, R., Cirach, M., Hoek, G., Jerrett, M., Brunekreef, B., Nieuwenhuijsen, M.J., 2014. Residential proximity to major roads and term low birth weight: The roles of air pollution, heat, noise, and road-adjacent trees. *Epidemiology*. <https://doi.org/10.1097/EDE.0000000000000107>
- Damiani, G., Bragazzi, N.L., McCormick, T.S., Pigatto, P.D.M., Leone, S., Pacifico, A., Todorovic, D., Di Franco, S., Alfieri, A., Fiore, M., 2020. Gut microbiota and nutrient interactions with skin in psoriasis: A comprehensive review of animal and human studies. *World J. Clin. Cases* 8, 1002–1012. <https://doi.org/10.12998/wjcc.v8.i6.1002>
- de Onis, M., Onyango, A.W., Borghi, E., Siyam, A., Nishida, C., Siekmann, J., 2007. Development of a WHO growth reference for school-aged children and adolescents. *Bull. World Health Organ.* 85, 660–7.
- DeSocio, J.E., 2018. Epigenetics, maternal prenatal psychosocial stress, and infant mental health. *Arch. Psychiatr. Nurs.* 32, 901–906. <https://doi.org/10.1016/j.apnu.2018.09.001>
- Donoho, D., Jin, J., 2008. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. U. S. A.* 105, 14790–14795. <https://doi.org/10.1073/pnas.0807471105>
- Dunkel Schetter, C., Tanner, L., 2012. Anxiety, depression and stress in pregnancy: Implications for mothers, children, research, and practice. *Curr. Opin. Psychiatry*. <https://doi.org/10.1097/YCO.0b013e3283503680>
- Dunn, O.J., 1961. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* 56, 52–64. <https://doi.org/10.1080/01621459.1961.10482090>
- Efron, B., Tibshirani, R.J., 1993. An introduction to Bootstrap. *An Introd. to Bootstrap*.
- Elisseeff, A., 2005. Stability of Randomized Learning Algorithms. *J. Mach. Learn. Res.* 6, 55–79.
- Fan, J., Lv, J., 2010. A Selective Overview of Variable Selection in High Dimensional Feature Space. *Stat. Sin.* 20, 101–148.
- Fan, Y., Demirkaya, E., Lv, J., 2019. Nonuniformity of p-values can occur early in diverging dimensions. *J. Mach. Learn. Res.* 20, 1–33.
- Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, Mattias, Grankvist, K., Johansson, Mikael, Assumma, M.B., Naccarati, A., Chadeau-Hyam, M., Ala, U., Faltus, C., Kaaks, R., Risch, A., De Stavola, B., Hodge, A., Giles, G.G., Southey, M.C., Relton, C.L., Haycock, P.C., Lund, E., Polidoro, S., Sandanger, T.M., Severi, G., Vineis, P., 2015. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* 6, 10192. <https://doi.org/10.1038/ncomms10192>
- Feil, R., Fraga, M.F., 2012. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* 13, 97–109. <https://doi.org/10.1038/nrg3142>
- Finucane, M.M., Stevens, G.A., Cowan, M.J., Danaei, G., Lin, J.K., Paciorek, C.J., Singh, G.M., Gutierrez, H.R., Lu, Y., Bahalim, A.N., Farzadfar, F., Riley, L.M., Ezzati, M., 2011. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9·1 million participants. *Lancet* 377, 557–567. [https://doi.org/10.1016/S0140-6736\(10\)62037-5](https://doi.org/10.1016/S0140-6736(10)62037-5)
- Forns, J., Mandal, S., Iszatt, N., Polder, A., Thomsen, C., Lyche, J.L., Stigum, H., Vermeulen, R., Eggesbø, M., 2016. Novel application of statistical methods for analysis of multiple

toxicants identifies DDT as a risk factor for early child behavioral problems. Environ. Res. 151, 91–100. <https://doi.org/10.1016/j.envres.2016.07.014>

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. <https://doi.org/10.18637/jss.v033.i01>

Friedman, J., Hastie, T., Tibshirani, R., Simon, N., Narasimhan, B., Qian, J., 2019. Package ‘
glmnet’.

Gakidou, E., Afshin, A., Abajobir, A.A., Abate, K.H., Abbafati, C., Abbas, K.M., Abd-Allah, F., Abdulle, A.M., Abera, S.F., Aboyans, V., Abu-Raddad, L.J., Abu-Rmeileh, N.M.E., Abyu, G.Y., Adedeji, I.A., Adetokunboh, O., Afarideh, M., Agrawal, A., Agrawal, S., Ahmad Kiadaliri, A., Ahmadieh, H., Ahmed, M.B., Aichour, A.N., Aichour, I., Aichour, M.T.E., Akinyemi, R.O., Akseer, N., Alahdab, F., Al-Aly, Z., Alam, K., Alam, N., Alam, T., Alasfoor, D., Alene, K.A., Ali, K., Alizadeh-Navaei, R., Alkerwi, A., Alla, F., Allebeck, P., Al-Raddadi, R., Alsharif, U., Altirkawi, K.A., Alvis-Guzman, N., Amare, A.T., Amini, E., Ammar, W., Amoako, Y.A., Ansari, H., Antó, J.M., Antonio, C.A.T., Anwari, P., Arian, N., Ärnlöv, J., Artaman, A., Aryal, K.K., Asayesh, H., Asgedom, S.W., Atey, T.M., Avila-Burgos, L., Avokpaho, E.F.G.A., Awasthi, A., Azzopardi, P., Bacha, U., Badawi, A., Balakrishnan, K., Ballew, S.H., Barac, A., Barber, R.M., Barker-Collo, S.L., Bärnighausen, T., Barquera, S., Barregard, L., Barrero, L.H., Batis, C., Battle, K.E., Baune, B.T., Beardsley, J., Bedi, N., Beghi, E., Bell, M.L., Bennett, D.A., Bennett, J.R., Bensenor, I.M., Berhane, A., Berhe, D.F., Bernabé, E., Betsu, B.D., Beuran, M., Beyene, A.S., Bhansali, A., Bhutta, Z.A., Bikbov, B., Birungi, C., Biryukov, S., Blosser, C.D., Boneya, D.J., Bou-Orm, I.R., Brauer, M., Breitborde, N.J.K., Brenner, H., Brugha, T.S., Bulto, L.N.B., Baumgartner, B.R., Butt, Z.A., Cahuana-Hurtado, L., Cárdenas, R., Carrero, J.J., Castañeda-Orjuela, C.A., Catalá-López, F., Cercy, K., Chang, H.Y., Charlson, F.J., Chimed-Ochir, O., Chisumpa, V.H., Chittheer, A.A., Christensen, H., Christopher, D.J., Cirillo, M., Cohen, A.J., Comfort, H., Cooper, C., Coresh, J., Cornaby, L., Cortesi, P.A., Criqui, M.H., Crump, J.A., Dandona, L., Dandona, R., Das Neves, J., Davey, G., Davitoiu, D. V., Davletov, K., De Courten, B., Degenhardt, L., Deiparine, S., Dellavalle, R.P., Deribe, K., Deshpande, A., Dharmaratne, S.D., Ding, E.L., Djalalinia, S., Do, H.P., Dokova, K., Doku, D.T., Dorsey, E.R., Driscoll, T.R., Dubey, M., Duncan, B.B., Duncan, S., Ebert, N., Ebrahimi, H., El-Khatib, Z.Z., Enayati, A., Endries, A.Y., Ermakov, S.P., Erskine, H.E., Eshrat, B., Eskandarieh, S., Esteghamati, A., Estep, K., Faraon, E.J.A., Farinha, C.S.E.S., Faro, A., Farzadfar, F., Fay, K., Feigin, V.L., Fereshtehnejad, S.M., Fernandes, J.C., Ferrari, A.J., Feyissa, T.R., Filip, I., Fischer, F., Fitzmaurice, C., Flaxman, A.D., Foigt, N., Foreman, K.J., Frostad, J.J., Fullman, N., Fürst, T., Furtado, J.M., Ganji, M., Garcia-Basteiro, A.L., Gebrehiwot, T.T., Geleijnse, J.M., Geleto, A., Gemechu, B.L., Gesesew, H.A., Gething, P.W., Ghajar, A., Gibney, K.B., Gill, P.S., Gillum, R.F., Giref, A.Z., Gishu, M.D., Giussani, G., Godwin, W.W., Gona, P.N., Goodridge, A., Gopalani, S.V., Goryakin, Y., Goulart, A.C., Graetz, N., Gugnani, H.C., Guo, J., Gupta, R., Gupta, T., Gupta, V., Gutiérrez, R.A., Hachinski, V., Hafezi-Nejad, N., Hailu, G.B., Hamadeh, R.R., Hamidi, S., Hammami, M., Handal, A.J., Hankey, G.J., Harb, H.L., Hareri, H.A., Hassanvand, M.S., Havmoeller, R., Hawley, C., Hay, S.I., Hedayati, M.T., Hendrie, D., Heredia-Pi, I.B., Hoek, H.W., Horita, N., Hosgood, H.D., Hostiuc, S., Hoy, D.G., Hsairi, M., Hu, G., Huang, H., Huang, J.J., Iburg, K.M., Ikeda, C., Inoue, M., Irvine, C.M.S., Jackson, M.D., Jacobsen, K.H., Jahanmehr, N., Jakovljevic, M.B., Jauregui, A., Javanbakht, M., Jeemon, P., Johansson, L.R.K., Johnson, C.O., Jonas, J.B., Jürisson, M., Kabir, Z., Kadel, R., Kahsay, A., Kamal, R., Karch, A., Karema, C.K., Kasaeian, A., Kassebaum, N.J., Kastor, A., Katikireddi, S.V., Kawakami, N., Keiyoro, P.N., Kelbore, S.G., Kemmer, L., Kengne, A.P., Kesavachandran, C.N., Khader, Y.S., Khalil, I.A., Khan, E.A., Khang, Y.H., Khosravi, A., Khubchandani, J., Kieling, C., Kim, D., Kim, J.Y., Kim, Y.J., Kimokoti, R.W., Kinfu, Y., Kisa, A., Kissimova-Skarbek, K.A., Kivimaki, M., Knibbs, L.D.,

Knudsen, A.K., Kopec, J.A., Kosen, S., Koul, P.A., Koyanagi, A., Kravchenko, M., Krohn, K.J., Kromhout, H., Kuaté Defo, B., Kucuk Bicer, B., Kumar, G.A., Kutz, M., Kyu, H.H., Lal, D.K., Laloo, R., Lallukka, T., Lan, Q., Lansingh, V.C., Larsson, A., Lee, A., Lee, P.H., Leigh, J., Leung, J., Levi, M., Li, Yichong, Li, Yongmei, Liang, X., Liben, M.L., Linn, S., Liu, P., Lodha, R., Logroscino, G., Looker, K.J., Lopez, A.D., Lorkowski, S., Lotufo, P.A., Lozano, R., Lunevicius, R., Macarayan, E.R.K., Magdy Abd El Razek, H., Magdy Abd El Razek, M., Majdan, M., Majdzadeh, R., Majeed, A., Malekzadeh, R., Malhotra, R., Malta, D.C., Mamun, A.A., Manguerra, H., Mantovani, L.G., Mapoma, C.C., Martin, R. V., Martinez-Raga, J., Martins-Melo, F.R., Mathur, M.R., Matsushita, K., Matzopoulos, R., Mazidi, M., McAlinden, C., McGrath, J.J., Mehata, S., Mehndiratta, M.M., Meier, T., Melaku, Y.A., Memiah, P., Memish, Z.A., Mendoza, W., Mengesha, M.M., Mensah, G.A., Mensink, G.B.M., Mereta, S.T., Meretoja, A., Meretoja, T.J., Mezgebe, H.B., Micha, R., Millear, A., Miller, T.R., Minnig, S., Mirarefin, M., Mirrakhimov, E.M., Misganaw, A., Mishra, S.R., Mohammad, K.A., Mohammed, K.E., Mohammed, S., Mohamed Ibrahim, N., Mohan, M.B.V., Mokdad, A.H., Monasta, L., Montañez Hernandez, J.C., Montico, M., Moradi-Lakeh, M., Moraga, P., Morawska, L., Morrison, S.D., Mountjoy-Venning, C., Mueller, U.O., Mullany, E.C., Muller, K., Murthy, G.V.S., Musa, K.I., Naghavi, M., Naheed, A., Nangia, V., Natarajan, G., Negoi, I., Negoi, R.I., Nguyen, C.T., Nguyen, G., Nguyen, M., Nguyen, Q., Le, Nguyen, T.H., Nichols, E., Ningrum, D.N.A., Nomura, M., Nong, V.M., Norheim, O.F., Norrving, B., Noubiap, J.J.N., Obermeyer, C.M., Ogbo, F.A., Oh, I.H., Oladimeji, O., Olagunju, A.T., Olagunju, T.O., Olivares, P.R., Olsen, H.E., Olusanya, B.O., Olusanya, J.O., Opio, J.N., Oren, E., Ortiz, A., Ota, E., Owolabi, M.O., Pa, M., Pacella, R.E., Pana, A., Panda, B.K., Panda-Jonas, S., Pandian, J.D., Papachristou, C., Park, E.K., Parry, C.D., Patten, S.B., Patton, G.C., Pereira, D.M., Perico, N., Pesudovs, K., Petzold, M., Phillips, M.R., Pillay, J.D., Piradov, M.A., Pishgar, F., Plass, D., Pletcher, M.A., Polinder, S., Popova, S., Poulton, R.G., Pourmalek, F., Prasad, N., Purcell, C., Qorbani, M., Radfar, A., Rafay, A., Rahimi-Movaghar, A., Rahimi-Movaghar, V., Rahman, M., Rahman, M.H.U., Rahman, M.A., Rai, R.K., Rajsic, S., Ram, U., Rawaf, S., Rehm, C.D., Rehm, J., Reiner, R.C., Reitsma, M.B., Reynales-Shigematsu, L.M., Remuzzi, G., Renzaho, A.M.N., Resnikoff, S., Rezaei, S., Ribeiro, A.L., Rivera, J.A., Roba, K.T., Rojas-Rueda, D., Roman, Y., Room, R., Roshandel, G., Roth, G.A., Rothenbacher, D., Rubagotti, E., Rushton, L., Sadat, N., Safdarian, M., Safi, S., Safiri, S., Sahathevan, R., Salama, J., Salomon, J.A., Samy, A.M., Sanabria, J.R., Sanchez-Niño, M.D., Sánchez-Pimienta, T.G., Santomauro, D., Santos, I.S., Santric Milicevic, M.M., Sartorius, B., Satpathy, M., Sawhney, M., Saxena, S., Schaeffner, E., Schmidt, M.I., Schneider, I.J.C., Schutte, A.E., Schwebel, D.C., Schwendicke, F., Seedat, S., Sepanlou, S.G., Serdar, B., Servan-Mori, E.E., Shaddick, G., Shaheen, A., Shahraz, S., Shaikh, M.A., Shamah Levy, T., Shamsipour, M., Shamsizadeh, M., Shariful Islam, S.M., Sharma, J., Sharma, R., She, J., Shen, J., Shi, P., Shibuya, K., Shields, C., Shiferaw, M.S., Shigematsu, M., Shin, M.J., Shiri, R., Shirkoohi, R., Shishani, K., Shoman, H., Shrime, M.G., Sigfusdottir, I.D., Silva, D.A.S., Silva, J.P., Silveira, D.G.A., Singh, J.A., Singh, V., Sinha, D.N., Skiadaresi, E., Slepak, E.L., Smith, D.L., Smith, M., Sobaih, B.H.A., Sobngwi, E., Soneji, S., Sorensen, R.J.D., Sposato, L.A., Sreeramareddy, C.T., Srinivasan, V., Steel, N., Stein, D.J., Steiner, C., Steinke, S., Stokes, M.A., Strub, B., Subart, M., Sufiyan, M.B., Suliankatchi, R.A., Sur, P.J., Swaminathan, S., Sykes, B.L., Szoëke, C.E.I., Tabarés-Seisdedos, R., Tadakamadla, S.K., Takahashi, K., Takala, J.S., Tandon, N., Tanner, M., Tarekegn, Y.L., Tavakkoli, M., Tegegne, T.K., Tehrani-Banihashemi, A., Terkawi, A.S., Tessema, B., Thakur, J.S., Thamsuwan, O., Thankappan, K.R., Theis, A.M., Thomas, M.L., Thomson, A.J., Thrift, A.G., Tillmann, T., Tobe-Gai, R., Tobollik, M., Tollanes, M.C., Tonelli, M., Topor-Madry, R., Torre, A., Tortajada, M., Touvier, M., Tran, B.X., Truelsen, T., Tuem, K.B., Tuzcu, E.M., Tyrovolas, S., Ukwaja, K.N., Uneke, C.J., Updike, R., Uthman, O.A., Van Boven, J.F.M., Van Donkelaar, A., Varughese, S., Vasankari, T., Veerman, L.J., Venkateswaran, V.,

Venketasubramanian, N., Violante, F.S., Vladimirov, S.K., Vlassov, V.V., Vollset, S.E., Vos, T., Wadilo, F., Wakayo, T., Wallin, M.T., Wang, Y.P., Weichenthal, S., Weiderpass, E., Weintraub, R.G., Weiss, D.J., Werdecker, A., Westerman, R., Whiteford, H.A., Wiysonge, C.S., Woldeyes, B.G., Wolfe, C.D.A., Woodbrook, R., Workicho, A., Wulf Hanson, S., Xavier, D., Xu, G., Yadgir, S., Yakob, B., Yan, L.L., Yaseri, M., Yimam, H.H., Yip, P., Yonemoto, N., Yoon, S.J., Yotebieng, M., Younis, M.Z., Zaidi, Z., El Sayed Zaki, M., Zavala-Arciniega, L., Zhang, X., Zimsen, S.R.M., Zipkin, B., Zodpey, S., Lim, S.S., Murray, C.J.L., 2017. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 390, 1345–1422. [https://doi.org/10.1016/S0140-6736\(17\)32366-8](https://doi.org/10.1016/S0140-6736(17)32366-8)

Galhardi, C.M., Diniz, Y.S., Faine, L.A., Rodrigues, H.G., Burneiko, R.C.M., Ribas, B.O., Novelli, E.L.B., 2004. Toxicity of copper intake: Lipid profile, oxidative stress and susceptibility to renal dysfunction. *Food Chem. Toxicol.* <https://doi.org/10.1016/j.fct.2004.07.020>

Gängler, S., Waldenberger, M., Artati, A., Adamski, J., van Bolhuis, J.N., Sørgjerd, E.P., van Vliet-Ostaptchouk, J., Makris, K.C., 2019. Exposure to disinfection byproducts and risk of type 2 diabetes: a nested case-control study in the HUNT and Lifelines cohorts. *Metabolomics* 15. <https://doi.org/10.1007/s11306-019-1519-0>

Gascon, M., Sunyer, J., Casas, M., Martínez, D., Ballester, F., Basterrechea, M., Bonde, J.P., Chatzi, L., Chevrier, C., Eggesbø, M., Esplugues, A., Govarts, E., Hannu, K., Ibarluzea, J., Kasper-Sonnenberg, M., Klümper, C., Koppen, G., Nieuwenhuijsen, M.J., Palkovicova, L., Pelé, F., Polder, A., Schoeters, G., Torrent, M., Trnovec, T., Vassilaki, M., Vrijheid, M., 2014. Prenatal exposure to DDE and PCB 153 and respiratory health in early childhood: A meta-analysis. *Epidemiology* 25, 544–553. <https://doi.org/10.1097/EDE.0000000000000097>

Gascon, M., Vrijheid, M., Nieuwenhuijsen, M.J., 2016. The Built Environment and Child Health: An Overview of Current Evidence. *Curr. Environ. Heal. reports*. <https://doi.org/10.1007/s40572-016-0094-z>

Ghiara, V., 2019. Inferring causation from big data in the social sciences. Doctor of Philosophy (PhD) thesis, University of Kent,

Govarts, E., Iszatt, N., Trnovec, T., de Cock, M., Eggesbø, M., Palkovicova Murinova, L., van de Bor, M., Guxens, M., Chevrier, C., Koppen, G., Lamoree, M., Hertz-Pannier, I., Lopez-Espinosa, M.J., Lertxundi, A., Grimalt, J.O., Torrent, M., Goñi-Irigoyen, F., Vermeulen, R., Legler, J., Schoeters, G., 2018. Prenatal exposure to endocrine disrupting chemicals and risk of being born small for gestational age: Pooled analysis of seven European birth cohorts. *Environ. Int.* 115, 267–278. <https://doi.org/10.1016/j.envint.2018.03.017>

Govarts, E., Remy, S., Bruckers, L., Den Hond, E., Sioen, I., Nelen, V., Baeyens, W., Nawrot, T.S., Loots, I., Van Larebeke, N., Schoeters, G., 2016. Combined effects of prenatal exposures to environmental chemicals on birth weight. *Int. J. Environ. Res. Public Health* 13. <https://doi.org/10.3390/ijerph13050495>

Grazuleviciene, R., Danileviciute, A., Dedele, A., Vencloviene, J., Andrusaitite, S., Uždanaviciute, I., Nieuwenhuijsen, M.J., 2015. Surrounding greenness, proximity to city parks and pregnancy outcomes in Kaunas cohort study. *Int. J. Hyg. Environ. Health* 218, 358–365. <https://doi.org/10.1016/j.ijheh.2015.02.004>

Greene, N.D.E., Leung, K.-Y., Copp, A.J., 2017. Inositol, neural tube closure and the prevention of neural tube defects. *Birth Defects Res.* 109, 68–80. <https://doi.org/10.1002/bdra.23533>

- Guxens, M., Ballester, F., Espada, M., Fernández, M.F., Grimalt, J.O., Ibarluzea, J., Olea, N., Rebagliato, M., Tardón, A., Torrent, M., Vioque, J., Vrijheid, M., Sunyer, J., 2012. Cohort Profile: The INMA—INFancia y Medio Ambiente—(Environment and Childhood) Project. *Int. J. Epidemiol.* 41, 930–940. <https://doi.org/10.1093/ije/dyr054>
- Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* 1157–1182.
- Han, J.C., Lawlor, D.A., Kimm, S.Y., 2010. Childhood obesity. *Lancet.* [https://doi.org/10.1016/S0140-6736\(10\)60171-7](https://doi.org/10.1016/S0140-6736(10)60171-7)
- Hansen, S., Strøm, M., Olsen, S.F., Maslova, E., Rantakokko, P., Kiviranta, H., Rytter, D., Bech, B.H., Hansen, L. V., Halldorsson, T.I., 2014. Maternal concentrations of persistent organochlorine pollutants and the risk of asthma in offspring: Results from a prospective cohort with 20 years of follow-up. *Environ. Health Perspect.* 122, 93–99. <https://doi.org/10.1289/ehp.1206397>
- Haug, L.S., Sakhi, A.K., Cequier, E., Casas, M., Maitre, L., Basagana, X., Andrusaityte, S., Chalkiadaki, G., Chatzi, L., Coen, M., de Bont, J., Dedele, A., Ferrand, J., Grazuleviciene, R., Gonzalez, J.R., Gutzkow, K.B., Keun, H., McEachan, R., Meltzer, H.M., Petraviciene, I., Robinson, O., Saulnier, P.J., Slama, R., Sunyer, J., Urquiza, J., Vafeiadi, M., Wright, J., Vrijheid, M., Thomsen, C., 2018. In-utero and childhood chemical exposome in six European mother-child cohorts. *Environ. Int.* 121, 751–763. <https://doi.org/10.1016/j.envint.2018.09.056>
- Heindel, J.J., Balbus, J., Birnbaum, L., Brune-Drissé, M.N., Grandjean, P., Gray, K., Landrigan, P.J., Sly, P.D., Suk, W., Slezakta, D.C., Thompson, C., Hanson, M., 2015. Developmental origins of health and disease: Integrating environmental influences. *Endocrinology* 156, 3416–3421. <https://doi.org/10.1210/EN.2015-1394>
- Hernán, M.A., Hernández-Díaz, S., Robins, J.M., 2004. A structural approach to selection bias. *Epidemiology* 15, 615–625. <https://doi.org/10.1097/01.ede.0000135174.63482.43>
- Hernán, M.A., Hsu, J., Healy, B., 2019. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks, Chance. <https://doi.org/10.1080/09332480.2019.1579578>
- Heude, Barbara, Forhan, Anne, Slama, Rémy, Douhaud, L., Bedel, S., Saurel-Cubizolles, M.-J., Hankard, Régis, Thiebaugeorges, Olivier, De Agostini, Maria, Annesi-Maesano, Isabella, Kaminski, Monique, Charles, Marie-Aline, Annesi-Maesano, I., Bernard, J., Botton, J., Charles, M-A, Dargent-Molina, P., de Lauzon-Guillain, B., Ducimetière, P., de Agostini, M., Foliguet, B., Forhan, A, Fritel, X., Germa, A., Goua, V., Hankard, R, Heude, B, Kaminski, M, Larroque, B., Lelong, N., Lepeule, J., Magnin, G., Marchand, L., Nabet, C., Pierre, F., Slama, R, Saurel-Cubizolles, M., Schweitzer, M., Thiebaugeorges, O, 2016. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *Int. J. Epidemiol.* 45, 353–363. <https://doi.org/10.1093/ije/dyv151>
- Hill, A.B., Doll, R., 1950. Smoking and carcinoma of the lung preliminary report. *Br. Med. J.* 2, 739–748. <https://doi.org/10.1136/bmj.2.4682.739>
- Ho, S.-M., Johnson, A., Tarapore, P., Janakiram, V., Zhang, X., Leung, Y.-K., 2012. Environmental Epigenetics and Its Implication on Disease Risk and Health Outcomes. *ILAR J.* 53, 289–305. <https://doi.org/10.1093/ilar.53.3-4.289>
- Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Applications to Nonorthogonal Problems.

Technometrics 12, 69. <https://doi.org/10.2307/1267352>

Hofhuis, W., de Jongste, J.C., Merkus, P.J.F.M., 2003. Adverse health effects of prenatal and postnatal tobacco smoke exposure on children. *Arch. Dis. Child.* 88, 1086–1090. <https://doi.org/10.1136/adc.88.12.1086>

Holtcamp, W., 2012. Obesogens: an environmental link to obesity. *Environ. Health Perspect.* 120, a62-8. <https://doi.org/10.1289/ehp.120-a62>

Höskuldsson, A., 1988. PLS regression methods. *J. Chemom.* 2, 211–228. <https://doi.org/10.1002/cem.1180020306>

Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A., 2010. Can shared-neighbor distances defeat the curse of dimensionality?, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Berlin, Heidelberg, pp. 482–500. https://doi.org/10.1007/978-3-642-13818-8_34

Huang, R., Chen, Q., Zhang, L., Luo, K., Chen, L., Zhao, S., Feng, L., Zhang, J., 2019. Prenatal exposure to perfluoroalkyl and polyfluoroalkyl substances and the risk of hypertensive disorders of pregnancy. *Environ. Heal. A Glob. Access Sci. Source* 18, 5. <https://doi.org/10.1186/s12940-018-0445-3>

Hume, D., 1740. An Abstract of a Book lately Published: Entitled A Treatise of Human Nature.

Janesick, A., Blumberg, B., 2011. Endocrine disrupting chemicals and the developmental programming of adipogenesis and obesity. *Birth Defects Res. Part C Embryo Today Rev.* 93, 34–50. <https://doi.org/10.1002/bdrc.20197>

Jeong, A., Fiorito, G., Keski-Rahkonen, P., Imboden, M., Kiss, A., Robinot, N., Gmuender, H., Vlaanderen, J., Vermeulen, R., Kyrtopoulos, S., Herceg, Z., Ghantous, A., Lovison, G., Galassi, C., Ranzi, A., Krogh, V., Grioni, S., Agnoli, C., Sacerdote, C., Mostafavi, N., Naccarati, A., Scalbert, A., Vineis, P., Probst-Hensch, N., 2018. Perturbation of metabolic pathways mediates the association of air pollutants with asthma and cardiovascular diseases. *Environ. Int.* 119, 334–345. <https://doi.org/10.1016/J.ENVINT.2018.06.025>

Joubert, B.R., Felix, J.F., Yousefi, P., Bakulski, K.M., Just, A.C., Breton, C., Reese, S.E., Markunas, C.A., Richmond, R.C., Xu, C.J., Küpers, L.K., Oh, S.S., Hoyo, C., Gruzieva, O., Söderhäll, C., Salas, L.A., Baïz, N., Zhang, H., Lepeule, J., Ruiz, C., Ligthart, S., Wang, T., Taylor, J.A., Duijts, L., Sharp, G.C., Jankipersadsing, S.A., Nilsen, R.M., Vaez, A., Fallin, M.D., Hu, D., Litonjua, A.A., Fuemmeler, B.F., Huen, K., Kere, J., Kull, I., Munthe-Kaas, M.C., Gehring, U., Bustamante, M., Saurel-Coulibolles, M.J., Quraishi, B.M., Ren, J., Tost, J., Gonzalez, J.R., Peters, M.J., Häberg, S.E., Xu, Z., Van Meurs, J.B., Gaunt, T.R., Kerkhof, M., Corpeleijn, E., Feinberg, A.P., Eng, C., Baccarelli, A.A., Benjamin Neelon, S.E., Bradman, A., Merid, S.K., Bergström, A., Herceg, Z., Hernandez-Vargas, H., Brunekreef, B., Pinart, M., Heude, B., Ewart, S., Yao, J., Lemonnier, N., Franco, O.H., Wu, M.C., Hofman, A., McArdle, W., Van Der Vlies, P., Falahi, F., Gillman, M.W., Barcellos, L.F., Kumar, A., Wickman, M., Guerra, S., Charles, M.A., Holloway, J., Auffray, C., Tiemeier, H.W., Smith, G.D., Postma, D., Hivert, M.F., Eskenazi, B., Vrijheid, M., Arshad, H., Antó, J.M., Dehghan, A., Karmaus, W., Annesi-Maesano, I., Sunyer, J., Ghantous, A., Pershagen, G., Holland, N., Murphy, S.K., Demeo, D.L., Burchard, E.G., Ladd-Acosta, C., Snieder, H., Nystad, W., Koppelman, G.H., Relton, C.L., Jaddoe, V.W.V., Wilcox, A., Melén, E., London, S.J., 2016. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* 98, 680–696. <https://doi.org/10.1016/j.ajhg.2016.02.019>

- Khan, M.H.R., Bhadra, A., Howlader, T., 2016. Stability Selection for Lasso, Ridge and Elastic Net Implemented with AFT Models.
- Kindlund, K., Thomsen, S.F., Stensballe, L.G., Skytthe, A., Kyvik, K.O., Backer, V., Bisgaard, H., 2010. Birth weight and risk of asthma in 3 - 9-year-old twins: Exploring the fetal origins hypothesis. *Thorax* 65, 146–149. <https://doi.org/10.1136/thx.2009.117101>
- Kreiner-Møller, E., Bisgaard, H., Bønnelykke, K., 2014. Prenatal and postnatal genetic influence on lung function development. *J. Allergy Clin. Immunol.* 134, 1036-1042.e15. <https://doi.org/10.1016/j.jaci.2014.04.003>
- Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* 6, 1–15. <https://doi.org/10.1186/1758-2946-6-10>
- Küpers, L.K., Monnereau, C., Sharp, G.C., Yousefi, P., Salas, L.A., Ghantous, A., Page, C.M., Reese, S.E., Wilcox, A.J., Czamara, D., Starling, A.P., Novoloaca, A., Lent, S., Roy, R., Hoyo, C., Breton, C. V., Allard, C., Just, A.C., Bakulski, K.M., Holloway, J.W., Everson, T.M., Xu, C.J., Huang, R.C., van der Plaat, D.A., Wielscher, M., Merid, S.K., Ullemar, V., Rezwan, F.I., Lahti, J., van Dongen, J., Langie, S.A.S., Richardson, T.G., Magnus, M.C., Nohr, E.A., Xu, Z., Duijts, L., Zhao, S., Zhang, W., Plusquin, M., DeMeo, D.L., Solomon, O., Heimovaara, J.H., Jima, D.D., Gao, L., Bustamante, M., Perron, P., Wright, R.O., Hertz-Pannier, I., Zhang, H., Karagas, M.R., Gehring, U., Marsit, C.J., Beilin, L.J., Vonk, J.M., Jarvelin, M.R., Bergström, A., Örtqvist, A.K., Ewart, S., Villa, P.M., Moore, S.E., Willemse, G., Standaert, A.R.L., Häberg, S.E., Sørensen, T.I.A., Taylor, J.A., Räikkönen, K., Yang, I. V., Kechris, K., Nawrot, T.S., Silver, M.J., Gong, Y.Y., Richiardi, L., Kogevinas, M., Litonjua, A.A., Eskenazi, B., Huen, K., Mbarek, H., Maguire, R.L., Dwyer, T., Vrijheid, M., Bouchard, L., Baccarelli, A.A., Croen, L.A., Karmaus, W., Anderson, D., de Vries, M., Sebert, S., Kere, J., Karlsson, R., Arshad, S.H., Hämäläinen, E., Routledge, M.N., Boomsma, D.I., Feinberg, A.P., Newschaffer, C.J., Govarts, E., Moisse, M., Fallin, M.D., Melén, E., Prentice, A.M., Kajantie, E., Almqvist, C., Oken, E., Dabelea, D., Boezen, H.M., Melton, P.E., Wright, R.J., Koppelman, G.H., Trevisi, L., Hivert, M.F., Sunyer, J., Munthe-Kaas, M.C., Murphy, S.K., Corpeleijn, E., Wiemels, J., Holland, N., Herceg, Z., Binder, E.B., Davey Smith, G., Jaddoe, V.W.V., Lie, R.T., Nystad, W., London, S.J., Lawlor, D.A., Relton, C.L., Snieder, H., Felix, J.F., 2019. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nat. Commun.* 10, 1–11. <https://doi.org/10.1038/s41467-019-10967-3>
- Küpers, L.K., Xu, X., Jankipersadsing, S.A., Vaez, A., La Bastide-van Gemert, S., Scholtens, S., Nolte, I.M., Richmond, R.C., Relton, C.L., Felix, J.F., Duijts, L., Van Meurs, J.B., Tiemeier, H., Jaddoe, V.W., Wang, X., Corpeleijn, E., Snieder, H., 2015. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int. J. Epidemiol.* 44, 1224–1237. <https://doi.org/10.1093/ije/dyv048>
- Lachowycz, K., Jones, A.P., 2011. Greenspace and obesity: A systematic review of the evidence. *Obes. Rev.* 12. <https://doi.org/10.1111/j.1467-789X.2010.00827.x>
- Landeira, D., Bagci, H., Malinowski, A.R., Brown, K.E., Soza-Ried, J., Feytout, A., Webster, Z., Ndjetehe, E., Cantone, I., Asenjo, H.G., Brockdorff, N., Carroll, T., Merkenschlager, M., Fisher, A.G., 2015. Jarid2 coordinates Nanog expression and PCP/Wnt signaling required for efficient ESC differentiation and early embryo development. *Cell Rep.* 12, 573–586. <https://doi.org/10.1016/j.celrep.2015.06.060>
- Latzin, P., Röösli, M., Huss, A., Kuehni, C.E., Frey, U., 2009. Air pollution during pregnancy and lung function in newborns: A birth cohort study. *Eur. Respir. J.* 33, 594–603.

<https://doi.org/10.1183/09031936.00084008>

Lazarevic, N., Barnett, A.G., Sly, P.D., Knibbs, L.D., 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>

Lee, H.W., Lawton, C., Na, Y.J., Yoon, S., 2013. Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery. *Stat. Appl. Genet. Mol. Biol.* 12, 207–223. <https://doi.org/10.1515/sagmb-2012-0067>

Leng, C., Lin, Y., Wahba, G., 2006. a Note on the Lasso and Related Procedures. *Stat. Sin.* 16, 1273–1284.

Lenters, V., Portengen, L., Rignell-Hydbom, A., Jönsson, B.A.G., Lindh, C.H., Piersma, A.H., Toft, G., Bonde, J.P., Heederik, D., Rylander, L., Vermeulen, R., 2016. Prenatal phthalate, perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: Multi-pollutant models based on elastic net regression. *Environ. Health Perspect.* 124, 365–372. <https://doi.org/10.1289/ehp.1408933>

Lenters, V., Portengen, L., Smit, L.A.M., Jönsson, B.A.G., Giwercman, A., Rylander, L., Lindh, C.H., Spanò, M., Pedersen, H.S., Ludwicki, J.K., Chumak, L., Piersma, A.H., Toft, G., Bonde, J.P., Heederik, D., Vermeulen, R., 2015. Phthalates, perfluoroalkyl acids, metals and organochlorines and reproductive function: A multipollutant assessment in Greenlandic, Polish and Ukrainian men. *Occup. Environ. Med.* 72, 385–393. <https://doi.org/10.1136/oemed-2014-102264>

Lenters, V., Vermeulen, R., Portengen, L., 2018. Performance of variable selection methods for assessing the health effects of correlated exposures in case-control studies. *Occup. Environ. Med.* 75, 522–529. <https://doi.org/10.1136/oemed-2016-104231>

Leonelli, S., 2014. What difference does quantity make? On the epistemology of Big Data in biology. *Big Data Soc.* 1, 205395171453439. <https://doi.org/10.1177/2053951714534395>

Li, M.-X., Yeung, J.M.Y., Cherny, S.S., Sham, P.C., 2012. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* 131, 747–56. <https://doi.org/10.1007/s00439-011-1118-2>

Li, X., Hawkins, G.A., Ampleford, E.J., Moore, W.C., Li, H., Hastie, A.T., Howard, T.D., Boushey, H.A., Busse, W.W., Calhoun, W.J., Castro, M., Erzurum, S.C., Israel, E., Lemanske, R.F., Szeffler, S.J., Wasserman, S.I., Wenzel, S.E., Peters, S.P., Meyers, D.A., Bleecker, E.R., Bleecker, E.R., 2013. Genome-wide association study identifies TH1 pathway genes associated with lung function in asthmatic patients. *J. Allergy Clin. Immunol.* 132, 313–20.e15. <https://doi.org/10.1016/j.jaci.2013.01.051>

Lim, C., Yu, B., 2016. Estimation Stability With Cross-Validation (ESCV). *J. Comput. Graph. Stat.* 25, 464–492. <https://doi.org/10.1080/10618600.2015.1020159>

Little, R.E., 1977. Moderate alcohol use during pregnancy and decreased infant birth weight. *Am. J. Public Health* 67, 1154–1156. <https://doi.org/10.2105/AJPH.67.12.1154>

Louis, G.M.B., Sundaram, R., Schisterman, E.F., Sweeney, A.M., Lynch, C.D., Gore-Langton, R.E., Maisog, J., Kim, S., Chen, Z., Barr, D.B., 2013. Persistent Environmental Pollutants and Couple Fecundity: The LIFE Study. *Environ. Health Perspect.* 121, 231–236. <https://doi.org/10.1289/ehp.1205301>

- Lyon-Caen, S., Siroux, V., Lepeule, J., Lorimier, P., Hainaut, P., Mossuz, P., Quentin, J., Supernant, K., Meary, D., Chaperot, L., Bayat, S., Cassee, F., Valentino, S., Couturier-Tarrade, A., Rousseau-Ralliard, D., Chavatte-Palmer, P., Philippat, C., Pin, I., Slama, R., 2019. Deciphering the impact of early-life exposures to highly variable environmental factors on foetal and child health: Design of SEPAGES couple-child cohort. *Int. J. Environ. Res. Public Health* 16. <https://doi.org/10.3390/ijerph16203888>
- MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., Sheets, V., 2002. A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* 7, 83–104.
- Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A.K., Handal, M., Haugen, M., Høiseth, G., Knudsen, G.P., Paltiel, L., Schreuder, P., Tambs, K., Vold, L., Stoltenberg, C., 2016. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *Int. J. Epidemiol.* 45, 382–388. <https://doi.org/10.1093/ije/dyw029>
- Magurran, A., 2004. Measuring biological diversity. Blackwell Publ. Oxford, UK.
- Maitre, L., de Bont, J., Casas, M., Robinson, O., Aasvang, G.M., Agier, L., Andrušaitytė, S., Ballester, F., Basagaña, X., Borràs, E., Brochot, C., Bustamante, M., Carracedo, A., de Castro, M., Dedele, A., Donaire-Gonzalez, D., Estivill, X., Evandt, J., Fossati, S., Giorgis-Allemand, L., R Gonzalez, J., Granum, B., Grazulevičiene, R., Bjerve Gützkow, K., Småstuen Haug, L., Hernandez-Ferrer, C., Heude, B., Ibarluzea, J., Julvez, J., Karachaliou, M., Keun, H.C., Hjertager Krog, N., Lau, C.-H.E., Leventakou, V., Lyon-Caen, S., Manzano, C., Mason, D., McEachan, R., Meltzer, H.M., Petraciūnienė, I., Quentin, J., Roumeliotaki, T., Sabido, E., Saulnier, P.-J., Siskos, A.P., Siroux, V., Sunyer, J., Tamayo, I., Urquiza, J., Vafeiadi, M., van Gent, D., Vives-Usano, M., Waiblinger, D., Warembourg, C., Chatzi, L., Coen, M., van den Hazel, P., Nieuwenhuijsen, M.J., Slama, R., Thomsen, C., Wright, J., Vrijheid, M., 2018. Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. *BMJ Open* 8, e021311. <https://doi.org/10.1136/bmjopen-2017-021311>
- Manrai, A.K., Cui, Y., Bushel, P.R., Hall, M., Karakitsios, S., Mattingly, C.J., Ritchie, M., Schmitt, C., Sarigiannis, D.A., Thomas, D.C., Wishart, D., Balshaw, D.M., Patel, C.J., 2017. Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. *Annu. Rev. Public Heal.* 38, 279–94. <https://doi.org/10.1146/annurev-publhealth>
- Marioni, R.E., McRae, A.F., Bressler, J., Colicino, E., Hannon, E., Li, S., Prada, D., Smith, J.A., Trevisi, L., Tsai, P.-C., Vojinovic, D., Simino, J., Levy, D., Liu, C., Mendelson, M., Satizabal, C.L., Yang, Q., Jhun, M.A., Kardia, S.L.R., Zhao, W., Bandinelli, S., Ferrucci, L., Hernandez, D.G., Singleton, A.B., Harris, S.E., Starr, J.M., Kiel, D.P., McLean, R.R., Just, A.C., Schwartz, J., Spiro, A., Vokonas, P., Amin, N., Ikram, M.A., Uitterlinden, A.G., van Meurs, J.B.J., Spector, T.D., Steves, C., Baccarelli, A.A., Bell, J.T., van Duijn, C.M., Fornage, M., Hsu, Y.-H., Mill, J., Mosley, T.H., Seshadri, S., Deary, I.J., 2018. Meta-analysis of epigenome-wide association studies of cognitive abilities. *Mol. Psychiatry* 1. <https://doi.org/10.1038/s41380-017-0008-y>
- Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that Will Transform How We Live, Work and Think. London, UK John Murray.
- McAllister, E.J., Dhurandhar, N. V., Keith, S.W., Aronne, L.J., Barger, J., Baskin, M., Benca, R.M., Biggio, J., Boggiano, M.M., Eisenmann, J.C., Elobeid, M., Fontaine, K.R., Gluckman, P., Hanlon, E.C., Katzmarzyk, P., Pietrobelli, A., Redden, D.T., Ruden, D.M., Wang, C., Waterland, R.A., Wright, S.M., Allison, D.B., 2009. Ten putative contributors to the obesity epidemic. *Crit. Rev. Food Sci. Nutr.* 49, 868–913.

<https://doi.org/10.1080/10408390903372599>

Meinshausen, N., Bühlmann, P., 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>

Michalowsky, L.A., Jones, P.A., 1989. DNA methylation and differentiation. *Environ. Health Perspect.* 80, 189–197. <https://doi.org/10.1289/ehp.8980189>

Mills, J.L., Graubard, B.I., Harley, E.E., Rhoads, G.G., Berendes, H.W., 1984. Maternal Alcohol Consumption and Birth Weight: How Much Drinking During Pregnancy Is Safe? *JAMA J. Am. Med. Assoc.* 252, 1875–1879. <https://doi.org/10.1001/jama.1984.03350140021018>

Mubeen, S., Hoyt, C.T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H., Domingo-Fernández, D., 2019. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Front. Genet.* 10, 1203. <https://doi.org/10.3389/fgene.2019.01203>

Mustieles, V., Fernández, M.F., Martin-Olmedo, P., González-Alzaga, B., Fontalba-Navas, A., Hauser, R., Olea, N., Arrebola, J.P., 2017. Human adipose tissue levels of persistent organic pollutants and metabolic syndrome components: Combining a cross-sectional with a 10-year longitudinal study using a multi-pollutant approach. *Environ. Int.* 104, 48–57. <https://doi.org/10.1016/j.envint.2017.04.002>

Nieuwenhuijsen, M.J., Agier, L., Basagaña, X., Urquiza, J., Tamayo-Uria, I., Giorgis-Allemand, L., Robinson, O., Sioux, V., Maitre, L., de Castro, M., Valentin, A., Donaire, D., Dadvand, P., Aasvang, G.M., Krog, N.H., Schwarze, P.E., Chatzi, L., Grazuleviciene, R., Andrusaityte, S., Dedele, A., McEachan, R., Wright, J., West, J., Ibarluzea, J., Ballester, F., Vrijheid, M., Slama, R., 2019. Influence of the Urban Exposome on Birth Weight. *Environ. Health Perspect.* 127, 047007. <https://doi.org/10.1289/EHP3971>

Nogueira, S., Sechidis, K., Brown, G., 2017. On the Stability of Feature Selection Algorithms. *J. Mach. Learn. Res.* 18, 6345–6398. <https://doi.org/10.1162/153244320-947007>

Park, M.H., Falconer, C., Viner, R.M., Kinra, S., 2012. The impact of childhood obesity on morbidity and mortality in adulthood: A systematic review. *Obes. Rev.* 13, 985–1000. <https://doi.org/10.1111/j.1467-789X.2012.01015.x>

Park, S.S., Skaar, D.A., Jirtle, R.L., Hoyo, C., 2017. Epigenetics, obesity and early-life cadmium or lead exposure. *Epigenomics* 9, 57–75. <https://doi.org/10.2217/epi-2016-0047>

Parker, J.D., Woodruff, T.J., Basu, R., Schoendorf, K.C., 2005. Air pollution and birth weight among term infants in California. *Pediatrics* 115, 121–128. <https://doi.org/10.1542/peds.2004-0889>

Parkhomenko, E., Tritchler, D., Beyene, J., 2007. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.* 1, S119. <https://doi.org/10.1186/1753-6561-1-s1-s119>

Patel, C.J., 2017. Analytic Complexity and Challenges in Identifying Mixtures of Exposures Associated with Phenotypes in the Exposome Era. *Curr. Epidemiol. Reports* 4, 22–30. <https://doi.org/10.1007/s40471-017-0100-5>

Patel, C.J., Bhattacharya, J., Butte, A.J., Zeggini, E., Freathy, R., 2010. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. *PLoS One* 5, e10746. <https://doi.org/10.1371/journal.pone.0010746>

Patel, C.J., Rehkopf, D.H., Leppert, J.T., Bortz, W.M., Cullen, M.R., Chertow, G.M., Ioannidis, J.P.A., 2013. Systematic evaluation of environmental and behavioural factors associated with

- all-cause mortality in the united states national health and nutrition examination survey. *Int. J. Epidemiol.* 42, 1795–1810. <https://doi.org/10.1093/ije/dyt208>
- Patel, C.J., Sundaram, R., Buck Louis, G.M., 2017. A data-driven search for semen-related phenotypes in conception delay. *Andrology* 5, 95–102. <https://doi.org/10.1111/andr.12288>
- Pavlidis, P., Qin, J., Arango, V., Mann, J.J., Sibille, E., 2004. Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. *Neurochem. Res.* 29, 1213–1222. <https://doi.org/10.1023/B:NERE.0000023608.29741.45>
- Pearl, J., 2009. Causal inference in statistics: An overview. *Stat. Surv.* 3, 96–146. <https://doi.org/10.1214/09-ss057>
- Pearl, J., 1995. Causal diagrams for empirical research. *Biometrika* 82, 669–688. <https://doi.org/10.1093/biomet/82.4.669>
- Pedersen, M., Giorgis-Allemand, L., Bernard, C., Aguilera, I., Andersen, A.M.N., Ballester, F., Beelen, R.M.J., Chatzi, L., Cirach, M., Danileviciute, A., Dedele, A., Eijnsden, M. van, Estarlich, M., Fernández-Somoano, A., Fernández, M.F., Forastiere, F., Gehring, U., Grazuleviciene, R., Gruzieva, O., Heude, B., Hoek, G., Hoogh, K. de, van den Hooven, E.H., Håberg, S.E., Jaddoe, V.W.V., Klümper, C., Korek, M., Krämer, U., Lerchundi, A., Lepeule, J., Nafstad, P., Nystad, W., Patelarou, E., Porta, D., Postma, D., Raaschou-Nielsen, O., Rudnai, P., Sunyer, J., Stephanou, E., Sørensen, M., Thiering, E., Tuffnell, D., Varró, M.J., Vrijkotte, T.G.M., Wijga, A., Wilhelm, M., Wright, J., Nieuwenhuijsen, M.J., Pershagen, G., Brunekreef, B., Kogevinas, M., Slama, R., 2013. Ambient air pollution and low birthweight: A European cohort study (ESCAPE). *Lancet Respir. Med.* 1, 695–704. [https://doi.org/10.1016/S2213-2600\(13\)70192-9](https://doi.org/10.1016/S2213-2600(13)70192-9)
- Pereira, T.C.B., Campos, M.M., Bogo, M.R., 2016. Copper toxicology, oxidative stress and inflammation using zebrafish as experimental model. *J. Appl. Toxicol.* 36, 876–885. <https://doi.org/10.1002/jat.3303>
- Philippat, C., Heude, B., Botton, J., Alfaidy, N., Calafat, A.M., Slama, R., Group, the E.M.C.S., 2019. Prenatal Exposure to Select Phthalates and Phenols and Associations with Fetal and Placental Weight among Male Births in the EDEN Cohort (France). *Environ. Health Perspect.* 127, 017002. <https://doi.org/10.1289/EHP3523>
- Poggio, T., Rifkin, R., Mukherjee, S., Niyogi, P., 2004. General conditions for predictivity in learning theory. *Nature* 428, 419–422. <https://doi.org/10.1038/nature02341>
- Qin, X. Di, Qian, Z. (Min), Dharmage, S.C., Perret, J., Geiger, S.D., Rigdon, S.E., Howard, S., Zeng, X.W., Hu, L.W., Yang, B.Y., Zhou, Y., Li, M., Xu, S.L., Bao, W.W., Zhang, Y.Z., Yuan, P., Wang, J., Zhang, C., Tian, Y.P., Nian, M., Xiao, X., Chen, W., Lee, Y.L., Dong, G.H., 2017. Association of perfluoroalkyl substances exposure with impaired lung function in children. *Environ. Res.* 155, 15–21. <https://doi.org/10.1016/j.envres.2017.01.025>
- Quanjer, P.H., Stanojevic, S., Cole, T.J., Baur, X., Hall, G.L., Culver, B.H., Enright, P.L., Hankinson, J.L., Ip, M.S.M., Zheng, J., Stocks, J., Schindler, C., 2012. Multi-ethnic reference values for spirometry for the 3–95-yr age range: The global lung function 2012 equations. *Eur. Respir. J.* 40, 1324–1343. <https://doi.org/10.1183/09031936.00080312>
- Quek, Y.H., Tam, W.W.S., Zhang, M.W.B., Ho, R.C.M., 2017. Exploring the association between childhood and adolescent obesity and depression: a meta-analysis. *Obes. Rev.* <https://doi.org/10.1111/obr.12535>

- Rappaport, S.M., 2016. Genetic factors are not the major causes of chronic diseases. *PLoS One* 11. <https://doi.org/10.1371/journal.pone.0154387>
- Rappaport, S.M., 2012. Biomarkers intersect with the exposome. *Biomarkers* 17, 483–489. <https://doi.org/10.3109/1354750X.2012.691553>
- Ratti, E., 2015. Big data biology: Between eliminative inferences and exploratory experiments. *Philos. Sci.* 82, 198–218. <https://doi.org/10.1086/680332>
- Richmond, R.C., Sharp, G.C., Ward, M.E., Fraser, A., Lyttleton, O., McArdle, W.L., Ring, S.M., Gaunt, T.R., Lawlor, D.A., Smith, G.D., Relton, C.L., 2016. DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. *Diabetes* 65, 1231–1244. <https://doi.org/10.2337/DB15-0996>
- Roberts, S., Nowak, G., 2014. Stabilizing the lasso against cross-validation variability. *Comput. Stat. Data Anal.* 70, 198–211. <https://doi.org/10.1016/J.CSDA.2013.09.008>
- Rolland, M., Lyon-Caen, S., Sakhi, A.K., Pin, I., Sabaredzovic, A., Thomsen, C., Slama, R., Philippat, C., 2020. Exposure to phenols during pregnancy and the first year of life in a new type of couple-child cohort relying on repeated urine biospecimens. *Environ. Int.* 139, 105678. <https://doi.org/10.1016/j.envint.2020.105678>
- Rothman, K., Greenland, S., Lash, T., 2012. Modern Epidemiology [WWW Document]. 3rd ed Philadelphia Lippincott Williams Wilkins. URL https://books.google.fr/books?hl=fr&lr=&id=Z3vjT9ALxHUC&oi=fnd&pg=PR7&dq=+rothman+modern+epidemiology+greenland&ots=aROIbIUL8W&sig=etZ2wtoFYJAi2TnE5gYzeIj-uBY&redir_esc=y#v=onepage&q=rothman modern epidemiology greenland&f=false (accessed 7.22.20).
- Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., Sejdinovic, D., 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* 5. <https://doi.org/10.1126/sciadv.aau4996>
- Russo, F., Williamson, J., 2007. Interpreting Causality in the Health Sciences. *Int. Stud. Philos. Sci.* 21, 157–170. <https://doi.org/10.1080/02698590701498084>
- Samblas, M., Milagro, F.I., Mansego, M.L., Martí, A., Martínez, J.A., 2018. *PTPRS* and *PER3* methylation levels are associated with childhood obesity: results from a genome-wide methylation analysis. *Pediatr. Obes.* 13, 149–158. <https://doi.org/10.1111/ijpo.12224>
- Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C.M., Reichenberg, A., 2014. The familial risk of autism. *JAMA - J. Am. Med. Assoc.* 311, 1770–1777. <https://doi.org/10.1001/jama.2014.4144>
- Sang, Y., Zhang, R., Sun, L., Chen, K.K., Li, S.W., Xiong, L., Peng, Y., Zeng, L., Huang, G., 2019. MORF4L1 suppresses cell proliferation, migration and invasion by increasing p21 and E-cadherin expression in nasopharyngeal carcinoma. *Oncol. Lett.* 17, 294–302. <https://doi.org/10.3892/ol.2018.9588>
- Shao, W., Liu, Q., He, X., Liu, H., Gu, A., Jiang, Z., 2017. Association between level of urinary trace heavy metals and obesity among children aged 6–19 years: NHANES 1999–2011. *Environ. Sci. Pollut. Res.* 24, 11573–11581. <https://doi.org/10.1007/s11356-017-8803-1>
- Singh, A.S., Mulder, C., Twisk, J.W.R., Van Mechelen, W., Chinapaw, M.J.M., 2008. Tracking of childhood overweight into adulthood: A systematic review of the literature. *Obes. Rev.* <https://doi.org/10.1111/j.1467-789X.2008.00475.x>

- Sinisi, S.E., van der Laan, M.J., 2004. Deletion/Substitution/Addition Algorithm in Learning with Applications in Genomics. *Stat. Appl. Genet. Mol. Biol.* 3, 1–38. <https://doi.org/10.2202/1544-6115.1069>
- Sinisi, S.E., Van Der Laan, M.J., 2004. Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat. Appl. Genet. Mol. Biol.* 3, 1–38. <https://doi.org/10.2202/1544-6115.1069>
- Siroux, V., Agier, L., Slama, R., 2016. The exposome concept: a challenge and a potential driver for environmental health research. *Eur. Respir. Rev.* 25, 124–9. <https://doi.org/10.1183/16000617.0034-2016>
- Slama, R., Vrijheid, M., 2015. Some challenges of studies aiming to relate the Exposome to human health. *Occup. Environ. Med.* 72, 383–384. <https://doi.org/10.1136/oemed-2014-102546>
- Sohn, M.B., Li, H., 2019. Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* 13, 661–681. <https://doi.org/10.1214/18-AOAS1210>
- Steckling, N., Gotti, A., Bose-O'Reilly, S., Chapizanis, D., Costopoulou, D., De Vocht, F., Gari, M., Grimalt, J.O., Heath, E., Hiscock, R., Jagodic, M., Karakitsios, S.P., Kedikoglou, K., Kosjek, T., Leondiadis, L., Maggos, T., Mazej, D., Polańska, K., Povey, A., Rovira, J., Schoierer, J., Schuhmacher, M., Špirić, Z., Stajnko, A., Stierum, R., Tratnik, J.S., Vassiliadou, I., Annesi-Maesano, I., Horvat, M., Sarigiannis, D.A., 2018. Biomarkers of exposure in environment-wide association studies – Opportunities to decode the exposome using human biomonitoring data. *Environ. Res.* <https://doi.org/10.1016/j.envres.2018.02.041>
- Stein, A.D., Lumey, L.H., 2000. The Relationship between Maternal and Offspring Birth Weights after Maternal Prenatal Famine Exposure: The Dutch Famine Birth Cohort Study. *Hum. Biol.* <https://doi.org/10.2307/41465863>
- Stieb, D.M., Chen, L., Eshoul, M., Judek, S., 2012. Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environ. Res.* <https://doi.org/10.1016/j.envres.2012.05.007>
- Strand, L.B., Barnett, A.G., Tong, S., 2011. The influence of season and ambient temperature on birth outcomes: A review of the epidemiological literature. *Environ. Res.* <https://doi.org/10.1016/j.envres.2011.01.023>
- Strandberg-Larsen, K., Poulsen, G., Bech, B.H., Chatzi, L., Cordier, S., Dale, M.T.G., Fernandez, M., Henriksen, T.B., Jaddoe, V.W., Kogevinas, M., Kruithof, C.J., Lindhard, M.S., Magnus, P., Nohr, E.A., Richiardi, L., Rodriguez-Bernal, C.L., Rouget, F., Rusconi, F., Vrijheid, M., Andersen, A.M.N., 2017. Association of light-to-moderate alcohol drinking in pregnancy with preterm birth and birth weight: elucidating bias by pooling data from nine European cohorts. *Eur. J. Epidemiol.* 32, 751–764. <https://doi.org/10.1007/s10654-017-0323-2>
- Sur, P., Candès, E.J., 2019. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. U. S. A.* 116, 14516–14525. <https://doi.org/10.1073/pnas.1810420116>
- Tamayo-Uria, I., Maitre, L., Thomsen, C., Nieuwenhuijsen, M.J., Chatzi, L., Siroux, V., Aasvang, G.M., Agier, L., Andrusaityte, S., Casas, M., de Castro, M., Dedele, A., Haug, L.S., Heude, B., Grazuleviciene, R., Gutzkow, K.B., Krog, N.H., Mason, D., McEachan, R.R.C., Meltzer, H.M., Petraciciene, I., Robinson, O., Roumeliotaki, T., Sakhi, A.K., Urquiza, J., Vafeiadi, M., Waiblinger, D., Warembourg, C., Wright, J., Slama, R., Vrijheid, M., Basagaña, X., 2019. The early-life exposome: Description and patterns in six European countries. *Environ. Int.* 123,

- 189–200. <https://doi.org/10.1016/j.envint.2018.11.067>
- Tamayo, I., Maitre, L., Thomsen, C., Nieuwenhuijsen, M.J., Chatzi, L., 2018. The early-life exposome: description and patterns in six European countries. *Rev.*
- Tanabe, M., Kanehisa, M., 2012. Using the KEGG database resource. *Curr. Protoc. Bioinforma.* 11, 1.12.1-1.12.54. <https://doi.org/10.1002/0471250953.bi0112s38>
- Tenenhaus, M., Tenenhaus, A., Groenen, P.J.F., 2017. Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. *Psychometrika* 82, 737–777. <https://doi.org/10.1007/s11336-017-9573-x>
- Ternès, N., Rotolo, F., Michiels, S., 2016. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. *Stat. Med.* 35, 2561–2573. <https://doi.org/10.1002/sim.6927>
- Thayer, K.A., Heindel, J.J., Bucher, J.R., Gallo, M.A., 2012. Role of Environmental Chemicals in Diabetes and Obesity: A National Toxicology Program Workshop Review. *Environ. Health Perspect.* 120, 779. <https://doi.org/10.1289/EHP.1104597>
- Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., Teschendorff, A.E., 2017. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 33, 3982–3984. <https://doi.org/10.1093/bioinformatics/btx513>
- Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. <https://doi.org/10.2307/2346178>
- Titiunik, R., 2014. Can big data solve the fundamental problem of causal inference?, in: PS - Political Science and Politics. Cambridge University Press, pp. 75–79. <https://doi.org/10.1017/S1049096514001772>
- Tobi, E.W., Z wet, E.W. van, Lumey, L., Heijmans, B.T., 2018. Why mediation analysis trumps Mendelian randomization in population epigenomics studies of the Dutch Famine. *bioRxiv* 362392. <https://doi.org/10.1101/362392>
- Tong, V.T., Dietz, P.M., Morrow, B., D'Angelo, D. V., Farr, S.L., Rockhill, K.M., England, L.J., 2013. Trends in smoking before, during, and after pregnancy - Pregnancy risk assessment monitoring system, United States, 40 Sites, 2000-2010. *MMWR Surveill. Summ.* <https://doi.org/10.2307/24806088>
- Uriu-Adams, J.Y., Keen, C.L., 2005. Copper, oxidative stress, and human health. *Mol. Aspects Med.* 26, 268–98. <https://doi.org/10.1016/j.mam.2005.07.015>
- Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. *Ecol. Modell.* 203, 312–318. <https://doi.org/10.1016/j.ecolmodel.2006.11.033>
- van der Laan, M.J., Polley, E.C., Hubbard, A.E., 2007. Super Learner. *Stat. Appl. Genet. Mol. Biol.* 6. <https://doi.org/10.2202/1544-6115.1309>
- van der Laan, M.J., Starmans, R.J.C.M., 2014. Entering the Era of Data Science: Targeted Learning and the Integration of Statistics and Computational Data Analysis. *Adv. Stat.* 2014, 1–19. <https://doi.org/10.1155/2014/502678>
- Van Der Maaten, L.J.P., Postma, E.O., Van Den Herik, H.J., 2009. Dimensionality Reduction: A Comparative Review. *J. Mach. Learn. Res.* 10, 1–41. <https://doi.org/10.1080/13506280444000102>
- VanderWeele, T.J., 2011. Controlled direct and mediated effects: definition, identification and

- bounds. *Scand. Stat. Theory Appl.* 38, 551. <https://doi.org/10.1111/J.1467-9469.2010.00722.X>
- Vanderweele, T.J., Vansteelandt, S., 2009. Conceptual issues concerning mediation, interventions and composition. *Stat. Interface* 2, 457–468. <https://doi.org/10.4310/SII.2009.v2.n4.a7>
- Vermeulen, R., Schymanski, E.L., Barabási, A.-L., Miller, G.W., 2020. The exposome and health: Where chemistry meets biology. *Science* (80-). 367, 392–396. <https://doi.org/10.1126/science.aay3164>
- Vernet, C., Philippat, C., Agier, L., Calafat, A.M., Ye, X., Lyon-Caen, S., Hainaut, P., Siroux, V., Schisterman, E.F., Slama, R., 2019. An Empirical Validation of the Within-subject Biospecimens Pooling Approach to Minimize Exposure Misclassification in Biomarker-based Studies. *Epidemiology* 30, 756–767. <https://doi.org/10.1097/ede.0000000000001056>
- Vernet, C., Philippat, C., Calafat, A.M., Ye, X., Lyon-Caen, S., Siroux, V., Schisterman, E.F., Slama, R., 2018. Within-day, between-day, and between-week variability of urinary concentrations of phenol biomarkers in pregnant women. *Environ. Health Perspect.* 126, 037005. <https://doi.org/10.1289/EHP1994>
- Vernet, C., Pin, I., Giorgis-Allemand, L., Philippat, C., Benmerad, M., Quentin, J., Calafat, A.M., Ye, X., Annesi-Maesano, I., Siroux, V., Slama, Rémy, Botton, J., Charles, M.A., Dargent-Molina, P., de Lauzon-Guillain, B., Ducimetière, P., de Agostini, M., Foliguet, B., Forhan, A., Fritel, X., Germa, A., Goua, V., Hankard, R., Heude, B., Kaminski, M., Larroque, B., Lelong, N., Lepeule, J., Magnin, G., Marchand, L., Nabet, C., Slama, R., Saurel-Cubizolles, M.J., Schweitzer, M., Thiebaugeorge, O., 2017. In utero exposure to select phenols and phthalates and respiratory health in five-year-old boys: A prospective study. *Environ. Health Perspect.* 125, 097006. <https://doi.org/10.1289/EHP1015>
- Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., Kogevinas, M., Kyrtopoulos, S., Nieuwenhuijsen, M., Phillips, D.H.H., Probst-Hensch, N., Scalbert, A., Vermeulen, R., Wild, C.P.P., 2017. The exposome in practice: Design of the EXPOsOMICS project. *Int. J. Hyg. Environ. Health* 220, 142–151. <https://doi.org/10.1016/j.ijheh.2016.08.001>
- Vineis, P., Demetriou, C.A., Probst-Hensch, N., 2020. Long-term effects of air pollution: an exposome meet-in-the-middle approach. *Int. J. Public Health*. <https://doi.org/10.1007/s00038-019-01329-7>
- Vineis, P., Perera, F., 2007. Molecular epidemiology and biomarkers in etiologic cancer research: The new in light of the old. *Cancer Epidemiol. Biomarkers Prev.* <https://doi.org/10.1158/1055-9965.EPI-07-0457>
- Vineis, P., van Veldhoven, K., Chadeau-Hyam, M., Athersuch, T.J., 2013. Advancing the application of omics-based biomarkers in environmental epidemiology. *Environ. Mol. Mutagen.* 54, 461–467. <https://doi.org/10.1002/em.21764>
- Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J., 2012. Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24. <https://doi.org/10.1016/j.ajhg.2011.11.029>
- Von Kries, R., Toschke, A.M., Koletzko, B., Slikker, W., 2002. Maternal smoking during pregnancy and childhood obesity. *Am. J. Epidemiol.* 156, 954–961. <https://doi.org/10.1093/aje/kwf128>
- Vrijheid, M., 2014. The exposome: A new paradigm to study the impact of environment on health. *Thorax* 69, 876–878. <https://doi.org/10.1136/thoraxjnl-2013-204949>

- Vrijheid, M., Casas, M., Gascon, M., Valvi, D., Nieuwenhuijsen, M., 2016. Environmental pollutants and child health—A review of recent concerns. *Int. J. Hyg. Environ. Health* 219, 331–342. <https://doi.org/10.1016/j.ijheh.2016.05.001>

Vrijheid, M., Fossati, S., Maitre, L., Márquez, S., Roumeliotaki, T., Agier, L., Andrusaityte, S., Cadiou, S., Casas, M., Dedele, A., Donaire-gonzalez, D., Grazuleviciene, R., Haug, L.S., McEachan, R., Meltzer, H.M., Papadopoulou, E., Robinson, O., Sakhi, A.K., Siroux, V., Sunyer, J., Schwarze, P.E., Tamayo-uria, I., Urquiza, J., Vafeiadi, M., Valentín, A., Warembourg, C., Wright, J., Nieuwenhuijsen, M.J., Thomsen, C., Basagaña, X., Slama, R., Chatzi, L., 2020. Early-Life Environmental Exposures and Childhood Obesity : An Exposome-Wide Approach. *Environ. Health Perspect.* 128. <https://doi.org/https://ehp.niehs.nih.gov/doi/10.1289/EHP5975>

Vrijheid, M., Slama, R., Robinson, O., Chatzi, L., Coen, M., van den Hazel, P., Thomsen, C., Wright, J., Athersuch, T.J., Avellana, N., Basagaña, X., Brochot, C., Bucchini, L., Bustamante, M., Carracedo, A., Casas, M., Estivill, X., Fairley, L., van Gent, D., Gonzalez, J.R., Granum, B., Gražulevičienė, R., Gutzkow, K.B., Julvez, J., Keun, H.C., Kogevinas, M., McEachan, R.R.C., Meltzer, H.M., Sabidó, E., Schwarze, P.E., Siroux, V., Sunyer, J., Want, E.J., Zeman, F., Nieuwenhuijsen, M.J., Gražulevičienė, R., Gutzkow, K.B., Julvez, J., Keun, H.C., Kogevinas, M., McEachan, R.R.C., Meltzer, H.M., Sabidó, E., Schwarze, P.E., Siroux, V., Sunyer, J., Want, E.J., Zeman, F., Nieuwenhuijsen, M.J., Gražulevičienė, R., Gutzkow, K.B., Julvez, J., Keun, H.C., Kogevinas, M., McEachan, R.R.C., Meltzer, H.M., Sabidó, E., Schwarze, P.E., Siroux, V., Sunyer, J., Want, E.J., Zeman, F., Nieuwenhuijsen, M.J., Gražulevičienė, R., Gutzkow, K.B., Julvez, J., Keun, H.C., Kogevinas, M., McEachan, R.R.C., Meltzer, H.M., Sabidó, E., Schwarze, P.E., Siroux, V., Sunyer, J., Want, E.J., Zeman, F., Nieuwenhuijsen, M.J., Gražulevičienė, R., Gutzkow, K.B., Julvez, J., Keun, H.C., Kogevinas, M., McEachan, R.R.C., Meltzer, H.M., Sabidó, E., Schwarze, P.E., Siroux, V., Sunyer, J., Want, E.J., Zeman, F., Nieuwenhuijsen, M.J., 2014. The human early-life exposome (HELIX): project rationale and design., *Environmental Health Perspectives*. National Institute of Environmental Health Science. <https://doi.org/10.1289/ehp.1307204>

Wang, L., Pinkerton, K.E., 2008. Detrimental effects of tobacco smoke exposure during development on postnatal lung function and asthma. *Birth Defects Res. Part C - Embryo Today Rev.* 84, 54–60. <https://doi.org/10.1002/bdrc.20114>

Warembourg, C., Maitre, L., Tamayo-Uria, I., Fossati, S., Roumeliotaki, T., Aasvang, G.M., Andrusaityte, S., Casas, M., Cequier, E., Chatzi, L., Dedele, A., Gonzalez, J.-R., Gražulevičienė, R., Haug, L.S., Hernandez-Ferrer, C., Heude, B., Karachaliou, M., Krog, N.H., McEachan, R., Nieuwenhuijsen, M., Petraciūnė, I., Quentin, J., Robinson, O., Sakhi, A.K., Slama, R., Thomsen, C., Urquiza, J., Vafeiadi, M., West, J., Wright, J., Vrijheid, M., Basagaña, X., 2019. Early-Life Environmental Exposures and Blood Pressure in Children. *J. Am. Coll. Cardiol.* 74, 1317–1328. <https://doi.org/10.1016/j.jacc.2019.06.069>

Weyde, K.V., Krog, N.H., Oftedal, B., Magnus, P., White, R., Stansfeld, S., Øverland, S., Aasvang, G.M., 2018. A longitudinal study of road traffic noise and body mass index trajectories from birth to 8 Years. *Epidemiology* 29, 729–738. <https://doi.org/10.1097/EDE.0000000000000868>

Wild, C.P., 2012. The exposome: From concept to utility. *Int. J. Epidemiol.* 41, 24–32. <https://doi.org/10.1093/ije/dyr236>

Wild, C.P., 2005. Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol. Prev. Biomarkers* 14.

Winckelmans, E., Nawrot, T.S., Tsamou, M., Den Hond, E., Baeyens, W., Kleinjans, J., Lefebvre, W., Van Larebeke, N., Peusens, M., Plusquin, M., Reynders, H., Schoeters, G., Vanpoucke,

- C., de Kok, T.M., Vrijens, K., 2017a. Transcriptome-wide analyses indicate mitochondrial responses to particulate air pollution exposure. *Environ. Heal.* 16, 87.
<https://doi.org/10.1186/s12940-017-0292-7>
- Winckelmans, E., Vrijens, K., Tsamou, M., Janssen, B.G., Saenen, N.D., Roels, H.A., Kleinjans, J., Lefebvre, W., Vanpoucke, C., De Kok, T.M., Nawrot, T.S., 2017b. Newborn sex-specific transcriptome signatures and gestational exposure to fine particles: findings from the ENVIRONAGE birth cohort. *Environ. Heal. A Glob. Access Sci. Source* 16, 52.
<https://doi.org/10.1186/s12940-017-0264-y>
- Windham, G.C., Hopkins, B., Fenster, L., Swan, S.H., 2000. Prenatal active or passive tobacco smoke exposure and the risk of preterm delivery or low birth weight. *Epidemiology*.
<https://doi.org/10.1097/00001648-200007000-00011>
- Woods, M.M., Lanphear, B.P., Braun, J.M., McCandless, L.C., 2017. Gestational exposure to endocrine disrupting chemicals in relation to infant birth weight: A Bayesian analysis of the HOME Study. *Environ. Heal. A Glob. Access Sci. Source* 16.
<https://doi.org/10.1186/s12940-017-0332-3>
- Wright, J., Small, N., Raynor, P., Tuffnell, D., Bhopal, R., Cameron, N., Fairley, L., Lawlor, D.A., Parslow, R., Petherick, E.S., Pickett, K.E., Waiblinger, D., West, J., 2013. Cohort Profile: The Born in Bradford multi-ethnic family cohort study. *Int. J. Epidemiol.* 42, 978–991.
<https://doi.org/10.1093/ije/dys112>
- Yamada, H., Masuko, H., Yatagai, Y., Sakamoto, T., Kaneko, Y., Iijima, H., Naito, T., Noguchi, E., Konno, S., Nishimura, M., Hirota, T., Tamari, M., Hizawa, N., 2016. Role of Lung Function Genes in the Development of Asthma. *PLoS One* 11, e0145832.
<https://doi.org/10.1371/journal.pone.0145832>
- Zhou, W., Lo, S.H., 2018. Analysis of genotype by methylation interactions through sparsity-inducing regularized regression 06 Biological Sciences 0604 Genetics, in: *BMC Proceedings*. BioMed Central Ltd., p. 40. <https://doi.org/10.1186/s12919-018-0145-6>
- Zou, H., 2006. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101, 1418–1429.
<https://doi.org/10.1198/016214506000000735>
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

APPENDIX I: Oral communications and publications; curriculum vitae

- Scientific production

Oral communications:

- **Use of methylation marks to inform Association between Early-life Air pollution Exposures and child Body Mass Index: An analysis based on a priori selected pathways,** Solène Cadiou; Bustamante, Mariona; Agier, Lydiane; Maitre, Lea; Gonzalez, Juan R.; Hernandez-Ferrer, Carles; Carracedo, Angel; Vives, Marta; Wright, John; Chatzi, Leda Grazuleviciene, Regina; Meltzer, Helle M.; Nieuwenhuijsen, Mark; Vrijheid, Martine; Slama, Rémy, ISEE-ISES 2018 congress, Ottawa.
- **Using DNA methylation to characterize more efficiently associations between the exposome and child lung function,** Solène Cadiou, Lydiane Agier, Mariona Bustamante, Léa Maitre, Xavier Basagana, Martine Vrijheid, Valérie Siroux, Rémy Slama, ISEE 2019 congress, Utrecht

Manuscripts directly related to the PhD:

- Cadiou, S., Bustamante, M., Agier, L., Andrusaityte, S., Basagaña, X., Carracedo, A., Chatzi, L., Grazuleviciene, R., Gonzalez, J.R., Gutzkow, K.B., Maitre, L., Mason, D., Millot, F., Nieuwenhuijsen, M., Papadopoulou, E., Santorelli, G., Saulnier, P.-J., Vives, M., Wright, J., Vrijheid, M., Slama, R., 2020. “**Using methylome data to inform exposome-health association studies: An application to the identification of environmental drivers of child body mass index.**” *Environment International*. 138, 105622, <https://doi.org/10.1016/j.envint.2020.105622>
- Cadiou, S., Slama R., “**Some insights regarding the instability of variable-selection algorithms used for causal inference purposes in epidemiology**” (*in review*)
- Cadiou, S., Basagana, X., González, JR., Lepeule, J., Siroux, V., Vrijheid, M., Slama, R., “**Performance of approaches relying on multidimensional intermediary data to decipher causal relationships between the exposome and health: a simulation study under various causal structures**” (*in review*)

Other manuscripts:

- Blum, M., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., Slama, R., 2020, “**Challenges raised by mediation in a high dimension setting**”, *Environmental Health Perspectives* 128, 055001. <https://doi.org/10.1289/EHP6240>
- Calvo, B., Lau, CH., Gutzkow, KB., Siskos, A., Maitre, L., ... Cadiou, S....Coen, M., Vrijheid, M., Keun, H., Escaramis, G., Bustamante, M., “**Urinary metabolite quantitative trait loci in children and their interaction with dietary factors**” (*in review*)
- Maitre, L., Bustamante, M., Hernandez-Ferrer, C., Thiel, D., Lau, CH., Siskos, A., Vives-Usano, M., Robinson, O., Wright, J., ..., Cadiou, S., Slama, R., .. Sunyer, J., Casas. M., Gutzkow, KB., ..., Grazuleviciene, R., ..., Chatzi, L., Thomsen, C., Nieuwenhuijsen, M.,.., Basagana, X., Tamayo, I., Sabido, E., Borras, E., Carracedo, A., Quintela, I., Estivill, X., Urquiza, J., Coren, M., Gonzalez, JR, Keun, H., Vrijheid, M., “**The molecular signatures of the Human Early Life Exposome using multi-omics profiling**” (*in preparation*)
- De Paro-Bert, P., Ruiz-Arenas, C., Hernandez-Ferrer, C.,, Cadiou, S., Slama, R., ..., Vrijheid, M., Bustamante, M., “**The early-life exposome and epigenetic age acceleration in children**” (*in preparation*)
- Ruiz-Arenas, C., Hernandez-Ferrer, C., Vives-Usano, M., Quintela, I., Mason, D., Cadiou, S., Casas, M., Andrusaityte, S., Gutzkow, K.B., Vafeiadi, M., Wright, J., Lepeule, J., Grazuleviciene,

- R., Chatzi, L., Carracedo, A., Estivill, X., Martí, E., Escaramís, G., Vrijheid, M., González, JR., Bustamante, M., “**Identification of blood autosomal cis-expression quantitative trait methylation (cis-eQTM)s in children**” (in review) <https://www.biorxiv.org/content/10.1101/2020.11.05.368076v1>
- Vives-Usano, M., Hernandez-Ferrer, C., Maitre, L., Ruiz-Arena, C., Andrusaityte, S., Borràs, E., Cadiou, S., Carracedo, A., Casas, M., Chatzi, L., Coen, M., Estivill, X., González, JR., Grazuleviciene, R., Gutzkow, KB., Keun, H., Lau, CH., Lepeule, J., Mason, D., Quintela, I., Robinson, O., Sabidó, E., Santorelli, G., Schwarze, P., Siskos, A., Slama, R., Thomsen, C., Vafeiadi, M., Martí, E., Vrijheid, M., Bustamante, M., “**In utero and childhood exposure to tobacco smoke and multi-layer molecular signatures in children**”, BMC Medicine. 2020;18(1):1-19. <https://doi.org/10.1186/s12916-020-01686-8>
 - Vrijheid, M., Fossati, S., Maitre, L., Márquez, S., Roumeliotaki, T., Agier, L., Andrusaityte, S., Cadiou, S., Casas, M., de Castro, M., Dedele, A., Donaire, D., Grazuleviciene, R., Haug, L., McEachan, R., Meltzer, HM., Papadopoulou, E., Robinson, E., Sakhi, AK., Sioux, V., Sunyer, J., Schwarze PE., Tamayo-Uria, I., Urquiza, J., Vafeiadi, M., Valentin A., Warembourg, C., Wright, J., Nieuwenhuijsen, M., Thomsen, C., Basagaña, X., Slama R., Chatzi, L., 2020, “**Early-life Environmental Exposures and Childhood Obesity: an Exposome-wide Approach**” Environmental Health Perspectives 128. <https://doi/10.1289/EHP5975>

- **Curriculum Vitae**

Education:

2017-2020: Institute for Advanced Biosciences, Inserm U1209, (Grenoble, France)
Team of Environmental Epidemiology Applied to Reproduction and Respiratory Health
PhD Student under the Supervision of Remy Slama,

2016-2017 École Nationale des Ponts and Chaussées and AgroParisTech (Paris, France)
Master2 « Public Policies for Sustainable development ».

2015-2016 Engineering school AgroParisTech (Paris, France)

Master2: METATOX: « From Assessment to politic management of toxicological risks for human and environment ». Toxicology-statistics-public policies

Optional lessons from Master 2 Paris 5 University: Statistical methods for therapy assessment

2012-2015 Ecole Polytechnique (competitive scientific school) – Paris Saclay University (Palaiseau, France)

Master 2: Sciences for sustainable development: informatics, mathematics, economy, mechanics, human biology

Professional experience:

Since September 2020: French Ministry of Environment, Paris, France - in charge of data spaces (“Green Data Hub”) and data sciences expert;

2017-2020 (September-August): Institute for Advanced Biosciences, Grenoble, France - PhD in Environmental Epidemiology under the supervision of Rémy Slama;

2017 (March – July): Consulting mission for APHP (Paris Public Hospitals) to build a strategy to apply European Directive on falsified drugs;

2016 (September-December): Consulting mission for the French Ministry of Health and Agriculture on the governance of The National Council for Alimentation;

2016 (March –August): CRCHUM- Montreal University, Canada – Health Population Unit Research internship: statistical description of occupational exposures national database;

2015 (March -August): Imperial College London, United-Kingdom –Bioengineering Department Research internship: Development of new indicators to study local risks factors of atherosclerosis

2012: Compagnie de Gendarmerie Départementale de Lyon (Lyon, France) -Military service as a police officer;

- **Doctoral training (149 hours)**

Scientific training (54 hours)

XP Exposome short course series: Advanced OMICS profiling and integration in Exposome research M Chadeau-Hyam, London (28 hours)

Utilisation des données génétiques issues du séquençage haut débit dans l'étude des maladies : méthodologie statistique et applications INSERM, Bordeaux (14 hours)

Doctoral school annual scientific day (12 hours)

Professional integration (60 hours)

Equivalence: professional contract as a high civil servant (60 hours)

Transversal training (35 hours)

Coding in PYTHON, introduction, UGA (26 hours)

Linux for scientific computation, UGA (9 hours)

APPENDIX II: Prenatal exposures and birthweight in SEPAGES mother-child study, an adapted oriented Meet-in-the-Middle approach

We present here preliminary results from a study performed on the French SEPAGES mother-child cohort, relating a ‘small’ prenatal exposome to the birth weight of the child, relying on additional information from the mother blood methylome during pregnancy. We used a third variation of our oMITM design: as a first step, we chose to perform a supervised data-driven dimension reduction of the methylome using Partial Least Square (PLS) rather than an *a priori* preselection of relevant CpGs based on external databases as done in chapter II. Moreover, to test the association between the (reduced) exposome and the outcome, we used a multiple regression rather than ExWAS, as this may be more efficient in the context of the small exposome considered here.

1. Introduction

Low birth weight is known to be a predictor of later adverse health conditions (Belbasis et al., 2016). Environmental and behavioral determinants of decreased birth weight have thus been explored for many decades in epidemiology, leading to effective public health interventions, such as limiting the alcohol consumptions and tobacco exposure during pregnancy (Burling et al., 1991; Chersich et al., 2012; Tong et al., 2013). With the advent of the exposome paradigm (Wild, 2005), various other prenatal exposures are now studied simultaneously within ambitious exposome projects (Agier et al., 2020a). In the French Sepages cohort (Lyon-Caen et al., 2019), several phenolic compounds and phthalates exposures have been characterized during pregnancy, with an expected low measurement error due to reliance on the within-subject biospecimens pooling approach (Vernet et al., 2019, 2018). In this study, we aimed to perform an exposome wide analysis of the environmental drivers of the child birth weight in the Sepages cohort, relying on methylome data and on the oMITM design that we previously developed (Cadiou et al., 2020, see Chapter II), in an attempt to limit false positive signals. We adapted the implementation of oMITM chosen in (Cadiou et al., 2020) to the dimension of our different layers: due to the limited size of our exposome (see below), we chose to conduct multiple regression analysis instead of ExWAS for the tests of association relating directly the exposome to the health outcome. We also chose to perform a data-driven dimension reduction for the methylome, as the reliance on external heterogenous

database to preselect a part of the methylome could be questioned, as detailed in (Cadiou et al., 2020) and (Mubeen et al., 2019).

2. Methods

2.1. Study population and outcome

We relied on the SEPAGES parents-child cohort, in which the prenatal exposome, the mother DNA methylome (from peripheral blood during pregnancy) and birthweight were assessed in 438 mother–single child pair recruited around the Grenoble (France) metropolitan area before 19 weeks of pregnancy between July 2014 and July 2017, in ultrasound medical center or after having spontaneously contacted the study team (Lyon-Caen et al., 2019). Weight was measured at birth and various relevant covariates were assessed during pregnancy and at birth: maternal weight before pregnancy, maternal height, season of conception, maternal highest diploma, maternal smoking before pregnancy, maternal smoking during pregnancy, maternal parity and gestational duration.

2.2. Exposome assessment

As detailed elsewhere (Lyon-Caen et al., 2019), an extended personal exposure assessment was conducted during pregnancy in the 484 women of the Sepages study. In this study, we considered 34 variables, available in 338 women, for the exposome. These environmental factors are detailed in Table 0.1 and belong to 3 major groups: air pollutants, phenols and phthalates biomarkers and behavioral factors. Air pollutants (2 variables) were measured by personal dosimeter carried during the pregnancy by the mother during one follow up week taking place before the blood sampling used for methylome assessment (see below). Phenols and phthalates (26 variables) were assessed in pooled urine samples from three micturitions per day collecting during each follow-up week; values below quantification limit were imputed using the fill-in method as detailed in Rolland et al. (2020). Last, 5 parameters corresponding to behavior were built from the questionnaires. All exposures were corrected for relevant protocol covariables to remove batch effects.

Table 0.1: Exposome components assessed in Sepages cohort during pregnancy, with mean and standard deviation for quantitative variables and frequency for qualitative variables, and amount of missing data. (438 mothers recruited between 2014 and 2017).

Exposure group	Exposure	Modality	Mean + SD	n (%)	Missing (%)	Unit
Air pollutant	PM _{2.5} - trimester 1		13.9 (7.2)	132 (30.1)	3	ug/m ³
Air pollutant	NO ₂ - trimester 1		20.9 (7.4)	12 (2.74)	3	ug/m ³
Behaviour	Number of alcohol drinks per month during pregnancy	No alcohol	311 (71)	69 (15.753)		
		Less than one glass per month	49 (11)	69 (15.753)		
		More than one glass per month	9 (2)	69 (15.753)		
Behaviour	More than 3 alcohol drinks	Never	319 (73)	72 (16.438)		
		Less than one time per month	29 (7)	72 (16.438)		
		One time per month	14 (3)	72 (16.438)		
		One time per week or more	4 (1)	72 (16.438)		
Behaviour	Slimming diet during pregnancy	No	374 (85)	61 (13.927)		
		Yes	3 (1)	61 (13.927)		
Behaviour	Any slimming diet before pregnancy	No	300 (68)	60 (13.699)		
		Yes	78 (18)	60 (13.699)		
Behaviour	Self-assessed stress level	Nervous most of the time	6 (1)	82 (18.721)		
		Often nervous	59 (13)	82 (18.721)		
		Sometimes nervous	238 (54)	82 (18.721)		
		Never nervous	53 (12)	82 (18.721)		
Urinary phthalates	MnBP - Trimester 2		13.2 (10.8)	-		μg/l

Urinary phthalates	MEHHP - Trimester 2	9.6 (17.9)	-	µg/l
Urinary phthalates	oomph - Trimester 2	1.2 (1.9)	-	µg/l
Urinary phthalates	ohm INCH - Trimester 2	4.2 (10.8)	-	µg/l
Urinary phthalates	MiBP - Trimester 2	18.8 (13.3)	-	µg/l
Urinary phthalates	MEOHP - Trimester 2	6.9 (13.3)	-	µg/l
Urinary phthalates	omen - Trimester 2	8.9 (12.5)	-	µg/l
Urinary phthalates	oxo MINCH - Trimester 2	3 (6.6)	-	µg/l
Urinary phthalates	MBzP - Trimester 2	6.6 (8.4)	-	µg/l
Urinary phthalates	MECPP - Trimester 2	12.6 (19.4)	-	µg/l
Urinary phthalates	oxolin - Trimester 2	4.1 (9.5)	-	µg/l
Urinary phthalates	MEP - Trimester 2	38.3 (61.1)	-	µg/l
Urinary phthalates	MEHP - Trimester 2	3.3 (6.9)	-	µg/l
Urinary phthalates	MMCHP - Trimester 2	9.4 (12.5)	-	µg/l
Urinary phthalates	cumin - Trimester 2	8.2 (16)	-	µg/l
Urinary phthalates	MBzP - Trimester 2	6.6 (8.4)	-	µg/l
Urinary phenols	MEPA - Trimester 2	97.7 (530.6)	-	µg/l
Urinary phenols	BPA - Trimester 2	2.4 (3.2)	-	µg/l
Urinary phenols	ETPA total - Trimester 2	8.4 (37.4)	-	µg/l

Urinary phenols	OXBE total - Trimester 2		11.1 (80.3)	-	µg/l
Urinary phenols	PRPA total - Trimester 2		19.8 (95.1)	-	µg/l
Urinary phenols	Triclosan total - Trimester 2		32.9 (149.7)	-	µg/l
Urinary phenols	BUPA total - Trimester 2	<LOD	331 (76)	-	
		Between LOD and LOQ	61 (14)	-	
		>LOQ	46 (11)	-	
Urinary phenols	BPS total - Trimester 2	<LOD	326 (74)	-	
		Between LOD and LOQ	21 (5)	-	
		>LOQ	91 (21)	-	
Urinary phenols	BPF total - Trimester 2	<LOD	431 (98)	-	
		>LOQ	7 (2)	-	
Urinary phenols	Trichlorobenzene total - Trimester 2	<LOD	435 (99)	-	
		Between LOD and LOQ	2 (0)	-	
		>LOQ	1 (0)	-	

BPA: bisphenol A; BPF: bisphenol F; BPS: bisphenol S; BUPA: butyl paraben; cxMiNP: Mono-4-methyl-7-carboxyoctyl-phthalate; ETPA: ethyl paraben; LOD: Limit of Detection; LOQ: Limit of Quantification ; MBzP: mono benzyl phthalate; MECPP: mono-2-ethyl 5-carboxypentyl phthalate ; MEHHP: mono-2-ethyl-5-hydroxyhexyl phthalate; MEHP: Mono(2-éthylhexyl) phthalate; MEOHP: mono-2-ethyl-5-oxohexyl phthalate; MEP: monoethyl phthalate; MEPA: methyl paraben; MiBP: mono-iso-butyl phthalate; MMCHP: Mono-2-carboxymethyl hexyl-phthalate ; MnBP: mono-n-butyl phthalate; ohMiNCH: 2-(((Hydroxy-4-methyloctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid; OHMiNP: mono-4-methyl-7-hydroxyoctyl phthalate; ohMPHP: 6-Hydroxy Monopropylheptyl-phthalate; OXBE:oxybenzone/benzophenone-3; oxoMiNCH: 2-(((4-Methyl-7-oxooctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid; OXOMiNP: mono-4-methyl-7-oxooctyl phthalate; PRPA: propyl paraben

2.3. DNA methylation

Peripheral blood was collected during the study visit taking place at the clinic around 19 gestational weeks, using EDTA tubes. DNA extraction was performed at Quiagen, Heselberg, Germany). DNA was extracted from buffy coat; DNA methylation was assessed with the Infinium Human Methylation 850K beadchip. A first filtering of probes was performed using ChAMP protocol (Tian et al., 2017) to eliminate probes with detection p-values lower than 0.01, probes with a single-nucleotide polymorphism (SNP), probes with beadcount lower than 3 in at least 5% of the samples and unspecific probes, leading to 792,152 remaining CpG probes in 487 samples. Data were normalized using Beta-Mixture Quantile and imputed using ChAMP (Tian et al., 2017). A final filtering was performed to eliminate probes on sexual chromosomes and duplicate samples. Only the samples corresponding to individuals for which at least one exposure variable from each of the three groups of exposures assessed was available were considered. Finally, the remaining 774,172 CpGs were expressed as Beta-values for 438 subjects. Relevant “protocol” covariates (e.g. batch and plate) were identified and added to all models including methylome data to avoid batch effects. Correlation within the methylome was estimated by averaging the Pearson’s correlation within 1000 sets of 100 randomly selected CpGs to avoid computing all pairwise correlations between the 774,172 CpGs.

2.4. Statistical analyses

We followed the oriented Meet-in-the-Middle (oMITM) design proposed by Cadiou et al. (2020) using the methylome layer to reduce the dimension of the exposome. It follows three steps: a) supervised dimension reduction of the methylome using the outcome of interest; b) tests of association between the relevant dimension of the methylome defined at step a) and each component of the exposome conditionally on the outcome; c) test of association between each component of the exposome found associated with the reduced methylome at step b) (i.e. the reduced exposome) and the outcome. For step a), we removed the linear effect of relevant

covariates from the birthweight and used Partial-Least Square regression (Höskuldsson, 1988) to relate the whole methylome to the residual: this dimension reduction technique builds summary variables as linear combinations of the original set of variables, which are defined iteratively such that they explain as much of the remaining covariance between the predictors and the outcome as possible. The number of relevant components was determined by cross-validation using *plsgenomics* package (Boulesteix et al., 2018), which was also used for the PLS analyses itself. For step b), we used an ExWAS-type approach, i.e. multiple univariate regressions corrected for multiple comparisons using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), adjusted on the relevant covariates (see IV.2.1.) and on birth weight. For step c), we used a multiple regression model adjusted on the same covariates and including simultaneously the whole reduced exposome. Considering the choice of adjustment covariates, season of measure was added to the set of covariates specified in IV.2.1. in the models including air pollutants variables. Missing values for exposures were single-imputed using MICE package (Buuren and Groothuis-Oudshoorn, 2011),

2.5. Sensitivity analyses

We performed 3 sensitivity analyses:

- Sensitivity analysis I: an ‘agnostic’ multiple linear regression ignoring the methylome: the whole exposome was related to the birth weight using a single multiple regression model as in step c).
- Sensitivity analysis II: an oMITM approach using ExWAS-type analysis at each step a), b) and c).
- Sensitivity analysis III: an oMITM approach repeating the oMITM implemented in Cadiou et al., 2020, using ExWAS type at each step and an a priori reduced methylome of only 2004 CpGs, representing the intersection of the CpGs available in our study and the 2284 CpGs selected in a priori reduced methylome relevant for the child BMI in Cadiou et al. (2020).

3. Results

3.1. Study population

Among the 438 children for which both prenatal exposures and maternal methylome data were available, mean birth weight was 3285 g (CI: 460) with 4 missing data. Detailed information about the covariates and exposures levels in the study population are presented in Table 0.1 and Table 0.2.

Table 0.2: Characteristics of the 438 mother-child pairs included in the exposome analysis based on Sepages study.

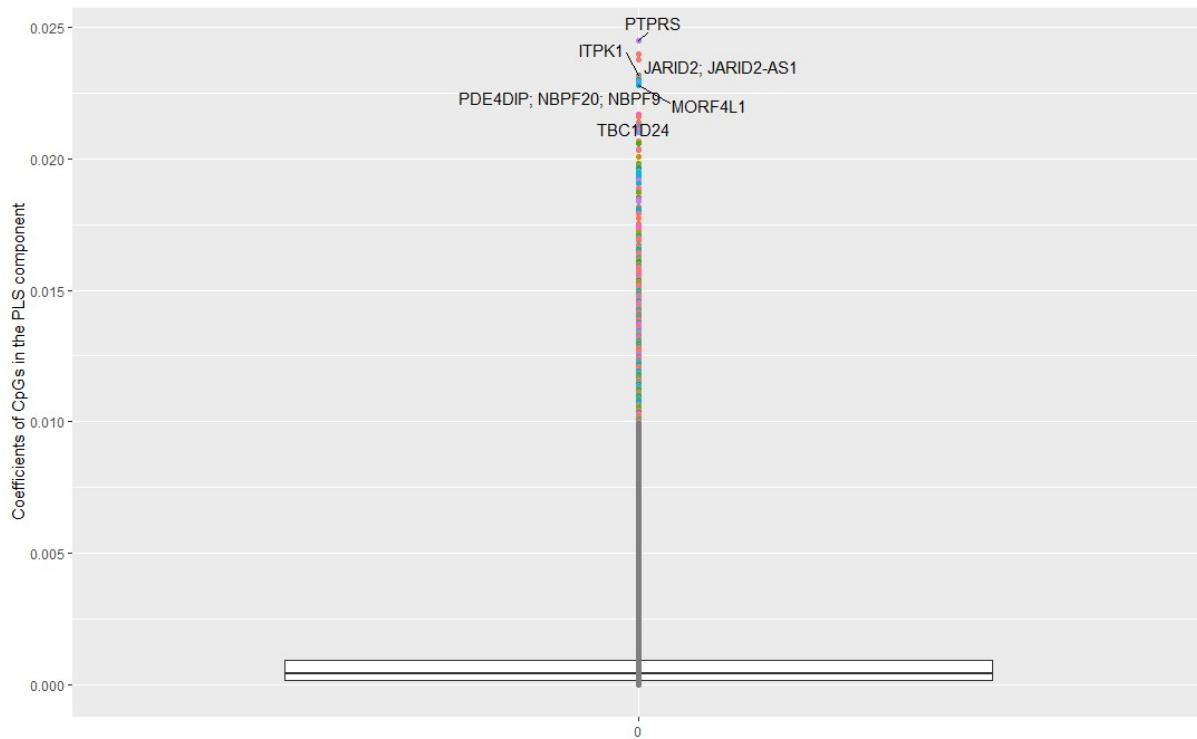
Characteristic	Modality	Mean +- SD	n (%)	Missing (%)	Unit
Preterm birth	No		418 (95)	3 (0.685)	
	Yes		17 (4)	3 (0.685)	
Season of conception	Jan-Feb-Mar		112 (26)		
	Apr-May-Jun		93 (21)		
	Jul-Aug-Sept		105 (24)		
	Oct-Nov-Dec		128 (29)		
Maternal age		32.6 (3.8)			year
Child sex	Male		229 (52)	4 (0.913)	
	Female		205 (47)	4 (0.913)	
Birth weight		3285 (460)		4 (0.913)	g
Gestational duration		39.7 (1.5)		3 (0.685)	week
Maternal height		165.3 (6)		4 (0.913)	cm
Maternal weight before pregnancy		61.2 (11)			kg
Maternal highest diploma	Before high school		25 (6)	2 (0.457)	
	High school before bachelor		49 (11)	2 (0.457)	
	Between bachelor and master		115 (26)	2 (0.457)	
	Master or higher		247 (56)	2 (0.457)	
Maternal smoking before pregnancy		0.5 (2.3)		43 (9.817)	cigarettes/day
Maternal smoking during any trimester of pregnancy	No		371 (85)	38 (8.676)	
	Yes		29 (7)	38 (8.676)	
	No child				
Parity	before		197 (45)		
	1 child		196 (45)		
	2 children or more		45 (10)		
Measurement season for air pollutants	Jan-Feb-Mar		121 (28)	9 (2.055)	
	Apr-May-Jun		122 (28)	9 (2.055)	
	Jul-Aug-Sept		101 (23)	9 (2.055)	
	Oct-Nov-Dec		85 (19)	9 (2.055)	

3.2. First step of the oMITM approach: relationship between methylome and birthweight

PLS analysis – step a) of the oMITM approach

One PLS component was enough to relate methylome to birth weight, as determined by cross validation. The Pearson correlation of this component with birth weight was 0.29. The distribution of the weights of the CpG in this component are presented in Figure 1. The 10 top CpGs belonged to the following genes: *PTPRS*; *ITPK1*; *JARID2*; *PDE4DIP-NBPF20-NBPF9*; *TBC1D24* and *MORF4L1*.

Figure 1: Distribution of the weight of CpGs in the selected PLS component. Colors indicate the corresponding gene for the top 500 CpGs and the top 10 are annotated.



Sensitivity analyses: ExWAS-type analyses on the methylome

No significant associations between the methylome and the birth weight were found when correcting for multiple comparisons (Sensitivity analysis II, lowest corrected p-value: 0.38). Without correction (lowest p-value: 1.11×10^{-6}), 54090 associations were significant; the top 10

CpGs belonged to the following genes: *TSSC4*, *GABRB3*, *ATP1A3*, *GIMD1*, *CCDC25*, *ZNF775*, *COPE*, *DUSP16*, *AGBL3* and *EPN2* (Sensitivity analysis II). Similarly, none of the 2004 CpGs of the restricted methylome was associated with the birth weight neither with correction for multiple testing (Sensitivity analysis III, lowest corrected p-value: 0.33) or without (lowest p-value: 3.23×10^{-4}).

3.3. Whole oMITM approach

None of the component of the exposome was associated with the PLS component with (lowest p-value: 0.715) adjustment for multiple testing (see Table 0.3). The reduced exposome thus did not contain any exposures and the whole oMITM approach did not point any association. Without the correction for multiple tests, the association between the PLS component with the stress level self-assessed by the mother was significant (p-value: 0.048).

3.4. Agnostic multiple regression

An agnostic multivariate analysis showed no significant associations between the exposome and birthweight (sensitivity analysis I, Table 0.4) with correction for multiple testing (lowest adjusted p-value: 0.67) or without (lowest p-value: 0.062).

Table 0.3: Step b) of the oMITM approach: estimates, confidence intervals, uncorrected and corrected for multiple comparisons p-values of the tests of association between exposome and PLS component adjusted on relevant covariates and birth weight (434 mother-child pairs from the Sepages cohort). *: for qualitative exposures, p-values by level and the reference level are indicated.

Exposure	Modality	Estimate	CI2.5	CI97.5	Global unadjusted p-value	Unadjusted p-value by level*	p-value corrected for multiple testing
Self-assessed stress level	Nervous most of the time	2.484	-2.844	7.811	0.048	0.360	0.715
	Often nervous	1.821	-0.032	3.673			
	Never nervous	2.316	0.324	4.308			
	Sometimes nervous						
MnBP - Trimester 2		0.053	-0.014	0.121	0.119		0.715
cxMiNP - Trimester 2		0.069	-0.020	0.159	0.129		0.715
ohMINCH - Trimester 2		0.298	-0.141	0.737	0.182		0.715
MEOHP - Trimester 2		0.446	-0.221	1.113	0.190		0.715
MEPA - Trimester 2		0.001	-6.913x10 ⁻⁴	0.003	0.192		0.715
Slimming diet during pregnancy	Yes	2.773	-1.422	6.968	0.195	0.195	0.715
Slimming diet during pregnancy	No					Reference level	
oxoMiNP - Trimester 2		-0.090	-0.228	0.048	0.202		0.715
oxoMINCH - Trimester 2		-0.445	-1.159	0.270	0.222		0.715
MMCHP - Trimester 2		-0.220	-0.574	0.135	0.224		0.715
MEP - Trimester 2		0.006	-0.005	0.017	0.279		0.812
ohMPHP - Trimester 2		-0.168	-0.528	0.193	0.362		0.846
TRCB total - Trimester 2	Between LOD and LOQ	6.583	-3.398	16.565	0.407	0.195	0.846
	>LOQ	2.443	-11.080	15.966		0.723	
TRCB total - Trimester 2	<LOD					Reference level	
MBzP - Trimester 2		-0.034	-0.120	0.052	0.431		0.846
PRPA total - Trimester 2		-0.004	-0.016	0.007	0.447		0.846

MEHHP - Trimester 2		-0.164	-0.590	0.262	0.451		0.846
BUPA total - Trimester 2	Between LOD and LOQ	0.446	-1.547	2.439	0.483	0.660	0.846
	>LOQ	-1.193	-3.413	1.026		0.291	
	<LOD					Reference level	
BPA - Trimester 2		-0.065	-0.271	0.140	0.533		0.846
MiBP - Trimester 2		-0.017	-0.072	0.038	0.539		0.846
ohMiNP - Trimester 2		-0.032	-0.136	0.072	0.544		0.846
TRCS total - Trimester 2		-0.001	-0.006	0.003	0.555		0.846
BPF total - Trimester 2	>LOQ	1.320	-3.998	6.638	0.626	0.626	0.910
	<LOD					Reference level	
OXBE total - Trimester 2		-0.002	-0.012	0.008	0.738		0.942
How many times more than 3 drinks	Less than one time per month	1.150	-1.406	3.707	0.763	0.377	0.942
	One time per month	0.456	-2.848	3.760		0.786	
	One time per week or more	-1.813	-8.240	4.614		0.579	
	Never					Reference level	
BPS total - Trimester 2	Between LOD and LOQ	-1.058	-4.247	2.130	0.765	0.514	0.942
	>LOQ	-0.336	-1.998	1.326		0.691	
	<LOD					Reference level	
ETPA total - Trimester 2		-0.003	-0.021	0.015	0.765		0.942
MEHP - Trimester 2		-0.054	-0.493	0.384	0.807		0.957
Any slimming diet before pregnancy	Yes	-0.161	-1.802	1.480	0.847	0.847	0.968
	No					Reference level	
MECPP - Trimester 2		0.017	-0.260	0.293	0.907		0.975
PM - Trimester 1		0.003	-0.095	0.100	0.959		0.975
NO2 - Trimester 1		-0.002	-0.099	0.094	0.961		0.975

Number of alcohol drinks per month during pregnancy	Less than one glass per month	-0.037	-1.967	1.893	0.975	0.970	0.975
	More than one glass per month	0.416	-3.311	4.144		0.826	
	No alcohol				Reference level		

BPA: bisphenol A; BPF: bisphenol F; BPS: bisphenol S; BUPA: butyl paraben; cxMiNP: Mono-4-methyl-7-carboxyoctyl-phthalate; ETPA: ethyl paraben; LOD : limit of detection; LOQ: limit of quantification; MBzP: mono benzyl phthalate; MECPP: mono-2-ethyl 5-carboxypentyl phthalate ; MEHHP: mono-2-ethyl-5-hydroxyhexyl phthalate; MEHP: Mono(2-éthylhexyl) phthalate; MEOHP: mono-2-ethyl-5-oxohexyl phthalate; MEP: monoethyl phthalate; MEPA: methyl paraben; MiBP: mono-iso-butyl phthalate; MMCHP: Mono-2-carboxymethyl hexyl-phthalate ; MnBP: mono-n-butyl phthalate; ohMiNCH: 2-(((Hydroxy-4-methyloctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid; OHMiNP: mono-4-methyl-7-hydroxyoctyl phthalate; ohMPHP: 6-Hydroxy Monopropylheptyl-phthalate; OXBE:oxybenzone/benzophenone-3; oxoMiNCH: 2-(((4-Methyl-7-oxyoctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid; OXOMiNP: mono-4-methyl-7-oxooctyl phthalate; PRPA: propyl paraben

Table 0.4: Agnostic multiple regression relating the exposome to the birth weight: estimates, confidence intervals, uncorrected and corrected for multiple comparisons p-values of the tests of association between exposome and birthweight adjusted on relevant covariates (434 mother-child pairs from the Sepages cohort).

Exposure	Modality	Estimate	CI2.5	CI97.5	Unadjusted p-value	Unadjusted p-value by level	p-value corrected for multiple testing
MEHP - Trimester 2		22.981	-1.162	47.124	0.062		0.672
MEHHP - Trimester 2		-20.868	-44.191	2.454	0.079		0.672
cxMiNP - Trimester 2		3.595	-1.258	8.448	0.146		0.672
BPS total - Trimester 2	Between LOD and LOQ	-40.122	-216.375	136.131	0.158	0.655	0.672
	>LOQ	83.226	-7.062	173.514	0.158	0.071	0.672
	<LOD				0.158		0.672
MEPA - Trimester 2		0.078	-0.035	0.191	0.174		0.672
Any slimming diet before pregnancy	Yes	62.935	-31.065	156.935	0.189	0.189	0.672
Any slimming diet before pregnancy	No				0.189		0.672
MECPP - Trimester 2		-9.893	-25.341	5.555	0.209		0.672
BPF total - Trimester 2	>LOQ	-182.968	-474.919	108.983	0.219	0.219	0.672
	<LOD				0.219		0.672
MiBP - Trimester 2		1.828	-1.148	4.803	0.228		0.672
MMCHP - Trimester 2		11.790	-7.935	31.515	0.241		0.672
OXBE total - Trimester 2		-0.321	-0.878	0.235	0.257		0.672
Number of alcohol drinks per month during pregnancy	Less than one glass per month	-10.993	-118.175	96.190	0.266	0.840	0.672
	More than one glass per month	168.556	-38.200	375.312	0.266	0.110	0.672
	No alcohol				0.266		0.672
BUPA total - Trimester 2	Between LOD and LOQ	69.449	-41.058	179.956	0.273	0.217	0.672
	>LOQ	-55.957	-178.807	66.894	0.273	0.371	0.672
	<LOD				0.273		0.672
Self-assessed stress level	Nervous most of the time	-279.230	-573.150	14.689	0.317	0.063	0.723

	Often nervous	6.631	-99.763	113.025	0.317	0.903	0.723
	Never nervous	-4.418	-117.370	108.534	0.317	0.939	0.723
	Sometimes nervous				0.317		0.723
MEOHP - Trimester 2		18.179	-19.164	55.522	0.339		0.723
ohMiNP - Trimester 2		-2.378	-8.038	3.282	0.409		0.794
PRPA total - Trimester 2		-0.249	-0.882	0.383	0.438		0.794
PM - Trimester 1		-2.183	-7.817	3.450	0.446		0.794
How many times more than 3 drinks	Less than one time per month	41.458	-101.156	184.073	0.606	0.568	0.935
	One time per month	-58.308	-248.316	131.700	0.606	0.547	0.935
	One time per week or more	191.093	-177.170	559.357	0.606	0.308	0.935
	Never				0.606		0.935
Slimming diet during pregnancy	Yes	-53.197	-286.875	180.480	0.655	0.655	0.935
	No				0.655		0.935
TRCB total - Trimester 2	Between LOD and LOQ	-176.696	-721.172	367.780	0.665	0.524	0.935
	>LOQ	241.056	-496.590	978.702	0.665	0.521	0.935
	<LOD				0.665		0.935
MnBP - Trimester 2		-0.747	-4.490	2.996	0.695		0.935
oxoMINCH - Trimester 2		-6.147	-46.086	33.793	0.762		0.935
ohMPHP - Trimester 2		2.760	-17.197	22.717	0.786		0.935
oxoMiNP - Trimester 2		-0.809	-8.355	6.738	0.833		0.935
ETPA total - Trimester 2		0.103	-0.889	1.095	0.838		0.935
TRCS total - Trimester 2		-0.025	-0.268	0.217	0.838		0.935
ohMINCH - Trimester 2		2.425	-22.059	26.909	0.846		0.935
BPA - Trimester 2		1.114	-10.248	12.476	0.847		0.935
MEP - Trimester 2		0.040	-0.564	0.643	0.898		0.957
NO2 - Trimester 1		-0.165	-5.618	5.289	0.953		0.974
MBzP - Trimester 2		-0.076	-4.770	4.617	0.974		0.974

BPA: bisphenol A; BPF: bisphenol F; BPS: bisphenol S; BUPA: butyl paraben; cxMiNP: Mono-4-methyl-7-carboxyoctyl-phthalate; ETPA: ethyl paraben; LOD : limit of detection; LOQ: limit of quantification; MBzP: mono benzyl phthalate; MECPP: mono-2-ethyl 5-carboxypentyl phthalate ; MEHHP: mono-2-ethyl-5-hydroxyhexyl phthalate; MEHP: Mono(2-éthylhexyl) phthalate; MEOHP: mono-2-ethyl-5-oxohexyl phthalate; MEP: monoethyl phthalate; MEPA: methyl paraben; MiBP: mono-iso-butyl phthalate; MMCHP: Mono-2-carboxymethyl hexyl-phthalate ; MnBP: mono-n-butyl phthalate; ohMiNCH: 2-(((Hydroxy-4-methyloctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid; OHMiNP:

APPENDIX II

mono-4-methyl-7-hydroxyoctyl phthalate; ohMPHP: 6-Hydroxy Monopropylheptyl-phthalate; OXBE:oxybenzone/benzophenone-3; oxoMiNCH: 2-(((4-Methyl-7-oxyoctyl)oxy)carbonyl)cyclohexanecarboxylic-Acid; OXOMiNP: mono-4-methyl-7-oxooctyl phthalate; PRPA: propyl paraben

4. Discussion

Neither the novel implementation of our oMITM using PLS nor the implementation repeating the oMITM applied in (Cadiou et al., 2020) did point any exposure of our small exposome as related to child birthweight. The first step of oMITM, the supervised dimension reduction of the methylome, gave plausible results: indeed, the genes on which the CpGs having the highest weight in the PLS component built to predict birth weight were located seem relevant for birth weight according to the literature. *JARID2* and *TBC1D24* methylation level have been linked to birth weight in a meta-analysis of genome-wide methylation association studies (Küpers et al., 2019); *JARID2* is a gene expected to have an important role in fetal growth and cell differentiation (Cervantes et al., 2017; Landeira et al., 2015). *PTPRS* methylation levels was found associated with child obesity (Samblas et al., 2018) and is part of inflammation regulation pathways (Bunin et al., 2015). Embryos with reduced expression of *ITPK1* are more likely to show growth retardation and adverse birth outcome such as neural tube defects (Chamberlain et al., 2007; Greene et al., 2017). *PDE4DIP* is involved in hippocampal differential gene expression, which has been associated with low birth weight (Buschdorf et al., 2016). Last, *MORF4L1* has a role in embryonic development via chromatin remodeling and transcriptional regulation (Sang et al., 2019). On the contrary, the two sensitivity analyses involving methylome-wide type association tests with birth weight did not point any specific CpG (sensitivity analysis II and III). Thus, PLS regression may be a relevant alternative way to perform dimension reduction when the sample size makes the classical methylome-wide analysis difficult to use due to a low power. Another possibility could have been to perform an a priori selection of CpG specific to the birth weight instead of using the selection made for child body mass index in (Cadiou et al., 2020)).

When correcting for multiple comparisons, no significant association was found between this PLS component and any pregnancy exposure. Without this correction, the average level of stress self-assessed by the mother (qualitative variable with 4 modalities) was the only exposure associated

with the PLS component. Overall, as the reduced exposome was empty, the oMITM approach did not point any association.

Regarding the agnostic analysis by multiple regression, no significant associations were found, either with or without correction for multiple testing. In particular, the self-assessed level of stress, which was the only exposure associated with the PLS component without correction for multiple testing, was not associated with birth weight (unadjusted p-value: 0.32); among the three modalities of the variable (which quantified the frequency of the feeling of stress) compared to the reference level (the mother “sometimes felt stressed”), the most significant association with birth weight was for the feeling of “almost always being stressed” (p-value: 0.06): the effect estimate was negative, corresponding to a lower birth weight with higher level of self-assessed stress. The two other modalities showed higher p-value (higher than 0.9). In past studies, high level of anxiety during pregnancy was found associated with lower birth weight (Dunkel Schetter and Tanner, 2012) with possible involvement of epigenetic mechanisms (DeSocio, 2018). This may be indicative of a lack of a sensitivity both in the oMITM and in the agnostic approaches, supported by the fact that only a very small number of women (7 among 434) declared a high level of stress. Similarly, in the agnostic approach, we did not find any association of low birth weight with alcohol consumption during pregnancy or air pollutants, which have been pointed by previous studies (Little, 1977; Mills et al., 1984; Stieb et al., 2012; Strandberg-Larsen et al., 2017). Overall, these results put a new perspective on our work on the oMITM design developed in (Cadiou et al., 2020) (see Chapter II). The oMITM design was designed as a way to face the challenge of false-positive discoveries in exposome study: it indeed allows to gain in specificity (see Chapter II and IV) and even in sensitivity (see Chapter III and IV) when the intermediate layer carries some information. However, the informed dimension reduction of the exposome that we proposed cannot go without a sample size sufficient for assessing exposures effects without correction for multiple testing: in particular, if no exposure is associated with the outcome without correction for multiple testing, the oMITM won’t be able to point any exposures. Indeed, if the reduced exposome is empty, the oMITM cannot

point any exposure, and if the reduced exposome contains one exposure or more, a test corrected for multiple testing of its association with the outcome would necessarily provide associations p-values equal or higher than the p-values of the agnostic analysis uncorrected for multiple testing. This is true when ExWAS or multiple regression corrected for multiple comparisons are used to relate the exposome to the health outcome. But this may also be true if more complex linear algorithms such as DSA or LASSO would be used, as they are expected to give the same results as a multiple regression when the dimension is low, which may be the case for the reduced exposome. This underlined that the improvement of exposome studies may not only require the improvement of methods, in the direction that we followed for this PhD, but also a sustained effort to build larger exposome cohorts.

Acknowledgment

We acknowledge the input of Stephan Gabet and Johanna Lepeule, from IAB Grenoble, for the processing of the methylation data.

The cohort was supported by the European Research Council (consolidator grant N°311765-EDOHaD, PI, R. Slama), by the European Community's Seventh Framework Programm (FP7/2007-206, grant N°308333-HELIX, PI, M. Vrijheid), by ANR, the French Research Agency (PAPER project ANR-12-PDOC-0029-01, PI, J. Lepeule; SHALCOH project, 14-CE21-0007-01, PI, R. Slama; GUMME project, PI, R. Slama; ETAPE ANR 18-CE36-005, PI, J. Lepeule; SYMER project, ANR-15-IDEX-02, PI, U. Schlattner, Mobil'Air project, ANR-15-IDEX, PI, S. Mathy, supported by University Grenoble-3Alpes), by ANSES (CNAP project, PI C. Philippat, PENDORE project, PI, V. Siroux), by Plan Cancer (Canc'Air project, PI, P. Guénel), by Association de Recherche sur le Cancer (ARC, PI, P. Guénel), by AGIR pour les maladies chroniques (PI, R. Slama and PRENAPAR project, V. Siroux), and Fonds de Recherche pour la Santé Respiratoire (FRSR, PI, I. Pin) and by Fondation de France (CLIMATHES—00081169, J. Lepeule).

APPENDIX III: Large supplementary materials

Supplementary Material II.10	263
Supplementary Material IV.2	312
Supplementary Table V.3	381
Supplementary Material V.1	409

Supplementary Material II.10: Sensitivity analysis III - adjusted associations between the whole methylome and zBMI in 1,173 children from the HELIX cohort (ExWAS model, step b of the Meet-in-the-Middle approach applied to the whole methylome). Results are presented only for CpGs with a (FDR - corrected for multiple hypothesis testing) p-value below 0.05 in ExWAS.

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
SFXN5	cg02032125	-6.27	-3.28	-4.77	5.27x10 ⁻¹⁰
	cg10040131	3.55	7.26	5.41	1.42x10 ⁻⁸
	cg16549957	3.27	7.08	5.18	1.15x10 ⁻⁷
CLK1	cg21990144	4.18	9.01	6.59	1.01x10 ⁻⁷
ARHGEF3	cg25799109	-3.49	-1.60	-2.55	1.64x10 ⁻⁷
FAM107B	cg04638265	2.76	6.02	4.39	1.49x10 ⁻⁷
TMEM126B	cg21787323	-5.42	-2.46	-3.94	2.02x10 ⁻⁷
CBFA2T3	cg27061485	-5.57	-2.52	-4.05	2.34x10 ⁻⁷
PIK3CD	cg15701170	2.93	6.66	4.79	5.31x10 ⁻⁷
FGR	cg09845000	-5.05	-2.20	-3.62	6.72x10 ⁻⁷
	cg11532433	-3.87	-1.72	-2.79	4.05x10 ⁻⁷
DUSP2	cg13725590	1.97	4.48	3.22	5.38x10 ⁻⁷
ATP2B2	cg09477124	4.21	9.63	6.92	6.02x10 ⁻⁷
ZNF620	cg07448319	-6.97	-3.02	-5.00	7.87x10 ⁻⁷
ACPL2	cg22893494	-8.57	-3.76	-6.16	5.80x10 ⁻⁷
SLC12A7	cg04467119	-9.54	-4.14	-6.84	7.55x10 ⁻⁷
TAP1	cg08818207	1.84	4.26	3.05	9.22x10 ⁻⁷
FOXK1	cg11931953	3.08	7.14	5.11	9.13x10 ⁻⁷
	cg06688910	-5.93	-2.59	-4.26	6.18x10 ⁻⁷
	cg13514049	-5.45	-2.36	-3.90	8.41x10 ⁻⁷
DTX1	cg19743522	3.98	9.21	6.59	8.64x10 ⁻⁷
	cg26438942	2.26	5.22	3.74	8.10x10 ⁻⁷
ZNF710	cg00624799	4.54	10.30	7.42	4.90x10 ⁻⁷
	cg09690072	-8.30	-3.63	-5.96	6.32x10 ⁻⁷
ITGB2	cg13315706	2.68	6.08	4.38	4.96x10 ⁻⁷
	cg20510033	1.94	4.55	3.25	1.15x10 ⁻⁶
	cg00024471	-5.35	-2.29	-3.82	1.12x10 ⁻⁶

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg07708521	-8.79	-3.74	-6.26	1.30x10 ⁻⁶
NCALD	cg19308132	-9.55	-4.07	-6.81	1.24x10 ⁻⁶
KIAA0748	cg01874152	1.68	3.92	2.80	1.10x10 ⁻⁶
ARMC7	cg02186444	-6.11	-2.61	-4.36	1.22x10 ⁻⁶
PLEKHF1	cg14945937	3.18	7.47	5.33	1.28x10 ⁻⁶
ZCCHC3	cg10107382	-4.33	-1.85	-3.09	1.24x10 ⁻⁶
	cg26033520	2.96	6.97	4.97	1.35x10 ⁻⁶
CPEB2	cg07809027	-4.49	-1.90	-3.20	1.40x10 ⁻⁶
PIK3CD	cg01943221	2.47	5.85	4.16	1.51x10 ⁻⁶
ZNF250	cg10639435	2.33	5.51	3.92	1.51x10 ⁻⁶
	cg02381820	65.99	27.75	-46.87	1.72x10 ⁻⁶
VEGFA	cg25343661	-4.45	-1.87	-3.16	1.68x10 ⁻⁶
PDE4D	cg09062288	2.81	6.70	4.75	1.77x10 ⁻⁶
ZFAT	cg16921643	3.64	8.67	6.15	1.82x10 ⁻⁶
CCDC88C	cg07112604	2.44	5.82	4.13	1.89x10 ⁻⁶
LGALS3BP	cg17836612	2.07	4.97	3.52	2.13x10 ⁻⁶
FAM100B	cg04912316	2.93	7.05	4.99	2.30x10 ⁻⁶
H2AFY	cg01874869	-5.42	-2.25	-3.84	2.40x10 ⁻⁶
LCN6	cg14611112	-4.89	-2.02	-3.46	2.57x10 ⁻⁶
PIK3CD	cg23098018	2.12	5.16	3.64	2.80x10 ⁻⁶
DHCR24	cg17901584	-2.98	-1.22	-2.10	2.97x10 ⁻⁶
ETS1	cg08452327	11.83	28.77	20.30	2.90x10 ⁻⁶
DGKA	cg06739462	2.19	5.33	3.76	2.86x10 ⁻⁶
DTX1	cg04456029	2.68	6.52	4.60	2.94x10 ⁻⁶
	cg26504467	1.81	4.41	3.11	2.97x10 ⁻⁶
ANKRD29	cg15393702	-5.09	-2.10	-3.60	2.70x10 ⁻⁶
CUTA	cg01300096	2.62	6.39	4.50	3.09x10 ⁻⁶
FGR	cg00404394	-6.17	-2.52	-4.34	3.40x10 ⁻⁶
DGKA	cg06915826	1.24	3.04	2.14	3.39x10 ⁻⁶
	cg10643916	-6.17	-2.52	-4.35	3.32x10 ⁻⁶
CDV3	cg21466315	1.57	3.85	2.71	3.54x10 ⁻⁶
PNPLA6	cg25025866	-9.38	-3.82	-6.60	3.54x10 ⁻⁶
FOXK1	cg11647481	3.22	7.92	5.57	3.67x10 ⁻⁶
PEAR1	cg13224583	1.58	3.89	2.73	3.82x10 ⁻⁶
FMNL3	cg04070601	2.23	5.49	3.86	3.86x10 ⁻⁶
	cg18232497	3.45	8.62	6.04	5.11x10 ⁻⁶
LOC284632	cg15571353	-7.04	-2.81	-4.93	5.23x10 ⁻⁶
	cg17945429	1.27	3.18	2.23	5.11x10 ⁻⁶

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
HPCAL1	cg04618171	3.46	8.62	6.04	4.83x10 ⁻⁶
	cg03765419	10.75	-4.32	-7.54	4.73x10 ⁻⁶
DUSP2	cg02431562	1.88	4.68	3.28	4.92x10 ⁻⁶
SPATA5	cg06695691	-5.70	-2.27	-3.99	5.65x10 ⁻⁶
PDLIM4	cg17852932	10.11	-4.06	-7.08	4.89x10 ⁻⁶
LY6G6E	cg13123009	3.28	8.14	5.71	4.30x10 ⁻⁶
FKBPL	cg18080528	15.07	-6.08	-10.58	4.33x10 ⁻⁶
LAMA4	cg11934419	-6.64	-2.65	-4.65	5.40x10 ⁻⁶
PHACTR2	cg10735015	-2.79	-1.11	-1.95	5.69x10 ⁻⁶
FOXK1	cg05066096	2.23	5.58	3.90	5.46x10 ⁻⁶
GIMAP8	cg12644845	1.79	4.47	3.13	5.19x10 ⁻⁶
LY96	cg13213009	1.27	3.18	2.22	5.37x10 ⁻⁶
EIF2C2	cg23731089	1.63	4.04	2.83	4.07x10 ⁻⁶
ALOX5	cg10909790	-5.62	-2.24	-3.93	5.62x10 ⁻⁶
	cg22648996	2.15	5.33	3.74	4.49x10 ⁻⁶
RNF141	cg07565042	-6.04	-2.42	-4.23	4.83x10 ⁻⁶
C11orf58	cg05512310	-7.56	-3.04	-5.30	4.54x10 ⁻⁶
ZNF259	cg12395125	15.68	-6.33	-11.01	4.26x10 ⁻⁶
DGKA	cg10782923	1.36	3.42	2.39	5.50x10 ⁻⁶
INTS6	cg21028785	3.73	9.28	6.50	4.81x10 ⁻⁶
TMX1	cg16565294	1.84	4.55	3.20	4.42x10 ⁻⁶
	cg23033518	-5.88	-2.34	-4.11	5.59x10 ⁻⁶
VAC14	cg00259097	2.15	5.33	3.74	4.64x10 ⁻⁶
COMP	cg15986030	-7.78	-3.12	-5.45	5.13x10 ⁻⁶
	cg00522276	10.86	-4.32	-7.59	5.81x10 ⁻⁶
	cg16522993	-6.81	-2.70	-4.75	6.01x10 ⁻⁶
GATS	cg23221013	2.30	5.79	4.05	6.11x10 ⁻⁶
C8orf73	cg11904266	-7.27	-2.89	-5.08	6.10x10 ⁻⁶
HADH	cg01150799	1.86	4.70	3.28	6.25x10 ⁻⁶
PSMB8	cg23750151	3.36	8.48	5.92	6.22x10 ⁻⁶
FGR	cg00065048	-3.51	-1.39	-2.45	6.52x10 ⁻⁶
PLEK	cg10812236	-5.81	-2.30	-4.05	6.54x10 ⁻⁶
CADPS	cg13716443	11.49	-4.54	-8.02	6.78x10 ⁻⁶
HNRNPAB	cg18428006	2.28	5.76	4.02	6.82x10 ⁻⁶

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
CUTA	cg14021478	1.97	4.98	3.47	6.68x10 ⁻⁶
	cg04326499	1.63	4.13	2.88	6.86x10 ⁻⁶
SLC35C1	cg05151185	-6.72	-2.65	-4.68	6.98x10 ⁻⁶
TBK1	cg06092729	-6.68	-2.64	-4.66	6.92x10 ⁻⁶
TMC6	cg07313882	3.52	8.92	6.22	6.95x10 ⁻⁶
SBNO2	cg18608055	1.75	4.43	3.09	6.81x10 ⁻⁶
FAM111A	cg18529243	-7.06	-2.78	-4.92	7.13x10 ⁻⁶
	cg03946955	-16.97	-6.67	-11.82	7.31x10 ⁻⁶
LMNA	cg19414383	1.73	4.40	3.06	7.53x10 ⁻⁶
	cg21366673	1.29	3.29	2.29	7.51x10 ⁻⁶
HLA-E	cg13036546	1.26	3.22	2.24	7.76x10 ⁻⁶
GPER	cg10055222	-9.93	-3.89	-6.91	7.74x10 ⁻⁶
CUX1	cg10657965	-12.44	-4.88	-8.66	7.64x10 ⁻⁶
	cg17945323	-5.14	-2.01	-3.57	7.74x10 ⁻⁶
ANKMY1	cg03978514	-7.26	-2.84	-5.05	8.21x10 ⁻⁶
CALHM2	cg23753748	-6.21	-2.43	-4.32	8.30x10 ⁻⁶
BNIP3	cg10206933	1.47	3.76	2.62	8.21x10 ⁻⁶
SLC48A1	cg16170936	-3.64	-1.42	-2.53	8.24x10 ⁻⁶
CHFR	cg21232015	1.57	4.01	2.79	8.30x10 ⁻⁶
SOCS1	cg05730996	-21.48	-8.39	-14.93	8.31x10 ⁻⁶
	cg18740175	2.17	5.53	3.85	7.92x10 ⁻⁶
TPST2	cg09856467	-5.19	-2.03	-3.61	8.12x10 ⁻⁶
	cg25918827	2.08	5.35	3.72	9.20x10 ⁻⁶
UXS1	cg00295485	1.92	4.94	3.43	8.98x10 ⁻⁶
CCRL2	cg16021018	-6.77	-2.63	-4.70	8.99x10 ⁻⁶
PPP1R2	cg27230534	1.96	5.04	3.50	8.76x10 ⁻⁶
NEU1	cg27645345	-8.25	-3.21	-5.73	9.03x10 ⁻⁶
TRIM8	cg06662132	7.75	19.91	13.83	8.92x10 ⁻⁶
GLB1L2	cg15868183	2.47	6.35	4.41	9.22x10 ⁻⁶
	cg12212198	-3.61	-1.41	-2.51	8.58x10 ⁻⁶
	cg13617812	-6.33	-2.46	-4.39	9.16x10 ⁻⁶
CORO1A	cg06038367	3.11	7.99	5.55	8.98x10 ⁻⁶
	cg15238325	1.62	4.15	2.88	8.55x10 ⁻⁶
CHD3	cg22695339	-18.62	-7.24	-12.93	9.01x10 ⁻⁶
	cg13700939	-4.90	-1.90	-3.40	9.34x10 ⁻⁶
PMEPA1	cg26681770	-3.12	-1.21	-2.16	9.36x10 ⁻⁶

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
C1orf126	cg05338397	2.04	5.36	3.70	1.31x10 ⁻⁵ 0.03
ECE1	cg13551754	2.13	5.54	3.83	1.17x10 ⁻⁵ 0.03
SLC9A1	cg17820878	1.73	4.55	3.14	1.32x10 ⁻⁵ 0.03
RNF220	cg13573582	-2.55	-0.97	-1.76	1.27x10 ⁻⁵ 0.03
LMO4	cg20633321	-4.82	-1.86	-3.34	1.11x10 ⁻⁵ 0.03
CREB3L4	cg09895920	-2.34	-0.89	-1.62	1.28x10 ⁻⁵ 0.03
	cg10421194	12.92	-4.95	-8.93	1.18x10 ⁻⁵ 0.03
PLEK	cg02861056	-4.77	-1.82	-3.30	1.22x10 ⁻⁵ 0.03
GLI2	cg08561286	3.14	8.16	5.65	1.09x10 ⁻⁵ 0.03
	cg20402826	1.32	3.45	2.39	1.16x10 ⁻⁵ 0.03
SF3B1	cg07204893	-6.84	-2.63	-4.74	1.09x10 ⁻⁵ 0.03
SP140	cg05564251	1.89	4.88	3.38	9.96x10 ⁻⁶ 0.03
SUSD5	cg13747145	-9.15	-3.49	-6.32	1.27x10 ⁻⁵ 0.03
MGLL	cg03781224	-8.38	-3.20	-5.79	1.29x10 ⁻⁵ 0.03
VWA5B2	cg00985388	-5.38	-2.07	-3.72	1.13x10 ⁻⁵ 0.03
NFKB1	cg25905215	2.88	7.57	5.23	1.31x10 ⁻⁵ 0.03
	cg13316433	1.73	4.50	3.12	1.13x10 ⁻⁵ 0.03
PAM	cg17259761	-8.30	-3.20	-5.75	1.08x10 ⁻⁵ 0.03
STARD4	cg11925263	-3.93	-1.51	-2.72	1.16x10 ⁻⁵ 0.03
NEU1	cg02776448	-9.57	-3.70	-6.63	1.02x10 ⁻⁵ 0.03
TCP11	cg09027493	2.41	6.28	4.35	1.16x10 ⁻⁵ 0.03
	cg09009770	2.67	6.96	4.81	1.14x10 ⁻⁵ 0.03
RAB11FIP1	cg19626138	-5.51	-2.12	-3.81	1.11x10 ⁻⁵ 0.03
CHD7	cg14719959	3.08	8.04	5.56	1.22x10 ⁻⁵ 0.03
EIF2C2	cg13157980	1.91	5.01	3.46	1.32x10 ⁻⁵ 0.03
ZC3H3	cg11848483	1.12	2.93	2.03	1.19x10 ⁻⁵ 0.03
BAT2L1	cg13637151	6.42	16.82	11.62	1.27x10 ⁻⁵ 0.03
SEC31B	cg23599026	2.41	6.30	4.35	1.21x10 ⁻⁵ 0.03
	cg12378753	-6.41	-2.45	-4.43	1.25x10 ⁻⁵ 0.03
TRIM66	cg08926040	28.49	10.86	-19.68	1.30x10 ⁻⁵ 0.03
	cg22843803	2.87	7.51	5.19	1.24x10 ⁻⁵ 0.03
PITPNM1	cg04383058	2.43	6.35	4.39	1.20x10 ⁻⁵ 0.03
ALDH3B1	cg05620821	-4.93	-1.90	-3.42	1.06x10 ⁻⁵ 0.03
DGKA	cg07679948	1.11	2.87	1.99	9.66x10 ⁻⁶ 0.03
	cg03670162	-8.22	-3.15	-5.68	1.18x10 ⁻⁵ 0.03
	cg26707718	1.92	4.96	3.44	9.79x10 ⁻⁶ 0.03
	cg25936358	-3.96	-1.51	-2.74	1.27x10 ⁻⁵ 0.03
MPP5	cg17327331	-4.86	-1.87	-3.37	1.06x10 ⁻⁵ 0.03

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
KIAA1199	cg17820039	-6.60	-2.53	-4.56	1.19x10 ⁻⁵	0.03
SLCO3A1	cg01794926	-3.62	-1.38	-2.50	1.23x10 ⁻⁵	0.03
	cg03854595	2.31	6.03	4.17	1.21x10 ⁻⁵	0.03
FAM38A	cg06007201	-6.10	-2.33	-4.22	1.28x10 ⁻⁵	0.03
	cg07435331	-3.96	-1.52	-2.74	1.13x10 ⁻⁵	0.03
TMEM49	cg01409343	1.61	4.17	2.89	1.05x10 ⁻⁵	0.03
SOCS3	cg18181703	2.59	6.70	4.64	1.03x10 ⁻⁵	0.03
TTC39C	cg05401069	1.63	4.21	2.92	9.61x10 ⁻⁶	0.03
NFATC1	cg13207250	2.51	6.58	4.55	1.30x10 ⁻⁵	0.03
	cg26853368	1.45	3.76	2.61	1.04x10 ⁻⁵	0.03
TMEM90B	cg24122922	-5.15	-1.99	-3.57	1.00x10 ⁻⁵	0.03
TGIF2	cg10566581	2.67	6.97	4.82	1.19x10 ⁻⁵	0.03
	cg12485727	1.96	5.14	3.55	1.32x10 ⁻⁵	0.03
PKIG	cg07097722	0.94	2.46	1.70	1.20x10 ⁻⁵	0.03
PMEPA1	cg08567517	-8.00	-3.07	-5.54	1.15x10 ⁻⁵	0.03
IL2RB	cg21307484	2.13	5.51	3.82	9.76x10 ⁻⁶	0.03
C11orf10	cg10669451	18.49	-7.04	-12.77	1.34x10 ⁻⁵	0.03
RERE	cg03610117	3.13	8.24	5.69	1.38x10 ⁻⁵	0.03
FDPS	cg06352803	-4.20	-1.59	-2.90	1.42x10 ⁻⁵	0.03
MAP4K4	cg13522882	1.30	3.41	2.36	1.40x10 ⁻⁵	0.03
ICOS	cg15344028	1.03	2.73	1.88	1.46x10 ⁻⁵	0.03
ARSK	cg09244244	-5.38	-2.04	-3.71	1.41x10 ⁻⁵	0.03
ITGB1	cg20545410	-4.12	-1.56	-2.84	1.46x10 ⁻⁵	0.03
SFXN3	cg15428620	-5.84	-2.22	-4.03	1.35x10 ⁻⁵	0.03
IFITM1	cg03038262	1.40	3.70	2.55	1.41x10 ⁻⁵	0.03
P4HA3	cg10408430	2.46	6.48	4.47	1.44x10 ⁻⁵	0.03
DGKA	cg25416125	1.49	3.92	2.70	1.45x10 ⁻⁵	0.03
ABCC4	cg05412028	-2.65	-1.01	-1.83	1.39x10 ⁻⁵	0.03
SLC25A29	cg14064024	-3.22	-1.22	-2.22	1.43x10 ⁻⁵	0.03
	cg06836102	1.60	4.23	2.92	1.45x10 ⁻⁵	0.03
CHD3	cg12353788	1.28	3.39	2.33	1.41x10 ⁻⁵	0.03
ORMDL3	cg08932654	-4.56	-1.73	-3.14	1.40x10 ⁻⁵	0.03
AXIN2	cg04239967	3.29	8.70	6.00	1.45x10 ⁻⁵	0.03
	cg01715745	-4.20	-1.59	-2.90	1.50x10 ⁻⁵	0.03
EPB49	cg03979241	-3.50	-1.32	-2.41	1.55x10 ⁻⁵	0.03
MIR1205	cg03611307	1.45	3.83	2.64	1.53x10 ⁻⁵	0.03
ASAP1	cg18822036	10.12	-3.82	-6.97	1.53x10 ⁻⁵	0.03
C11orf45	cg04468568	-6.60	-2.49	-4.55	1.54x10 ⁻⁵	0.03

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
LIMA1	cg06817772	-4.38	-1.65	-3.01	1.54x10 ⁻⁵
	cg06329392	1.40	3.70	2.55	1.52x10 ⁻⁵
HSPA2	cg20014063	-7.42	-2.80	-5.11	1.52x10 ⁻⁵
SLC10A1	cg04227591	-7.74	-2.93	-5.33	1.51x10 ⁻⁵
	cg00430895	-11.67	-4.41	-8.04	1.50x10 ⁻⁵
NFATC1	cg13608166	2.80	7.41	5.11	1.48x10 ⁻⁵
CTAGE5	cg18976418	-2.73	-1.03	-1.88	1.56x10 ⁻⁵
KCNG3	cg21129729	-4.25	-1.60	-2.93	1.58x10 ⁻⁵
ST3GAL5	cg09113070	-4.28	-1.61	-2.94	1.62x10 ⁻⁵
	cg20532925	1.98	5.26	3.62	1.60x10 ⁻⁵
PCCB	cg10109841	-5.16	-1.94	-3.55	1.60x10 ⁻⁵
CHD7	cg22822599	2.60	6.92	4.76	1.61x10 ⁻⁵
SH2D4B	cg14914442	-4.76	-1.79	-3.28	1.63x10 ⁻⁵
DPF2	cg09609051	1.64	4.37	3.01	1.62x10 ⁻⁵
PLA2G4E	cg03080639	-16.77	-6.31	-11.54	1.63x10 ⁻⁵
CLUAP1	cg09934892	1.77	4.70	3.24	1.61x10 ⁻⁵
CLDND2	cg22065498	-5.84	-2.20	-4.02	1.64x10 ⁻⁵
ABCA5	cg09530790	-7.69	-2.89	-5.29	1.68x10 ⁻⁵
	cg08160085	-5.91	-2.22	-4.06	1.70x10 ⁻⁵
DTX1	cg18449739	2.17	5.79	3.98	1.74x10 ⁻⁵
VPS37B	cg05298628	-7.12	-2.67	-4.89	1.75x10 ⁻⁵
CHST11	cg16618104	-4.23	-1.58	-2.90	1.76x10 ⁻⁵
C1orf187	cg12403889	-7.15	-2.67	-4.91	1.83x10 ⁻⁵
SRGAP3	cg12426802	-54.08	20.18	-37.13	1.88x10 ⁻⁵
	cg12593793	-4.86	-1.82	-3.34	1.84x10 ⁻⁵
	cg27228986	1.50	4.01	2.75	1.87x10 ⁻⁵
PTGER4	cg13475822	1.62	4.34	2.98	1.87x10 ⁻⁵
	cg17448336	-6.25	-2.34	-4.29	1.77x10 ⁻⁵
PHF15	cg05476182	-9.75	-3.65	-6.70	1.81x10 ⁻⁵
NEU1	cg26747317	-5.54	-2.07	-3.80	1.83x10 ⁻⁵
PTPRN2	cg09365094	1.70	4.55	3.12	1.83x10 ⁻⁵
	cg17841267	1.69	4.52	3.10	1.86x10 ⁻⁵
BATF	cg24673600	2.46	6.58	4.52	1.83x10 ⁻⁵
ZC3H18	cg04476070	-2.82	-1.05	-1.93	1.86x10 ⁻⁵
MAFG	cg09253473	-5.34	-1.99	-3.67	1.86x10 ⁻⁵
PLK1S1	cg23666170	-5.50	-2.05	-3.78	1.88x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
NAGLU	cg12280065	-9.08	-3.38	-6.23	1.89x10 ⁻⁵	0.03
PHACTR3	cg24334029	-3.80	-1.42	-2.61	1.91x10 ⁻⁵	0.03
MACF1	cg00352106	-3.78	-1.41	-2.60	1.91x10 ⁻⁵	0.03
LAMP3	cg18030943	1.29	3.47	2.38	1.94x10 ⁻⁵	0.03
C6orf27	cg05565809	11.98	-4.46	-8.22	1.93x10 ⁻⁵	0.03
ZNF589	cg05070273	1.58	4.24	2.91	1.94x10 ⁻⁵	0.03
	cg18146927	1.72	4.62	3.17	1.97x10 ⁻⁵	0.03
PEBP4	cg09423126	1.88	5.06	3.47	1.97x10 ⁻⁵	0.03
SLC22A8	cg16146033	-6.94	-2.58	-4.76	1.97x10 ⁻⁵	0.03
	cg10393508	3.61	9.72	6.66	1.98x10 ⁻⁵	0.03
NOSIP	cg21292909	3.44	9.26	6.35	1.98x10 ⁻⁵	0.03
LOC100130872- SPON2	cg25002788	-5.21	-1.94	-3.57	2.00x10 ⁻⁵	0.03
MXRA8	cg23973885	2.17	5.85	4.01	2.08x10 ⁻⁵	0.03
C1orf122	cg04299080	-6.32	-2.34	-4.33	2.17x10 ⁻⁵	0.03
CHI3L2	cg14414943	1.57	4.23	2.90	2.12x10 ⁻⁵	0.03
CD58	cg21039631	16.71	-6.19	-11.45	2.09x10 ⁻⁵	0.03
GALNT2	cg06801857	1.63	4.39	3.01	2.07x10 ⁻⁵	0.03
EML4	cg04741211	1.19	3.23	2.21	2.18x10 ⁻⁵	0.03
HECW2	cg22765299	1.12	3.03	2.07	2.05x10 ⁻⁵	0.03
TGFBR2	cg07285675	2.98	8.03	5.50	2.07x10 ⁻⁵	0.03
RAB43	cg05859076	1.72	4.63	3.17	2.13x10 ⁻⁵	0.03
RFC4	cg03852656	10.25	27.64	18.95	2.08x10 ⁻⁵	0.03
HES1	cg14475875	-5.22	-1.93	-3.57	2.14x10 ⁻⁵	0.03
C5orf56	cg05452391	11.55	-4.28	-7.91	2.15x10 ⁻⁵	0.03
SIT1	cg13876315	1.90	5.12	3.51	2.02x10 ⁻⁵	0.03
AKAP2	cg14488390	10.36	-3.84	-7.10	2.06x10 ⁻⁵	0.03
ABCC2	cg03381359	-3.96	-1.46	-2.71	2.15x10 ⁻⁵	0.03
	cg03558688	2.03	5.46	3.75	2.03x10 ⁻⁵	0.03
PTPRCAP	cg02685484	1.74	4.69	3.22	2.17x10 ⁻⁵	0.03
	cg13133835	-7.20	-2.66	-4.93	2.17x10 ⁻⁵	0.03
SNUPN	cg20630655	-5.53	-2.05	-3.79	2.14x10 ⁻⁵	0.03
	cg00103299	1.37	3.70	2.53	2.12x10 ⁻⁵	0.03
	cg26676129	1.40	3.77	2.58	2.09x10 ⁻⁵	0.03
	cg10334416	-3.75	-1.39	-2.57	2.09x10 ⁻⁵	0.03
YTHDF1	cg12589387	-8.40	-3.11	-5.76	2.12x10 ⁻⁵	0.03

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
LIMK2	cg07713946	1.07	2.90	1.98	2.18x10 ⁻⁵
SPATA6	cg16965552	-2.59	-0.96	-1.77	2.19x10 ⁻⁵
FNIP1	cg21016699	-7.94	-2.93	-5.43	2.20x10 ⁻⁵
DENND2D	cg09826895	1.31	3.56	2.43	2.23x10 ⁻⁵
PHACTR4	cg00956142	-2.59	-0.96	-1.78	2.24x10 ⁻⁵
NEU1	cg19070159	-8.05	-2.97	-5.51	2.25x10 ⁻⁵
	cg15584070	-6.67	-2.46	-4.56	2.28x10 ⁻⁵
TP53INP2	cg15636879	1.65	4.47	3.06	2.29x10 ⁻⁵
SH3BP5	cg07078958	-5.58	-2.06	-3.82	2.30x10 ⁻⁵
CMKLR1	cg06933965	-7.24	-2.67	-4.95	2.35x10 ⁻⁵
PSMB8	cg25693349	1.92	5.21	3.56	2.37x10 ⁻⁵
	cg14343017	1.55	4.21	2.88	2.39x10 ⁻⁵
LOC349114	cg07437464	-6.79	-2.50	-4.64	2.38x10 ⁻⁵
CORO1B	cg17690322	2.75	7.47	5.11	2.37x10 ⁻⁵
CD7	cg11294761	2.66	7.25	4.96	2.39x10 ⁻⁵
PIK3CD	cg09706586	1.81	4.93	3.37	2.46x10 ⁻⁵
	cg12587087	-4.60	-1.69	-3.14	2.45x10 ⁻⁵
	cg22626683	-3.43	-1.26	-2.35	2.42x10 ⁻⁵
	cg03040292	1.16	3.18	2.17	2.53x10 ⁻⁵
PFKFB2	cg02233614	-3.99	-1.46	-2.73	2.49x10 ⁻⁵
	cg17163425	-4.65	-1.70	-3.17	2.53x10 ⁻⁵
CPS1	cg21967368	-3.06	-1.12	-2.09	2.49x10 ⁻⁵
OXNAD1	cg02255107	1.72	4.69	3.20	2.50x10 ⁻⁵
	cg23279756	-3.23	-1.19	-2.21	2.45x10 ⁻⁵
SMURF1	cg08486903	2.05	5.58	3.81	2.47x10 ⁻⁵
PEBP4	cg17836177	1.78	4.86	3.32	2.52x10 ⁻⁵
CDC42BPB	cg21057323	1.17	3.20	2.19	2.47x10 ⁻⁵
RAP1GAP2	cg12061069	-13.88	-5.09	-9.48	2.50x10 ⁻⁵
	cg03834031	-4.92	-1.80	-3.36	2.46x10 ⁻⁵
RCAN3	cg01768001	2.02	5.53	3.78	2.59x10 ⁻⁵
ZNF619	cg12176605	-4.17	-1.52	-2.85	2.57x10 ⁻⁵
STX1A	cg01712428	-5.61	-2.05	-3.83	2.60x10 ⁻⁵
ARPC1B	cg02257708	3.23	8.83	6.03	2.60x10 ⁻⁵
CPEB3	cg24238409	-2.77	-1.01	-1.89	2.60x10 ⁻⁵
	cg26130090	-6.42	-2.35	-4.39	2.60x10 ⁻⁵
RAI1	cg06685437	2.15	5.88	4.02	2.59x10 ⁻⁵
PTPN7	cg07577934	-11.95	-4.36	-8.16	2.67x10 ⁻⁵
C2orf68	cg24222817	-8.43	-3.08	-5.75	2.67x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
ARID5B	cg16389209	1.74	4.78	3.26	2.67x10 ⁻⁵
	cg15227994	1.69	4.64	3.17	2.66x10 ⁻⁵
	cg02353048	-4.40	-1.60	-3.00	2.65x10 ⁻⁵
AZU1	cg24101283	-5.40	-1.97	-3.68	2.68x10 ⁻⁵
BLCAP	cg22378853	-6.94	-2.53	-4.74	2.64x10 ⁻⁵
	cg05945608	1.88	5.15	3.51	2.71x10 ⁻⁵
PIK3CD	cg19267205	2.50	6.87	4.68	2.75x10 ⁻⁵
SESN2	cg06754224	-6.07	-2.21	-4.14	2.74x10 ⁻⁵
RALY	cg19110982	3.07	8.43	5.75	2.76x10 ⁻⁵
DAND5	cg07782285	2.06	5.67	3.87	2.77x10 ⁻⁵
PIK3CD	cg11947782	1.63	4.48	3.05	2.80x10 ⁻⁵
NRM	cg11784631	-6.51	-2.37	-4.44	2.80x10 ⁻⁵
	cg05280814	-4.38	-1.59	-2.99	2.79x10 ⁻⁵
SLC35C1	cg24303076	-8.12	-2.95	-5.54	2.83x10 ⁻⁵
UBASH3A	cg27111890	1.01	2.77	1.89	2.83x10 ⁻⁵
MAN1C1	cg10555744	1.40	3.86	2.63	2.97x10 ⁻⁵
FBLN7	cg01154505	1.39	3.83	2.61	2.90x10 ⁻⁵
PASK	cg25488284	1.97	5.44	3.71	2.92x10 ⁻⁵
	cg15969804	-5.48	-1.99	-3.73	2.94x10 ⁻⁵
	cg06434490	-5.96	-2.16	-4.06	2.91x10 ⁻⁵
	cg01198111	-5.55	-2.02	-3.78	2.90x10 ⁻⁵
FLJ23834	cg00156506	15.72	-5.70	-10.71	2.97x10 ⁻⁵
CHD7	cg25011252	1.74	4.81	3.28	2.93x10 ⁻⁵
EFEMP2	cg20051715	-6.36	-2.31	-4.33	2.95x10 ⁻⁵
BCKDK	cg05299836	1.95	5.35	3.65	2.86x10 ⁻⁵
C17orf108	cg19001909	1.44	3.96	2.70	2.89x10 ⁻⁵
COL1A1	cg18618815	-5.25	-1.90	-3.58	2.96x10 ⁻⁵
TMEM49	cg18942579	1.59	4.38	2.98	2.95x10 ⁻⁵
SH2D3A	cg15055101	1.79	4.92	3.35	2.94x10 ⁻⁵
PGLYRP2	cg07408456	-5.59	-2.03	-3.81	2.94x10 ⁻⁵
	cg13556767	13.15	-4.77	-8.96	2.99x10 ⁻⁵
CHST11	cg06647068	-3.54	-1.28	-2.41	3.04x10 ⁻⁵
PTPRM	cg25279586	-4.84	-1.75	-3.30	3.04x10 ⁻⁵
SYDE1	cg22027399	-5.04	-1.82	-3.43	3.04x10 ⁻⁵
INPP5J	cg26373518	-6.07	-2.20	-4.13	3.05x10 ⁻⁵
PMEPA1	cg00138126	-5.12	-1.85	-3.49	3.06x10 ⁻⁵
IFFO2	cg09012001	-4.91	-1.77	-3.34	3.19x10 ⁻⁵
TMEM222	cg10239563	-6.21	-2.24	-4.22	3.27x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
PTGFRN	cg08870281	1.64	4.57	3.11	3.32x10 ⁻⁵
	cg03388247	-7.01	-2.53	-4.77	3.08x10 ⁻⁵
	cg01100951	-6.98	-2.52	-4.75	3.22x10 ⁻⁵
PPARGC1B	cg26876242	-2.83	-1.02	-1.92	3.31x10 ⁻⁵
SLIT3	cg06007761	-10.37	-3.74	-7.05	3.19x10 ⁻⁵
	cg12694224	-5.69	-2.06	-3.87	3.16x10 ⁻⁵
RNASET2	cg17991206	1.22	3.37	2.29	3.10x10 ⁻⁵
C7orf50	cg08973950	-3.03	-1.09	-2.06	3.33x10 ⁻⁵
SCRN1	cg10113101	-6.10	-2.20	-4.15	3.32x10 ⁻⁵
DDX56	cg13810644	-4.07	-1.47	-2.77	3.10x10 ⁻⁵
NCALD	cg27110054	-5.75	-2.07	-3.91	3.28x10 ⁻⁵
KIFC2	cg04794268	-5.05	-1.82	-3.44	3.15x10 ⁻⁵
SURF4	cg13660174	1.78	4.93	3.36	3.25x10 ⁻⁵
ELOVL3	cg18288462	-5.43	-1.96	-3.70	3.14x10 ⁻⁵
	cg14816825	2.00	5.56	3.78	3.24x10 ⁻⁵
	cg06640997	1.08	2.98	2.03	3.10x10 ⁻⁵
FAM113B	cg06547285	-6.04	-2.17	-4.11	3.31x10 ⁻⁵
MGC14436	cg00050496	2.48	6.88	4.68	3.28x10 ⁻⁵
DTX1	cg18567954	2.47	6.83	4.65	3.15x10 ⁻⁵
C12orf27	cg10969412	-7.62	-2.75	-5.18	3.25x10 ⁻⁵
	cg07899411	1.15	3.20	2.18	3.18x10 ⁻⁵
	cg25015038	-3.61	-1.30	-2.45	3.18x10 ⁻⁵
EVL	cg14245199	1.11	3.09	2.10	3.22x10 ⁻⁵
AKAP13	cg05239225	-2.51	-0.90	-1.71	3.31x10 ⁻⁵
UNKL	cg04928129	-14.02	-5.06	-9.54	3.16x10 ⁻⁵
SF3B3	cg00864954	1.24	3.43	2.33	3.32x10 ⁻⁵
CUEDC1	cg08885409	-11.73	-4.23	-7.98	3.17x10 ⁻⁵
	cg13470125	-6.15	-2.22	-4.18	3.18x10 ⁻⁵
HCN2	cg11638347	-6.11	-2.20	-4.15	3.29x10 ⁻⁵
GALNT2	cg01554316	4.20	11.69	7.94	3.34x10 ⁻⁵
	cg15723468	1.48	4.13	2.81	3.42x10 ⁻⁵
CLNK	cg02424103	-9.22	-3.31	-6.27	3.42x10 ⁻⁵
	cg05582979	-9.79	-3.51	-6.65	3.42x10 ⁻⁵
FER	cg16174681	-4.11	-1.48	-2.80	3.43x10 ⁻⁵
STK17A	cg05415936	1.77	4.91	3.34	3.40x10 ⁻⁵
NEDD1	cg21721825	1.26	3.51	2.39	3.38x10 ⁻⁵
MAFG	cg07855221	-4.46	-1.60	-3.03	3.43x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
SMARCA4	cg15001636	2.31	6.42	4.37	3.38x10 ⁻⁵
ARID5B	cg07520810	2.24	6.25	4.25	3.45x10 ⁻⁵
MAP2K5	cg27219399	-5.77	-2.07	-3.92	3.45x10 ⁻⁵
MICA	cg02726263	17.67	-6.34	-12.00	3.47x10 ⁻⁵
ALDOA	cg00583733	-5.95	-2.13	-4.04	3.47x10 ⁻⁵
GRB2	cg27110374	1.41	3.94	2.68	3.48x10 ⁻⁵
SKI	cg24481594	1.52	4.25	2.88	3.50x10 ⁻⁵
CACNA1C	cg10031793	-3.56	-1.28	-2.42	3.51x10 ⁻⁵
	cg19061690	-7.98	-2.86	-5.42	3.51x10 ⁻⁵
PRDM2	cg24715767	-6.63	-2.37	-4.50	3.58x10 ⁻⁵
IL19	cg15392364	0.95	2.66	1.81	3.59x10 ⁻⁵
SPATS2L	cg08153883	-6.18	-2.21	-4.20	3.56x10 ⁻⁵
	cg10125599	-7.66	-2.74	-5.20	3.60x10 ⁻⁵
SHC3	cg13351583	6.02	16.83	11.43	3.59x10 ⁻⁵
LOC286359	cg21177384	14.88	-5.32	-10.10	3.60x10 ⁻⁵
EHD1	cg07019638	1.47	4.11	2.79	3.59x10 ⁻⁵
ITPK1	cg09257735	-4.48	-1.60	-3.04	3.53x10 ⁻⁵
VAC14	cg08329113	1.53	4.28	2.91	3.59x10 ⁻⁵
DLG4	cg12228229	14.63	-5.24	-9.94	3.58x10 ⁻⁵
RHEB	cg06603923	-3.98	-1.42	-2.70	3.63x10 ⁻⁵
ASAP1IT1	cg03295852	-5.57	-1.99	-3.78	3.64x10 ⁻⁵
C9orf139	cg02971262	-8.98	-3.21	-6.10	3.65x10 ⁻⁵
FAM107B	cg14295960	1.56	4.36	2.96	3.61x10 ⁻⁵
SDCBP2	cg05007126	1.84	5.14	3.49	3.63x10 ⁻⁵
RORA	cg23411013	-4.03	-1.44	-2.74	3.67x10 ⁻⁵
ORMDL3	cg19511844	-7.64	-2.73	-5.19	3.67x10 ⁻⁵
	cg21241675	13.51	-4.82	-9.17	3.69x10 ⁻⁵
LATS2	cg18151858	15.58	-5.56	-10.57	3.73x10 ⁻⁵
	cg03598938	-4.71	-1.68	-3.20	3.74x10 ⁻⁵
HLA-E	cg23235965	1.60	4.48	3.04	3.75x10 ⁻⁵
	cg11323439	-6.19	-2.21	-4.20	3.75x10 ⁻⁵
ADAR	cg00048381	2.03	5.70	3.87	3.77x10 ⁻⁵
	cg06699555	3.29	9.24	6.27	3.79x10 ⁻⁵
WDR82	cg12048331	1.10	3.09	2.10	3.80x10 ⁻⁵
PEBP4	cg01211283	1.79	5.03	3.41	3.81x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
SLC7A7	cg24165921	-7.70	-2.74	-5.22	3.81x10 ⁻⁵
	cg10853566	-6.57	-2.34	-4.46	3.81x10 ⁻⁵
UACA	cg01918043	-4.11	-1.46	-2.79	3.80x10 ⁻⁵
	cg11945474	-3.07	-1.09	-2.08	3.84x10 ⁻⁵
LRRC33	cg02067535	-5.59	-1.99	-3.79	3.87x10 ⁻⁵
	MIR429	2.05	5.78	3.92	4.05x10 ⁻⁵
S100A6	cg24375627	-4.22	-1.50	-2.86	4.13x10 ⁻⁵
	cg10154812	1.13	3.18	2.15	3.95x10 ⁻⁵
PSMB8	cg12048225	1.26	3.56	2.41	4.09x10 ⁻⁵
	cg16597079	10.73	-3.80	-7.27	4.13x10 ⁻⁵
SND1	cg02538772	1.26	3.55	2.40	3.90x10 ⁻⁵
	LZTS1	cg10121113	-7.17	-2.55	-4.86
RHOBTB1	cg05171195	-7.35	-2.61	-4.98	4.01x10 ⁻⁵
	CD81	cg10504000	-4.38	-1.56	-2.97
KCNQ1	cg05438727	-6.25	-2.22	-4.23	4.07x10 ⁻⁵
	cg16981685	1.14	3.22	2.18	4.13x10 ⁻⁵
KCNJ1	cg25893560	1.39	3.90	2.64	3.95x10 ⁻⁵
	cg19369955	-2.70	-0.96	-1.83	4.10x10 ⁻⁵
NCOR2	cg02666020	-5.74	-2.03	-3.89	4.12x10 ⁻⁵
	SLC25A30	cg20558112	-5.05	-1.79	-3.42
STARD5	cg10877241	-3.24	-1.15	-2.19	3.94x10 ⁻⁵
	cg01176433	-3.51	-1.25	-2.38	4.01x10 ⁻⁵
RRN3P2	cg05150327	-8.07	-2.86	-5.47	4.12x10 ⁻⁵
	cg08166214	1.96	5.52	3.74	3.94x10 ⁻⁵
CCL17	cg00525931	-5.22	-1.85	-3.54	4.03x10 ⁻⁵
	cg01100208	11.90	-4.22	-8.06	4.06x10 ⁻⁵
CMIP	cg03766174	-4.38	-1.56	-2.97	4.01x10 ⁻⁵
	MSI2	cg21139312	1.88	5.31	3.60
AXIN2	cg26921093	1.21	3.42	2.32	4.11x10 ⁻⁵
	SEPT9	cg27627381	2.19	6.15	4.17
NPLOC4	cg10954938	-7.08	-2.52	-4.80	3.96x10 ⁻⁵
	BLVRB	cg06663644	10.31	-3.66	-6.98
PATZ1	cg19344545	1.24	3.51	2.38	4.10x10 ⁻⁵
	cg16369835	11.52	32.48	22.00	4.11x10 ⁻⁵
EIF4E3	cg22888463	1.46	4.12	2.79	4.15x10 ⁻⁵
	EIF2C2	cg19352830	1.81	5.12	3.47
	cg05262463	1.34	3.79	2.56	4.19x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
NMNAT3	cg03891318	-6.66	-2.36	-4.51	4.20x10 ⁻⁵	0.03
PIK3CD	cg03971555	1.47	4.15	2.81	4.33x10 ⁻⁵	0.03
ABCB6	cg19412109	-6.56	-2.32	-4.44	4.33x10 ⁻⁵	0.03
	cg05705140	1.01	2.86	1.94	4.30x10 ⁻⁵	0.03
DEPDC1B	cg08795084	-5.94	-2.10	-4.02	4.27x10 ⁻⁵	0.03
	cg11935041	-5.05	-1.79	-3.42	4.26x10 ⁻⁵	0.03
TTYH3	cg03590328	-6.85	-2.42	-4.64	4.30x10 ⁻⁵	0.03
CHRNA6	cg19352832	1.56	4.43	2.99	4.31x10 ⁻⁵	0.03
	cg14832378	0.87	2.47	1.67	4.23x10 ⁻⁵	0.03
SLC25A29	cg02540736	-4.10	-1.45	-2.78	4.26x10 ⁻⁵	0.03
GPR114	cg07877900	-6.75	-2.39	-4.57	4.27x10 ⁻⁵	0.03
	cg10371414	-2.15	-0.76	-1.46	4.33x10 ⁻⁵	0.03
TRAPPC1	cg13021301	-4.81	-1.70	-3.26	4.31x10 ⁻⁵	0.03
KSR1	cg10334489	1.72	4.88	3.30	4.26x10 ⁻⁵	0.03
MMP9	cg10505873	-5.55	-1.96	-3.76	4.30x10 ⁻⁵	0.03
THEMIS	cg11610626	1.21	3.41	2.31	4.35x10 ⁻⁵	0.03
	cg06771839	1.76	4.97	3.36	4.36x10 ⁻⁵	0.03
	cg15013544	-8.77	-3.10	-5.93	4.39x10 ⁻⁵	0.03
SH3PXD2A	cg00277384	-9.44	-3.33	-6.38	4.39x10 ⁻⁵	0.03
	cg07920381	-2.52	-0.89	-1.71	4.39x10 ⁻⁵	0.03
ACOT7	cg25165880	-4.70	-1.66	-3.18	4.46x10 ⁻⁵	0.03
FGGY	cg00581017	-18.54	-6.54	-12.54	4.42x10 ⁻⁵	0.03
	cg06566627	2.26	6.41	4.34	4.45x10 ⁻⁵	0.03
CCND3	cg17159550	-7.82	-2.76	-5.29	4.47x10 ⁻⁵	0.03
HACE1	cg03002526	1.16	3.29	2.23	4.48x10 ⁻⁵	0.03
	cg13540411	1.98	5.61	3.79	4.44x10 ⁻⁵	0.03
	cg09488090	-5.28	-1.86	-3.57	4.44x10 ⁻⁵	0.03
RAB37	cg03469804	-3.19	-1.12	-2.15	4.46x10 ⁻⁵	0.03
C21orf57	cg11715092	-4.74	-1.67	-3.20	4.47x10 ⁻⁵	0.03
MTA3	cg02679503	1.59	4.52	3.05	4.51x10 ⁻⁵	0.03
KIAA1949	cg13357602	-4.14	-1.46	-2.80	4.54x10 ⁻⁵	0.03
UNKL	cg26728422	-5.72	-2.01	-3.87	4.52x10 ⁻⁵	0.03
TTYH2	cg23632849	-7.04	-2.48	-4.76	4.55x10 ⁻⁵	0.03
TTC39C	cg18719665	1.15	3.26	2.20	4.55x10 ⁻⁵	0.03
AIRE	cg26472802	1.19	3.38	2.28	4.53x10 ⁻⁵	0.03
	cg25989526	-2.99	-1.05	-2.02	4.57x10 ⁻⁵	0.04
CCHCR1	cg12044213	-4.67	-1.64	-3.15	4.59x10 ⁻⁵	0.04
PNPLA2	cg14689532	-5.12	-1.80	-3.46	4.64x10 ⁻⁵	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg09142405	-4.99	-1.75	-3.37	4.62x10 ⁻⁵
	cg19515108	1.98	5.62	3.80	4.64x10 ⁻⁵
ADPGK	cg12691651	2.02	5.74	3.88	4.62x10 ⁻⁵
IL24	cg16417028	1.39	3.97	2.68	4.66x10 ⁻⁵
JARID2	cg06487194	-2.26	-0.79	-1.52	4.67x10 ⁻⁵
BCAT1	cg07479001	10.98	-3.86	-7.42	4.68x10 ⁻⁵
CD58	cg26118759	-7.24	-2.54	-4.89	4.72x10 ⁻⁵
RFC1	cg12062995	25.65	-9.01	-17.33	4.71x10 ⁻⁵
CUEDC1	cg00968616	-4.02	-1.41	-2.71	4.72x10 ⁻⁵
HDAC4	cg05570654	-7.17	-2.52	-4.85	4.75x10 ⁻⁵
	cg21360910	-2.83	-0.99	-1.91	4.77x10 ⁻⁵
	cg02888518	-8.67	-3.04	-5.85	4.77x10 ⁻⁵
	cg12073436	0.92	2.61	1.76	4.80x10 ⁻⁵
DENND2C	cg11395799	-8.17	-2.86	-5.52	4.84x10 ⁻⁵
PDE4DIP	cg22712955	1.40	4.01	2.71	4.92x10 ⁻⁵
SLC27A3	cg21279955	-6.96	-2.44	-4.70	4.95x10 ⁻⁵
	cg12697139	-8.29	-2.89	-5.59	5.16x10 ⁻⁵
MRPL55	cg08158976	-4.35	-1.52	-2.94	4.89x10 ⁻⁵
	cg13693136	-4.84	-1.69	-3.27	5.11x10 ⁻⁵
	cg04335293	-6.39	-2.24	-4.31	4.96x10 ⁻⁵
PDCD1	cg25890838	1.83	5.24	3.54	5.21x10 ⁻⁵
CTNNB1	cg09678212	-5.45	-1.90	-3.67	5.22x10 ⁻⁵
SEC22C	cg25686812	4.01	11.47	7.74	5.03x10 ⁻⁵
FRMD4B	cg07164133	13.89	-4.85	-9.37	5.12x10 ⁻⁵
	cg10599446	1.17	3.36	2.26	5.17x10 ⁻⁵
	cg06012804	1.26	3.60	2.43	5.23x10 ⁻⁵
KIAA1530	cg25505570	3.86	11.04	7.45	5.06x10 ⁻⁵
N4BP2	cg21410980	2.85	8.18	5.52	5.19x10 ⁻⁵
TCF7	cg18338046	1.16	3.33	2.25	5.25x10 ⁻⁵
	cg17328407	1.55	4.45	3.00	5.19x10 ⁻⁵
ATXN1	cg03680932	7.93	22.74	15.33	5.18x10 ⁻⁵
ETV7	cg20447038	-4.67	-1.63	-3.15	5.25x10 ⁻⁵
PTK7	cg08558323	-6.78	-2.37	-4.58	5.09x10 ⁻⁵
MTHFD1L	cg22954978	12.63	-4.41	-8.52	5.04x10 ⁻⁵
JAZF1	cg22938901	-5.85	-2.04	-3.95	5.20x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
TMEM140	cg12577010	-11.51	-4.02 -7.76	5.04x10 ⁻⁵	0.04
	cg00459992	1.50	4.28 2.89	4.98x10 ⁻⁵	0.04
NSMAF	cg25874782	-4.16	-1.46 -2.81	4.90x10 ⁻⁵	0.04
ZC3H3	cg24561207	3.03	8.65 5.84	4.82x10 ⁻⁵	0.04
	cg14554880	-5.70	-1.99 -3.85	5.21x10 ⁻⁵	0.04
INPP5A	cg13521018	0.77	2.21 1.49	5.18x10 ⁻⁵	0.04
APBB1	cg16230121	1.05	3.00 2.02	4.86x10 ⁻⁵	0.04
	cg24068972	1.77	5.06 3.42	4.97x10 ⁻⁵	0.04
BSCL2	cg23916044	-6.71	-2.35 -4.53	4.91x10 ⁻⁵	0.04
SNORD22	cg18159646	2.18	6.23 4.20	4.90x10 ⁻⁵	0.04
EHD1	cg14378925	1.48	4.25 2.87	5.25x10 ⁻⁵	0.04
SIK3	cg06928797	-4.88	-1.70 -3.29	5.18x10 ⁻⁵	0.04
HDAC7	cg15820955	1.89	5.42 3.66	5.11x10 ⁻⁵	0.04
CAB39L	cg02210115	-3.24	-1.13 -2.19	4.93x10 ⁻⁵	0.04
	cg11879741	-4.81	-1.68 -3.25	5.06x10 ⁻⁵	0.04
LTK	cg21867850	-18.16	-6.33 -12.24	5.22x10 ⁻⁵	0.04
	cg21024495	-4.09	-1.43 -2.76	5.07x10 ⁻⁵	0.04
KSR1	cg05784862	1.31	3.74 2.52	5.11x10 ⁻⁵	0.04
RAB34	cg12873610	-5.33	-1.87 -3.60	4.94x10 ⁻⁵	0.04
TMC8	cg26003388	1.87	5.37 3.62	5.16x10 ⁻⁵	0.04
GNG7	cg22459924	1.28	3.68 2.48	5.04x10 ⁻⁵	0.04
FAM129C	cg12550597	-5.55	-1.94 -3.75	5.11x10 ⁻⁵	0.04
ITPKC	cg21202522	-7.89	-2.76 -5.32	4.91x10 ⁻⁵	0.04
OXT	cg13285174	-2.65	-0.92 -1.79	5.17x10 ⁻⁵	0.04
C21orf96	cg16071713	-10.36	-3.62 -6.99	5.10x10 ⁻⁵	0.04
YPEL1	cg02774935	-3.71	-1.30 -2.50	5.05x10 ⁻⁵	0.04
TAP1	cg17626301	0.95	2.72 1.83	5.26x10 ⁻⁵	0.04
LGALS2	cg14711067	-8.76	-3.05 -5.91	5.28x10 ⁻⁵	0.04
FAM63A	cg21937128	-3.99	-1.38 -2.68	5.48x10 ⁻⁵	0.04
FAM179A	cg04304121	-3.26	-1.13 -2.19	5.52x10 ⁻⁵	0.04
SRBD1	cg12939390	-2.96	-1.03 -1.99	5.50x10 ⁻⁵	0.04
PLEK	cg04872689	-6.91	-2.40 -4.66	5.40x10 ⁻⁵	0.04
	cg24407086	-10.67	-3.70 -7.18	5.61x10 ⁻⁵	0.04
ECEL1P2	cg09490371	1.32	3.81 2.56	5.62x10 ⁻⁵	0.04
	cg16033723	-7.87	-2.73 -5.30	5.62x10 ⁻⁵	0.04
ACAP2	cg05141059	-4.39	-1.53 -2.96	5.46x10 ⁻⁵	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
PDGFC	cg18688704	-5.05	-1.75	-3.40	5.59x10 ⁻⁵	0.04
PALLD	cg08947774	-3.59	-1.25	-2.42	5.36x10 ⁻⁵	0.04
SLC6A18	cg12913587	13.83	-4.80	-9.31	5.57x10 ⁻⁵	0.04
	cg04505252	-5.78	-2.00	-3.89	5.57x10 ⁻⁵	0.04
DAXX	cg07607074	2.54	7.34	4.94	5.56x10 ⁻⁵	0.04
PRPH2	cg14517133	-8.71	-3.02	-5.86	5.52x10 ⁻⁵	0.04
	cg10794973	-6.15	-2.14	-4.15	5.31x10 ⁻⁵	0.04
BACH2	cg03035849	1.23	3.56	2.40	5.63x10 ⁻⁵	0.04
	cg22277485	1.14	3.29	2.22	5.49x10 ⁻⁵	0.04
LOC100287834	cg26948064	16.52	-5.74	-11.13	5.41x10 ⁻⁵	0.04
TPST1	cg03193328	-3.94	-1.37	-2.65	5.46x10 ⁻⁵	0.04
FGF17	cg08095204	-8.25	-2.86	-5.55	5.59x10 ⁻⁵	0.04
PPP3CC	cg19694519	1.09	3.14	2.12	5.58x10 ⁻⁵	0.04
CUGBP2	cg15777781	2.14	6.19	4.16	5.64x10 ⁻⁵	0.04
MGMT	cg04876500	1.77	5.10	3.43	5.38x10 ⁻⁵	0.04
AIP	cg20973735	1.31	3.77	2.54	5.60x10 ⁻⁵	0.04
UCP2	cg25429672	1.03	2.96	1.99	5.65x10 ⁻⁵	0.04
	cg26585452	1.29	3.72	2.51	5.32x10 ⁻⁵	0.04
	cg09232358	1.16	3.33	2.24	5.47x10 ⁻⁵	0.04
TTC7B	cg06937409	-6.21	-2.16	-4.18	5.37x10 ⁻⁵	0.04
PAK6	cg01965476	3.19	9.16	6.17	5.39x10 ⁻⁵	0.04
NTAN1	cg26629184	-6.55	-2.28	-4.41	5.44x10 ⁻⁵	0.04
CORO1A	cg02767068	2.98	8.61	5.80	5.65x10 ⁻⁵	0.04
MIR21	cg15759721	1.42	4.08	2.75	5.41x10 ⁻⁵	0.04
SLC16A5	cg08434692	3.15	9.05	6.10	5.38x10 ⁻⁵	0.04
FAM100B	cg16429214	2.16	6.22	4.19	5.56x10 ⁻⁵	0.04
TTC39C	cg12639429	1.43	4.12	2.78	5.58x10 ⁻⁵	0.04
	cg04638374	3.62	10.41	7.02	5.37x10 ⁻⁵	0.04
TIAM1	cg26648185	1.44	4.14	2.79	5.50x10 ⁻⁵	0.04
	cg05949181	1.12	3.21	2.16	5.51x10 ⁻⁵	0.04
KCNQ1	cg07556018	-3.41	-1.18	-2.29	5.68x10 ⁻⁵	0.04
LIMD2	cg20797699	33.61	11.64	-22.62	5.68x10 ⁻⁵	0.04
RNF186	cg17135423	2.63	7.61	5.12	5.81x10 ⁻⁵	0.04
INTS7	cg02451670	-5.98	-2.07	-4.02	5.82x10 ⁻⁵	0.04
ZNF620	cg14039773	27.12	-9.38	-18.25	5.79x10 ⁻⁵	0.04
GABBR1	cg09740560	-3.42	-1.18	-2.30	5.71x10 ⁻⁵	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
PRKAG2	cg17192599	-3.85	-1.33	-2.59	5.79x10 ⁻⁵
LY6E	cg12110437	0.84	2.44	1.64	5.82x10 ⁻⁵
IFITM1	cg01886988	1.94	5.62	3.78	5.82x10 ⁻⁵
SNORD30	cg09448652	1.54	4.45	3.00	5.70x10 ⁻⁵
OAF	cg01146320	-6.39	-2.21	-4.30	5.79x10 ⁻⁵
SSH1	cg26663772	1.52	4.40	2.96	5.79x10 ⁻⁵
ZFP36L1	cg01424562	1.16	3.37	2.27	5.81x10 ⁻⁵
MYO5C	cg06192883	2.21	6.40	4.31	5.84x10 ⁻⁵
RORA	cg14720274	3.27	9.46	6.37	5.77x10 ⁻⁵
WDR7	cg26024874	1.23	3.57	2.40	5.80x10 ⁻⁵
ELOF1	cg23760945	0.95	2.76	1.86	5.84x10 ⁻⁵
PHLDB3	cg15664905	-5.75	-1.99	-3.87	5.83x10 ⁻⁵
ADORA2A	cg27381549	1.55	4.49	3.02	5.77x10 ⁻⁵
RERE	cg19679865	1.25	3.62	2.43	6.10x10 ⁻⁵
H6PD	cg09025327	2.15	6.23	4.19	6.06x10 ⁻⁵
C2orf62	cg27367952	16.68	-5.75	-11.21	6.06x10 ⁻⁵
PASK	cg10648547	3.85	11.15	7.50	6.02x10 ⁻⁵
	cg21413797	10.13	-3.50	-6.81	5.96x10 ⁻⁵
	cg15504461	-7.15	-2.47	-4.81	5.90x10 ⁻⁵
PLCH1	cg11932158	-6.99	-2.41	-4.70	6.08x10 ⁻⁵
HAVCR2	cg19646897	1.57	4.56	3.06	6.01x10 ⁻⁵
ZNF193	cg15407162	-3.65	-1.26	-2.45	5.88x10 ⁻⁵
SYNCRIP	cg00937359	1.78	5.15	3.46	5.93x10 ⁻⁵
HECA	cg08943714	-3.92	-1.35	-2.64	5.92x10 ⁻⁵
FOXK1	cg01974478	1.90	5.52	3.71	6.03x10 ⁻⁵
PTPRN2	cg01235172	1.74	5.04	3.39	6.10x10 ⁻⁵
PTGS1	cg00501774	-7.27	-2.50	-4.89	6.09x10 ⁻⁵
	cg07807026	-6.37	-2.20	-4.29	5.94x10 ⁻⁵
	cg24730612	-3.68	-1.27	-2.47	6.06x10 ⁻⁵
GCHFR	cg19433807	2.04	5.90	3.97	5.88x10 ⁻⁵
TRAPPC1	cg01837574	-3.75	-1.29	-2.52	6.01x10 ⁻⁵
ORMDL3	cg10444806	-3.55	-1.22	-2.39	6.07x10 ⁻⁵
LGALS3BP	cg04927537	1.45	4.20	2.83	5.92x10 ⁻⁵
CSNK1D	cg01807946	-4.99	-1.72	-3.35	6.09x10 ⁻⁵
APBA3	cg20366831	1.28	3.72	2.50	6.04x10 ⁻⁵
OXT	cg04731988	-4.37	-1.51	-2.94	5.91x10 ⁻⁵
OXT	cg26267561	-3.82	-1.32	-2.57	6.00x10 ⁻⁵
PYGB	cg04348305	1.54	4.47	3.01	6.05x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
WISP2	cg01750375	-5.03	-1.73	-3.38	5.99x10 ⁻⁵
	cg05941631	-12.25	-4.22	-8.23	6.17x10 ⁻⁵
CHFR	cg02519218	1.28	3.71	2.49	6.19x10 ⁻⁵
ERC1	cg06708720	-4.61	-1.59	-3.10	6.23x10 ⁻⁵
	cg17098093	-6.13	-2.11	-4.12	6.27x10 ⁻⁵
ARAP2	cg22518157	2.32	6.73	4.52	6.26x10 ⁻⁵
CSPP1	cg24190223	-3.33	-1.14	-2.24	6.27x10 ⁻⁵
	cg07673979	-7.48	-2.57	-5.02	6.27x10 ⁻⁵
JARID2	cg13027183	-3.07	-1.05	-2.06	6.31x10 ⁻⁵
WIBG	cg08338281	0.69	2.01	1.35	6.34x10 ⁻⁵
TMEM53	cg10808027	-13.83	-4.75	-9.29	6.40x10 ⁻⁵
LRRC33	cg15726700	-6.77	-2.32	-4.54	6.41x10 ⁻⁵
GPSM3	cg26301953	-5.40	-1.85	-3.62	6.41x10 ⁻⁵
C9orf167	cg13274149	1.39	4.05	2.72	6.37x10 ⁻⁵
BIRC2	cg26207239	1.32	3.84	2.58	6.41x10 ⁻⁵
ATP11A	cg27391403	-9.58	-3.29	-6.44	6.40x10 ⁻⁵
TNFRSF1B	cg15526535	1.18	3.44	2.31	6.48x10 ⁻⁵
	cg12646386	3.67	10.71	7.19	6.60x10 ⁻⁵
MAN1C1	cg04631202	-6.63	-2.27	-4.45	6.54x10 ⁻⁵
SPAG16	cg18859776	-4.77	-1.64	-3.21	6.45x10 ⁻⁵
	cg18727895	-12.75	-4.37	-8.56	6.56x10 ⁻⁵
	cg10920316	4.30	12.56	8.43	6.50x10 ⁻⁵
PSMB8	cg11381564	1.30	3.78	2.54	6.55x10 ⁻⁵
	cg22273555	-7.04	-2.41	-4.73	6.59x10 ⁻⁵
NR4A3	cg13655635	-5.57	-1.91	-3.74	6.58x10 ⁻⁵
MYST4	cg07410339	-2.82	-0.97	-1.89	6.56x10 ⁻⁵
BTG4	cg09148270	-4.70	-1.61	-3.15	6.56x10 ⁻⁵
FLI1	cg02536065	-5.16	-1.77	-3.47	6.44x10 ⁻⁵
	cg25780735	-10.16	-3.48	-6.82	6.59x10 ⁻⁵
LGALS3BP	cg25178683	1.49	4.36	2.93	6.45x10 ⁻⁵
PODNL1	cg26969888	-5.41	-1.85	-3.63	6.54x10 ⁻⁵
ASXL1	cg24727216	1.44	4.20	2.82	6.58x10 ⁻⁵
	cg01772743	1.00	2.92	1.96	6.51x10 ⁻⁵
TEF	cg13228442	-9.35	-3.21	-6.28	6.51x10 ⁻⁵
LYRM4	cg24686957	-5.79	-1.98	-3.88	6.62x10 ⁻⁵
	cg03984055	-5.47	-1.87	-3.67	6.68x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
WDR82	cg25236791	1.41	4.13	2.77	6.69x10 ⁻⁵
ITGAM	cg00833777	-6.08	-2.08	-4.08	6.70x10 ⁻⁵
	cg03324138	1.46	4.26	2.86	6.76x10 ⁻⁵
FOXN2	cg00032884	1.02	2.99	2.01	6.73x10 ⁻⁵
ADK	cg09299075	-3.21	-1.10	-2.15	6.76x10 ⁻⁵
IFIT1L	cg18044543	-2.59	-0.89	-1.74	6.74x10 ⁻⁵
	cg25514148	2.28	6.67	4.47	6.77x10 ⁻⁵
	cg07684519	-4.01	-1.37	-2.69	6.76x10 ⁻⁵
HIST1H3G	cg02481934	-7.48	-2.56	-5.02	6.82x10 ⁻⁵
VWA5B2	cg25447717	-5.57	-1.90	-3.74	6.85x10 ⁻⁵
	cg06875162	-3.82	-1.30	-2.56	6.87x10 ⁻⁵
SPTBN1	cg03017850	1.70	4.99	3.34	6.90x10 ⁻⁵
IFITM1	cg23570810	1.14	3.33	2.24	6.90x10 ⁻⁵
DST	cg15599182	1.03	3.01	2.02	6.93x10 ⁻⁵
UBAC2	cg24509225	1.71	5.00	3.35	6.95x10 ⁻⁵
SKI	cg01852476	2.43	7.14	4.79	7.06x10 ⁻⁵
RPS6KA1	cg09900893	-6.78	-2.31	-4.54	7.10x10 ⁻⁵
GBP1	cg07970007	2.30	6.74	4.52	7.03x10 ⁻⁵
C1orf9	cg00452835	-6.71	-2.29	-4.50	7.08x10 ⁻⁵
MRPL55	cg09462576	-3.75	-1.28	-2.52	6.99x10 ⁻⁵
	cg14915462	-8.53	-2.91	-5.72	7.03x10 ⁻⁵
CCDC12	cg00906812	-7.95	-2.71	-5.33	6.98x10 ⁻⁵
	cg16735495	-4.34	-1.48	-2.91	7.09x10 ⁻⁵
LY96	cg23732024	0.69	2.03	1.36	7.05x10 ⁻⁵
ENTPD7	cg12608633	-5.34	-1.82	-3.58	7.04x10 ⁻⁵
FLT3	cg20227511	-2.33	-0.79	-1.56	7.11x10 ⁻⁵
EPSTI1	cg09843907	1.25	3.66	2.45	7.07x10 ⁻⁵
ACSF3	cg06419212	7.11	20.89	14.00	7.11x10 ⁻⁵
HMHA1	cg16239536	1.20	3.52	2.36	7.03x10 ⁻⁵
WISP2	cg02481642	-3.69	-1.26	-2.48	7.06x10 ⁻⁵
MDS2	cg22613968	0.94	2.77	1.86	7.20x10 ⁻⁵
	cg27565650	11.55	-3.93	-7.74	7.13x10 ⁻⁵
	cg08056629	1.47	4.32	2.89	7.20x10 ⁻⁵
ZBTB43	cg05986288	-3.38	-1.15	-2.26	7.17x10 ⁻⁵
	cg14408428	1.58	4.65	3.11	7.20x10 ⁻⁵
	cg04384031	-5.23	-1.78	-3.51	7.18x10 ⁻⁵
STAT3	cg08652441	1.37	4.04	2.70	7.20x10 ⁻⁵
DNAH17	cg00735218	-7.52	-2.56	-5.04	7.21x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
IL4I1	cg09767736	-6.35	-2.16	-4.25	7.18x10 ⁻⁵	0.04
MFSD2A	cg01869896	-3.59	-1.22	-2.41	7.24x10 ⁻⁵	0.04
	cg17850055	1.41	4.16	2.78	7.23x10 ⁻⁵	0.04
KSR1	cg17585031	2.07	6.10	4.09	7.24x10 ⁻⁵	0.04
ZNF697	cg15513620	-5.05	-1.71	-3.38	7.26x10 ⁻⁵	0.04
	cg08889234	20.64	-7.01	-13.83	7.28x10 ⁻⁵	0.04
IMMP2L	cg17631454	-6.30	-2.14	-4.22	7.27x10 ⁻⁵	0.04
	cg02729030	-6.96	-2.37	-4.66	7.28x10 ⁻⁵	0.04
TRIP4	cg18711535	1.11	3.26	2.19	7.29x10 ⁻⁵	0.04
SFXN5	cg17329648	-2.46	-0.83	-1.65	7.37x10 ⁻⁵	0.04
RAPGEF4	cg23622047	1.50	4.43	2.96	7.42x10 ⁻⁵	0.04
PDCD1	cg18096388	2.50	7.38	4.94	7.44x10 ⁻⁵	0.04
SLC6A6	cg16032408	-7.19	-2.44	-4.82	7.45x10 ⁻⁵	0.04
CD200R1	cg10708271	0.92	2.70	1.81	7.37x10 ⁻⁵	0.04
ADAMTS2	cg00329441	-5.83	-1.98	-3.90	7.45x10 ⁻⁵	0.04
	cg11630152	-3.45	-1.17	-2.31	7.38x10 ⁻⁵	0.04
ADRBK1	cg19905880	-9.39	-3.19	-6.29	7.42x10 ⁻⁵	0.04
	cg02259081	-3.74	-1.27	-2.51	7.39x10 ⁻⁵	0.04
TTC23	cg07470275	14.12	-4.79	-9.45	7.43x10 ⁻⁵	0.04
	cg03835547	3.55	10.45	7.00	7.42x10 ⁻⁵	0.04
GNB1L	cg15198068	-5.99	-2.03	-4.01	7.35x10 ⁻⁵	0.04
	cg18141622	-2.51	-0.85	-1.68	7.53x10 ⁻⁵	0.04
TADA2B	cg01220768	12.15	-4.12	-8.13	7.52x10 ⁻⁵	0.04
	cg16401465	1.68	4.97	3.33	7.54x10 ⁻⁵	0.04
PRF1	cg23364656	1.61	4.76	3.18	7.53x10 ⁻⁵	0.04
AIP	cg13603599	1.00	2.95	1.97	7.50x10 ⁻⁵	0.04
	cg19697042	-5.12	-1.73	-3.43	7.51x10 ⁻⁵	0.04
HSH2D	cg02025270	-5.48	-1.86	-3.67	7.54x10 ⁻⁵	0.04
C19orf51	cg12018403	16.60	-5.63	-11.11	7.52x10 ⁻⁵	0.04
	cg08422803	1.90	5.62	3.76	7.52x10 ⁻⁵	0.04
TGFB3	cg08848903	-6.39	-2.16	-4.28	7.68x10 ⁻⁵	0.04
	cg22534374	-3.52	-1.19	-2.36	7.69x10 ⁻⁵	0.04
ULK4	cg06061536	12.54	-4.24	-8.39	7.66x10 ⁻⁵	0.04
	cg04451880	-3.52	-1.19	-2.36	7.64x10 ⁻⁵	0.04
SKIL	cg14612335	-4.16	-1.41	-2.79	7.60x10 ⁻⁵	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
KIAA0922	cg20367304	1.16	3.43	2.29	7.57x10 ⁻⁵
ATP6V0E1	cg05201300	-5.03	-1.70	-3.37	7.67x10 ⁻⁵
INTS1	cg06942110	2.81	8.31	5.56	7.69x10 ⁻⁵
LRRN3	cg09837977	1.34	3.94	2.64	7.57x10 ⁻⁵
JPH1	cg05134476	-13.53	-4.58	-9.05	7.64x10 ⁻⁵
DGKA	cg02330507	1.88	5.56	3.72	7.59x10 ⁻⁵
C14orf64	cg00263248	0.92	2.71	1.81	7.61x10 ⁻⁵
BCL11B	cg16452866	0.81	2.40	1.61	7.69x10 ⁻⁵
ISG20	cg00854314	2.03	5.98	4.00	7.62x10 ⁻⁵
C2orf88	cg05969591	-4.74	-1.60	-3.17	7.71x10 ⁻⁵
TNIP3	cg24215459	0.95	2.81	1.88	7.72x10 ⁻⁵
FBXO34	cg04853218	-3.70	-1.25	-2.47	7.74x10 ⁻⁵
CAPZB	cg16694480	1.59	4.71	3.15	7.78x10 ⁻⁵
NDUFS2	cg14382215	-6.74	-2.28	-4.51	7.79x10 ⁻⁵
TNNT3	cg02556649	4.92	14.56	9.74	7.77x10 ⁻⁵
MT2A	cg06663317	2.35	6.97	4.66	7.80x10 ⁻⁵
AXIN2	cg23475474	1.64	4.85	3.24	7.81x10 ⁻⁵
	cg02849956	-3.96	-1.34	-2.65	7.80x10 ⁻⁵
	cg17648076	1.05	3.11	2.08	7.83x10 ⁻⁵
PLEK	cg13060970	-4.32	-1.46	-2.89	7.87x10 ⁻⁵
PAM	cg04227789	-58.87	19.88	-39.37	7.88x10 ⁻⁵
SLC9A9	cg25945642	-2.24	-0.76	-1.50	7.89x10 ⁻⁵
PVT1	cg23896695	1.05	3.10	2.07	7.92x10 ⁻⁵
FNBP1L	cg22803510	-6.85	-2.31	-4.58	7.95x10 ⁻⁵
	cg07680505	2.21	6.55	4.38	7.95x10 ⁻⁵
FAM102B	cg21805788	-4.07	-1.37	-2.72	8.01x10 ⁻⁵
RNASET2	cg11301670	1.50	4.44	2.97	8.01x10 ⁻⁵
EXOC4	cg06070229	-12.40	-4.18	-8.29	7.99x10 ⁻⁵
CCDC155	cg03100209	-3.26	-1.10	-2.18	7.98x10 ⁻⁵
	cg15569630	-6.59	-2.22	-4.41	8.05x10 ⁻⁵
	cg00058291	-3.94	-1.33	-2.64	8.06x10 ⁻⁵
	cg20155035	-4.24	-1.43	-2.84	8.07x10 ⁻⁵
LRRFIP1	cg12251803	-5.15	-1.74	-3.45	8.11x10 ⁻⁵
ZNF620	cg13461273	-6.12	-2.06	-4.09	8.09x10 ⁻⁵
UBA7	cg19381811	-4.41	-1.48	-2.95	8.08x10 ⁻⁵
ZBTB9	cg14999947	-2.90	-0.98	-1.94	8.11x10 ⁻⁵
ZMIZ1	cg11961495	-5.34	-1.80	-3.57	8.03x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
C17orf75	cg17055821	-4.19	-1.41	-2.80	8.11x10 ⁻⁵	0.04
TNFAIP3	cg08919597	-3.98	-1.34	-2.66	8.12x10 ⁻⁵	0.04
HES2	cg00644814	-17.35	-5.84	-11.59	8.17x10 ⁻⁵	0.04
PLXND1	cg15629311	-10.52	-3.54	-7.03	8.17x10 ⁻⁵	0.04
ZNF589	cg01458495	1.64	4.87	3.25	8.19x10 ⁻⁵	0.04
TNFRSF8	cg22123711	-4.78	-1.61	-3.19	8.39x10 ⁻⁵	0.04
IFFO2	cg05626128	-5.37	-1.81	-3.59	8.40x10 ⁻⁵	0.04
C1orf21	cg22039357	-3.26	-1.10	-2.18	8.30x10 ⁻⁵	0.04
FAIM3	cg17100176	1.30	3.88	2.59	8.32x10 ⁻⁵	0.04
INPP1	cg15635368	-6.83	-2.29	-4.56	8.40x10 ⁻⁵	0.04
DOCK10	cg04168675	1.44	4.29	2.87	8.43x10 ⁻⁵	0.04
UBE2E1	cg19719475	-4.87	-1.64	-3.25	8.30x10 ⁻⁵	0.04
GNL3	cg18595196	-18.54	-6.23	-12.39	8.43x10 ⁻⁵	0.04
FRMD4B	cg11377213	-2.84	-0.95	-1.89	8.43x10 ⁻⁵	0.04
EPGN	cg19747465	1.15	3.41	2.28	8.32x10 ⁻⁵	0.04
	cg21960364	1.69	5.04	3.37	8.34x10 ⁻⁵	0.04
	cg16230626	-3.33	-1.12	-2.22	8.29x10 ⁻⁵	0.04
LSM11	cg27117005	-9.72	-3.27	-6.49	8.28x10 ⁻⁵	0.04
	cg23127986	-6.26	-2.10	-4.18	8.42x10 ⁻⁵	0.04
	cg15758240	-4.23	-1.42	-2.83	8.23x10 ⁻⁵	0.04
	cg11107430	-8.19	-2.75	-5.47	8.32x10 ⁻⁵	0.04
KCNQ1	cg12141659	-5.50	-1.85	-3.68	8.39x10 ⁻⁵	0.04
	cg20907456	-5.51	-1.85	-3.68	8.41x10 ⁻⁵	0.04
	cg11465943	1.34	3.99	2.67	8.27x10 ⁻⁵	0.04
SERPINA1	cg01606800	-7.39	-2.48	-4.94	8.38x10 ⁻⁵	0.04
	cg08884974	-7.86	-2.64	-5.25	8.20x10 ⁻⁵	0.04
SYNGR3	cg26787199	-8.81	-2.97	-5.89	8.21x10 ⁻⁵	0.04
MIR193B	cg06273075	-4.41	-1.48	-2.95	8.42x10 ⁻⁵	0.04
C17orf75	cg04936619	-3.88	-1.30	-2.59	8.37x10 ⁻⁵	0.04
ERAP1	cg01142811	-5.62	-1.89	-3.76	8.48x10 ⁻⁵	0.04
CTNNBIP1	cg08934126	-5.87	-1.97	-3.92	8.56x10 ⁻⁵	0.04
PLK3	cg27583604	1.55	4.64	3.10	8.72x10 ⁻⁵	0.04
PIGC	cg16177739	-11.03	-3.70	-7.36	8.60x10 ⁻⁵	0.04
ADCY3	cg26752663	1.06	3.15	2.10	8.58x10 ⁻⁵	0.04
	cg12542656	-4.44	-1.49	-2.97	8.60x10 ⁻⁵	0.04
DUSP2	cg04118190	1.98	5.89	3.93	8.53x10 ⁻⁵	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
CX3CR1	cg24130739	-4.44	-1.49	-2.96	8.69x10 ⁻⁵
EIF4G1	cg05677133	1.86	5.56	3.71	8.68x10 ⁻⁵
LEF1	cg12623364	2.38	7.08	4.73	8.55x10 ⁻⁵
SNX18	cg26118943	-3.41	-1.14	-2.28	8.64x10 ⁻⁵
LHFPL2	cg14114546	-2.70	-0.90	-1.80	8.73x10 ⁻⁵
C6orf154	cg01316923	-7.32	-2.45	-4.89	8.73x10 ⁻⁵
SMARCA2	cg13586038	2.79	8.33	5.56	8.73x10 ⁻⁵
GRK5	cg09337049	2.10	6.26	4.18	8.51x10 ⁻⁵
FAM53B	cg07081759	-5.21	-1.75	-3.48	8.59x10 ⁻⁵
SNORD30	cg03431111	1.25	3.72	2.49	8.64x10 ⁻⁵
ALDH3B1	cg00414166	11.93	-4.00	-7.96	8.57x10 ⁻⁵
CUX2	cg16721194	-6.71	-2.25	-4.48	8.60x10 ⁻⁵
	cg05899984	1.41	4.20	2.80	8.72x10 ⁻⁵
	cg00478326	1.28	3.81	2.55	8.73x10 ⁻⁵
RAB37	cg08661469	-6.21	-2.08	-4.14	8.69x10 ⁻⁵
	cg05850338	-2.78	-0.93	-1.85	8.73x10 ⁻⁵
IGLL1	cg02415431	-5.96	-2.00	-3.98	8.71x10 ⁻⁵
	cg14085060	-3.88	-1.30	-2.59	8.77x10 ⁻⁵
GATA3	cg00463367	2.61	7.80	5.20	8.77x10 ⁻⁵
	cg15514918	1.09	3.26	2.18	8.80x10 ⁻⁵
	cg10855531	-6.85	-2.29	-4.57	8.80x10 ⁻⁵
CEACAM3	cg15094920	-9.00	-3.01	-6.01	8.80x10 ⁻⁵
AGTRAP	cg21826784	-3.20	-1.07	-2.13	9.07x10 ⁻⁵
DDOST	cg14519150	1.62	4.83	3.23	8.91x10 ⁻⁵
FBXO11	cg09664186	-4.00	-1.33	-2.67	9.19x10 ⁻⁵
	cg27614178	-2.98	-0.99	-1.99	9.33x10 ⁻⁵
GPR55	cg19827923	3.10	9.27	6.18	8.93x10 ⁻⁵
FOXP1	cg21379733	1.14	3.41	2.28	9.23x10 ⁻⁵
	cg18772838	-5.63	-1.87	-3.75	9.34x10 ⁻⁵
BDH2	cg11469321	-9.43	-3.14	-6.29	9.12x10 ⁻⁵
C4orf39	cg00007239	13.78	-4.60	-9.19	8.97x10 ⁻⁵
SH3TC2	cg26474043	-5.24	-1.75	-3.49	9.16x10 ⁻⁵
	cg20892245	-4.76	-1.59	-3.17	9.23x10 ⁻⁵
JARID2	cg08757670	-7.21	-2.40	-4.80	9.23x10 ⁻⁵
C6orf27	cg15501231	-6.21	-2.07	-4.14	9.01x10 ⁻⁵
SYNGAP1	cg01468420	-6.26	-2.09	-4.18	9.08x10 ⁻⁵
ETV1	cg20909656	19.23	-6.42	-12.82	9.08x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg15812586	-3.33	-1.11	-2.22	9.21x10 ⁻⁵
ZNF862	cg19507256	3.12	9.35	6.23	9.07x10 ⁻⁵
FAM167A	cg06551905	2.17	6.50	4.34	9.14x10 ⁻⁵
PGCP	cg11536940	12.62	-4.22	-8.42	8.95x10 ⁻⁵
NCALD	cg24023489	-2.98	-0.99	-1.99	9.27x10 ⁻⁵
	cg11263617	-7.53	-2.52	-5.02	9.05x10 ⁻⁵
NCOA4	cg13859860	-6.96	-2.32	-4.64	9.32x10 ⁻⁵
DLG5	cg19836471	-9.02	-3.01	-6.01	8.95x10 ⁻⁵
ZMIZ1	cg23865980	-3.22	-1.07	-2.15	9.28x10 ⁻⁵
HPSE2	cg24134845	40.41	13.49	-26.95	9.03x10 ⁻⁵
TNNI2	cg25623459	-6.05	-2.02	-4.04	8.89x10 ⁻⁵
KCNQ1	cg13428066	-3.77	-1.26	-2.52	8.88x10 ⁻⁵
UEVLD	cg15846482	-4.25	-1.42	-2.83	9.08x10 ⁻⁵
	cg24119798	2.41	7.25	4.83	9.31x10 ⁻⁵
RPS6KB2	cg03559915	1.42	4.26	2.84	8.96x10 ⁻⁵
TMEM133	cg06965373	1.17	3.50	2.33	9.18x10 ⁻⁵
FBXL14	cg17150898	1.66	5.00	3.33	9.30x10 ⁻⁵
RAPGEF3	cg14854517	-3.06	-1.02	-2.04	9.24x10 ⁻⁵
DGKA	cg26477856	1.33	3.97	2.65	9.04x10 ⁻⁵
	cg18038207	-3.63	-1.21	-2.42	9.21x10 ⁻⁵
LMO7	cg00902815	2.42	7.25	4.83	9.18x10 ⁻⁵
RASA3	cg03399609	-6.11	-2.04	-4.08	8.88x10 ⁻⁵
	cg26271776	1.29	3.87	2.58	9.15x10 ⁻⁵
MMP14	cg10418289	-5.65	-1.89	-3.77	8.98x10 ⁻⁵
BCL11B	cg08129129	1.33	3.97	2.65	8.93x10 ⁻⁵
EVL	cg17813891	0.90	2.69	1.80	9.04x10 ⁻⁵
SH2D7	cg02279625	-3.89	-1.30	-2.59	9.02x10 ⁻⁵
AXIN2	cg26308530	-6.16	-2.06	-4.11	8.91x10 ⁻⁵
LGALS3BP	cg11105610	1.50	4.50	3.00	9.29x10 ⁻⁵
NAPA	cg24009074	2.13	6.38	4.26	8.92x10 ⁻⁵
C20orf141	cg15013617	-3.15	-1.05	-2.10	9.30x10 ⁻⁵
EIF2S2	cg01562356	-8.43	-2.81	-5.62	9.28x10 ⁻⁵
MICAL3	cg04520704	1.33	3.98	2.65	9.25x10 ⁻⁵
ADSL	cg25636159	1.43	4.27	2.85	8.86x10 ⁻⁵
ADSL	cg13370427	2.48	7.43	4.96	8.90x10 ⁻⁵
	cg09978533	-3.43	-1.14	-2.29	9.18x10 ⁻⁵
HHLA2	cg02059214	2.12	6.36	4.24	9.36x10 ⁻⁵
RERE	cg17442683	1.11	3.32	2.22	9.43x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
FYCO1	cg18682028	1.29	3.88	2.59	9.41x10 ⁻⁵
RAB6B	cg00448875	-5.31	-1.77	-3.54	9.40x10 ⁻⁵
TNRC18	cg09022230	-4.45	-1.48	-2.97	9.39x10 ⁻⁵
	cg14517126	-3.14	-1.04	-2.09	9.43x10 ⁻⁵
TSPO	cg00343092	-4.80	-1.60	-3.20	9.44x10 ⁻⁵
NUP210	cg06143615	1.60	4.82	3.21	9.46x10 ⁻⁵
	cg07974367	1.64	4.93	3.29	9.46x10 ⁻⁵
	cg06287548	-2.98	-0.99	-1.98	9.50x10 ⁻⁵
	cg24043628	-5.57	-1.85	-3.71	9.53x10 ⁻⁵
CHI3L2	cg09516523	1.16	3.49	2.33	9.56x10 ⁻⁵
GNL1	cg15789723	10.24	-3.40	-6.82	9.56x10 ⁻⁵
MAPKAPK2	cg06419268	-4.51	-1.50	-3.01	9.59x10 ⁻⁵
STRADA	cg20157577	1.33	4.01	2.67	9.59x10 ⁻⁵
PSMD13	cg05600342	1.43	4.31	2.87	9.62x10 ⁻⁵
	cg02033302	-4.36	-1.45	-2.90	9.62x10 ⁻⁵
	cg26600461	1.43	4.30	2.87	9.63x10 ⁻⁵
CA12	cg03036214	-3.89	-1.29	-2.59	9.67x10 ⁻⁵
EPN1	cg27573549	3.26	9.80	6.53	9.66x10 ⁻⁵
IFITM1	cg21686213	1.07	3.22	2.15	9.68x10 ⁻⁵
KIF26B	cg21301514	-9.05	-3.00	-6.03	9.71x10 ⁻⁵
PRPSAP1	cg00549574	-5.84	-1.94	-3.89	9.74x10 ⁻⁵
LGALS3BP	cg14870271	1.71	5.15	3.43	9.73x10 ⁻⁵
	cg08636638	-5.72	-1.90	-3.81	9.78x10 ⁻⁵
C12orf27	cg07545743	-2.33	-0.77	-1.55	9.78x10 ⁻⁵
FGR	cg09370867	-3.48	-1.15	-2.31	9.81x10 ⁻⁵
	cg05483875	15.74	-5.22	-10.48	9.81x10 ⁻⁵
IL6R	cg25135018	-4.20	-1.39	-2.80	9.83x10 ⁻⁵
	cg19634252	-5.38	-1.78	-3.58	9.83x10 ⁻⁵
CBX4	cg13475704	-6.75	-2.24	-4.50	9.85x10 ⁻⁵
IKZF1	cg07103517	1.39	4.20	2.79	9.89x10 ⁻⁵
IL21R	cg02656594	-2.70	-0.89	-1.79	9.88x10 ⁻⁵
SMG6	cg23698124	1.19	3.58	2.39	9.89x10 ⁻⁵
SLC9A8	cg02003117	-6.25	-2.07	-4.16	9.88x10 ⁻⁵
APBB1	cg27216937	1.95	5.88	3.91	9.91x10 ⁻⁵
ZNF608	cg26413942	-3.13	-1.04	-2.08	9.93x10 ⁻⁵
PTBP1	cg17357561	24.23	-8.03	-16.13	9.93x10 ⁻⁵
CYTH3	cg10950593	-4.20	-1.39	-2.79	9.98x10 ⁻⁵

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
DGKA	cg21587469	0.99	2.98	1.98	9.98x10 ⁻⁵	
CHD3	cg04674060	1.20	3.61	2.40	9.97x10 ⁻⁵	
NDUFA4L2	cg02948264	-9.05	-2.99	-6.02	1.00x10 ⁻⁴	
TERT	cg16429735	1.09	3.29	2.19	1.00x10 ⁻⁴	
TUBA1C	cg16881676	-5.83	-1.93	-3.88	1.00x10 ⁻⁴	
FAIM3	cg22945467	2.59	7.81	5.20	1.01x10 ⁻⁴	
MAGEF1	cg26152983	-4.23	-1.40	-2.82	1.01x10 ⁻⁴	
MFN2	cg09820729	-5.21	-1.72	-3.47	1.01x10 ⁻⁴	
	cg16704158	-3.68	-1.22	-2.45	1.01x10 ⁻⁴	
	cg17384323	3.09	9.34	6.21	1.02x10 ⁻⁴	
C4orf41	cg21127184	-6.35	-2.10	-4.23	1.01x10 ⁻⁴	
	cg16223546	-4.82	-1.59	-3.21	1.02x10 ⁻⁴	
N4BP3	cg17343167	-7.29	-2.41	-4.85	1.02x10 ⁻⁴	
PRDM1	cg17143179	1.03	3.13	2.08	1.02x10 ⁻⁴	
	cg07276621	2.14	6.48	4.31	1.02x10 ⁻⁴	
LRP6	cg00046744	-8.06	-2.66	-5.36	1.02x10 ⁻⁴	
RAB21	cg02611240	-6.65	-2.20	-4.42	1.01x10 ⁻⁴	
TESC	cg25598890	-6.04	-2.00	-4.02	1.01x10 ⁻⁴	
NCOR2	cg11639950	-4.92	-1.63	-3.27	1.02x10 ⁻⁴	
	cg11162888	1.09	3.31	2.20	1.02x10 ⁻⁴	
ZSWIM1	cg13780718	2.42	7.30	4.86	1.01x10 ⁻⁴	
	cg04957307	1.68	5.08	3.38	1.03x10 ⁻⁴	
NAPSA	cg19901005	1.71	5.19	3.45	1.03x10 ⁻⁴	
RREB1	cg12407685	-7.95	-2.63	-5.29	1.03x10 ⁻⁴	
TREX1	cg21788755	1.99	6.02	4.00	1.03x10 ⁻⁴	
	cg04892766	1.38	4.19	2.79	1.03x10 ⁻⁴	
ABLIM1	cg00950497	-3.85	-1.27	-2.56	1.03x10 ⁻⁴	
SEL1L	cg16379999	-3.09	-1.02	-2.06	1.03x10 ⁻⁴	
MCTP2	cg02700491	-2.84	-0.94	-1.89	1.04x10 ⁻⁴	
PAFAH1B1	cg08696470	-4.13	-1.36	-2.75	1.04x10 ⁻⁴	
IQCE	cg04109092	-4.93	-1.63	-3.28	1.04x10 ⁻⁴	
TRRAP	cg23151309	-	48.64	16.05	-32.34	1.04x10 ⁻⁴
PRDM11	cg24904788	-3.50	-1.15	-2.32	1.04x10 ⁻⁴	
FLJ23834	cg04215126	-8.22	-2.71	-5.46	1.04x10 ⁻⁴	
TMEM49	cg16936953	1.22	3.71	2.47	1.05x10 ⁻⁴	
TMEM49	cg12054453	1.23	3.72	2.47	1.05x10 ⁻⁴	
RERE	cg21774136	1.06	3.22	2.14	1.05x10 ⁻⁴	
MPHOSPH9	cg07732037	1.16	3.51	2.33	1.06x10 ⁻⁴	

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
ZSWIM4	cg25751484	-7.00	-2.31	-4.66	1.06x10 ⁻⁴	0.04
RGS1	cg06357908	1.42	4.30	2.86	1.06x10 ⁻⁴	0.04
SKAP2	cg13081213	-6.46	-2.13	-4.29	1.06x10 ⁻⁴	0.04
RABL4	cg06235224	-7.45	-2.45	-4.95	1.06x10 ⁻⁴	0.04
FBP1	cg01210663	-6.64	-2.19	-4.41	1.06x10 ⁻⁴	0.04
ELN	cg25729687	11.01	-3.63	-7.32	1.06x10 ⁻⁴	0.04
	cg03412153	1.15	3.49	2.32	1.06x10 ⁻⁴	0.04
KCTD19	cg05762671	1.28	3.89	2.58	1.06x10 ⁻⁴	0.04
	cg26840889	1.66	5.04	3.35	1.07x10 ⁻⁴	0.04
TGIF1	cg07529654	-4.23	-1.39	-2.81	1.07x10 ⁻⁴	0.04
	cg26579986	-3.84	-1.26	-2.55	1.07x10 ⁻⁴	0.04
	cg00265360	11.36	-3.74	-7.55	1.08x10 ⁻⁴	0.04
BCL9	cg23014759	1.19	3.64	2.42	1.10x10 ⁻⁴	0.04
	cg06647600	-6.36	-2.09	-4.22	1.08x10 ⁻⁴	0.04
	cg11531339	2.13	6.50	4.32	1.11x10 ⁻⁴	0.04
	cg03834116	-3.97	-1.30	-2.63	1.11x10 ⁻⁴	0.04
	cg08009902	1.39	4.24	2.81	1.09x10 ⁻⁴	0.04
PLCL1	cg06734510	3.06	9.30	6.18	1.08x10 ⁻⁴	0.04
	cg17328631	23.38	-7.67	-15.52	1.11x10 ⁻⁴	0.04
SH3BP2	cg23746574	22.31	-7.32	-14.82	1.11x10 ⁻⁴	0.04
	cg04131890	1.15	3.51	2.33	1.11x10 ⁻⁴	0.04
	cg02381853	-3.28	-1.08	-2.18	1.08x10 ⁻⁴	0.04
ODZ2	cg01227558	-9.05	-2.98	-6.01	1.09x10 ⁻⁴	0.04
LRRC16A	cg09110394	-4.59	-1.51	-3.05	1.09x10 ⁻⁴	0.04
ARPC5L	cg14287565	16.31	-5.36	-10.84	1.10x10 ⁻⁴	0.04
	cg25500001	-5.95	-1.96	-3.95	1.08x10 ⁻⁴	0.04
	cg05482603	1.93	5.87	3.90	1.10x10 ⁻⁴	0.04
	cg18434560	1.23	3.74	2.48	1.08x10 ⁻⁴	0.04
	cg08214423	21.27	-6.98	-14.13	1.11x10 ⁻⁴	0.04
UBXN1	cg20140054	-5.72	-1.88	-3.80	1.10x10 ⁻⁴	0.04
AIP	cg19528654	1.11	3.39	2.25	1.09x10 ⁻⁴	0.04
	cg05592911	28.67	-9.41	-19.04	1.11x10 ⁻⁴	0.04
MLL	cg14242895	-4.89	-1.61	-3.25	1.10x10 ⁻⁴	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
RPS26	cg25095951	-7.10	-2.33	-4.72	1.09x10 ⁻⁴	0.04
OBFC2B	cg10654546	-3.13	-1.03	-2.08	1.09x10 ⁻⁴	0.04
	cg09251317	-6.40	-2.10	-4.25	1.10x10 ⁻⁴	0.04
AKAP13	cg09874482	-3.50	-1.15	-2.32	1.11x10 ⁻⁴	0.04
SLCO3A1	cg25508605	1.89	5.77	3.83	1.10x10 ⁻⁴	0.04
LONP2	cg09539739	-5.00	-1.64	-3.32	1.10x10 ⁻⁴	0.04
WDR16	cg26976042	14.66	-4.82	-9.74	1.09x10 ⁻⁴	0.04
FCER2	cg12261095	-4.83	-1.59	-3.21	1.10x10 ⁻⁴	0.04
ZFP36	cg05209306	1.11	3.37	2.24	1.08x10 ⁻⁴	0.04
ZNF296	cg17769442	10.25	-3.36	-6.81	1.10x10 ⁻⁴	0.04
OXT	cg19592472	-2.50	-0.82	-1.66	1.09x10 ⁻⁴	0.04
	cg18908845	11.76	-3.86	-7.81	1.11x10 ⁻⁴	0.04
FAM38A	cg02610723	-6.51	-2.13	-4.32	1.12x10 ⁻⁴	0.04
AXIN2	cg18956355	2.49	7.61	5.05	1.12x10 ⁻⁴	0.04
CD44	cg19292760	1.12	3.41	2.26	1.12x10 ⁻⁴	0.04
IRS2	cg12085119	1.63	4.99	3.31	1.12x10 ⁻⁴	0.04
LPP	cg16464007	1.52	4.63	3.07	1.13x10 ⁻⁴	0.04
CAST	cg02291010	-2.86	-0.94	-1.90	1.12x10 ⁻⁴	0.04
RBPMS	cg26122129	-5.16	-1.69	-3.43	1.13x10 ⁻⁴	0.04
DCLK3	cg21113478	1.52	4.63	3.07	1.13x10 ⁻⁴	0.04
TMEM180	cg00122347	1.04	3.18	2.11	1.13x10 ⁻⁴	0.04
PRR5L	cg20934799	16.93	-5.55	-11.24	1.13x10 ⁻⁴	0.04
HLA-E	cg17615629	0.68	2.07	1.38	1.14x10 ⁻⁴	0.04
DUSP10	cg26562772	-2.85	-0.93	-1.89	1.14x10 ⁻⁴	0.04
RCSD1	cg01793445	2.69	8.20	5.44	1.14x10 ⁻⁴	0.04
KIAA1688	cg02340818	1.12	3.43	2.27	1.14x10 ⁻⁴	0.04
GUCY1B2	cg24013213	-3.35	-1.10	-2.22	1.14x10 ⁻⁴	0.04
	cg02393107	1.28	3.91	2.59	1.14x10 ⁻⁴	0.04
EGFL7	cg21184800	10.52	-3.44	-6.98	1.14x10 ⁻⁴	0.04
ITM2C	cg04293526	-8.53	-2.79	-5.66	1.15x10 ⁻⁴	0.04
ATP10B	cg13629358	1.65	5.03	3.34	1.15x10 ⁻⁴	0.04
TMEM8B	cg11825899	-3.45	-1.13	-2.29	1.15x10 ⁻⁴	0.04
DAPK1	cg08719486	-4.68	-1.53	-3.11	1.15x10 ⁻⁴	0.04
GRN	cg23570245	-3.66	-1.20	-2.43	1.15x10 ⁻⁴	0.04
PNPLA1	cg22027204	-3.71	-1.21	-2.46	1.16x10 ⁻⁴	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
RAB32	cg12872357	-3.33	-1.09	-2.21	1.16x10 ⁻⁴
SLMO1	cg16498391	-4.15	-1.36	-2.76	1.15x10 ⁻⁴
ACSL3	cg23325384	-4.12	-1.35	-2.73	1.16x10 ⁻⁴
VIT	cg09154591	-3.50	-1.14	-2.32	1.16x10 ⁻⁴
	cg09333584	-7.97	-2.60	-5.29	1.17x10 ⁻⁴
CD34	cg21697512	-2.34	-0.76	-1.55	1.17x10 ⁻⁴
MARCH1	cg19400238	-7.32	-2.39	-4.85	1.17x10 ⁻⁴
	cg16994941	1.90	5.83	3.87	1.18x10 ⁻⁴
SLIT1	cg05265607	-5.13	-1.68	-3.41	1.18x10 ⁻⁴
COQ9	cg00247269	-4.70	-1.54	-3.12	1.17x10 ⁻⁴
MLC1	cg10854441	-4.59	-1.50	-3.05	1.18x10 ⁻⁴
PRKCZ	cg07693617	-5.67	-1.84	-3.76	1.23x10 ⁻⁴
	cg14217990	-8.71	-2.84	-5.77	1.21x10 ⁻⁴
	cg04991639	-7.79	-2.53	-5.16	1.23x10 ⁻⁴
KIAA0319L	cg10959672	1.11	3.40	2.26	1.19x10 ⁻⁴
FAM69A	cg22128645	1.48	4.52	3.00	1.19x10 ⁻⁴
KIRREL	cg11525252	1.31	4.04	2.68	1.23x10 ⁻⁴
RAB7L1	cg26418147	1.19	3.65	2.42	1.23x10 ⁻⁴
TRAF5	cg09825327	1.42	4.36	2.89	1.23x10 ⁻⁴
CPSF3	cg12057242	-3.47	-1.13	-2.30	1.23x10 ⁻⁴
GYPC	cg08129583	-2.94	-0.96	-1.95	1.22x10 ⁻⁴
TMEFF2	cg02288301	-16.55	-5.39	-10.97	1.22x10 ⁻⁴
DGKD	cg00711795	2.45	7.50	4.98	1.20x10 ⁻⁴
	cg22856834	-4.10	-1.34	-2.72	1.18x10 ⁻⁴
	cg25616514	-4.87	-1.59	-3.23	1.21x10 ⁻⁴
PRICKLE2	cg18450254	-3.64	-1.19	-2.42	1.22x10 ⁻⁴
RBM47	cg20435896	-6.13	-2.00	-4.06	1.19x10 ⁻⁴
LNX1	cg06495586	-3.13	-1.02	-2.08	1.23x10 ⁻⁴
	cg01563714	-32.30	10.54	-21.42	1.18x10 ⁻⁴
JARID2	cg16362595	1.20	3.67	2.43	1.19x10 ⁻⁴
	cg19612068	-7.11	-2.31	-4.71	1.22x10 ⁻⁴
	cg03875496	-8.85	-2.88	-5.86	1.21x10 ⁻⁴
NOD1	cg09579281	-4.35	-1.42	-2.88	1.23x10 ⁻⁴
SLC25A13	cg27185510	1.46	4.47	2.97	1.18x10 ⁻⁴
GAL3ST4	cg24616382	1.02	3.15	2.09	1.22x10 ⁻⁴
FLJ23834	cg21586203	-8.76	-2.85	-5.81	1.21x10 ⁻⁴
EPHB6	cg12599168	1.38	4.24	2.81	1.19x10 ⁻⁴
GSR	cg06049177	-5.52	-1.80	-3.66	1.24x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
SLC25A25	cg14944923	-3.69	-1.20	-2.44	1.23x10 ⁻⁴ 0.04
	cg14522718	0.97	2.97	1.97	1.22x10 ⁻⁴ 0.04
LCN6	cg13544075	-4.55	-1.48	-3.02	1.20x10 ⁻⁴ 0.04
	cg14480266	-7.50	-2.44	-4.97	1.22x10 ⁻⁴ 0.04
DGKZ	cg18765542	2.85	8.75	5.80	1.20x10 ⁻⁴ 0.04
MTCH2	cg22321327	-3.86	-1.26	-2.56	1.22x10 ⁻⁴ 0.04
	cg25248213	-7.73	-2.52	-5.12	1.19x10 ⁻⁴ 0.04
PTPRO	cg12601353	-9.00	-2.93	-5.97	1.20x10 ⁻⁴ 0.04
BTBD11	cg01686739	-3.29	-1.07	-2.18	1.19x10 ⁻⁴ 0.04
	cg19627549	1.68	5.15	3.42	1.20x10 ⁻⁴ 0.04
CRIP1	cg07065217	-7.22	-2.35	-4.78	1.23x10 ⁻⁴ 0.04
	cg14082893	14.56	-4.74	-9.65	1.21x10 ⁻⁴ 0.04
CHD2	cg12644285	-2.95	-0.96	-1.96	1.20x10 ⁻⁴ 0.04
SOCS1	cg03014241	1.11	3.40	2.25	1.20x10 ⁻⁴ 0.04
MIR193B	cg03295417	-3.27	-1.06	-2.17	1.22x10 ⁻⁴ 0.04
DLG4	cg03508063	-5.66	-1.84	-3.75	1.23x10 ⁻⁴ 0.04
MSI2	cg05347965	1.21	3.72	2.46	1.21x10 ⁻⁴ 0.04
	cg27112972	-4.97	-1.62	-3.29	1.23x10 ⁻⁴ 0.04
RPTOR	cg01432609	-2.92	-0.95	-1.94	1.22x10 ⁻⁴ 0.04
MAFG	cg00926657	-5.09	-1.66	-3.37	1.22x10 ⁻⁴ 0.04
LAMA3	cg13270625	1.35	4.14	2.75	1.21x10 ⁻⁴ 0.04
MBP	cg06773488	-4.70	-1.53	-3.12	1.19x10 ⁻⁴ 0.04
FBXO46	cg09277709	-4.40	-1.43	-2.92	1.20x10 ⁻⁴ 0.04
SIRPG	cg11061975	2.22	6.83	4.52	1.21x10 ⁻⁴ 0.04
LOC100271722	cg23564243	-5.87	-1.91	-3.89	1.23x10 ⁻⁴ 0.04
IL16	cg14898127	1.10	3.39	2.24	1.24x10 ⁻⁴ 0.04
LOC100133612	cg09275693	0.83	2.56	1.69	1.25x10 ⁻⁴ 0.04
	cg10582860	-4.93	-1.60	-3.27	1.24x10 ⁻⁴ 0.04
ENTPD7	cg11690884	-3.02	-0.98	-2.00	1.24x10 ⁻⁴ 0.04
	cg11736228	-5.58	-1.81	-3.70	1.24x10 ⁻⁴ 0.04
LTB4R	cg05302095	1.13	3.48	2.31	1.24x10 ⁻⁴ 0.04
	cg07042014	-4.22	-1.37	-2.80	1.24x10 ⁻⁴ 0.04
KCNIP1	cg13152690	0.99	3.04	2.02	1.24x10 ⁻⁴ 0.04
	cg07363543	-5.07	-1.65	-3.36	1.25x10 ⁻⁴ 0.04
CUTA	cg02640147	1.03	3.17	2.10	1.25x10 ⁻⁴ 0.04
	cg14114297	-3.07	-1.00	-2.03	1.25x10 ⁻⁴ 0.04
ICAM2	cg05352838	1.43	4.42	2.93	1.25x10 ⁻⁴ 0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg18077387	3.40	10.47	6.93	1.26x10 ⁻⁴
FOXP1	cg15958828	-6.54	-2.12	-4.33	1.26x10 ⁻⁴
ATF6B	cg26539631	-8.57	-2.78	-5.67	1.26x10 ⁻⁴
KDM2B	cg21340621	-3.31	-1.07	-2.19	1.26x10 ⁻⁴
RSBN1	cg03469607	1.11	3.41	2.26	1.27x10 ⁻⁴
	cg07367601	-4.68	-1.52	-3.10	1.27x10 ⁻⁴
	cg03106207	2.53	7.79	5.16	1.27x10 ⁻⁴
RASGEF1A	cg24020157	-3.22	-1.04	-2.13	1.27x10 ⁻⁴
ENTPD7	cg26114100	-6.21	-2.01	-4.11	1.27x10 ⁻⁴
	cg09228327	-16.65	-5.40	-11.02	1.27x10 ⁻⁴
REM2	cg24681208	-6.89	-2.23	-4.56	1.27x10 ⁻⁴
DOK4	cg16547186	-3.78	-1.23	-2.51	1.27x10 ⁻⁴
GNG7	cg19477361	1.68	5.18	3.43	1.27x10 ⁻⁴
	cg22726155	1.09	3.38	2.24	1.27x10 ⁻⁴
	cg07350631	-14.75	-4.78	-9.76	1.28x10 ⁻⁴
PACS2	cg19769147	-6.79	-2.20	-4.49	1.28x10 ⁻⁴
MAD1L1	cg08972190	1.52	4.70	3.11	1.28x10 ⁻⁴
AXIN2	cg09231741	1.29	3.99	2.64	1.28x10 ⁻⁴
	cg09277256	-6.15	-1.99	-4.07	1.28x10 ⁻⁴
TBC1D4	cg00810292	0.81	2.50	1.66	1.28x10 ⁻⁴
	cg22490722	-11.47	-3.71	-7.59	1.29x10 ⁻⁴
EHMT2	cg25202877	-13.77	-4.46	-9.11	1.29x10 ⁻⁴
HYOU1	cg11962566	-5.50	-1.78	-3.64	1.29x10 ⁻⁴
MPND	cg25631746	-20.48	-6.63	-13.56	1.28x10 ⁻⁴
TGFA	cg21346154	-3.59	-1.16	-2.38	1.29x10 ⁻⁴
CXCR6	cg25226014	0.95	2.93	1.94	1.29x10 ⁻⁴
SCARB1	cg01663970	-4.56	-1.48	-3.02	1.29x10 ⁻⁴
RORA	cg09879458	1.13	3.48	2.30	1.29x10 ⁻⁴
SNORA59A	cg22866430	1.32	4.08	2.70	1.31x10 ⁻⁴
FASLG	cg00071250	1.03	3.18	2.11	1.31x10 ⁻⁴
	cg08914730	-6.43	-2.08	-4.25	1.31x10 ⁻⁴
	cg07203767	1.08	3.35	2.21	1.31x10 ⁻⁴
RNF13	cg24105729	-8.47	-2.74	-5.61	1.30x10 ⁻⁴
SPATA5	cg05805445	1.29	4.00	2.64	1.31x10 ⁻⁴
F2RL2	cg00415993	-2.88	-0.93	-1.91	1.30x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
DOCK4	cg11884089	-4.64	-1.50	-3.07	1.31x10 ⁻⁴	0.04
IL2RA	cg11127249	4.14	12.80	8.47	1.31x10 ⁻⁴	0.04
RPS6KB2	cg17375396	1.41	4.36	2.89	1.30x10 ⁻⁴	0.04
	cg21927426	-6.24	-2.02	-4.13	1.30x10 ⁻⁴	0.04
PACS2	cg06783197	2.31	7.14	4.72	1.31x10 ⁻⁴	0.04
PRPSAP1	cg24146586	12.77	-4.13	-8.45	1.30x10 ⁻⁴	0.04
KIAA1683	cg01087239	1.10	3.40	2.25	1.31x10 ⁻⁴	0.04
ZBTB32	cg09231418	0.94	2.90	1.92	1.31x10 ⁻⁴	0.04
ETFB	cg17110052	0.97	3.01	1.99	1.31x10 ⁻⁴	0.04
MTF1	cg13567986	-4.64	-1.50	-3.07	1.32x10 ⁻⁴	0.04
	cg01829632	-3.05	-0.99	-2.02	1.32x10 ⁻⁴	0.04
	cg04958236	-3.49	-1.13	-2.31	1.32x10 ⁻⁴	0.04
UNCX	cg17294725	12.28	-3.97	-8.12	1.32x10 ⁻⁴	0.04
MARK2	cg01102854	-5.27	-1.70	-3.49	1.32x10 ⁻⁴	0.04
SYT1	cg25193867	2.53	7.83	5.18	1.32x10 ⁻⁴	0.04
GPR81	cg22534509	1.11	3.43	2.27	1.32x10 ⁻⁴	0.04
LY9	cg01367992	0.79	2.44	1.61	1.35x10 ⁻⁴	0.04
	cg15711902	24.54	-7.91	-16.23	1.35x10 ⁻⁴	0.04
	cg17819732	1.33	4.14	2.74	1.37x10 ⁻⁴	0.04
SLC26A9	cg05122040	-3.96	-1.27	-2.62	1.37x10 ⁻⁴	0.04
C2orf29	cg21044433	-7.22	-2.33	-4.77	1.36x10 ⁻⁴	0.04
	cg06826886	17.59	-5.67	-11.63	1.36x10 ⁻⁴	0.04
HDAC4	cg11308319	1.44	4.47	2.95	1.36x10 ⁻⁴	0.04
CXCR6	cg08450017	0.76	2.36	1.56	1.37x10 ⁻⁴	0.04
PTH1R	cg16016506	13.13	-4.24	-8.68	1.34x10 ⁻⁴	0.04
	cg04322363	-9.27	-2.99	-6.13	1.35x10 ⁻⁴	0.04
	cg18157353	14.20	-4.57	-9.39	1.38x10 ⁻⁴	0.04
QRFPR	cg21631428	-3.17	-1.02	-2.10	1.37x10 ⁻⁴	0.04
MAML3	cg03180134	-5.62	-1.81	-3.72	1.33x10 ⁻⁴	0.04
CSNK1G3	cg27054084	1.43	4.42	2.92	1.34x10 ⁻⁴	0.04
ACSL6	cg14841483	1.37	4.24	2.80	1.35x10 ⁻⁴	0.04
TAP1	cg26234900	0.96	2.97	1.96	1.35x10 ⁻⁴	0.04
MCM7	cg22940022	10.15	-3.27	-6.71	1.35x10 ⁻⁴	0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
CUX1	cg15755348	2.62	8.15	5.39	1.37x10 ⁻⁴ 0.04
TNFRSF10D	cg14918744	0.94	2.91	1.92	1.34x10 ⁻⁴ 0.04
RAB2A	cg14839134	-3.05	-0.98	-2.01	1.37x10 ⁻⁴ 0.04
SIT1	cg15518883	1.11	3.46	2.29	1.36x10 ⁻⁴ 0.04
NAA35	cg13727957	-6.70	-2.16	-4.43	1.35x10 ⁻⁴ 0.04
NCRNA00092	cg05138681	-6.49	-2.09	-4.29	1.37x10 ⁻⁴ 0.04
LZTS2	cg03271893	-4.81	-1.55	-3.18	1.34x10 ⁻⁴ 0.04
TCF7L2	cg00831931	-12.82	-4.13	-8.47	1.37x10 ⁻⁴ 0.04
STK32C	cg13432205	-9.41	-3.03	-6.22	1.36x10 ⁻⁴ 0.04
LGALS12	cg21964800	-2.38	-0.77	-1.57	1.34x10 ⁻⁴ 0.04
PTPRO	cg13554714	1.46	4.53	3.00	1.37x10 ⁻⁴ 0.04
GXYLT1	cg13431800	-5.17	-1.67	-3.42	1.35x10 ⁻⁴ 0.04
ADCY6	cg02423534	-4.08	-1.31	-2.69	1.37x10 ⁻⁴ 0.04
NCOR2	cg15085899	-2.49	-0.80	-1.64	1.35x10 ⁻⁴ 0.04
	cg20749792	1.10	3.41	2.25	1.36x10 ⁻⁴ 0.04
ZFP36L1	cg10099732	0.76	2.37	1.56	1.34x10 ⁻⁴ 0.04
	cg22582617	-2.37	-0.76	-1.57	1.36x10 ⁻⁴ 0.04
LMF1	cg08018572	2.11	6.56	4.34	1.38x10 ⁻⁴ 0.04
ABCC1	cg16702014	-3.23	-1.04	-2.13	1.38x10 ⁻⁴ 0.04
IL21R	cg02787852	0.86	2.66	1.76	1.37x10 ⁻⁴ 0.04
PYCARD	cg12100791	-5.22	-1.69	-3.46	1.35x10 ⁻⁴ 0.04
	cg03259887	2.81	8.70	5.75	1.33x10 ⁻⁴ 0.04
GAS7	cg05558046	1.88	5.81	3.85	1.33x10 ⁻⁴ 0.04
SLFN5	cg05897169	1.13	3.50	2.31	1.35x10 ⁻⁴ 0.04
HEXIM2	cg08351131	-3.08	-0.99	-2.04	1.36x10 ⁻⁴ 0.04
MSI2	cg10723617	-3.96	-1.27	-2.62	1.38x10 ⁻⁴ 0.04
AZI1	cg05371735	3.38	10.47	6.93	1.33x10 ⁻⁴ 0.04
MAPRE2	cg05735765	-14.12	-4.55	-9.33	1.37x10 ⁻⁴ 0.04
	cg12759166	2.14	6.65	4.40	1.36x10 ⁻⁴ 0.04
CD97	cg08239297	-8.74	-2.82	-5.78	1.36x10 ⁻⁴ 0.04
THEMIS	cg17113883	1.43	4.45	2.94	1.38x10 ⁻⁴ 0.04
METTL10	cg16832958	-19.63	-6.32	-12.97	1.38x10 ⁻⁴ 0.04
LAMA3	cg01152726	1.32	4.09	2.70	1.39x10 ⁻⁴ 0.04
	cg17756730	1.05	3.25	2.15	1.39x10 ⁻⁴ 0.04
ZNF740	cg00078245	1.45	4.50	2.97	1.39x10 ⁻⁴ 0.04
SKI	cg15007228	-9.58	-3.08	-6.33	1.39x10 ⁻⁴ 0.04
PRDM16	cg10493186	-4.24	-1.36	-2.80	1.39x10 ⁻⁴ 0.04

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
ADAP1	cg20705627	-5.36	-1.72	-3.54	1.39x10 ⁻⁴
	cg26763394	0.96	2.99	1.98	1.39x10 ⁻⁴
SLC22A8	cg15572436	-8.90	-2.86	-5.88	1.40x10 ⁻⁴
OLFML3	cg13393917	-7.77	-2.50	-5.14	1.41x10 ⁻⁴
QSOX1	cg09505809	-4.36	-1.40	-2.88	1.40x10 ⁻⁴
CCRL2	cg13070763	-38.53	12.39	-25.46	1.40x10 ⁻⁴
INPP4B	cg03476007	-8.08	-2.60	-5.34	1.41x10 ⁻⁴
CUX1	cg20253855	-3.65	-1.17	-2.41	1.41x10 ⁻⁴
VPS37B	cg01832672	1.50	4.65	3.07	1.41x10 ⁻⁴
	cg09351315	-4.61	-1.48	-3.04	1.41x10 ⁻⁴
PLCG1	cg06493829	5.64	17.56	11.60	1.41x10 ⁻⁴
CLTCL1	cg24911827	1.83	5.70	3.77	1.40x10 ⁻⁴
MEI1	cg05023013	-9.07	-2.92	-5.99	1.41x10 ⁻⁴
CDC42BPB	cg04189326	-5.27	-1.69	-3.48	1.41x10 ⁻⁴
PHYH	cg19018267	-3.45	-1.11	-2.28	1.41x10 ⁻⁴
SSBP3	cg09417011	1.52	4.74	3.13	1.44x10 ⁻⁴
ITPKB	cg23717186	1.17	3.66	2.42	1.43x10 ⁻⁴
	cg16704560	-3.33	-1.07	-2.20	1.44x10 ⁻⁴
IP6K2	cg16180552	-6.93	-2.22	-4.57	1.43x10 ⁻⁴
DHX29	cg26464796	-2.87	-0.92	-1.90	1.42x10 ⁻⁴
LHFPL2	cg06759890	-3.30	-1.06	-2.18	1.43x10 ⁻⁴
HLA-E	cg27486585	0.93	2.91	1.92	1.43x10 ⁻⁴
CLIC1	cg11093373	-7.35	-2.36	-4.86	1.42x10 ⁻⁴
RGL2	cg05876591	-14.79	-4.75	-9.77	1.43x10 ⁻⁴
MAMDC2	cg13870494	2.45	7.64	5.05	1.44x10 ⁻⁴
VAV2	cg14308466	-5.52	-1.77	-3.65	1.42x10 ⁻⁴
PFKFB3	cg14038949	1.34	4.17	2.75	1.43x10 ⁻⁴
	cg12999267	1.07	3.32	2.19	1.42x10 ⁻⁴
	cg04716447	1.53	4.78	3.15	1.44x10 ⁻⁴
TOMM20L	cg04728863	-5.17	-1.66	-3.41	1.44x10 ⁻⁴
PLEKHG3	cg11802553	-10.62	-3.41	-7.01	1.42x10 ⁻⁴
C15orf61	cg22706883	-3.21	-1.03	-2.12	1.43x10 ⁻⁴
CREBBP	cg01312837	-6.27	-2.01	-4.14	1.44x10 ⁻⁴
MIR193B	cg07665535	-4.48	-1.44	-2.96	1.43x10 ⁻⁴
ABCC11	cg02938045	-3.13	-1.01	-2.07	1.42x10 ⁻⁴
TMEM93	cg13345380	-6.59	-2.12	-4.36	1.42x10 ⁻⁴
GNG7	cg23463608	1.04	3.25	2.15	1.43x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
POP4	cg06738169	2.21	6.88	4.54	1.44x10 ⁻⁴
	cg04447445	-11.92	-3.82	-7.87	1.44x10 ⁻⁴
GRK5	cg13898290	-16.01	-5.13	-10.57	1.45x10 ⁻⁴
	cg00021275	-4.04	-1.30	-2.67	1.45x10 ⁻⁴
	cg01894846	-5.84	-1.87	-3.86	1.45x10 ⁻⁴
	cg22627800	5.44	16.98	11.21	1.46x10 ⁻⁴
HDAC4	cg00144180	1.27	3.95	2.61	1.46x10 ⁻⁴
	cg18432806	-5.57	-1.78	-3.68	1.46x10 ⁻⁴
C11orf58	cg20474675	-7.51	-2.41	-4.96	1.46x10 ⁻⁴
SLC9A3R1	cg13896106	-6.14	-1.97	-4.05	1.46x10 ⁻⁴
SYT2	cg11071448	-3.53	-1.13	-2.33	1.47x10 ⁻⁴
	cg07586956	0.82	2.56	1.69	1.47x10 ⁻⁴
	cg03604731	2.81	8.78	5.79	1.47x10 ⁻⁴
	cg00873919	-7.44	-2.38	-4.91	1.46x10 ⁻⁴
SLMAP	cg25235205	1.40	4.36	2.88	1.47x10 ⁻⁴
	cg20378690	-3.09	-0.99	-2.04	1.47x10 ⁻⁴
XPO6	cg00468395	1.30	4.06	2.68	1.47x10 ⁻⁴
HDAC5	cg19163395	1.04	3.26	2.15	1.47x10 ⁻⁴
MIR27A	cg02990289	-7.53	-2.41	-4.97	1.47x10 ⁻⁴
IGLL1	cg14902267	-4.16	-1.33	-2.75	1.47x10 ⁻⁴
	cg15321306	1.21	3.77	2.49	1.47x10 ⁻⁴
KLRG1	cg26806779	-2.68	-0.86	-1.77	1.48x10 ⁻⁴
	cg20244489	0.92	2.88	1.90	1.48x10 ⁻⁴
	cg12426870	-7.05	-2.25	-4.65	1.48x10 ⁻⁴
NEK7	cg12750917	-4.26	-1.36	-2.81	1.49x10 ⁻⁴
HHAT	cg22332066	-9.57	-3.06	-6.31	1.49x10 ⁻⁴
	cg03936870	-4.10	-1.31	-2.71	1.49x10 ⁻⁴
	cg17839366	-4.18	-1.34	-2.76	1.49x10 ⁻⁴
CD80	cg13458803	1.82	5.68	3.75	1.48x10 ⁻⁴
SNX25	cg04716478	-5.10	-1.63	-3.37	1.48x10 ⁻⁴
BRAF	cg12750675	-3.33	-1.07	-2.20	1.48x10 ⁻⁴
	cg02661764	1.20	3.76	2.48	1.49x10 ⁻⁴
LGALS12	cg11183156	-9.25	-2.96	-6.10	1.49x10 ⁻⁴
	cg19529621	-7.08	-2.26	-4.67	1.49x10 ⁻⁴
	cg25252598	-10.43	-3.34	-6.89	1.48x10 ⁻⁴
GUCY2D	cg06079742	-13.00	-4.16	-8.58	1.49x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
PLEKHG5	cg26460530	2.23	6.98	4.60	1.49x10 ⁻⁴
	cg21012139	1.12	3.49	2.31	1.50x10 ⁻⁴
	cg15681737	16.15	-5.16	-10.66	1.50x10 ⁻⁴
DICER1	cg24130561	1.10	3.45	2.28	1.50x10 ⁻⁴
ABCC2	cg17044311	-3.48	-1.11	-2.30	1.50x10 ⁻⁴
SPSB1	cg06779253	1.41	4.41	2.91	1.51x10 ⁻⁴
	cg19075787	-3.00	-0.96	-1.98	1.51x10 ⁻⁴
	cg20105257	1.16	3.65	2.41	1.51x10 ⁻⁴
HLA-E	cg14003265	1.33	4.15	2.74	1.51x10 ⁻⁴
	cg22277154	-4.01	-1.28	-2.65	1.51x10 ⁻⁴
	cg19010490	-5.19	-1.66	-3.43	1.51x10 ⁻⁴
MEIS2	cg01231183	-2.35	-0.75	-1.55	1.51x10 ⁻⁴
	cg21565496	1.21	3.78	2.49	1.52x10 ⁻⁴
	cg25130381	2.16	6.77	4.46	1.52x10 ⁻⁴
SLC9A1	cg11236515	1.33	4.16	2.74	1.52x10 ⁻⁴
	cg04425710	1.01	3.17	2.09	1.52x10 ⁻⁴
	cg22435313	0.97	3.04	2.00	1.52x10 ⁻⁴
FAT1	cg02998591	14.31	-4.56	-9.44	1.53x10 ⁻⁴
ARSB	cg10864794	-3.91	-1.24	-2.58	1.53x10 ⁻⁴
TACC1	cg14773619	1.67	5.23	3.45	1.53x10 ⁻⁴
BCL11B	cg07025989	1.56	4.90	3.23	1.54x10 ⁻⁴
APBA2	cg08908089	1.04	3.26	2.15	1.53x10 ⁻⁴
GMEB2	cg10170269	-6.30	-2.01	-4.15	1.53x10 ⁻⁴
PDCD1	cg03889044	1.27	3.98	2.62	1.54x10 ⁻⁴
CCR1	cg10499974	-3.03	-0.97	-2.00	1.54x10 ⁻⁴
NACC2	cg14126392	-6.94	-2.21	-4.58	1.54x10 ⁻⁴
CORO1B	cg01525879	1.35	4.24	2.80	1.54x10 ⁻⁴
GIPR	cg02942825	-3.40	-1.08	-2.24	1.54x10 ⁻⁴
	cg01901101	-6.42	-2.05	-4.23	1.55x10 ⁻⁴
	cg18938313	-5.73	-1.83	-3.78	1.55x10 ⁻⁴
MAD1L1	cg26785220	-3.40	-1.08	-2.24	1.55x10 ⁻⁴
	cg15089567	1.62	5.08	3.35	1.55x10 ⁻⁴
	cg23934731	-3.53	-1.12	-2.32	1.55x10 ⁻⁴
MICAL3	cg12923994	1.94	6.09	4.02	1.55x10 ⁻⁴
FAM102A	cg13558912	1.78	5.60	3.69	1.55x10 ⁻⁴
ULK1	cg21649013	-5.21	-1.66	-3.44	1.55x10 ⁻⁴
SELO	cg24993400	-4.03	-1.28	-2.66	1.56x10 ⁻⁴
	cg20956594	-2.05	-0.65	-1.35	1.56x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
ACSL6	cg09035699	1.23	3.86	2.55	1.56x10 ⁻⁴
ABHD4	cg06813554	-4.52	-1.44	-2.98	1.56x10 ⁻⁴
LDLR	cg07960944	17.02	-5.41	-11.22	1.56x10 ⁻⁴
GRB10	cg13879047	-6.96	-2.21	-4.58	1.57x10 ⁻⁴
S100A10	cg18348690	-6.58	-2.09	-4.34	1.57x10 ⁻⁴
LOC285780	cg12030031	1.11	3.48	2.29	1.57x10 ⁻⁴
MDFI	cg06688989	-9.06	-2.88	-5.97	1.57x10 ⁻⁴
MEIS2	cg25447144	-6.69	-2.13	-4.41	1.57x10 ⁻⁴
SH2D2A	cg09888330	2.34	7.35	4.85	1.58x10 ⁻⁴
	cg16173109	1.16	3.64	2.40	1.58x10 ⁻⁴
ANKRD33B	cg10581071	-5.96	-1.89	-3.93	1.58x10 ⁻⁴
HSD11B1	cg20941184	-3.78	-1.20	-2.49	1.58x10 ⁻⁴
ZHX2	cg26427777	1.44	4.53	2.99	1.58x10 ⁻⁴
LMO4	cg26801613	-3.98	-1.26	-2.62	1.58x10 ⁻⁴
	cg14426268	10.86	-3.45	-7.15	1.58x10 ⁻⁴
ACACB	cg08866695	-3.29	-1.05	-2.17	1.58x10 ⁻⁴
	cg07930673	0.90	2.82	1.86	1.59x10 ⁻⁴
KIFC1	cg13199639	12.14	-3.86	-8.00	1.59x10 ⁻⁴
KCNE3	cg18838431	-3.21	-1.02	-2.11	1.60x10 ⁻⁴
TAOK3	cg16301036	-2.53	-0.80	-1.67	1.60x10 ⁻⁴
ZBTB47	cg06399164	-4.48	-1.42	-2.95	1.60x10 ⁻⁴
EPHA3	cg04515667	10.33	-3.28	-6.81	1.60x10 ⁻⁴
C4BPB	cg19707677	1.28	4.04	2.66	1.62x10 ⁻⁴
TTC7A	cg25550753	-4.82	-1.53	-3.17	1.61x10 ⁻⁴
	cg08722695	12.17	-3.86	-8.02	1.61x10 ⁻⁴
TMEM185B	cg01109047	2.80	8.83	5.81	1.62x10 ⁻⁴
AGXT	cg17461448	-6.06	-1.92	-3.99	1.60x10 ⁻⁴
	cg02295156	-9.28	-2.94	-6.11	1.61x10 ⁻⁴
TXNDC15	cg14069412	2.22	6.99	4.60	1.62x10 ⁻⁴
	cg19930737	-4.08	-1.30	-2.69	1.61x10 ⁻⁴
BAIAP2L1	cg17341174	-2.66	-0.84	-1.75	1.62x10 ⁻⁴
FAM102A	cg13992008	0.94	2.97	1.95	1.62x10 ⁻⁴
	cg02845870	10.22	-3.24	-6.73	1.62x10 ⁻⁴
RARG	cg09107344	1.28	4.03	2.65	1.61x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg26345203	1.27	3.99	2.63	1.61x10 ⁻⁴
KCNK13	cg02136132	-36.51	-11.59	-24.05	1.60x10 ⁻⁴
BCL11B	cg07015803	0.78	2.44	1.61	1.61x10 ⁻⁴
UBASH3A	cg27280688	0.79	2.50	1.64	1.61x10 ⁻⁴
C17orf44	cg25228625	-5.99	-1.90	-3.94	1.62x10 ⁻⁴
	cg11314827	1.66	5.24	3.45	1.62x10 ⁻⁴
HCP5	cg01082299	1.16	3.65	2.40	1.62x10 ⁻⁴
C8orf73	cg24467290	-6.83	-2.16	-4.49	1.63x10 ⁻⁴
VPS53	cg06500079	-4.40	-1.39	-2.90	1.63x10 ⁻⁴
FCGRT	cg15528736	-4.23	-1.34	-2.78	1.63x10 ⁻⁴
	cg24527636	0.80	2.53	1.66	1.63x10 ⁻⁴
SULT1C2	cg25570328	-3.11	-0.99	-2.05	1.63x10 ⁻⁴
TOLLIP	cg01517832	-3.95	-1.25	-2.60	1.63x10 ⁻⁴
ARID2	cg23192604	-6.17	-1.96	-4.06	1.63x10 ⁻⁴
	cg25729350	-3.58	-1.13	-2.35	1.63x10 ⁻⁴
FHOD1	cg26719831	-4.63	-1.47	-3.05	1.64x10 ⁻⁴
KIAA1949	cg10145196	1.06	3.33	2.19	1.64x10 ⁻⁴
	cg03185794	1.15	3.64	2.40	1.64x10 ⁻⁴
CNR2	cg03611151	-4.05	-1.28	-2.67	1.64x10 ⁻⁴
CDC42BPA	cg03890680	-6.58	-2.08	-4.33	1.64x10 ⁻⁴
PITPNC1	cg15797314	1.32	4.15	2.74	1.64x10 ⁻⁴
EIF4E3	cg23333146	-2.57	-0.81	-1.69	1.65x10 ⁻⁴
	cg19107511	-3.65	-1.15	-2.40	1.65x10 ⁻⁴
SLC25A46	cg02061660	0.91	2.88	1.90	1.66x10 ⁻⁴
	cg13572592	-7.95	-2.52	-5.23	1.66x10 ⁻⁴
METTL9	cg06257110	0.86	2.73	1.79	1.66x10 ⁻⁴
KIAA1310	cg22057050	-5.42	-1.72	-3.57	1.66x10 ⁻⁴
PLCH1	cg18623216	-4.91	-1.55	-3.23	1.66x10 ⁻⁴
	cg16565528	-3.01	-0.95	-1.98	1.66x10 ⁻⁴
DDX6	cg05587627	0.87	2.76	1.81	1.66x10 ⁻⁴
CCR7	cg13504059	1.65	5.20	3.42	1.66x10 ⁻⁴
	cg19100292	-8.37	-2.65	-5.51	1.66x10 ⁻⁴
	cg06760077	1.83	5.79	3.81	1.67x10 ⁻⁴
CNR2	cg07967717	-3.69	-1.17	-2.43	1.68x10 ⁻⁴
ZCCHC11	cg05931551	2.30	7.27	4.79	1.67x10 ⁻⁴
LRPAP1	cg24789434	-11.21	-3.54	-7.38	1.68x10 ⁻⁴
DAB2	cg25105652	-3.69	-1.17	-2.43	1.67x10 ⁻⁴
LOC728264	cg12675571	-7.49	-2.37	-4.93	1.68x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
ZNF212	cg15320998	-3.40	-1.07	-2.24	1.68x10 ⁻⁴ 0.05
SIT1	cg13832290	1.89	6.00	3.95	1.68x10 ⁻⁴ 0.05
MON2	cg19781863	10.80	-3.41	-7.11	1.68x10 ⁻⁴ 0.05
PPP2R5C	cg17904575	0.72	2.27	1.50	1.67x10 ⁻⁴ 0.05
ZNF592	cg21253742	0.78	2.46	1.62	1.67x10 ⁻⁴ 0.05
SEPT9	cg19654743	1.07	3.38	2.23	1.67x10 ⁻⁴ 0.05
MARCH3	cg08975164	-2.59	-0.82	-1.70	1.68x10 ⁻⁴ 0.05
	cg02368583	57.88	18.29	-38.08	1.68x10 ⁻⁴ 0.05
	cg27332104	-4.60	-1.45	-3.03	1.69x10 ⁻⁴ 0.05
	cg20554353	-4.72	-1.49	-3.11	1.70x10 ⁻⁴ 0.05
PLD3	cg20513206	-7.01	-2.21	-4.61	1.70x10 ⁻⁴ 0.05
DNAJB6	cg00114346	-8.50	-2.69	-5.59	1.70x10 ⁻⁴ 0.05
CD3E	cg24612198	0.89	2.82	1.85	1.70x10 ⁻⁴ 0.05
	cg22800884	1.90	6.03	3.97	1.71x10 ⁻⁴ 0.05
	cg19619014	-2.40	-0.76	-1.58	1.71x10 ⁻⁴ 0.05
RRAS2	cg11645556	-4.27	-1.35	-2.81	1.71x10 ⁻⁴ 0.05
	cg05191839	1.16	3.67	2.42	1.71x10 ⁻⁴ 0.05
PITPNC1	cg01599633	-5.96	-1.88	-3.92	1.71x10 ⁻⁴ 0.05
FAIM3	cg23088126	1.09	3.46	2.27	1.73x10 ⁻⁴ 0.05
PRKCE	cg22284398	-4.53	-1.43	-2.98	1.73x10 ⁻⁴ 0.05
GTDC1	cg14768164	0.80	2.55	1.68	1.72x10 ⁻⁴ 0.05
	cg15050753	10.85	-3.42	-7.13	1.73x10 ⁻⁴ 0.05
F2RL1	cg18586277	-2.89	-0.91	-1.90	1.72x10 ⁻⁴ 0.05
MICAL1	cg13206063	-4.27	-1.35	-2.81	1.72x10 ⁻⁴ 0.05
TMEM200A	cg09301462	15.39	-4.85	-10.12	1.72x10 ⁻⁴ 0.05
GNA12	cg03081478	-3.00	-0.94	-1.97	1.71x10 ⁻⁴ 0.05
PSAT1	cg13740985	-3.26	-1.03	-2.15	1.73x10 ⁻⁴ 0.05
NEK6	cg13505631	-4.57	-1.44	-3.01	1.73x10 ⁻⁴ 0.05
FBXL14	cg17824906	1.30	4.14	2.72	1.72x10 ⁻⁴ 0.05
CACNA1C	cg05824594	-3.52	-1.11	-2.31	1.72x10 ⁻⁴ 0.05
	cg16340767	-5.04	-1.59	-3.32	1.72x10 ⁻⁴ 0.05
LAMA3	cg26485825	1.03	3.28	2.16	1.71x10 ⁻⁴ 0.05
CTSZ	cg12831034	-4.60	-1.45	-3.02	1.71x10 ⁻⁴ 0.05
RUNX1	cg26360881	22.97	-7.24	-15.11	1.73x10 ⁻⁴ 0.05
GPR157	cg20284891	0.89	2.84	1.87	1.74x10 ⁻⁴ 0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg09337254	-5.23	-1.65	-3.44	1.74x10 ⁻⁴
ITPR1	cg02105211	-2.39	-0.75	-1.57	1.74x10 ⁻⁴
RAB11FIP1	cg07313709	-7.35	-2.32	-4.83	1.74x10 ⁻⁴
C10orf25	cg00899659	-2.76	-0.87	-1.82	1.74x10 ⁻⁴
	cg09287328	-2.61	-0.82	-1.72	1.74x10 ⁻⁴
PCCA	cg01165817	1.31	4.16	2.73	1.74x10 ⁻⁴
	cg04284196	1.39	4.43	2.91	1.73x10 ⁻⁴
LYRM1	cg08224563	-2.91	-0.92	-1.91	1.74x10 ⁻⁴
	cg13223043	-3.88	-1.22	-2.55	1.74x10 ⁻⁴
COL11A2	cg18651026	-7.40	-2.33	-4.87	1.75x10 ⁻⁴
FOXK1	cg10365743	2.10	6.66	4.38	1.75x10 ⁻⁴
	cg13424302	-2.77	-0.87	-1.82	1.75x10 ⁻⁴
	cg01191058	4.23	13.43	8.83	1.75x10 ⁻⁴
SLC10A7	cg05477920	15.66	-4.93	-10.29	1.75x10 ⁻⁴
	cg15335297	-7.17	-2.26	-4.71	1.75x10 ⁻⁴
LRP2	cg02361027	-9.00	-2.83	-5.92	1.76x10 ⁻⁴
UTP23	cg14355654	-5.84	-1.84	-3.84	1.76x10 ⁻⁴
FLJ42289	cg10933959	-2.34	-0.74	-1.54	1.76x10 ⁻⁴
	cg17581104	1.69	5.37	3.53	1.76x10 ⁻⁴
DOHH	cg03597940	1.98	6.28	4.13	1.76x10 ⁻⁴
CD37	cg07418126	1.65	5.23	3.44	1.76x10 ⁻⁴
ST6GAL1	cg25599673	1.17	3.73	2.45	1.77x10 ⁻⁴
FAM83D	cg04071118	-6.29	-1.98	-4.13	1.77x10 ⁻⁴
NOD1	cg12999366	-3.39	-1.06	-2.23	1.77x10 ⁻⁴
HSPB6	cg01080592	20.70	-6.51	-13.60	1.77x10 ⁻⁴
CTNNBIP1	cg16578549	-5.83	-1.83	-3.83	1.77x10 ⁻⁴
	cg20611272	0.93	2.96	1.95	1.77x10 ⁻⁴
KLRG1	cg14913610	-3.68	-1.16	-2.42	1.77x10 ⁻⁴
TMEM156	cg02131853	1.15	3.65	2.40	1.78x10 ⁻⁴
C20orf117	cg07556261	-6.37	-2.00	-4.18	1.78x10 ⁻⁴
NAALADL2	cg01379656	1.25	3.96	2.61	1.78x10 ⁻⁴
	cg09841889	-1.85	-0.58	-1.22	1.78x10 ⁻⁴
CLEC3B	cg10324158	-4.47	-1.40	-2.94	1.78x10 ⁻⁴
	cg06193328	1.05	3.33	2.19	1.79x10 ⁻⁴
EXT1	cg23554164	-7.31	-2.30	-4.80	1.79x10 ⁻⁴
LCOR	cg13918042	-5.22	-1.64	-3.43	1.78x10 ⁻⁴
KIF2B	cg10158151	11.40	-3.58	-7.49	1.79x10 ⁻⁴

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
FLJ90757	cg23959115	-7.12	-2.24	-4.68	1.79x10 ⁻⁴	0.05
C10orf128	cg00142482	-3.97	-1.25	-2.61	1.79x10 ⁻⁴	0.05
HADH	cg01573501	10.57	-3.32	-6.94	1.80x10 ⁻⁴	0.05
ZC3H12D	cg04494800	-3.38	-1.06	-2.22	1.80x10 ⁻⁴	0.05
CORO1B	cg16408081	1.81	5.76	3.79	1.80x10 ⁻⁴	0.05
TRAPPC1	cg01667702	-3.93	-1.24	-2.59	1.80x10 ⁻⁴	0.05
FOSL2	cg25784219	-3.13	-0.98	-2.05	1.80x10 ⁻⁴	0.05
PARL	cg25686905	15.06	-4.73	-9.90	1.80x10 ⁻⁴	0.05
CNR2	cg00660272	10.68	-3.35	-7.01	1.80x10 ⁻⁴	0.05
CDYL	cg11935248	-3.65	-1.14	-2.40	1.81x10 ⁻⁴	0.05
GNG7	cg24723883	1.21	3.87	2.54	1.81x10 ⁻⁴	0.05
RHOU	cg07838098	3.91	12.47	8.19	1.81x10 ⁻⁴	0.05
	cg19362478	-3.79	-1.19	-2.49	1.81x10 ⁻⁴	0.05
RBMS1	cg17360854	-7.79	-2.44	-5.12	1.82x10 ⁻⁴	0.05
	cg03982897	-6.55	-2.06	-4.30	1.82x10 ⁻⁴	0.05
MICAL3	cg02610425	-7.42	-2.33	-4.88	1.82x10 ⁻⁴	0.05
DIP2C	cg23844018	15.04	47.97	31.51	1.82x10 ⁻⁴	0.05
BANP	cg05166473	2.36	7.51	4.93	1.82x10 ⁻⁴	0.05
SCRN1	cg11855325	14.25	-4.47	-9.36	1.83x10 ⁻⁴	0.05
CALCR	cg13916255	-5.73	-1.79	-3.76	1.83x10 ⁻⁴	0.05
TMCO3	cg25840538	-4.20	-1.32	-2.76	1.83x10 ⁻⁴	0.05
ZNF385A	cg17465423	-3.70	-1.16	-2.43	1.83x10 ⁻⁴	0.05
LZTS2	cg19499884	-2.76	-0.86	-1.81	1.84x10 ⁻⁴	0.05
	cg26200118	-3.45	-1.08	-2.26	1.84x10 ⁻⁴	0.05
PECAM1	cg20830994	1.21	3.86	2.53	1.84x10 ⁻⁴	0.05
WDR25	cg13529217	1.17	3.74	2.46	1.84x10 ⁻⁴	0.05
	cg16788865	-1.78	-0.56	-1.17	1.84x10 ⁻⁴	0.05
GYLTL1B	cg03994942	2.52	8.06	5.29	1.85x10 ⁻⁴	0.05
UBE2Q2	cg04346683	-4.07	-1.28	-2.67	1.85x10 ⁻⁴	0.05
IL21R	cg10454258	-3.58	-1.12	-2.35	1.85x10 ⁻⁴	0.05
FAM134A	cg03878190	-3.24	-1.01	-2.13	1.86x10 ⁻⁴	0.05
TSPAN13	cg11384744	-3.40	-1.06	-2.23	1.86x10 ⁻⁴	0.05
UBAC2	cg19635644	2.05	6.56	4.31	1.86x10 ⁻⁴	0.05
	cg18745279	10.80	-3.38	-7.09	1.86x10 ⁻⁴	0.05
RAP1GAP2	cg16037981	-6.34	-1.99	-4.16	1.86x10 ⁻⁴	0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
ARL5C	cg00808555	-5.94	-1.86	-3.90	1.86x10 ⁻⁴ 0.05
	cg26611328	1.95	6.23	4.09	1.86x10 ⁻⁴ 0.05
	cg10595031	1.42	4.54	2.98	1.87x10 ⁻⁴ 0.05
EXPH5	cg10022155	11.47	-3.59	-7.53	1.87x10 ⁻⁴ 0.05
TSC2	cg08404702	1.66	5.32	3.49	1.87x10 ⁻⁴ 0.05
KIAA1257	cg01223011	0.95	3.05	2.00	1.87x10 ⁻⁴ 0.05
	cg20657864	-4.69	-1.47	-3.08	1.87x10 ⁻⁴ 0.05
KATNAL1	cg02788021	-2.47	-0.77	-1.62	1.87x10 ⁻⁴ 0.05
NLRC5	cg16411857	1.32	4.21	2.76	1.87x10 ⁻⁴ 0.05
MPZL1	cg04846203	-3.39	-1.06	-2.23	1.89x10 ⁻⁴ 0.05
	cg04920761	-8.25	-2.58	-5.41	1.89x10 ⁻⁴ 0.05
ZFP36L2	cg21394171	-4.62	-1.44	-3.03	1.89x10 ⁻⁴ 0.05
	cg02099877	-4.68	-1.46	-3.07	1.89x10 ⁻⁴ 0.05
LETM1	cg08299791	11.53	-3.60	-7.56	1.89x10 ⁻⁴ 0.05
MYOZ3	cg15699693	-2.39	-0.75	-1.57	1.88x10 ⁻⁴ 0.05
FAM8A1	cg03068319	-5.31	-1.66	-3.48	1.89x10 ⁻⁴ 0.05
	cg14560699	-5.62	-1.76	-3.69	1.89x10 ⁻⁴ 0.05
UPP1	cg01092213	-3.03	-0.95	-1.99	1.88x10 ⁻⁴ 0.05
IL23A	cg19951006	1.01	3.24	2.13	1.89x10 ⁻⁴ 0.05
DZIP1	cg24328944	-5.28	-1.65	-3.46	1.88x10 ⁻⁴ 0.05
	cg05859308	1.07	3.41	2.24	1.89x10 ⁻⁴ 0.05
RIOK3	cg05812269	11.71	-3.66	-7.69	1.89x10 ⁻⁴ 0.05
DOK5	cg09177519	23.26	-7.27	-15.26	1.89x10 ⁻⁴ 0.05
ZBTB7B	cg01782486	1.58	5.07	3.32	1.91x10 ⁻⁴ 0.05
ZIC1	cg05371578	-7.71	-2.41	-5.06	1.90x10 ⁻⁴ 0.05
LHFPL2	cg04286455	-2.30	-0.72	-1.51	1.91x10 ⁻⁴ 0.05
EGFL8	cg22101249	-6.57	-2.05	-4.31	1.91x10 ⁻⁴ 0.05
	cg03957124	1.18	3.76	2.47	1.91x10 ⁻⁴ 0.05
ZFAT	cg26464586	-5.93	-1.85	-3.89	1.91x10 ⁻⁴ 0.05
	cg21672855	5.65	18.09	11.87	1.90x10 ⁻⁴ 0.05
TET1	cg17817532	-5.97	-1.86	-3.92	1.89x10 ⁻⁴ 0.05
	cg02345399	1.53	4.90	3.22	1.90x10 ⁻⁴ 0.05
ETS1	cg25578781	-1.97	-0.62	-1.30	1.90x10 ⁻⁴ 0.05
	cg17458693	0.85	2.73	1.79	1.91x10 ⁻⁴ 0.05
C14orf45	cg18638434	-3.88	-1.21	-2.55	1.90x10 ⁻⁴ 0.05
ANKRD11	cg02930721	1.39	4.44	2.91	1.90x10 ⁻⁴ 0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
PDE6G	cg16592596	-6.07	-1.90	-3.98	1.91x10 ⁻⁴ 0.05
ID3	cg27518976	-2.63	-0.82	-1.72	1.93x10 ⁻⁴ 0.05
	cg08468732	1.41	4.52	2.96	1.92x10 ⁻⁴ 0.05
SERPINI1	cg18123677	-4.18	-1.30	-2.74	1.92x10 ⁻⁴ 0.05
	cg04314624	-8.52	-2.66	-5.59	1.92x10 ⁻⁴ 0.05
FBXL13	cg02508830	-3.14	-0.98	-2.06	1.93x10 ⁻⁴ 0.05
PDIA4	cg02892367	2.03	6.52	4.28	1.93x10 ⁻⁴ 0.05
C13orf15	cg23357533	1.01	3.22	2.11	1.93x10 ⁻⁴ 0.05
LBXCOR1	cg26545918	-3.93	-1.23	-2.58	1.93x10 ⁻⁴ 0.05
	cg16973527	14.29	-4.46	-9.37	1.92x10 ⁻⁴ 0.05
ACACA	cg17939040	1.09	3.50	2.30	1.93x10 ⁻⁴ 0.05
RIN2	cg19327615	-3.46	-1.08	-2.27	1.92x10 ⁻⁴ 0.05
	cg27338353	-3.58	-1.12	-2.35	1.92x10 ⁻⁴ 0.05
CCDC19	cg00950718	-3.74	-1.17	-2.45	1.94x10 ⁻⁴ 0.05
LRIG1	cg25111284	-7.08	-2.21	-4.64	1.93x10 ⁻⁴ 0.05
TIAL1	cg26164773	5.00	16.04	10.52	1.94x10 ⁻⁴ 0.05
LRFN3	cg25289028	1.00	3.21	2.10	1.93x10 ⁻⁴ 0.05
	cg15114607	1.09	3.49	2.29	1.94x10 ⁻⁴ 0.05
ERBB2	cg26041593	12.15	-3.79	-7.97	1.94x10 ⁻⁴ 0.05
	cg00122659	0.87	2.81	1.84	1.94x10 ⁻⁴ 0.05
C4orf21	cg23692997	-4.87	-1.52	-3.19	1.95x10 ⁻⁴ 0.05
MNAT1	cg13695076	-4.07	-1.27	-2.67	1.95x10 ⁻⁴ 0.05
TGFBR3	cg21382567	-5.33	-1.66	-3.50	1.95x10 ⁻⁴ 0.05
LGALS3BP	cg11202345	1.79	5.76	3.78	1.95x10 ⁻⁴ 0.05
ADORA2A	cg04990420	1.86	5.98	3.92	1.95x10 ⁻⁴ 0.05
SMOC2	cg16053651	10.65	-3.31	-6.98	1.96x10 ⁻⁴ 0.05
PTPRN2	cg01462349	1.40	4.49	2.94	1.96x10 ⁻⁴ 0.05
SYNPO2	cg12973294	-3.20	-1.00	-2.10	1.96x10 ⁻⁴ 0.05
AHCY	cg05674199	-5.79	-1.80	-3.80	1.96x10 ⁻⁴ 0.05
TRAF3IP2	cg15931839	1.91	6.15	4.03	1.97x10 ⁻⁴ 0.05
GPR3	cg13380502	-3.82	-1.19	-2.51	1.97x10 ⁻⁴ 0.05
LEPRE1	cg04118124	-4.36	-1.36	-2.86	1.97x10 ⁻⁴ 0.05
ZC3H3	cg04180114	0.98	3.14	2.06	1.97x10 ⁻⁴ 0.05
	cg16847800	12.40	-3.86	-8.13	1.97x10 ⁻⁴ 0.05
SPOCK2	cg01861509	1.13	3.65	2.39	1.97x10 ⁻⁴ 0.05
	cg11887996	3.51	11.27	7.39	1.97x10 ⁻⁴ 0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
NKD1	cg16750777	1.87	6.01	3.94	1.97x10 ⁻⁴ 0.05
INADL	cg13321661	-6.69	-2.08	-4.39	1.98x10 ⁻⁴ 0.05
LTB	cg02249390	1.36	4.38	2.87	1.99x10 ⁻⁴ 0.05
NEU1	cg13195955	-5.66	-1.76	-3.71	1.99x10 ⁻⁴ 0.05
C6orf129	cg05758489	-35.95	-11.18	-23.56	1.98x10 ⁻⁴ 0.05
SORBS1	cg03190891	1.54	4.95	3.24	1.98x10 ⁻⁴ 0.05
HSDL1	cg04006133	1.26	4.05	2.65	1.99x10 ⁻⁴ 0.05
KCNJ4	cg20674521	2.73	8.77	5.75	1.99x10 ⁻⁴ 0.05
CACNA1C	cg18806606	3.70	11.89	7.79	1.99x10 ⁻⁴ 0.05
	cg14841601	-2.84	-0.88	-1.86	2.00x10 ⁻⁴ 0.05
CEND1	cg09946142	3.89	12.52	8.21	2.00x10 ⁻⁴ 0.05
GAB2	cg07780377	-3.03	-0.94	-1.98	2.00x10 ⁻⁴ 0.05
RERE	cg00786138	2.06	6.62	4.34	2.01x10 ⁻⁴ 0.05
RERE	cg13104185	1.24	4.01	2.63	2.01x10 ⁻⁴ 0.05
	cg21417130	1.39	4.47	2.93	2.02x10 ⁻⁴ 0.05
ZFAND2B	cg06753787	-4.17	-1.29	-2.73	2.01x10 ⁻⁴ 0.05
GP5	cg14880079	-4.04	-1.26	-2.65	2.01x10 ⁻⁴ 0.05
CDYL	cg27307183	-2.34	-0.73	-1.53	2.01x10 ⁻⁴ 0.05
LY6G6E	cg14258501	-4.31	-1.34	-2.83	2.01x10 ⁻⁴ 0.05
NELL1	cg22307471	-9.24	-2.87	-6.05	2.01x10 ⁻⁴ 0.05
SECTM1	cg26312191	1.20	3.86	2.53	2.01x10 ⁻⁴ 0.05
SBF1	cg01843272	1.37	4.41	2.89	2.01x10 ⁻⁴ 0.05
PWWP2B	cg08751508	-3.45	-1.07	-2.26	2.02x10 ⁻⁴ 0.05
PCNXL2	cg16371229	-7.80	-2.42	-5.11	2.02x10 ⁻⁴ 0.05
PVT1	cg03481855	1.01	3.24	2.13	2.02x10 ⁻⁴ 0.05
TBX21	cg06927323	-9.27	-2.88	-6.08	2.02x10 ⁻⁴ 0.05
MTA3	cg25783099	2.54	8.17	5.36	2.02x10 ⁻⁴ 0.05
CAMK2A	cg05004855	-5.86	-1.82	-3.84	2.03x10 ⁻⁴ 0.05
H6PD	cg02276314	-3.33	-1.03	-2.18	2.03x10 ⁻⁴ 0.05
	cg07873290	-42.32	-13.13	-27.73	2.03x10 ⁻⁴ 0.05
SEMA4B	cg20246851	-4.36	-1.35	-2.86	2.03x10 ⁻⁴ 0.05
TBC1D16	cg02348119	-4.59	-1.42	-3.01	2.03x10 ⁻⁴ 0.05
ALPK2	cg25407077	1.84	5.92	3.88	2.03x10 ⁻⁴ 0.05
UBASH3A	cg09354050	0.88	2.84	1.86	2.03x10 ⁻⁴ 0.05
RCAN3	cg01519464	0.97	3.14	2.05	2.05x10 ⁻⁴ 0.05
	cg11550550	-8.66	-2.69	-5.67	2.04x10 ⁻⁴ 0.05
KIAA1949	cg23672659	1.23	3.96	2.59	2.04x10 ⁻⁴ 0.05
MAD2L1BP	cg02447620	1.48	4.77	3.12	2.04x10 ⁻⁴ 0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
ZMIZ1	cg17065712	-10.11	-3.13 -6.62	2.04x10 ⁻⁴	0.05
BATF	cg21048162	1.26	4.05 2.65	2.04x10 ⁻⁴	0.05
	cg23444610	-8.32	-2.58 -5.45	2.04x10 ⁻⁴	0.05
TMEM62	cg12413346	-3.73	-1.16 -2.44	2.05x10 ⁻⁴	0.05
NSMCE1	cg02654449	-4.54	-1.41 -2.98	2.04x10 ⁻⁴	0.05
TMC8	cg05637296	1.66	5.37 3.52	2.05x10 ⁻⁴	0.05
TBRG4	cg26720010	1.05	3.38 2.22	2.05x10 ⁻⁴	0.05
TNFSF12	cg00031162	1.23	3.98 2.60	2.05x10 ⁻⁴	0.05
ID3	cg18150584	-2.58	-0.80 -1.69	2.08x10 ⁻⁴	0.05
PTP4A2	cg21545720	1.12	3.64 2.38	2.10x10 ⁻⁴	0.05
	cg03879180	-4.24	-1.31 -2.77	2.08x10 ⁻⁴	0.05
	cg09019635	-8.61	-2.66 -5.64	2.08x10 ⁻⁴	0.05
B3GNT7	cg06574960	-4.36	-1.35 -2.86	2.07x10 ⁻⁴	0.05
	cg21096907	-6.01	-1.86 -3.93	2.10x10 ⁻⁴	0.05
C4orf33	cg17707295	-4.84	-1.50 -3.17	2.06x10 ⁻⁴	0.05
	cg00525964	1.02	3.31 2.17	2.07x10 ⁻⁴	0.05
TTC33	cg15528722	-3.19	-0.99 -2.09	2.07x10 ⁻⁴	0.05
ACSL6	cg11010552	-3.72	-1.15 -2.44	2.08x10 ⁻⁴	0.05
HLA-E	cg09326440	4.56	14.75 9.66	2.09x10 ⁻⁴	0.05
	cg17423416	-7.72	-2.39 -5.05	2.09x10 ⁻⁴	0.05
REV3L	cg09365147	-1.80	-0.56 -1.18	2.09x10 ⁻⁴	0.05
C7orf36	cg01835368	-5.16	-1.60 -3.38	2.07x10 ⁻⁴	0.05
CDK6	cg09653641	-2.55	-0.79 -1.67	2.07x10 ⁻⁴	0.05
	cg22728904	-3.60	-1.11 -2.36	2.10x10 ⁻⁴	0.05
CCDC136	cg07611843	-3.80	-1.17 -2.48	2.08x10 ⁻⁴	0.05
EPB49	cg04662594	-5.73	-1.77 -3.75	2.08x10 ⁻⁴	0.05
NDRG1	cg08691775	-3.04	-0.94 -1.99	2.10x10 ⁻⁴	0.05
GSN	cg01136942	-7.36	-2.28 -4.82	2.08x10 ⁻⁴	0.05
SH2D4B	cg14104280	-4.32	-1.34 -2.83	2.10x10 ⁻⁴	0.05
LCOR	cg25589001	-2.89	-0.89 -1.89	2.09x10 ⁻⁴	0.05
UCP3	cg10859442	0.74	2.38 1.56	2.09x10 ⁻⁴	0.05
ARRB1	cg12041266	-6.14	-1.90 -4.02	2.08x10 ⁻⁴	0.05
NCOR2	cg17187521	-3.44	-1.06 -2.25	2.09x10 ⁻⁴	0.05
	cg15899743	1.19	3.84 2.52	2.08x10 ⁻⁴	0.05
	cg03035162	0.98	3.16 2.07	2.08x10 ⁻⁴	0.05
FBLN5	cg02082843	0.98	3.16 2.07	2.09x10 ⁻⁴	0.05
GRIN2C	cg19965023	-8.80	-2.72 -5.76	2.08x10 ⁻⁴	0.05
KIAA0427	cg18530324	-4.95	-1.53 -3.24	2.09x10 ⁻⁴	0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
MBP	cg26457248	-6.32	-1.96	-4.14	2.06x10 ⁻⁴	0.05
C19orf35	cg01129847	-2.89	-0.89	-1.89	2.07x10 ⁻⁴	0.05
	cg20281375	-3.30	-1.02	-2.16	2.09x10 ⁻⁴	0.05
NOSIP	cg18734095	1.18	3.81	2.50	2.08x10 ⁻⁴	0.05
	cg26314722	2.31	7.49	4.90	2.11x10 ⁻⁴	0.05
SPARC	cg21877464	-2.99	-0.92	-1.96	2.11x10 ⁻⁴	0.05
FARS2	cg19689330	-3.31	-1.02	-2.17	2.11x10 ⁻⁴	0.05
C10orf99	cg04126866	1.59	5.16	3.38	2.11x10 ⁻⁴	0.05
	cg20626983	1.02	3.28	2.15	2.11x10 ⁻⁴	0.05
PIK3R5	cg16672810	2.28	7.37	4.82	2.11x10 ⁻⁴	0.05
MGST1	cg00609333	-4.46	-1.38	-2.92	2.11x10 ⁻⁴	0.05
NIN	cg08189198	1.19	3.84	2.51	2.12x10 ⁻⁴	0.05
	cg07820189	-3.81	-1.18	-2.49	2.12x10 ⁻⁴	0.05
CMTM8	cg01617750	-5.10	-1.58	-3.34	2.13x10 ⁻⁴	0.05
CDS1	cg21096050	11.77	-3.64	-7.71	2.13x10 ⁻⁴	0.05
DDAH2	cg03608520	-5.70	-1.76	-3.73	2.14x10 ⁻⁴	0.05
SRF	cg00537673	1.59	5.14	3.36	2.13x10 ⁻⁴	0.05
FAM91A1	cg20283107	-3.90	-1.20	-2.55	2.13x10 ⁻⁴	0.05
GPSM1	cg14164080	-7.02	-2.17	-4.60	2.13x10 ⁻⁴	0.05
GOLGA3	cg01230386	0.83	2.68	1.75	2.13x10 ⁻⁴	0.05
GMFB	cg02051941	-6.28	-1.94	-4.11	2.13x10 ⁻⁴	0.05
MPI	cg07630255	0.95	3.06	2.01	2.13x10 ⁻⁴	0.05
SMAD7	cg14283454	-4.91	-1.52	-3.21	2.13x10 ⁻⁴	0.05
RBM38	cg17799760	1.89	6.11	4.00	2.12x10 ⁻⁴	0.05
PMEPA1	cg04628369	-4.52	-1.39	-2.96	2.13x10 ⁻⁴	0.05
LIME1	cg12413156	1.05	3.41	2.23	2.13x10 ⁻⁴	0.05
LHFPL2	cg26951839	-6.12	-1.89	-4.00	2.14x10 ⁻⁴	0.05
SEC14L1	cg26477169	-6.60	-2.04	-4.32	2.14x10 ⁻⁴	0.05
AKAP8L	cg25533247	-4.96	-1.53	-3.25	2.14x10 ⁻⁴	0.05
DEPDC1	cg19916364	-3.59	-1.11	-2.35	2.15x10 ⁻⁴	0.05
CD1D	cg18234111	-2.69	-0.83	-1.76	2.15x10 ⁻⁴	0.05
	cg00017826	1.03	3.33	2.18	2.16x10 ⁻⁴	0.05
CCR1	cg11589536	-4.94	-1.52	-3.23	2.15x10 ⁻⁴	0.05
MED12L	cg19010441	1.17	3.78	2.47	2.16x10 ⁻⁴	0.05
	cg19118972	-3.80	-1.17	-2.49	2.16x10 ⁻⁴	0.05
TERT	cg13390570	1.10	3.58	2.34	2.16x10 ⁻⁴	0.05
WDFY4	cg07345108	-4.59	-1.41	-3.00	2.16x10 ⁻⁴	0.05
AMBRA1	cg01968525	-4.33	-1.33	-2.83	2.16x10 ⁻⁴	0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
TBCD	cg15787636	0.84	2.72	1.78	2.16x10 ⁻⁴ 0.05
MRPL16	cg08066376	-3.05	-0.94	-1.99	2.17x10 ⁻⁴ 0.05
	cg12759387	1.34	4.34	2.84	2.18x10 ⁻⁴ 0.05
	cg01479187	-2.60	-0.80	-1.70	2.18x10 ⁻⁴ 0.05
	cg22496809	-15.79	-4.86	-10.33	2.17x10 ⁻⁴ 0.05
NACC2	cg03587907	-6.69	-2.06	-4.38	2.18x10 ⁻⁴ 0.05
AMBRA1	cg20090290	-1.94	-0.60	-1.27	2.18x10 ⁻⁴ 0.05
PHACTR4	cg23847017	-3.19	-0.98	-2.09	2.19x10 ⁻⁴ 0.05
TRAPPC3	cg01747664	-5.25	-1.62	-3.43	2.19x10 ⁻⁴ 0.05
NPY2R	cg06812991	-17.80	-5.48	-11.64	2.19x10 ⁻⁴ 0.05
CARD11	cg21516162	-20.48	-6.31	-13.39	2.19x10 ⁻⁴ 0.05
PRMT3	cg05914060	-3.50	-1.08	-2.29	2.19x10 ⁻⁴ 0.05
ATP8B4	cg23671196	-3.03	-0.93	-1.98	2.19x10 ⁻⁴ 0.05
OSGIN1	cg06190046	-2.60	-0.80	-1.70	2.19x10 ⁻⁴ 0.05
AKAP8L	cg20222376	-4.93	-1.52	-3.22	2.19x10 ⁻⁴ 0.05
TCIRG1	cg08932343	-19.83	-6.11	-12.97	2.20x10 ⁻⁴ 0.05
SKI	cg01938025	2.34	7.59	4.96	2.20x10 ⁻⁴ 0.05
GP5	cg05323324	-2.52	-0.78	-1.65	2.20x10 ⁻⁴ 0.05
APLP2	cg25354657	1.04	3.38	2.21	2.20x10 ⁻⁴ 0.05
	cg05892030	-3.88	-1.19	-2.54	2.20x10 ⁻⁴ 0.05
FAM38A	cg27004870	-5.91	-1.82	-3.86	2.20x10 ⁻⁴ 0.05
SEPT9	cg10755077	-5.14	-1.58	-3.36	2.21x10 ⁻⁴ 0.05
NFIX	cg25556035	1.52	4.96	3.24	2.21x10 ⁻⁴ 0.05
NOL12	cg14398464	-6.64	-2.04	-4.34	2.21x10 ⁻⁴ 0.05
	cg17511731	-7.64	-2.35	-5.00	2.21x10 ⁻⁴ 0.05
FANCI	cg22813622	-3.86	-1.19	-2.52	2.21x10 ⁻⁴ 0.05
BRD9	cg18277507	-13.35	-4.11	-8.73	2.22x10 ⁻⁴ 0.05
ARMC2	cg04642300	-2.98	-0.92	-1.95	2.22x10 ⁻⁴ 0.05
PWWP2B	cg25303150	-4.27	-1.31	-2.79	2.22x10 ⁻⁴ 0.05
GUK1	cg07366503	1.12	3.66	2.39	2.23x10 ⁻⁴ 0.05
LOC100189589	cg06224721	-11.19	-3.44	-7.32	2.23x10 ⁻⁴ 0.05
	cg09122442	-5.28	-1.62	-3.45	2.24x10 ⁻⁴ 0.05
GRPEL2	cg22725986	-94.76	-29.12	-61.94	2.23x10 ⁻⁴ 0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value	
SLIT3	cg10165167	-6.96	-2.14	-4.55	2.23x10 ⁻⁴	0.05
BAT2	cg16278747	-13.65	-4.20	-8.92	2.23x10 ⁻⁴	0.05
MAD1L1	cg18044113	1.25	4.08	2.67	2.22x10 ⁻⁴	0.05
MFHAS1	cg02049979	2.48	8.08	5.28	2.23x10 ⁻⁴	0.05
SNX16	cg17516635	-5.90	-1.81	-3.85	2.22x10 ⁻⁴	0.05
BLNK	cg07397033	-5.65	-1.74	-3.69	2.23x10 ⁻⁴	0.05
GPR81	cg17346246	1.31	4.25	2.78	2.23x10 ⁻⁴	0.05
	cg10414881	1.05	3.41	2.23	2.23x10 ⁻⁴	0.05
	cg26854588	0.87	2.82	1.85	2.23x10 ⁻⁴	0.05
LAIR1	cg06238491	-10.48	-3.22	-6.85	2.23x10 ⁻⁴	0.05
	cg15138109	-6.58	-2.02	-4.30	2.24x10 ⁻⁴	0.05
ATP2B2	cg20235117	-11.42	-3.51	-7.46	2.24x10 ⁻⁴	0.05
KIAA1161	cg14013337	-11.51	-3.54	-7.52	2.24x10 ⁻⁴	0.05
FTO	cg10227678	1.23	4.01	2.62	2.24x10 ⁻⁴	0.05
TMIE	cg10775230	-7.30	-2.24	-4.77	2.25x10 ⁻⁴	0.05
CUX1	cg06294954	-11.35	-3.48	-7.42	2.26x10 ⁻⁴	0.05
GSS	cg00352780	-4.84	-1.49	-3.17	2.26x10 ⁻⁴	0.05
	cg09356672	-11.96	-3.67	-7.81	2.27x10 ⁻⁴	0.05
	cg14983236	-3.89	-1.19	-2.54	2.27x10 ⁻⁴	0.05
MTAP	cg25162921	-4.01	-1.23	-2.62	2.26x10 ⁻⁴	0.05
RASAL1	cg01616876	-3.22	-0.99	-2.10	2.27x10 ⁻⁴	0.05
	cg01207974	2.19	7.15	4.67	2.26x10 ⁻⁴	0.05
DAD1	cg07066907	-2.69	-0.83	-1.76	2.26x10 ⁻⁴	0.05
ZFP36L1	cg06617636	0.86	2.80	1.83	2.26x10 ⁻⁴	0.05
TGM5	cg25875049	-3.48	-1.07	-2.28	2.26x10 ⁻⁴	0.05
BLCAP	cg13591710	-6.27	-1.92	-4.10	2.27x10 ⁻⁴	0.05
CCT3	cg05532239	-4.15	-1.27	-2.71	2.28x10 ⁻⁴	0.05
CNIH4	cg18883472	3.32	10.84	7.08	2.28x10 ⁻⁴	0.05
ZAP70	cg14537825	1.36	4.43	2.89	2.28x10 ⁻⁴	0.05
ACTB	cg14145074	1.55	5.05	3.30	2.28x10 ⁻⁴	0.05
GRID2IP	cg07281318	-4.32	-1.32	-2.82	2.28x10 ⁻⁴	0.05
ARID5B	cg00928816	1.44	4.70	3.07	2.28x10 ⁻⁴	0.05
	cg11067407	-3.21	-0.98	-2.10	2.28x10 ⁻⁴	0.05
PLEKHA6	cg21581873	-8.39	-2.57	-5.48	2.29x10 ⁻⁴	0.05

Gene	CpG site	CI- 95%	Effect estimate	Uncorrected p-Value	FDR corrected p -Value
	cg14752089	-3.16	-0.97	-2.06	2.29x10 ⁻⁴
ABLIM1	cg15294435	12.36	-3.79	-8.08	2.29x10 ⁻⁴
C11orf16	cg14402562	-4.98	-1.53	-3.25	2.29x10 ⁻⁴
C3	cg11959387	13.59	-4.16	-8.88	2.29x10 ⁻⁴
GPR77	cg16734795	-3.89	-1.19	-2.54	2.29x10 ⁻⁴
	cg11963409	13.21	-4.05	-8.63	2.30x10 ⁻⁴
LRIG1	cg25742035	10.34	-3.17	-6.76	2.30x10 ⁻⁴
KIF26B	cg19368911	0.90	2.93	1.91	2.30x10 ⁻⁴
NECAB3	cg23687971	-3.29	-1.01	-2.15	2.31x10 ⁻⁴
F11R	cg20012024	-3.58	-1.10	-2.34	2.31x10 ⁻⁴
CLIP4	cg07285673	10.34	-3.17	-6.75	2.31x10 ⁻⁴
	cg20661080	0.80	2.63	1.72	2.31x10 ⁻⁴
PTPRCAP	cg20792833	1.88	6.14	4.01	2.31x10 ⁻⁴

Supplementary Material IV.2: Simulation commented script

This script will be available on github (<https://github.com/SoCadiou>) once the corresponding draft will be published.

- R script for causal structures A, B and C

```
##this code allows to perform a simulation to assess performance in terms of
## sensitivity and specificity of prespecified statistical methods
##used to find the true predictors of an health among the exposome under diverse
##causal structures involving a causal link from the exposome to an outcome.
##Some of them use an intermediary layer, some not. It contains 5 ##parts:
```

```
##1. defining the functions allowing to generate a realistic dataset of exposome
##intermediate layer and outcome. The three layers(E, M and Y) can be linearly
##related to simulate various causal structures.
```

```
#It needs real datasets (exposome/intermediate layer) as inputs, as well
##as parameters allowing to define the association within the three layers (number of
##predictors, variability explained, correlation..)
```

```
##2. defining the methods assessed
```

```
##3. defining some functions used to assess methods performance
```

```
##4. defining the simulation function, which, for a given scenario, generates
```

```

##the datasets, applies the methods and assess their performance. This function
##allows to parallelize the simulation.

##5. running the simulation itself with parallelization, repeating X times the
##function defined in 4. for each scenario and saving the results.
#####
#####

##load packages
library(mvtnorm)
library(boot)
library(parallel)
library(reshape)
library(glmnet)
library(DSA)

#####
##1. define the generating functions
#####
simulator<-
  function(E_true, ##real exposome
    M_true, ## real intermediate layer, eg methylome
    R2_tot = 0.5, ##total variability of the outcome explained by
    #the predictors of E and M
    propmE = 0, ##proportion of variables of M affected by E
    ##without affecting Y
    propmEY = 0.1, ##proportion of variables of M acting as mediators
    ##between E and Y
    propmY = 0, ##proportion of variables of M not affected by E
    ##but affecting Y.
    BetamEY = 0.1, ##coefficient of the effect of M on Y for variables
    ##of M acting as mediators between E and Y. It can be an unique value
    ##or a vector of length n_mEY
    BetamY = 0.1, ##coefficient of the effect of M on Y for variables
    ##of M acting as mediators between E and Y. It can be an unique value
    ##or a vector of length n_mY
    n_mE = NULL, ##alternative way to specify the three previous set
    ##of predictors: #directly giving the number of predictors
    n_mEY = NULL,
    n_mY = NULL,
    n_EmE = 0, ##number of exposures affecting variables of M without
    ##acting through M on the outcome
    n_EmEY = 3,##number of exposures having an effect on Y through M
    n_Ey = 0, ##number of exposures acting directly on M
    ##the 4 next variables specify the intersection between the different
    ## sets of predictors in E
    n_EmE_U_n_EmEY = 0,
    n_Ey_U_n_EmE = 0,

```

```

n_Ey_U_n_EmEY = 0,
n_Ey_U_n_EmE_U_n_EmEY = 0,
##the three next variables specify the coefficients of effects for
##the different sets of predictors in E. They can be an unique value
##or a vector of size respectively n_EmE, n_EmEY, n_Ey...
BetaEmE = 0.1,
BetaEmEY = 0.1,
BetaEy = 0.1,
test_and_training = TRUE ##generating only a training set or
##alternatively also a test set of same size
) {
##check of consistency within the input
if ((BetaEmEY == 0 &
n_EmEY != 0)) {
stop("error: BetaEmEY and n_EmEY not compatible")
}
if ((n_mEY == 0 &
n_EmEY != 0)) {
stop("error: n_mEY and n_EmEY not compatible")
}
if ((n_mEY != 0 &
n_EmEY == 0)) {
stop("error: n_mEY and n_EmEY not compatible")
}
##sampling with replacement the real data for exposome
data.X <- as.data.frame(dataExp_true)
names_row <- rownames(data.X)
data.X <-
  data.X[sample(1:nrow(data.X), 2 * nrow(data.X), replace = TRUE),]
rownames(data.X) <- c(names_row, sprintf('boot%0s', names_row))
dataExp <- data.X
remove(data.X)
##sampling with replacement the real intermediate data
data.X <- as.data.frame(M1_true)
names_row <- rownames(data.X)
data.X <-
  data.X[sample(1:nrow(data.X), 2 * nrow(data.X), replace = TRUE),]
rownames(data.X) <- c(names_row, sprintf('boot%0s', names_row))
M1 <- data.X

##setting linear relationship between E and M

##E on variables on M having no effect on Y

##converting if necessary proportion of predictors to number of predictors
if (is.null(n_mE)) {
  n_mE <- floor(propmE * nrow(M1))
}

```

```

##multiplicity constraint: check if number of variables n_mE is a multiple of
##n_EmE
if (n_mE != 0) {
  n_pat_mE <- (n_mE / n_EmE)
  if (trunc(n_pat_mE) != n_pat_mE) {
    stop("error: number of cpgs explained by E is not a multiple of
      the number of predictors from E")
  }
}
##creating the vector of predictors
mE_predictors <- list()
##generating vector of effect coefficients
Npred_mE_E <- rep(list(n_EmE), n_mE)
Betapred_mE_E <- list()
for (t in seq(length.out = n_EmE)) {
  list_temp <- as.list(rep(0, n_EmE))
  if (length(BetaEmE) == 1) {
    list_temp[[t]] <- BetaEmE
  } else{
    if (length(BetaEmE) != n_EmE) {
      stop("error: Betas for E explaining M are not consistent
        with the number of predictors")
    }
    list_temp[[t]] <- BetaEmE[[t]]
  }
  list_Beta_pred_mE_E_by_pat <- rep(list(list_temp), n_pat_mE)
  Betapred_mE_E <- c(Betapred_mE_E, list_Beta_pred_mE_E_by_pat)
}
##random sampling of causal exposures
ind_mE_E <- sample(ncol(dataExp), n_EmE)
##random sampling of variables of M affected
ind_mE_M <- sample(ncol(M1), n_mE)
##adding a linear effect from E on each variable of M
for (k in seq(length.out = n_mE)) {
  list_temp <-
    list(ind_mE_M[k],
      colnames(M1)[ind_mE_M[k]],
      Npred_mE_E[[k]],
      Betapred_mE_E[[k]],
      ind_mE_E)
  list_temp <- c(list_temp, list(colnames(dataExp)[list_temp[[5]]]))
  names(list_temp) <-
    c(
      "indice_cpg",
      "cpg",
      "nb_exp_predictors",
      "Beta_exp_predictors",
      "ind_exp_predictors",

```

```

  "name_exp_predictors"
)
cpg_temp <-
  simResponseSimple(
    met = dataExp,
    Nmet = list_temp[[3]],
    beta = unlist(list_temp[[4]]),
    list_temp[[5]]
  )
M1[, list_temp[[1]]] <-
  as.numeric(M1[, list_temp[[1]]] + cpg_temp$resp)
list_temp[[7]] <-
  estimatedR2(dataExp, list_temp[[6]], M1[, list_temp[[1]]], drop = FALSE)
names(list_temp)[[7]] <- "R2"
mE_predictors <- c(mE_predictors, list(list_temp))
remove(list_temp)
remove(cpg_temp)
}
##empirical estimation of mean R2 (mean variability of M affected by E
##explained by E)
if (n_mE != 0) {
  R2_mean_mE <-
    mean(unlist(lapply(mE_predictors, function(X)
      X$R2$r.squared)))
} else{
  R2_mean_mE = 0
}

##E on variables on M mediating effect on Y
##converting if necessary proportion of predictors to number of predictors
if (is.null(n_mEY)) {
  n_mEY <- floor(propmEY * nrow(M1))
}
##multiplicity constraint: check if number of variables n_mEY is a multiple of
##n_EmEY
if (n_mEY != 0){
  n_pat_mEY <- (n_mEY / n_EmEY)
  if (trunc(n_pat_mEY) != n_pat_mEY) {
    stop("error: number of cpgs explained by E explaining Y is not a
         multiple of the number of their predictors from E")
  }
}
###creating the vector of predictors
mEY_predictors <- list()
##generating vector of effect coefficients

```

```

Npred_mEY_E <- rep(list(n_EmEY), n_mEY)
Betapred_mEY_E <- list()
for (t in seq(length.out = n_EmEY)) {
  list_temp <- as.list(rep(0, n_EmEY))
  if (length(BetaEmEY) == 1) {
    list_temp[[t]] <- BetaEmEY
  } else{
    if (length(BetaEmEY) != n_EmEY) {
      stop("error: Betas for E explaining M are not consistent
            with the number of predictors")
    }
    list_temp[[t]] <- BetaEmEY[[t]]
  }
  list_Beta_pred_mEY_E_by_pat <- rep(list(list_temp), n_pat_mEY)
  Betapred_mEY_E <- c(Betapred_mEY_E, list_Beta_pred_mEY_E_by_pat)
}
##random sampling of causal exposures
if ((n_mE) != 0) {
  ind_mE_E_in_pred_mE = sample(ind_mE_E, n_mE_U_n_mEY)
  ind_mE_E <-
    c(ind_mE_E_in_pred_mE, sample((1:ncol(dataExp))[-ind_mE_E], n_mEY -
n_mE_U_n_mEY))
} else{
  ind_mE_E <- sample(ncol(dataExp), n_mEY)
}
##random sampling of variables of M
if (n_mE != 0) {
  ind_mE_M <-
    sample((1:ncol(M1))[-ind_mE_M], n_mEY)
} else{
  ind_mE_M <- sample(ncol(M1), n_mEY)
}

##adding a linear effect from E on each variable of M
for (k in seq(length.out = n_mEY)) {
  list_temp <-
    list(ind_mE_M[k],
         colnames(M1)[ind_mE_M[k]],
         Npred_mEY_E[[k]],
         Betapred_mEY_E[[k]],
         ind_mE_E)
  list_temp <- c(list_temp, list(colnames(dataExp)[list_temp[[5]]]))
  names(list_temp) <-
    c(
      "indice_cpg",
      "cpg",
      "nb_exp_predictors",
      "Beta_exp_predictors",

```

```

"ind_exp_predictors",
"name_exp_predictors"
)
cpg_temp <-
simResponseSimple(
  met = dataExp,
  Nmet = list_temp[[3]],
  beta = unlist(list_temp[[4]]),
  list_temp[[5]]
)
M1[, list_temp[[1]]] <-
  as.numeric(M1[, list_temp[[1]]] + cpg_temp$resp)
list_temp[[7]] <-
  estimatedR2(dataExp, list_temp[[6]], M1[, list_temp[[1]]], drop = FALSE)
names(list_temp)[[7]] <- "R2"
mEY_predictors <- c(mEY_predictors, list(list_temp))
remove(list_temp)
remove(cpg_temp)
}
##empirical estimation of mean R2 (mean variability of M affected by E
##explained by E)
if (n_mEY != 0) {
  R2_mean_mEY <-
    mean(unlist(lapply(mEY_predictors, function(X)
      X$R2$r.squared)))
} else{
  R2_mean_mEY = 0
}

##variables of M not affected by E having an effect on Y
if (is.null(n_mY)) {
  n_mY <- floor(propmY * nrow(M1))
}
##generating vector of effect coefficients
if (length(BetamY) == 1) {
  Betapred_yM_M <- rep(BetamY, n_mY)
} else{
  if (length(BetamY) != n_mY) {
    stop(
      "error: Betas for M explaining Y not explained by E are not consistent with the number of
predictors"
    )
  }
  Betapred_yM_M <- BetamY
}
##if there is not effect of mY, generating an empty yM
if (n_mY == 0) {
  yM = list(as.matrix(rep(0, nrow(M1))), ncol = 1, NULL, NULL)
}

```

```

names(yM) <- c("resp", "beta")
} else{
  ##if there is an effect, generating yM: part of the outcome which is a
  ##linear combination of variables of M not affected by E
  if (n_mEY == 0 & n_mE == 0) {
    ind_yM_M <- sample(ncol(M1), n_mY)
  } else{
    ind_yM_M <- sample((1:ncol(M1))[-c(ind_mEY_M, ind_mE_M)], n_mY)
  }
  yM <-
    simResponseSimple(
      met = M1,
      Nmet = length(ind_yM_M),
      beta = Betapred_yM_M,
      cpg = ind_yM_M
    )
}

##variables of M mediating an effect of E on Y
##generating vector of effect coefficients
if (length(BetamEY) == 1) {
  Betapred_yME_M <- rep(BetamEY, n_mEY)
} else{
  if (length(BetamEY) != n_mEY) {
    stop(
      "error: Betas for M explaining Y explained by E are not consistent with the number of
      predictors"
    )
  }
  Betapred_yM_M <- BetamY
}

##if there is an effect, generating yME: part of the outcome which is a
##linear combination of variables of M affected by E
ind_yME_M <- ind_mEY_M
if (n_mEY != 0) {
  yME <-
    simResponseSimple(
      met = M1,
      Nmet = length(ind_yME_M),
      beta = Betapred_yME_M,
      cpg = ind_yME_M
    )
} else{
  yME = list(as.matrix(rep(0, nrow(M1))), ncol = 1), NULL, NULL)
  names(yME) <- c("resp", "beta")
}

```

```

##direct effect of E on Y
##random sampling of exposures with respect to the specification of
#intersections between the different groups of exposures having different
##effects
if (n_EmE == 0 & n_EmEY == 0) {
  ind_yE_E <- sample(ncol(dataExp), n_Ey)
  ind_yE_E_shared_mE <- integer(0)
  ind_yE_E_shared_mE <- integer(0)
  ind_yE_E_shared_mE_mEY <- integer(0)
} else{
  ind_yE_E <-
    sample((1:ncol(dataExp))[-unique(c(ind_mE_E, ind_mEY_E))],
    n_Ey - n_Ey_U_n_EmE - n_Ey_U_n_EmEY + n_Ey_U_n_EmE_U_n_EmEY)
  if (n_Ey_U_n_EmE_U_n_EmEY != 0) {
    ind_yE_E_shared_mE_mEY <-
      sample(intersect(ind_mE_E, ind_mEY_E), n_Ey_U_n_EmE_U_n_EmEY)
  } else{
    ind_yE_E_shared_mE_mEY <- integer(0)
  }
  if (n_Ey_U_n_EmE != 0) {
    ind_yE_E_shared_mE <-
      sample(ind_mE_E[ind_mE_E %in% intersect(ind_mE_E, ind_mEY_E)],
      n_Ey_U_n_EmE -
        n_Ey_U_n_EmE_U_n_EmEY)
  } else{
    ind_yE_E_shared_mE <- integer(0)
  }
  if (n_Ey_U_n_EmEY != 0) {
    ind_yE_E_shared_mEY <-
      sample(ind_mEY_E[ind_mEY_E %in% intersect(ind_mE_E, ind_mEY_E)],
      n_Ey_U_n_EmEY -
        n_Ey_U_n_EmE_U_n_EmEY)
  } else{
    ind_yE_E_shared_mEY <- integer(0)
  }
  ind_yE_E <-
    c(ind_yE_E,
      ind_yE_E_shared_mE,
      ind_yE_E_shared_mEY,
      ind_yE_E_shared_mE_mEY)
}
##generating vector of effects coefficients
if (length(BetaEy) == 1) {
  Betapred_yE_E <- rep(BetaEy, n_Ey)
} else{
  if (length(BetaEy) != n_Ey) {
    stop(

```

```

"error: Betas for M explaining Y not explained by E are not
consistent with the number of predictors"
)
}
Betapred_yE_E <- BetaEy
}
##generating yE: part of the outcome which is a
##linear combination of exposures
yE <-
  simResponseSimple(
    met = dataExp,
    Nmet = length(ind_yE_E),
    beta = Betapred_yE_E,
    cpg = ind_yE_E
  )
##Creating the final Y by adding a gaussian according to the variability
##wanted to the different parts of Y already created
Y <- yE$resp + yME$resp + yM$resp
if (!is.na(R2_tot)) {
  if (((R2_tot) != 0)) {
    sigma <- var(Y) * (1 / R2_tot - 1)
  } else{
    R2 = 0.00000001
    sigma <- var(Y) * (1 / R2_tot - 1)
  }
  Y <- as.matrix(Y + rnorm(length(Y), mean(Y), sqrt(sigma)), ncol = 1)
}
##extracting all indirect predictors of Y from E
##and computing the corresponding betas
datapred <- data.frame(exp = character(0), beta = numeric(0))
for (k in seq(length.out = n_mEY)) {
  datapred_temp <-
    cbind(
      exp = unlist(mEY_predictors[[k]]$name_exp_predictors),
      beta = unlist(mEY_predictors[[k]]$Beta_exp_predictors) * yME$beta[k]
    )
  datapred <- rbind(datapred, datapred_temp)
}
class(datapred$beta) <- "numeric"
yME_E <-
  list(
    resp = yME$resp,
    beta = sapply(unique(datapred$exp), function(X)
      sum(datapred[datapred$exp == X, ]$beta)),
    predictors = unique(datapred$exp)
  )

```

```

##extracting all predictors of Y from E (direct and indirect)
##and computing the corresponding betas
datapred <-
  data.frame(cbind(
    exp = c(
      as.character(yE$predictors),
      as.character(yME_E$predictors)
    ),
    beta = c(as.numeric(yE$beta), as.numeric(yME_E$beta))
  ))
datapred$beta <- as.numeric(as.character(datapred$beta))
yE_ME <-
  list(
    resp = Y,
    beta = sapply(unique(datapred$exp), function(X)
      sum(datapred[datapred$exp == X, ]$beta)),
    predictors = unique(datapred$exp)
  )
remove(datapred)
##extracting all predictors of Y from M
##and computing the corresponding betas
datapred <-
  data.frame(cbind(
    cpg = c(as.character(yM$predictors), as.character(yME$predictors)),
    beta = c(as.numeric(yM$beta), as.numeric(yME$beta))
  ))
datapred$beta <- as.numeric(as.character(datapred$beta))
yM_ME <-
  list(
    resp = Y,
    beta = sapply(unique(datapred$cpg), function(X)
      sum(datapred[datapred$cpg == X, ]$beta)),
    predictors = unique(datapred$cpg)
  )
remove(datapred)

##computing the effective R2
if ((n_mEY + n_mE) != 0) {
  R2_mean_M_E <-
    (n_mEY * R2_mean_mEY + n_mE * R2_mean_mE) / ((n_mEY + n_mE))
} else{
  R2_mean_M_E = 0
}
R2 <-
  list(
    BMI_all_exp = estimatedR2(dataExp, yE_ME$predictors, Y)$r.squared,
    BMI_all_M = estimatedR2(M1, yM_ME$predictors, Y)$r.squared,
    mean_M_E = R2_mean_M_E
  )

```

```

)
##creating a list to return

resultats <-
list(
  Y_train = Y[1:(nrow(M1) / 2), , drop = FALSE],
  E_train = dataExp[1:(nrow(M1) / 2), , drop = FALSE],
  M_train = M1[1:(nrow(M1) / 2), , drop = FALSE],
  Y_test = Y[(nrow(M1) / 2):nrow(M1), , drop = FALSE],
  E_test = dataExp[(nrow(M1) / 2):nrow(M1), , drop = FALSE],
  M_test = M1[(nrow(M1) / 2):nrow(M1), , drop = FALSE],
  yM_E = yME_E,
  y_E = yE_ME,
  yE_E = yE,
  yM_M = yM_ME,
  R2 = R2,
  list_mY_predictor = as.character(yM$predictors),
  list_mE_Y_predictor = as.character(yME$predictors)
)
return(resultats)
}

##function used to create a linear response
simResponseSimple <- function(met, ##matrix of potential predictors
  Nmet = NA, ##number of predictors
  beta = NULL, ##vector of effects
  cpg = NULL) { ##optionnal: directly specifying
  ##some of the indexes of predictors
  if (all(c(is.na(Nmet), is.na(cpg))) == TRUE) {
    return (list(
      resp = as.matrix(rep(0, nrow(met)), ncol = 1),
      beta = NA,
      predictors = NA
    ))
  }
  temp <- Nmet - length(cpg)
  if (temp != 0) {
    wh <- sample((1:ncol(met)[-cpg]), temp)
    wh <- c(cpg, wh)
  } else{
    wh <- cpg
  }
  CovMat <- as.matrix(met[, wh])
  colnames(CovMat) <- colnames(met)[wh]
  # computing the response
  mean <- CovMat %*% matrix(beta, ncol = 1)
  rownames(mean)<-rownames(met)
}

```

```

names(beta) <- colnames(CovMat)
return (list(
  resp = mean,
  beta = beta,
  predictors = colnames(met)[wh]
))
}

##fonction to estimate R2 from a datafram of potential predictors, a vector of
##predictors names and the outcome
estimatedR2 <- function(X, truepred, Y) {
  if ("y" %in% truepred) {
    stop("error: one of the true predictors is named y")
  }
  if (ncol(Y) != 1) {
    stop("error:Y is multidimensionnal")
  }
  if (nrow(X) != nrow(Y)) {
    stop("error not the same number of rows")
  }
  if (isTRUE(all.equal(rownames(X), rownames(Y))) == FALSE) {
    stop("error individuals are not ordered similarly in X and Y")
  }
  if (all(truepred %in% colnames(X))) {
    data <- X[, colnames(X) %in% truepred, drop = FALSE]
    data <- cbind(Y, data)
    colnames(data)[1] <- "y"
    mod <- lm(y ~ ., as.data.frame(data))
    toselect.x <- summary(mod)$coeff[-1, 4]
    r <-
    list(summary(mod)$r.squared,
         summary(mod)$adj.r.squared,
         names(toselect.x)[toselect.x == 'TRUE'])
    names(r) <- c("r.squared", "adj.r.squared", "pred")
    return(r)
  } else{
    stop("error: X does not countain all true predictors")
  }
}

#####
##2. defining the methods to test
#####

###agnostic methods

```

```

##function to compute residuals of a linear model if covariates are specified
getresiduals_2df<-function(data_Y_in,data_covar_in,name_Y,covar){
  data_covar<-data_covar_in[,colnames(data_covar_in)%in%covar,drop=FALSE]
  data_Y<-
  data_Y_in[rownames(data_Y_in)%in%rownames(data_covar),colnames(data_Y_in)==name_Y,drop=FALSE]
  data_covar<-data_covar[rownames(data_covar)%in%rownames(data_Y),,drop=FALSE]
  data_covar<-data_covar[rownames(data_Y),,drop=FALSE]
  data_output<-data_Y
  data<-cbind(data_Y,data_covar)
  mod<-lm(data=data)
  data_output[,1]<-as.data.frame(residuals(mod))
  return(data_output)
}

####ExWAS
ewas<-
function(data_Xs_in=NULL,data_Y_in=NULL,name_Y,data_covar_in=NULL,covar=character(0),corr="BH"){
  require(parallel)
  if (length(covar)>0){
    data_covar<-
    data_covar_in[rownames(data_covar_in)%in%rownames(data_Y_in)&rownames(data_covar_in)%in%rownames(data_Xs_in),colnames(data_covar_in)%in%covar,drop=FALSE]
    data_Y<-
    data_Y_in[rownames(data_Y_in)%in%rownames(data_covar)&rownames(data_Y_in)%in%rownames(data_Xs_in),colnames(data_Y_in)==name_Y,drop=FALSE]
    data_Xs<-
    data_Xs_in[rownames(data_Xs_in)%in%rownames(data_covar)&rownames(data_Xs_in)%in%rownames(data_Y),,drop=FALSE]
    data_covar<-data_covar[rownames(data_Y),,drop=FALSE]
    data_Xs<-data_Xs_in[rownames(data_Xs_in)%in%rownames(data_Y_in),,drop=FALSE]
    data_Y<-getresiduals_2df(data_Y,data_covar,name_Y,covar)
  }else{
    data_Y<-
    data_Y_in[rownames(data_Y_in)%in%rownames(data_Xs_in),colnames(data_Y_in)==name_Y,drop=FALSE]
    data_Xs<-data_Xs_in[rownames(data_Xs_in)%in%rownames(data_Y_in),,drop=FALSE]
    data_Xs<-data_Xs_in[rownames(data_Y),,drop=FALSE]
  }
  if (is.null(data_Y)==TRUE |is.null(data_Xs)==TRUE |!(name_Y%in%colnames(data_Y))) {
    stop("Inconsistent data")
  }
}

##computing p.values
p.values <- lapply(1:ncol(data_Xs), function(x,data_Xs){
  c(colnames(data_Xs)[x],confint(lm(Y~var1,

```

```

data=data.frame(cbind(var1=data_Xs[,x],Y=data_Y[,1])))[2,],
summary(lm(Y~var1,
data=data.frame(cbind(var1=data_Xs[,x],Y=data_Y[,1])))$coefficients[2,]),
data_Xs)
if (length(p.values)>1){
  p.values <- cbind(matrix(unlist(p.values), ncol = 7, byrow = TRUE)[-6])
}

}else{
  p.values <- as.data.frame(t(as.data.frame(unlist(p.values)))[-6,drop=FALSE])
}

p.values<-as.data.frame(p.values)
colnames(p.values) <- c("var","conf - 2.5%","conf - 97.5%","Est","Sd","pVal")
p.values <- p.values[p.values$var!="Intercept"]
p.values$pVal<-as.numeric(as.character(p.values$pVal))
p.values.adj<-p.values
pVal <- as.numeric(as.character(p.values$pVal))
##add correction for multiple testing
if(corr=="None"){
  wh <- which(pVal<=0.05)
  p.values.adj$pVal_adj<-pVal}
if(corr=="Bon"){ wh <- which(pVal<=0.05/nrow(p.values))
p.values.adj$pVal_adj<-pVal*nrow(p.values)}
if(corr=="BH") {wh <- which(p.adjust(pVal,"BH")<=0.05)
p.values.adj$pVal_adj<-p.adjust(pVal,"BH")}
if(corr=="BY") {wh <- which(p.adjust(pVal,"BY")<=0.05)
p.values.adj$pVal_adj<-p.adjust(pVal,"BY")}
if(!corr%in%c("Bon","BH","BY","","None")) stop("Please specify a known correction method
for
multiple testing")

wh <- p.values$var[wh]
a<-list(wh,p.values.adj)
names(a)<-c("selected","pval")
return(a)
}

####LASSO
lasso <-
function(data_Xs_in,
  data_Y_in,
  name_Y,
  data_covar_in = NULL,
  covar = character(0)) {
  if (length(covar) > 0) {
    data_covar <-
      data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
        rownames(data_covar_in) %in% rownames(data_Xs_in),
        colnames(data_covar_in) %in%
        covar, drop = FALSE]
  }
}

```

```

data_Y <-
  data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
             rownames(data_Y_in) %in% rownames(data_Xs_in),
             colnames(data_Y_in) == name_Y, drop =
             FALSE]
data_Xs <-
  data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
             rownames(data_Xs_in) %in% rownames(data_Y_in), ,
             drop = FALSE]
data_covar <- data_covar[rownames(data_Y), ]
data_Xs <- data_Xs[rownames(data_Y), ]
data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
              colnames(data_Y_in) ==
              name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
model.enet <- cv.glmnet(data_Xs, data_Y, family = "gaussian",
                        alpha = 1)
cvfit <- model.enet

##Calcul Y_predit
Y_predit<-predict(cvfit,newx=data_Xs, s = "lambda.min")
Y_predit<-Y_predit[rownames(Y_predit),]

##liste des CPG selectionnés
tmp_coeffs <- coef(cvfit, s ="lambda.min")
cg_select<-data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1], coefficient =
tmp_coeffs@x)

cg_select<-cg_select$name[cg_select$name!="(Intercept)"]
a<-list()
if (length(cg_select)!=0){
  a<-list("selected"=cg_select,"prediction"=Y_predit)
} else{
  a<-list("selected"=character(),"prediction"="no_prediction")
}
return(a)

}

```

```

##DSA
DSAreg <- function(Exp,resp, family = gaussian,maxsize = 15, maxsumofpow = 2,
                     maxorderint = 2){
  Exp <- data.frame(cbind(data.frame(Exp), resp=resp))
  res <- DSA(resp ~ 1, data = Exp, family = family, maxsize = maxsize, maxsumofpow
             = maxsumofpow, maxorderint = maxorderint ,nsplits=1,usersplits = NULL)
  form <- gsub("I[()", "",colnames(coefficients(res)))
  form <- gsub("[*]",":",gsub("D]","",gsub("[^:]1","",form)))
  if(length(grep(":",form))>0){
    nam <- strsplit(form[grep("[:]",form)],":")
    for(j in 1:length(nam)){
      nam[[j]] <- gsub("[[:space:]]","",nam[[j]])
      name <- nam[[j]][1]
      for(k in 2:length(nam[[j]])){
        name <- paste(name,":",nam[[j]][k],sep="")
      }
      Exp <- cbind(Exp,name=apply(Exp[,nam[[j]]],1,prod))
    }
  }
  form2 <- "resp~1"
  if(length(form)>1)for(i in 2:length(form)) form2 <- paste(form2,"+",form[i])
  res2 <- lm(form2, data=data.frame(Exp))
  ##decomment next line and change "prediction" to pred in the return line
  ##if outcome predicted by DSA is needed (not used presently)
  #pred <- predict(res2,Exp)
  coef <- summary(res2)$coefficients
  coef <- as.character(rownames(coef)[rownames(coef)!="Intercept"])

  return(list(selected=coef[coef!="(Intercept)"], pred="prediction"))
}

#####
#####3. defining some functions used to assess methods performance
#####

sensitivity<-function(truepred, predfound){
  return(length(truepred%in%predfound))/length(truepred)
}
fdp<-function(truepred, predfound) { ##false discovery proportion
  if (length(predfound)==0) {return(0)}
  else{
    return(length(predfound[!predfound%in%truepred])/length(predfound))}
}
specificity<-function(truepred, predfound,n_base){
  return(
  (n_base-length(truepred)-length(predfound[!predfound%in%truepred]))/(n_base-
  length(truepred)))
}

```

```
}
```

```
#####
#####
```

```
##4. defining the simulation function which will be parallelized
```

```
#####
#####
```

```
##it first generates datasets, then applies methods and then assessed
```

```
##their performance
```

```
f0<-function(x){
```

```
##important: the parallelization is done on the seed
```

```
set.seed(x)
```

```
##generating datasets
```

```
simu <-
```

```
simulator(
```

```
  E_true = dataExp_true,
```

```
  M_true = M1_true,
```

```
  R2_tot = R2_fixed,
```

```
  propmE = 0,
```

```
  propmEY = 0.1,
```

```
  propmY = 0,
```

```
  BetamEY = BetamEY,
```

```
  BetamY = 0,
```

```
  n_mE = 0,
```

```
  n_mEY = n_mEY,
```

```
  n_mY = 0,
```

```
  n_EmE = 0,
```

```
  n_EmEY = n_EmEY,
```

```
  n_Ey = n_Ey,
```

```
  n_EmE_U_n_EmEY = 0,
```

```
  n_Ey_U_n_EmE = 0,
```

```
  n_Ey_U_n_EmEY = n_Ey_U_n_EmEY,
```

```
  n_Ey_U_n_EmE_U_n_EmEY = 0,
```

```
  BetaEmE = 0,
```

```
  BetaEmEY = BetaEmEY,
```

```
  BetaEy = BetaEy,
```

```
  test_and_training = TRUE
```

```
)
```

```
#####
#####
```

```
##applying methods
```

```
#####
#####
```

```
##ExWas on the intermediate layer
```

```
predBMI_M <-
```

```
list(ewas_BH = ewas(
```

```
  as.data.frame(simu$M_train),
```

```
  as.data.frame(simu$Y_train),
```

```

colnames(as.data.frame(simu$Y_train)),
corr = "BH"
))
##ExWAS on E
predBMI_E <-
list(
  ewas_BH = ewas(
    as.data.frame(simu$E_train),
    as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    corr = "BH"
  ),
  ewas_Bon = ewas(
    as.data.frame(simu$E_train),
    as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    corr = "Bon"
  )
)
print("ewas done")

#####
#oMITM
if (length(predBMI_M$ewas_BH$selected) != 0) {
  select_M <-
    as.data.frame(simu$M_train[,
      colnames(simu$M_train) %in% predBMI_M$ewas_BH$selected,
      drop = FALSE])
  print(ncol(select_M))
  rownames(select_M) <- rownames(simu$M_train)
  colnames(select_M) <- predBMI_M$ewas_BH$selected
  list <- list()
  list_exp <- list()
  list_nom <- list()
  list_ewas_signif <- list()
  ##step b using the exwas performed on M as step a
  for (i in (1:ncol(simu$E_train))) {
    predE_select_M <-
      ewas(
        as.data.frame(select_M),
        (simu$E_train[, i, drop = FALSE]),
        colnames(simu$E_train[, i, drop = FALSE]),
        corr = "None",
        data_covar_in = as.data.frame(simu$Y_train),
        covar = colnames(as.data.frame(simu$Y_train)[1])
      )
    list <- c(list, list(predE_select_M))
  }
}

```

```

list_nom <- c(list_nom, list(colnames(simu$E_train)[i]))

list_exp <- c(list_exp, list(colnames(simu$E_train)[i]))
temp_ewas <-
  cbind(predE_select_M$pval, rep(colnames(simu$E_train)[i],
                                   nrow(predE_select_M$pval)))
list_ewas_signif <- c(list_ewas_signif, list(temp_ewas))
remove(temp_ewas)
remove(predE_select_M)
}
df_all_ewas <- do.call("rbind", list_ewas_signif)
if (!is.null(df_all_ewas)) {
  df_all_ewas$pVal_adj <- p.adjust(df_all_ewas$pVal, "BH")
  colnames(df_all_ewas)[8] <- "exposures"
  names(list) <- as.vector(unlist(list_nom))
}
exp <- df_all_ewas$exposures[df_all_ewas$pVal_adj <= 0.05]
n_exp_select <- length(unique(exp))
##step c
if (length(exp) != 0) {
  select_E <- simu$E_train[, colnames(simu$E_train) %in% exp, drop = FALSE]
  ##ExWAS implementation for step c
  predBMI_E_MITM <-
    ewas(
      as.data.frame(select_E),
      as.data.frame(simu$Y_train),
      colnames(as.data.frame(simu$Y_train)),
      corr = "BH"
    )
  ##DSA implementation for step c
  predBMI_E_MITMds <-
    DSReg(
      Exp = as.data.frame(select_E),
      resp = simu$Y_train,
      maxsize = floor(ncol(simu$E_train) / 10),
      maxsumofpow = 1,
      maxorderint = 1
    )
  predReducedExp <- list(selected = unique(exp), pred = "NULL")
} else{
  predReducedExp <- list(vector(), vector())
  names(predReducedExp) <- c("selected", "pred")
}
if (exists("predBMI_E_MITM")) {
}
else{
  predBMI_E_MITM <- list(vector(), vector())
}

```

```

  names(predBMI_E_MITM) <- c("selected", "pval")
}
if (exists("predBMI_E_MITMds")) {

} else{
  predBMI_E_MITMds <- list(vector(), vector())
  names(predBMI_E_MITM) <- c("selected", "pred")
}
} else{
  predBMI_E_MITM <- list(vector(), vector())
  names(predBMI_E_MITM) <- c("selected", "pval")
  n_exp_select = 0
  predReducedExp <- list(vector(), vector())
  names(predReducedExp) <- c("selected", "pred")
  predBMI_E_MITMds <- list(vector(), vector())
  names(predBMI_E_MITMds) <- c("selected", "pred"
  )
}

##storing results in a list
predBMI_E <-
c(
  predBMI_E,
  MITM = list(predBMI_E_MITM),
  MITMds = list(predBMI_E_MITMds),
  ReducedExp = list(predReducedExp)
)
print("oMITM done")

####Control method : random sampling on a random set of exposures of same
##dimension as the reduced exposome of oMITM
if (n_exp_select > 0) {
  tirage <-
  ewas(
    as.data.frame(simu$E_train)[, sample(colnames(as.data.frame(simu$E_train)),
                                           n_exp_select), drop =
      FALSE],
    as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    corr = "BH"
  )
} else{
  tirage <- list(selected = character(0), null = "null")
}
##storing results in the same list
predBMI_E <- c(predBMI_E, random_sampling = list(tirage))
print(n_exp_select)

```

```

#####
##mediation
if (length(predBMI_E$ewas_BH$selected) != 0) {
  select_E <-
  as.data.frame(simu$E_train[,
    colnames(simu$E_train) %in% predBMI_E$ewas_BH$selected,
    drop = FALSE])
  rownames(select_E) <- rownames(simu$E_train)
  colnames(select_E) <- predBMI_E$ewas_BH$selected
  #step a
  list_temp_ewas_med <- list()
  for (i in 1:ncol(simu$M_train)) {
    exp_affecting_M_all <-
    ewas(as.data.frame(select_E),
      (simu$M_train[, i, drop = FALSE]),
      colnames(simu$M_train[, i, drop = FALSE]),
      corr = "None")
    temp_ewas_med <-
    cbind(exp_affecting_M_all$pval, rep(colnames(simu$M_train)[i],
      nrow(exp_affecting_M_all$pval)))
    list_temp_ewas_med <- c(list_temp_ewas_med, list(temp_ewas_med))
  }
  ewas_med <- do.call("rbind", list_temp_ewas_med)
  if (!is.null(ewas_med)) {
    ewas_med$pVal_adj_1 <- p.adjust(ewas_med$pVal, "BH")
    colnames(ewas_med)[8] <- "cpg"
    colnames(ewas_med)[1] <- "exp"
  }
  #step b
  list_temp_ewas_med_2 <- list()
  for (i in 1:ncol(select_E)) {
    M_affecting_Y_all <-
    ewas(
      as.data.frame(simu$M_train),
      as.data.frame(simu$Y_train),
      colnames(as.data.frame(simu$Y_train)),
      corr = "None",
      data_covar_in = (select_E[, i, drop = FALSE]),
      covar = colnames(select_E[, i, drop = FALSE]))
    )
    temp_ewas_med_2 <-
    cbind(M_affecting_Y_all$pval, rep(colnames(select_E)[i],
      nrow(M_affecting_Y_all$pval)))
    list_temp_ewas_med_2 <-
    c(list_temp_ewas_med_2, list(temp_ewas_med_2))
  }
  ewas_med_2 <- do.call("rbind", list_temp_ewas_med_2)
  if (!is.null(ewas_med_2)) {

```

```

ewas_med_2$pVal_adj_2 <- p.adjust(ewas_med_2$pVal, "BH")
colnames(ewas_med_2)[8] <- "exp"
colnames(ewas_med_2)[1] <- "cpg"
}

ewas_med_tot <-
  merge(
    ewas_med,
    ewas_med_2,
    by.x = c("exp", "cpg"),
    by.y = c("exp", "cpg")
  )
exp_med <-
  unique(ewas_med_tot$exp[ewas_med_tot$pVal_adj_2 <= 0.05 &
    ewas_med_tot$pVal_adj_1 <= 0.05])
exp_med <- exp_med[!is.na(exp_med)]
if (length(exp_med) != 0) {
  predMediation <- list(selected = exp_med, pred = NULL)
} else{
  predMediation <- list(selected = vector(), pred = vector())
}

} else{
  predMediation <- list(selected = vector(), pred = vector())
}
predBMI_E <- c(predBMI_E, mediation = list(predMediation))

#####
##applying agnostic methods
## lasso
predlasso <-
  lasso(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train))
  )
predBMI_E <- c(predBMI_E, lasso_CV = list(predlasso))
print("lasso_ done")
##DSA
predDSA <-
  DSAreg(
    Exp = simu$E_train,
    resp = simu$Y_train,
    maxsize = floor(ncol(simu$E_train) / 10),
    maxsumofpow = 1,
    maxorderint = 1
  )
predBMI_E <- c(predBMI_E, DSA = list(predDSA))

```

```

print("DSA done")

#####
##assessing performance for ExWAS on M
truepred<-simu$yM_M$predictors
for (k1 in 1:length(predBMI_M)){
  truepred<-simu$yM_M$predictors
  predfound<-predBMI_M[[k1]]$selected
  if (exists("predfound")&exists("truepred")){
    if (length(predfound)==0) {print("no predictors found")}
    a<-sensitivity(truepred,predfound)
    b<-specificity(truepred,predfound,ncol(simu$M_train))
    c<-fdp(truepred,predfound)
    d<-estimatedR2(simu$M_test,predfound,simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)
    remove(predfound,truepred)
  }else{
    if(!exists("predfound")&exists("truepred")){
      a<-0
      b<-specificity(truepred,numerical(0),ncol(simu$M_train))
      c<-0
      d<-0
      remove(truepred)
    }else{
      if(!exists("truepred")&exists("predfound")){
        a<-1
        b<-specificity(numerical(0),predfound,ncol(simu$M_train))
        c<-fdp(numerical(0),predfound)
        d<-estimatedR2(simu$M_test,predfound,simu$Y_test)$r.squared
        remove(predfound)
      }else{
        if(!exists("truepred")&!exists("predfound")){
          a<-1
          b<-1
          c<-0
          d<-0
        }}}
  }
  predBMI_M[[k1]]<-c(predBMI_M[[k1]],sens=a,spec=b,fdp=c,R2_test=d)
  remove(a)
  remove(b)
  remove(c)
  remove(d)
}

}

```

```

##assessing performance for all methods on E
truepred<-simu$y_E$predictors
for (k1 in 1:length(predBMI_E)){
  truepred<-simu$y_E$predictors
  truepred<-simu$y_E$predictors
  predfound<-predBMI_E[[k1]]$selected
  if (exists("predfound")&exists("truepred")){
    if (length(predfound)==0) {print("no predictors found")}
    a<-sensitivity(truepred,predfound)
    b<-specificity(truepred,predfound,ncol(simu$E_train))
    c<-fdp(truepred,predfound)
    d<-estimatedR2(simu$E_test,predfound,simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)
    remove(predfound,truepred)
  }else{
    if(!exists("predfound")&exists("truepred")){
      a<-0
      b<-specificity(truepred,numeric(0),ncol(simu$E_train))
      c<-0
      d<-0
      remove(truepred)
    }else{
      if(!exists("truepred")&exists("predfound")){
        a<-1
        b<-specificity(numeric(0),predfound,ncol(simu$E_train))
        c<-fdp(numeric(0),predfound)
        d<-estimatedR2(simu$E_test,predfound,simu$Y_test)$r.squared
        remove(predfound)
      }else{
        if(!exists("truepred")&!exists("predfound")){
          a<-1
          b<-1
          c<-0
          d<-0
        }}}
    }
  predBMI_E[[k1]]<-c(predBMI_E[[k1]],sens=a,spec=b,fdp=c,R2_test=d)
  remove(a)
  remove(b)
  remove
  remove(d)
}

##assessing the performance considering only the predictors having an indirect

```

```

##effect
truepred<-simu$yM_E$predictors
for (k1 in 1:length(predBMI_E)){
  truepred<-simu$yM_E$predictors
  predfound<-predBMI_E[[k1]]$selected
  if (exists("predfound")&exists("truepred")){
    if (length(predfound)==0) {print("no predictors found")}
    a<-sensitivity(truepred,predfound)
    b<-specificity(truepred,predfound,ncol(simu$E_train))
    c<-fdp(truepred,predfound)
    d<-estimatedR2(simu$E_test,predfound,simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)

    remove(predfound,truepred)
  }else{
    if(!exists("predfound")&exists("truepred")){
      a<-0
      b<-specificity(truepred,numeric(0),ncol(simu$E_train))
      c<-0
      d<-0
      remove(truepred)
    }else{
      if(!exists("truepred")&exists("predfound")){
        a<-1
        b<-specificity(numeric(0),predfound,ncol(simu$E_train))
        c<-fdp(numeric(0),predfound)
        d<-estimatedR2(simu$E_test,predfound,simu$Y_test)$r.squared
        remove(predfound)
      }else{
        if(!exists("truepred")&!exists("predfound")){
          a<-1
          b<-1
          c<-0
          d<-0
        }}}
    }
  predBMI_E[[k1]]<-c(predBMI_E[[k1]],sens=a,spec=b,fdp=c,R2_test=d)
  remove(a)
  remove(b)
  remove
  remove(d)
}
print("performance characterized")
##building the list with datasets generated + results of methods +
##performance to return

```

```

A<-list(simu=simu,predBMI_E=predBMI_E,predBMI_M=predBMI_M,
        nl_exp_select=n_exp_select)
remove(simu)
remove(select_M)
remove(predBMI_M)
remove(predBMI_E)
remove(predBMI_E_MITM_WM)
remove(predBMI_E_MITM)
remove(predlasso)
remove(n_exp_select)
gc()
return(A)

}

#####
##5. Running simulations
#####

##loading real datasets
dataExp_true <- readRDS("20190205 Exposome simu borne.rds")
M1_true <- readRDS("20191129 Methylome simu.Rds")
M1_true <- scale(M1_true)

##initialization
list_simulated_data <- list()
list_list_predBMI_E <- list()
list_list_predBMI_M <- list()
list_list_nl_exp_select <- list()

##setting simulations parameters
n_iter <- 100 ##number of iterations for one scenarios
##parameters for generating datasets
##all combinations will be tested (each combination allows to build a scenario)
##(adapt the code of the loop if multiple values instead of single values for
##some parameters)
c_n_my <- c(10, 18, 25, 100)
c_n_R2_fixed <- c(0.01, 0.05, 0.1, 0.4)
BetamEY = 0.01
c_BetaEy <- c(0.0001, 0.001, 0.01, 0.1, 0.5)
c_n_Ey <- c(1, 3, 10, 25)
c_BetaEmEY <- c(0.0001, 0.001, 0.01, 0.1, 0.5)
n_mE <- 0
n_mEY <- 0

##initialization of table of results
comp_method <-

```

```

data.frame(
  Methods = vector(),
  Association_tested = vector(),
  Nb_true_predictors_of_BMI_in_M = numeric(0),
  Nb_true_predictors_of_BMI_in_E = numeric(0),
  Total_variability_of_BMI_explained_by_EandM = numeric(0),
  Total_variability_of_BMI_explained_by_E = numeric(0),
  Total_variability_of_BMI_explained_by_M = numeric(0),
  Mean_variability_of_M_explained_by_E_for_Mey = numeric(0),
  Number_iterations = numeric(0),
  Mean_number_exp_selected_to_be_randomly_tested =
    numeric(0),
  Mean_number_predictors_found = numeric(0),
  Mean_sensitivity = numeric(0),
  Mean_specificity = numeric(0),
  Mean_fdp = numeric(0),
  Mean_R2_test = numeric(0),
  Mean-mediated_sensitivity = numeric(0),
  Mean-mediated_specificity = numeric(0),
  Mean-mediated_fdp = numeric(0),
  Mean_R2_test = numeric(0),
  SD_number_predictors_found = numeric(0),
  SD_sensitivity = numeric(0),
  SD_specificity = numeric(0),
  SD_fdp = numeric(0),
  SD_R2_test = numeric(0),
  SD-mediated_sensitivity = numeric(0),
  SD-mediated_specificity = numeric(0),
  SD-mediated_fdp = numeric(0),
  SD_R2_test = numeric(0),
  Which_iteration = numeric(0)
)

```

```

##looping on the different vectors of parameters
n = 1
for (i2 in 1:length(c_n_R2_fixed)) {
  R2_fixed <- c_n_R2_fixed[i2]
  for (i3 in 1:length(c_BetaEy)) {
    BetaEy <- c_BetaEy[i3]
    for (i4 in 1:length(c_n_Ey)) {
      n_Ey <- c_n_Ey[i4]
      n_Ey_U_n_EmEY <- n_Ey
      n_EmEY <- n_Ey
      for (i1 in 1:length(c_n_my)) {
        n_mEY <- c_n_my[i1]
        for (i5 in 1:length(c_BetaEmEY)) {
          BetaEmEY <- c_BetaEmEY[i5]

```

```

print(n)
iteration_OK <- TRUE

if (n < 1501) {
  n <- n + 1
} else{
  if (n_mE != 0) {
    n_pat_mE <- (n_mE / n_EmE)
    if (trunc(n_pat_mE) != n_pat_mE) {
      iteration_OK <- FALSE
    }
  }
  if (n_mE != 0) {
    n_pat_mE <- (n_mE / n_EmE)
    if (trunc(n_pat_mE) != n_pat_mE) {
      iteration_OK <- FALSE
    }
  }
}
if (iteration_OK == TRUE) {
  n_row = nrow(comp_method)
  simulated_data <- list()
  list_predBMI_E <- list()
  list_predBMI_M <- list()
  list_nl_exp_select <- list()
  start_time <- Sys.time()
  ##parallelization of f0
  cl <- makeClustergetOption("cl.cores", round(detectCores())))
  clusterExport(
    cl,
    list(
      "simulator",
      "simResponseSimple",
      "estimatedR2",
      "getresiduals_2df",
      "ewas",
      "lasso",
      "lasso_stab",
      "DSAreg",
      "sensitivity",
      "fdp",
      "specificity",
      "f0",
      "dataExp_true",
      "M1_true",
      "R2_fixed",
      "n_Ey",
      "n_mEY",
      "BetaEy",
    )
}

```

```

  "BetamEY",
  "n_Ey_U_n_EmEY",
  "n_EmEY",
  "BetaEmEY"
)
)
clusterEvalQ(cl, list(library("boot"), library("reshape"),
  library("glmnet"), library("DSA")))
results_1_jeu <- clusterApply(cl, 1:n_iter, f0)
stopCluster(cl)
simulated_data <- lapply(results_1_jeu, function(x)
  x$simu)

##structure of results prioritized by methods and not anymore
##prioritized by datasets
list_predBMI_E <- lapply(results_1_jeu, function(x)
  x$predBMI_E)
list_predBMI_M <- lapply(results_1_jeu, function(x)
  x$predBMI_M)
list_nl_exp_select <-
  lapply(results_1_jeu, function(x)
    x$nl_exp_select)
remove(results_1_jeu)

####compilation of results for this scenario

##table describing the empirical characteristics of
##the simulated datasets
param_simu <-
  data.frame(
    Parameters = vector(),
    Fixed_or_measured = vector(),
    Value = numeric(0)
  )
param_simu[1, ] <-
  c(
    as.character("Nb_true predictors of BMI in M"),
    as.character("Fixed"),
    mean(unlist(
      lapply(simulated_data, function(X)
        length(X[[10]]$beta)))
  )))
param_simu[2, ] <-
  c("Nb_true predictors of BMI in E",
    "Measured",
    mean(unlist(

```

```

lapply(simulated_data, function(X)
  length(X[[8]]$beta))
)))
param_simu[3, ] <-
  c("Total variability of BMI explained by (E+M)",
  "Fixed",
  "R2_fixed")
param_simu[4, ] <-
  c("Mean variability of BMI explained by E",
  "Measured",
  mean(unlist(
    lapply(simulated_data, function(X)
      (X[[11]]$BMI_all_exp))
  )))
param_simu[5, ] <-
  c("Mean variability of BMI explained by M",
  "Measured",
  mean(unlist(
    lapply(simulated_data, function(X)
      (X[[11]]$BMI_all_M))
  )))
param_simu[6, ] <-
  c(
    "Mean variability of part of M explained by E explained by E",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X[[11]]$mean_M_E))
    )))
  )
param_simu[7, ] <- c("Number_iterations", "Fixed", n_iter)
param_simu[8, ] <-
  c(
    "Mean_number_exp_selected_to_be_randomly_tested",
    "Measured",
    mean(unlist(list_nl_exp_select))
  )

##summarizing each method performance by a line in comp_method datadrame within
##this scenario

for (k1 in (1:length(list_predBMI_M[[1]]))) {
  comp_method[n_row + k1, ] <-
    c(
      names(list_predBMI_M[[1]][k1]),
      "BMI - M",
      param_simu[1, 3],

```

```

param_simu[2, 3],
R2_fixed,
param_simu[4, 3],
param_simu[5, 3],
param_simu[6, 3],
n_iter,
NA,
mean(unlist(
  lapply(list_predBMI_M, function(X)
    length(X[[k1]][[1]])))
)),
mean(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[3]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[4]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[5]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[6]]))
), na.rm = TRUE),
NA,
NA,
NA,
NA,
sd(unlist(
  lapply(list_predBMI_M, function(X)
    length(X[[k1]][[1]])))
)),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[3]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[4]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[5]]))
), na.rm = TRUE),
sd(unlist(

```

```

lapply(list_predBMI_M, function(X)
  (X[[k1]][[6]]))
), na.rm = TRUE),
NA,
NA,
NA,
NA,
NA,
n
)
}

for (k2 in (1:length(list_predBMI_E[[1]]))) {
  comp_method[n_row + k2 + k1, ] <-
  c(
    names(list_predBMI_E[[1]][k2]),
    "BMI - E",
    param_simu[1, 3],
    param_simu[2, 3],
    R2_fixed,
    param_simu[4, 3],
    param_simu[5, 3],
    param_simu[6, 3],
    n_iter,
    param_simu[8, 3],
    mean(unlist(
      lapply(list_predBMI_E, function(X)
        length(X[[k2]][[1]]))
    )),
    mean(unlist(
      lapply(list_predBMI_E, function(X)
        (X[[k2]][[3]]))
    )),
    mean(unlist(
      lapply(list_predBMI_E, function(X)
        (X[[k2]][[4]]))
    )),
    mean(unlist(
      lapply(list_predBMI_E, function(X)
        (X[[k2]][[5]]))
    )),
    mean(unlist(
      lapply(list_predBMI_E, function(X)
        (X[[k2]][[6]]))
    )),
    mean(unlist(
      lapply(list_predBMI_E, function(X)
        (X[[k2]][[7]]))
    ))
  )
}

```

```

)),  

mean(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[8]]))  

)),  

mean(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[9]]))  

)),  

mean(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[10]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    length(X[[k2]][[1]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[3]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[4]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[5]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[6]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[7]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[8]]))  

)),  

sd(unlist(  

  lapply(list_predBMI_E, function(X)  

    (X[[k2]][[9]]))  

)),  

sd(unlist(

```

```

lapply(list_predBMI_E, function(X)
  (X[[k2]][[10]]))
)),
n
)
}

##storing generated datasets and methods results for this scenario
list_list_predBMI_E <- c(list_list_predBMI_E, list(list_predBMI_E))
list_list_predBMI_M <- c(list_list_predBMI_M, list(list_predBMI_M))
list_list_nl_exp_select <-
  c(list_list_nl_exp_select,
    list(list_nl_exp_select))
end_time <- Sys.time()
end_time - start_time
##saving
saveRDS(comp_method,
  "comp_method_meditation_and_direct.Rds")
saveRDS(
  simulated_data,
  file = paste(
    n,
    '_simulated_data_scenario_iteration_meditation_and_direct.Rds'
  )
)
saveRDS(
  list_list_predBMI_E,
  "list_list_predBMI_E_meditation_and_direct.rds"
)
saveRDS(
  list_list_predBMI_M,
  "list_list_predBMI_M_meditation_and_direct.rds"
)
saveRDS(
  list_list_nl_exp_select,
  "list_list_nl_exp_select_meditation_and_direct.rds"
)

remove(simulated_data)
remove(list_predBMI_E)
remove(list_predBMI_M)
remove(list_nl_exp_select)
}
n = n + 1
}
}
}

```

```

  }
}
}

```

- R script for causal structure D and E

```

##this code allows to perform a simulation to assess performance in terms of
## sensitivity and specificity of prespecified statistical methods
##used to find the true predictors of an health among the exposome. The causal
##structures considered involve a reverse causative likn from the outcome on
##the exposome Some of them use an intermediary layer, some not.
##it contains 5 ##parts:

```

```

##1. defining the functions allowing to generate a realistic dataset of exposome
##intermediate layer and outcome. The three layers(E, M and Y) can be linearly
##related to simulate various causal structures.

```

```

#It needs real datasets (exosome/intermediate layer/outcome) as inputs, as well
##as parameters allowing to define the association within the three layers (number of
#predictors, #variability explained, correlation..)

```

```

##2. defining the methods assessed

```

```

##3. defining some functions used to assess methods performance

```

```

##4. defining the simulation function, which, for a given scenario, generates
##the datasets, applies the methods and assess their performance. This function
##allows to parallelize the simulation.

```

```

##5. runnning the simulation itself with parallelization, repeating X times the
##function defined in 4. for each scenario and saving the results.
#####
#####
```

```

##load packages
library(mvtnorm)
library(boot)
library(parallel)
library(reshape)
library(glmnet)
library(DSA)

```

```

#####
##1. define the generating functions
#####

```

```

simulator <-
  function(E_true,
    ##real exposome
    M_true,
    ## real intermediate layer, eg methylome
    Y_true,
    ##real outcome
    n_mY = 0,
    #variables of M not affected by E
    ##but affecting Y.
    R2_mY = 0,
    ##variability of M not affected by E affecting Y
    BetamY = 0.1,
    ##corresponding effect coefficient
    n_yM = 50,
    ##number of variables of M affected by Y
    Beta_yM = 0.001,
    ##corresponding effect coefficient
    n_yE = 5,
    ##number of variables of E affected by Y
    Beta_yE = 0.01,
    ##corresponding effect coefficient
    test_and_training = TRUE)#generating only a training set or
    ##alternatively also a test set of same size)
    {
    ##sampling with replacement the real data for exposome
    data.X <- as.data.frame(dataExp_true)
    names_row <- rownames(data.X)
    data.X <-
      data.X[sample(1:nrow(data.X), 2 * nrow(data.X), replace = TRUE),]
    rownames(data.X) <- c(names_row, sprintf('boot%0s', names_row))
    dataExp <- data.X
    remove(data.X)
    ##sampling with replacement the real intermediate data
    data.X <- as.data.frame(M1_true)
    names_row <- rownames(data.X)
    data.X <-
      data.X[sample(1:nrow(data.X), 2 * nrow(data.X), replace = TRUE),]
    rownames(data.X) <- c(names_row, sprintf('boot%0s', names_row))
    M1 <- data.X
    ##sampling with replacement the real outcome data
    data.X <- as.data.frame(Y_true)
    names_row <- rownames(data.X)
    data.X <-
      data.X[sample(1:nrow(data.X), 2 * nrow(data.X), replace = TRUE),
      , drop =
        FALSE]
  }

```

```

rownames(data.X) <- c(names_row, sprintf('boot%0s', names_row))
Y_boot <- data.X
remove(data.X)

##setting if necessary linear relationship from M to Y
if (n_mY != 0) {
  if (R2_mY == 0) {
    stop("error: n_MY and R2_mY non consistent")
  }
  ##generating vector of effect coefficients
  if (length(BetamY) == 1) {
    Betapred_mY <- rep(BetamY, n_mY)
  } else{
    if (length(BetamY) != n_mY) {
      stop("error: Betas for M explaining Y are not consistent
      with the number of predictors")
    }
    Betapred_mY <- BetamY
  }
  ##random sampling of variables of M affecting Y
  ind_mY <- sample(ncol(M1), n_mY)
  ##generating yM: part of the outcome which is a
  ##linear combination of variables of M not affected by E
  yM <-
    simResponseSimple(
      met = M1,
      Nmet = length(ind_mY),
      beta = Betapred_mY,
      cpg = ind_mY
    )
  if (is.na(R2_mY)) {
    if (((R2_mY) != 0)) {
      sigma <- var(yM$resp) * (1 / R2_mY - 1)
    } else{
      warning("R2 was not specified, automatic value")
      R2_mY = 0.00000001
      sigma <- var(Y$resp) * (1 / R2_mY - 1)
    }
    Y <-
      as.matrix(Y$resp + rnorm(length(Y$resp), mean(Y$resp),
      sqrt(sigma)), ncol = 1)
    #standardization
    Y <- as.data.frame(scale(Y))
  }
  ##empirical estimation of R2 ( variability of Y explained by M)
  R2_mY_measured <- estimatedR2(M1, ind_mY, Y)$r.squared
} else{

```

```

##if no effect, creating an empty yM
yM <- list(resp = Y_boot,
            beta = NULL,
            predictors = NULL)
Y <- as.data.frame(Y_boot)
R2_mY_measured <- 0
ind_mY <- integer()
Betapred_mY = NULL
}

##setting if necessary linear relationship from Y to M
if (n_yM != 0) {
  ##random sampling of variables of M affected by Y
  if (n_mY == 0) {
    ind_yM_M <- sample(ncol(M1), n_yM)
  } else{
    ind_yM_M <- sample((1:ncol(M1))[ind_mY], n_yM)
  }
  ##generating vector of effect coefficients
  if (length(BetayM) == 1) {
    Betapred_yM_M <- rep(BetayM, n_yM)
  } else{
    if (length(BetayM) != n_yM) {
      stop(
        "error: Betas for Y explaining M are not consistent with
        the number of predicted cpgs"
      )
    }
    Betapred_yM_M <- BetayM
  }
  ##adding a linear effect of Y on selected variables of M
  list_R2_M <- list()
  for (i in 1:n_yM) {
    M1[, ind_yM_M[i]] <- as.numeric(M1[, ind_yM_M[i]] +
      simResponseSimple(
        met = Y,
        Nmet = 1,
        beta = Betapred_yM_M[i],
        cpg = 1
      )$resp)
  }
  list_R2_M <-
    c(list_R2_M, list(estimatedR2(Y, colnames(Y)[1], M1[, ind_yM_M[i]], drop = FALSE)))
}
##empirical estimation of mean R2 (mean variability of M affected by Y)
mean_R2_M <-
  mean(unlist(lapply(list_R2_M, function(x)

```

```

x$r.squared)), na.rm = T)
SD_R2_M <-
  sd(unlist(lapply(list_R2_M, function(x)
    x$r.squared)), na.rm = T)
} else{
  ind_yM_M <- integer()
  Betapred_yM_M = NULL
  list_R2_M = NULL
  mean_R2_M = 0
}
#setting if necessary linear relationship from Y to E
list_R2_E <- list()
if (n_yE != 0) {
  ##random sampling of variables of e affected by Y
  ind_yE_E <- sample(ncol(dataExp), n_yE)

  ##generating vector of effect coefficients
  if (length(BetayE) == 1) {
    Betapred_yE_E <- rep(BetayE, n_yE)
  } else{
    if (length(BetayE) != n_yE) {
      stop(
        "error: Betas for Y explaining E are not consistent with the
        number of predicted exposures"
      )
    }
    Betapred_yE_E <- BetayE
  }
  ##adding a linear effect of Y on selected variables of e
  for (i in 1:n_yE) {
    dataExp[, ind_yE_E] <- as.numeric(dataExp[, ind_yE_E[i]] +
      simResponseSimple(
        met = Y,
        Nmet = 1,
        beta = Betapred_yE_E[i],
        cpg = 1
      )$resp)
  }
  list_R2_E <-
    c(list_R2_E, list(estimatedR2(Y, colnames(Y)[1], dataExp[, ind_yE_E[i], drop = FALSE])))
}
##empirical estimation of mean R2 (mean variability of e affected by Y)
mean_R2_E <-
  mean(unlist(lapply(list_R2_E, function(x)
    x$r.squared)), na.rm = T)
SD_R2_E <-
  sd(unlist(lapply(list_R2_E, function(x)

```

```

  x$r.squared)), na.rm = T)
} else{
  ind_yE_E <- integer()
  Betapred_yE_E = NULL
  list_R2_E = NULL
  mean_R2_E = 0
}

##Building a result object with generated datesets;
##vector of predictors and effects
results <-

list(
  Y_train = Y[1:(nrow(M1) / 2), , drop = FALSE],
  E_train = dataExp[1:(nrow(M1) / 2), , drop = FALSE],
  M_train = M1[1:(nrow(M1) / 2), , drop = FALSE],
  Y_test = Y[(nrow(M1) / 2):nrow(M1), , drop = FALSE],
  E_test = dataExp[(nrow(M1) / 2):nrow(M1), , drop = FALSE],
  M_test = M1[(nrow(M1) / 2):nrow(M1), , drop = FALSE],
  cpg_predictors = list(
    name = colnames(M1)[ind_mY],
    indices = ind_mY,
    betas = Betapred_mY,
    R2 = NULL
  ),
  cpg_predicted = list(
    name = colnames(M1)[ind_yM_M],
    indices = ind_yM_M,
    betas = Betapred_yM_M,
    R2 = list_R2_M
  ),
  exp_predicted = list(
    name = colnames(dataExp)[ind_yE_E],
    indices = ind_yE_E,
    betas = Betapred_yE_E,
    R2 = list_R2_E
  ),
  R2_mY_true = R2_mY,
  R2_mY_measured = R2_mY_measured,
  R2_yM_mean = mean_R2_M,
  R2_yM_SD = SD_R2_M,
  R2_yE_mean = mean_R2_E,
  R2_yE_SD = SD_R2_E
)

return(results)
}

```

```

##function used to create a linear response
simResponseSimple <- function(met,
                                ##matrix of potential predictors
                                Nmet = NA,
                                ##number of predictors
                                beta = NULL,
                                ##vector of effects
                                cpg = NULL) {
  ##optionnal: directly specifying
  ##some of the indexes of predictors
  if (all(c(is.na(Nmet), is.na(cpg))) == TRUE) {
    return (list(
      resp = as.matrix(rep(0, nrow(met)), ncol = 1),
      beta = NA,
      predictors = NA
    ))
  }
  temp <- Nmet - length(cpg)
  if (temp != 0) {
    wh <- sample((1:ncol(met)[-cpg]), temp)
    wh <- c(cpg, wh)
  } else{
    wh <- cpg
  }
  CovMat <- as.matrix(met[, wh])
  colnames(CovMat) <- colnames(met)[wh]
  # computing the response
  mean <- CovMat %*% matrix(beta, ncol = 1)
  rownames(mean) <- rownames(met)
  names(beta) <- colnames(CovMat)
  return (list(
    resp = mean,
    beta = beta,
    predictors = colnames(met)[wh]
  )))
}

##fonction to estimate R2 from a datafram of potential predictors, a vector of
##predictors names and the outcome
estimatedR2 <- function(X, truepred, Y) {
  if ("y" %in% truepred) {
    stop("error: one of the true predictors is named y")
  }
  if (ncol(Y) != 1) {
    stop("error:Y is multidimensionnal")
  }
}

```

```

if (nrow(X) != nrow(Y)) {
  stop("error not the same number of rows")
}
if (isTRUE(all.equal(rownames(X), rownames(Y))) == FALSE) {
  stop("error individuals are not ordered similarly in X and Y")
}
if (all(truepred %in% colnames(X))) {
  data <- X[, colnames(X) %in% truepred, drop = FALSE]
  data <- cbind(Y, data)
  colnames(data)[1] <- "y"
  mod <- lm(y ~ ., as.data.frame(data))
  toselect.x <- summary(mod)$coeff[-1, 4]
  r <-
    list(summary(mod)$r.squared,
         summary(mod)$adj.r.squared,
         names(toselect.x)[toselect.x == TRUE])
  names(r) <- c("r.squared", "adj.r.squared", "pred")
  return(r)
} else{
  stop("error: X does not contain all true predictors")
}

#####
##2. defining the methods to test
#####

###agnostic methods

##function to compute residuals of a linear model if covariates are specified
getresiduals_2df <- function(data_Y_in, data_covar_in, name_Y, covar) {
  data_covar <-
    data_covar_in[, colnames(data_covar_in) %in% covar, drop = FALSE]
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_covar), colnames(data_Y_in) == name_Y, drop = FALSE]
  data_covar <-
    data_covar[rownames(data_covar) %in% rownames(data_Y), , drop = FALSE]
  data_covar <- data_covar[rownames(data_Y), , drop = FALSE]
  data_output <- data_Y
  data <- cbind(data_Y, data_covar)
  mod <- lm(data = data)
  data_output[, 1] <- as.data.frame(residuals(mod))
  return(data_output)
}

###ExWAS

```

```

ewas <-
function(data_Xs_in = NULL,
        data_Y_in = NULL,
        name_Y,
        data_covar_in = NULL,
        covar = character(0),
        corr = "BH") {
  require(parallel)
  if (length(covar) > 0) {
    data_covar <-
      data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
                    rownames(data_covar_in) %in% rownames(data_Xs_in),
                    colnames(data_covar_in) %in% covar, drop = FALSE]
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
                    rownames(data_Y_in) %in% rownames(data_Xs_in),
                    colnames(data_Y_in) == name_Y, drop =
                    FALSE]
    data_Xs <-
      data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
                    rownames(data_Xs_in) %in% rownames(data_Y), , drop = FALSE]
    data_covar <- data_covar[rownames(data_Y), , drop = FALSE]
    data_Xs <- data_Xs[rownames(data_Y), , drop = FALSE]
    data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
  } else{
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in), colnames(data_Y_in) ==
                    name_Y, drop = FALSE]
    data_Xs <-
      data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
    data_Xs <- data_Xs[rownames(data_Y), , drop = FALSE]
  }
  if (is.null(data_Y) == TRUE |
      is.null(data_Xs) == TRUE | !(name_Y %in% colnames(data_Y))) {
    stop("Inconsistent data")
  }

##computing p.values
p.values <- lapply(1:ncol(data_Xs), function(x, data_Xs) {
  c(colnames(data_Xs)[x],
    confint(lm(Y ~ var1,
      data = data.frame(
        cbind(var1 = data_Xs[, x], Y = data_Y[, 1])
      )))[2, ],
    summary(lm(Y ~ var1, data = data.frame(
      cbind(var1 = data_Xs[, x], Y = data_Y[, 1])
    )))$coefficients[2, ])
})

```

```

  },
  data_Xs)
if (length(p.values) > 1) {
  p.values <-
    cbind(matrix(unlist(p.values), ncol = 7, byrow = TRUE)[, -6])

} else{
  p.values <-
    as.data.frame(t(as.data.frame(unlist(p.values)))[, -6, drop = FALSE])
}
p.values <- as.data.frame(p.values)
colnames(p.values) <-
  c("var", "conf - 2.5%", "conf - 97.5%", "Est", "Sd", "pVal")
p.values <- p.values[p.values$var != "Intercept", ]
p.values$pVal <- as.numeric(as.character(p.values$pVal))
p.values.adj <- p.values
pVal <- as.numeric(as.character(p.values$pVal))
##add correction for multiple testing
if (corr == "None") {
  wh <- which(pVal <= 0.05)
  p.values.adj$pVal_adj <- pVal
}
if (corr == "Bon") {
  wh <- which(pVal <= 0.05 / nrow(p.values))
  p.values.adj$pVal_adj <- pVal * nrow(p.values)
}
if (corr == "BH") {
  wh <- which(p.adjust(pVal, "BH") <= 0.05)
  p.values.adj$pVal_adj <- p.adjust(pVal, "BH")
}
if (corr == "BY") {
  wh <- which(p.adjust(pVal, "BY") <= 0.05)
  p.values.adj$pVal_adj <- p.adjust(pVal, "BY")
}
if (!corr %in% c("Bon", "BH", "BY", "", "None"))
  stop("Please specify a known correction method for
       multiple testing")
wh <- p.values$var[wh]
a <- list(wh, p.values.adj)
names(a) <- c("selected", "pval")
return(a)
}

####LASSO
lasso <-
function(data_Xs_in,
        data_Y_in,
        name_Y,

```

```

data_covar_in = NULL,
covar = character(0)) {
if (length(covar) > 0) {
  data_covar <-
    data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
      rownames(data_covar_in) %in% rownames(data_Xs_in),
      colnames(data_covar_in) %in%
        covar, drop = FALSE]
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
      rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop =
        FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
      rownames(data_Xs_in) %in% rownames(data_Y_in), ,
      drop = FALSE]
  data_covar <- data_covar[rownames(data_Y),]
  data_Xs <- data_Xs[rownames(data_Y),]
  data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) ==
        name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y),]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
model.enet <- cv.glmnet(data_Xs, data_Y, family = "gaussian",
  alpha = 1)
cvfit <- model.enet

##Calcul Y_predit
Y_predit <- predict(cvfit, newx = data_Xs, s = "lambda.min")
Y_predit <- Y_predit[rownames(Y_predit),]

##liste des CPG selectionnés
tmp_coeffs <- coef(cvfit, s = "lambda.min")
cg_select <-
  data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1], coefficient =
tmp_coeffs@x)

cg_select <- cg_select$name[cg_select$name != "(Intercept)"]
a <- list()
if (length(cg_select) != 0) {

```

```

a <- list("selected" = cg_select, "prediction" = Y_predit)
} else{
  a <- list("selected" = character(), "prediction" = "no_prediction")
}
return(a)

}

##DSA
DSAreg <-
  function(Exp,
    resp,
    family = gaussian,
    maxsize = 15,
    maxsumofpow = 2,
    maxorderint = 2) {
  Exp <- data.frame(cbind(data.frame(Exp), resp = resp))
  res <-
    DSA(
      resp ~ 1,
      data = Exp,
      family = family,
      maxsize = maxsize,
      maxsumofpow
      = maxsumofpow,
      maxorderint = maxorderint ,
      nsplits = 1,
      usersplits = NULL
    )
  form <- gsub("I[()", "", colnames(coefficients(res)))
  form <- gsub("[*]", ":", gsub("]", "", gsub("[:^]1", "", form)))
  if (length(grep(":", form)) > 0) {
    nam <- strsplit(form[grep(":", form)], ":")
    for (j in 1:length(nam)) {
      nam[[j]] <- gsub("[[:space:]]", "", nam[[j]])
      name <- nam[[j]][1]
      for (k in 2:length(nam[[j]]))
        name <- paste(name, ":", nam[[j]][k], sep = "")
      Exp <- cbind(Exp, name = apply(Exp[, nam[[j]]], 1, prod))
    }
  }
  form2 <- "resp~1"
  if (length(form) > 1)
    for(i in 2:length(form))
      form2 <- paste(form2, "+", form[i])
  res2 <- lm(form2, data = data.frame(Exp))
  ##decomment next line and change "prediction" to pred in the return line
}

```

```

##if outcome predicted by DSA is needed (not used presently)
#pred <- predict(res2,Exp)
coef <- summary(res2)$coefficients
coef <- as.character(rownames(coef)[rownames(coef) != "Intercept"])

return(list(selected = coef[coef != "(Intercept)"], pred = "prediction"))
}

#####
####3. defining some functions used to assess methods performance
#####

sensitivity <- function(truepred, predfound) {
  return(length(truepred[truepred %in% predfound]) / length(truepred))
}

fdp <- function(truepred, predfound) {
  ##false discovery proportion
  if (length(predfound) == 0) {
    return(0)
  } else{
    return(length(predfound[!predfound %in% truepred]) / length(predfound))
  }
}

specificity <- function(truepred, predfound, n_base) {
  return((n_base - length(truepred) - length(predfound[!predfound %in% truepred])) /
    (n_base - length(truepred)))
}

####

##4. defining the simulation function which will be parallelized
#####

##it first generates datasets, then applies methods and then assessed
##their performance
##simulation d'un jeu, application des méthodes, évaluation des méthodes
f0 <- function(x) {
  ##important: the parallelization is made on the seed
  set.seed(x)
  ##generating datasets
  simu <- simulator(
    E_true = dataExp_true,
    M_true = M1_true,
    Y_true = Y_true,

```

```

n_mY = n_mY,
R2_mY = R2_mY,
BetamY = BetamY,
n_yM = n_yM,
Beta_yM = Beta_yM,
n_yE = n_yE,
Beta_yE = Beta_yE,
test_and_training = TRUE
)

#simulated_data<-c(simulated_data,list(simu))

#####
##### applying methods
#####
##ExWas on the intermediate layer
predBMI_M <-
  list(ewas_BH = ewas(
    as.data.frame(simu$M_train),
    as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    corr = "BH"
  ))
predBMI_E <-
  list(
    ewas_BH = ewas(
      as.data.frame(simu$E_train),
      as.data.frame(simu$Y_train),
      colnames(as.data.frame(simu$Y_train)),
      corr = "BH"
    ),
    ewas_Bon = ewas(
      as.data.frame(simu$E_train),
      as.data.frame(simu$Y_train),
      colnames(as.data.frame(simu$Y_train)),
      corr = "Bon"
    )
  )
print("ewas done")

#####
#oMITM
if (length(predBMI_M$ewas_BH$selected) != 0) {
  select_M <-
    as.data.frame(simu$M_train[, colnames(simu$M_train) %in% predBMI_M$ewas_BH$selected,
      drop =
        FALSE])
}

```

```

print(ncol(select_M))
rownames(select_M) <- rownames(simu$M_train)
colnames(select_M) <- predBMI_M$ewas_BH$selected
list <- list()
list_exp <- list()
list_nom <- list()
list_ewas_signif <- list()
##step b using the exwas performed on M as step a
for (i in (1:ncol(simu$E_train))) {
  predE_select_M <-
    ewas(
      as.data.frame(select_M),
      (simu$E_train[, i, drop = FALSE]),
      colnames(simu$E_train[, i, drop = FALSE]),
      corr = "None",
      data_covar_in = as.data.frame(simu$Y_train),
      covar = colnames(as.data.frame(simu$Y_train)[1])
    )
  list <- c(list, list(predE_select_M))
  list_nom <- c(list_nom, list(colnames(simu$E_train)[i]))
}

list_exp <- c(list_exp, list(colnames(simu$E_train)[i]))
temp_ewas <-
  cbind(predE_select_M$pval, rep(colnames(simu$E_train)[i], nrow(predE_select_M$pval)))
list_ewas_signif <- c(list_ewas_signif, list(temp_ewas))
remove(temp_ewas)
remove(predE_select_M)
}
df_all_ewas <- do.call("rbind", list_ewas_signif)
if (!is.null(df_all_ewas)) {
  df_all_ewas$pVal_adj <- p.adjust(df_all_ewas$pVal, "BH")
  colnames(df_all_ewas)[8] <- "exposures"
  names(list) <- as.vector(unlist(list_nom))
}
exp <- df_all_ewas$exposures[df_all_ewas$pVal_adj <= 0.05]
##step c
n_exp_select <-
  length(unique(exp)) ##nb of exposures in reduced exposome

if (length(exp) != 0) {
  select_E <- simu$E_train[, colnames(simu$E_train) %in% exp, drop = FALSE]
  ##ExWAS implementation for step c
  predBMI_E_MITM <-
    ewas(
      as.data.frame(select_E),
      as.data.frame(simu$Y_train),
      colnames(as.data.frame(simu$Y_train)),
      corr = "BH"
    )
}

```

```

)
##DSA implementation for step c
predBMI_E_MITMds <-
  DSAreg(
    Exp = as.data.frame(select_E),
    resp = simu$Y_train,
    maxsize = floor(ncol(simu$E_train) / 10),
    maxsumofpow = 1,
    maxorderint = 1
  )
  predReducedExp <- list(selected = unique(exp), pred = "NULL")
} else{
  predReducedExp <- list(vector(), vector())
  names(predReducedExp) <- c("selected", "pred")
}
if (exists("predBMI_E_MITM")) {

} else{
  predBMI_E_MITM <- list(vector(), vector())
  names(predBMI_E_MITM) <- c("selected", "pval")
}
if (exists("predBMI_E_MITMds")) {

} else{
  predBMI_E_MITMds <- list(vector(), vector())
  names(predBMI_E_MITMds) <- c("selected", "pred")
}
} else{
  predBMI_E_MITM <- list(vector(), vector())
  names(predBMI_E_MITM) <- c("selected", "pval")
  n_exp_select = 0
  predReducedExp <- list(vector(), vector())
  names(predReducedExp) <- c("selected", "pred")
  predBMI_E_MITMds <- list(vector(), vector())
  names(predBMI_E_MITMds) <- c("selected", "pred")
}

}
##storing results in a list
predBMI_E <-
  c(
    predBMI_E,
    MITM = list(predBMI_E_MITM),
    MITMds = list(predBMI_E_MITMds),
    ReducedExp = list(predReducedExp)
  )
print("oMITM")
####Control method : random sampling on a random set of exposures of same
##dimension as the reduced exposome of oMITM

```

```

if (n_exp_select > 0) {
  tirage <-
  ewas(
    as.data.frame(simu$E_train)[, sample(colnames(as.data.frame(simu$E_train)), n_exp_select),
  drop =
    FALSE],
    as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    corr = "BH"
  )
} else{
  tirage <- list(selected = character(0), null = "null")
}
##storing results in the same list
predBMI_E <- c(predBMI_E, random_sampling = list(tirage))
print(n_exp_select)

#####
##mediation
if (length(predBMI_E$ewas_BH$selected) != 0) {
  select_E <-
  as.data.frame(simu$E_train[, colnames(simu$E_train) %in% predBMI_E$ewas_BH$selected,
  drop =
    FALSE])
  rownames(select_E) <- rownames(simu$E_train)
  colnames(select_E) <- predBMI_E$ewas_BH$selected
  #step a
  list_temp_ewas_med <- list()
  for (i in 1:ncol(simu$M_train)) {
    exp_affecting_M_all <-
    ewas(as.data.frame(select_E),
      (simu$M_train[, i, drop = FALSE]),
      colnames(simu$M_train[, i, drop = FALSE]),
      corr = "None")
    temp_ewas_med <-
    cbind(exp_affecting_M_all$pval, rep(
      colnames(simu$M_train)[i],
      nrow(exp_affecting_M_all$pval)
    ))
    list_temp_ewas_med <- c(list_temp_ewas_med, list(temp_ewas_med))
  }
  ewas_med <- do.call("rbind", list_temp_ewas_med)
  if (!is.null(ewas_med)) {
    ewas_med$pVal_adj_1 <- p.adjust(ewas_med$pVal, "BH")
    colnames(ewas_med)[8] <- "cpg"
    colnames(ewas_med)[1] <- "exp"
  }
}

```

```

#step b
list_temp_ewas_med_2 <- list()
for (i in 1:ncol(select_E)) {
  M_affecting_Y_all <-
  ewas(
    as.data.frame(simu$M_train),
    as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    corr = "None",
    data_covar_in = (select_E[, i, drop = FALSE]),
    covar = colnames(select_E[, i, drop = FALSE])
  )
  temp_ewas_med_2 <-
  cbind(M_affecting_Y_all$pval, rep(colnames(select_E)[i], nrow(M_affecting_Y_all$pval)))
  list_temp_ewas_med_2 <-
  c(list_temp_ewas_med_2, list(temp_ewas_med_2))
}
ewas_med_2 <- do.call("rbind", list_temp_ewas_med_2)
if (!is.null(ewas_med_2)) {
  ewas_med_2$pVal_adj_2 <- p.adjust(ewas_med_2$pVal, "BH")
  colnames(ewas_med_2)[8] <- "exp"
  colnames(ewas_med_2)[1] <- "cpg"
}

ewas_med_tot <-
merge(
  ewas_med,
  ewas_med_2,
  by.x = c("exp", "cpg"),
  by.y = c("exp", "cpg")
)
exp_med <-
ewas_med_tot$exp[ewas_med_tot$pVal_adj_2 <= 0.05 &
  ewas_med_tot$pVal_adj_1 <= 0.05]
exp_med <- exp_med[!is.na(exp_med)]
if (length(exp_med) != 0) {
  predMediation <- list(selected = exp_med, pred = NULL)
} else{
  predMediation <- list(selected = vector(), pred = vector())
}

} else{
  predMediation <- list(selected = vector(), pred = vector())
}
predBMI_E <- c(predBMI_E, mediation = list(predMediation))

#####
##applying agnostic methods

```

```

## lasso
predlasso <-
  lasso(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)))
)
predBMI_E <- c(predBMI_E, lasso_CV = list(predlasso))
print("lasso done")
##DSA
predDSA <-
  DSAreg(
    Exp = simu$E_train,
    resp = as.data.frame(simu$Y_train),
    maxsize = floor(ncol(simu$E_train) / 10),
    maxsumofpow = 1,
    maxorderint = 1
)
predBMI_E <- c(predBMI_E, DSA = list(predDSA))
print("DSA done")

#####
##assessing performance linked to reverse causality between M and Y
##(sensitivity to predicted exposures) for all methods
truepred <- simu$cpg_predicted$name
for (k1 in 1:length(predBMI_M)) {
  truepred <- as.character(simu$cpg_predicted$name)
  predfound <- as.character(predBMI_M[[k1]]$selected)
  if (exists("predfound") & exists("truepred")) {
    if (length(predfound) == 0) {
      print("no predictors found")
    }
    a <- sensitivity(truepred, predfound)
    b <- specificity(truepred, predfound, ncol(simu$M_train))
    c <- fdp(truepred, predfound)
    d <- estimatedR2(simu$M_test, predfound, simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)
    remove(predfound, truepred)
  } else{
    if (!exists("predfound") & exists("truepred")) {
      a <- 0
      b <- specificity(truepred, numeric(0), ncol(simu$M_train))
      c <- 0
      d <- 0
      remove(truepred)
    }
  }
}

```

```

} else{
  if (!exists("truepred") & exists("predfound")) {
    a <- 1
    b <- specificity(numeric(0), predfound, ncol(simu$M_train))
    c <- fdp(numeric(0), predfound)
    d <-
      estimatedR2(simu$M_test, predfound, simu$Y_test)$r.squared
    remove(predfound)
  } else{
    if (!exists("truepred") & !exists("predfound")) {
      a <- 1
      b <- 1
      c <- 0
      d <- 0
    }
  }
}
predBMI_M[[k1]] <-
c(
  predBMI_M[[k1]],
  sens_rev_caus = a,
  spec_rev_caus = b,
  fdp_rev_caus = c,
  R2_test_rev_caus = d
)
remove(a)
remove(b)
remove(c)
remove(d)

}

#####
##assessing performance for ExWAS on M
truepred <- simu$cpg_predictors$name
for (k1 in 1:length(predBMI_M)) {
  truepred <- as.character(simu$cpg_predictors$name)
  predfound <- as.character(predBMI_M[[k1]]$selected)
  if (exists("predfound") & exists("truepred")) {
    if (length(predfound) == 0) {
      print("no predictors found")
    }
    a <- sensitivity(truepred, predfound)
    b <- specificity(truepred, predfound, ncol(simu$M_train))
    c <- fdp(truepred, predfound)
    d <- estimatedR2(simu$M_test, predfound, simu$Y_test)$r.squared
    # print(a)
  }
}

```

```

# print(b)
# print(c)
# print(d)
remove(predfound, truepred)
} else{
  if (!exists("predfound") & exists("truepred")) {
    a <- 0
    b <- specificity(truepred, numeric(0), ncol(simu$M_train))
    c <- 0
    d <- 0
    remove(truepred)
  } else{
    if (!exists("truepred") & exists("predfound")) {
      a <- 1
      b <- specificity(numeric(0), predfound, ncol(simu$M_train))
      c <- fdp(numeric(0), predfound)
      d <-
        estimatedR2(simu$M_test, predfound, simu$Y_test)$r.squared
      remove(predfound)
    } else{
      if (!exists("truepred") & !exists("predfound")) {
        a <- 1
        b <- 1
        c <- 0
        d <- 0
      }
    }
  }
}
predBMI_M[[k1]] <-
c(
  predBMI_M[[k1]],
  sens = a,
  spec = b,
  fdp = c,
  R2_test = d
)
remove(a)
remove(b)
remove(c)
remove(d)

}

#####
##assessing performance linked to reverse causality between E and Y
##( sensitivity to predicted exposures) for all methods
truepred <- simu$exp_predicted$name

```

```

for (k1 in 1:length(predBMI_E)) {
  truepred <- as.character(simu$exp_predicted$name)
  predfound <- as.character(predBMI_E[[k1]]$selected)
  if (exists("predfound") & exists("truepred")) {
    if (length(predfound) == 0) {
      print("no predictors found")
    }
    a <- sensitivity(truepred, predfound)
    b <- specificity(truepred, predfound, ncol(simu$E_train))
    c <- fdp(truepred, predfound)
    d <- estimatedR2(simu$E_test, predfound, simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)
    remove(predfound, truepred)
  } else{
    if (!exists("predfound") & exists("truepred")) {
      a <- 0
      b <- specificity(truepred, numeric(0), ncol(simu$E_train))
      c <- 0
      d <- 0
      remove(truepred)
    } else{
      if (!exists("truepred") & exists("predfound")) {
        a <- 1
        b <- specificity(numeric(0), predfound, ncol(simu$E_train))
        c <- fdp(numeric(0), predfound)
        d <-
          estimatedR2(simu$E_test, predfound, simu$Y_test)$r.squared
        remove(predfound)
      } else{
        if (!exists("truepred") & !exists("predfound")) {
          a <- 1
          b <- 1
          c <- 0
          d <- 0
        }
      }
    }
  }
  predBMI_E[[k1]] <-
  c(
    predBMI_E[[k1]],
    sens_rev_caus = a,
    spec_rev_caus = b,
    fdp_rev_caus = c,
    R2_test_rev_caus = d
  )
}

```

```

)
remove(a)
remove(b)
remove
remove(d)
}

#####
##assessing performance between E and Y (false detection)
truepred <- character()
for (k1 in 1:length(predBMI_E)) {
  truepred <- character()
  predfound <- as.character(predBMI_E[[k1]]$selected)
  if (exists("predfound") & exists("truepred")) {
    if (length(predfound) == 0) {
      print("no predictors found")
    }
    a <- sensitivity(truepred, predfound)
    b <- specificity(truepred, predfound, ncol(simu$E_train))
    c <- fdp(truepred, predfound)
    d <- estimatedR2(simu$E_test, predfound, simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)
    remove(predfound, truepred)
  } else{
    if (!exists("predfound") & exists("truepred")) {
      a <- 0
      b <- specificity(truepred, numeric(0), ncol(simu$E_train))
      c <- 0
      d <- 0
      remove(truepred)
    } else{
      if (!exists("truepred") & exists("predfound")) {
        a <- 1
        b <- specificity(numeric(0), predfound, ncol(simu$E_train))
        c <- fdp(numeric(0), predfound)
        d <-
        estimatedR2(simu$E_test, predfound, simu$Y_test)$r.squared
        remove(predfound)
      } else{
        if (!exists("truepred") & !exists("predfound")) {
          a <- 1
          b <- 1
          c <- 0
          d <- 0
        }
      }
    }
  }
}

```

```

        }
    }
}

predBMI_E[[k1]] <-
  c(
    predBMI_E[[k1]],
    sens = a,
    spec = b,
    fdp = c,
    R2_test = d
  )
remove(a)
remove(b)
remove(c)
remove(d)

}

print("performance characterized")

##building the list with datasets generated + results of methods +
##performance to return
A <-
  list(
    simu = simu,
    predBMI_E = predBMI_E,
    predBMI_M = predBMI_M,
    nl_exp_select = n_exp_select
  )
remove(simu)
remove(select_M)
remove(predBMI_M)
remove(predBMI_E)
remove(predBMI_E_MITM_WM)
remove(predBMI_E_MITM)
remove(predlasso)
remove(n_exp_select)

gc()
return(A)
}

#####
##5. Running simulations
#####

```

```

##loading real datasets

dataExp_true <- readRDS("20190205 Exposome simu borne.rds")
M1_true <- readRDS("20191129 Methylome simu.Rds")
Y_true <- readRDS("20190612_ZBMI_scaled_a_utiliser_pour_simu.rds")
M1_true <- scale(M1_true)

##initialization
list_simulated_data <- list()
list_list_predBMI_E <- list()
list_list_predBMI_M <- list()
list_list_nl_exp_select <- list()

##setting simulations parameters
n_iter <- 100 ##number of iterations for one scenarios
##parameters for generating datasets
##all combinations will be tested (each combination allows to build a scenario)
##(adapt the code of the loop if multiple values instead of single values for
##some parameters)
c_n_yM <- c(10, 18, 25, 100)

c_BetayM <- c(0.0001, 0.001, 0.01, 0.1, 0.5, 2)
c_BetayE <- c(0.0001, 0.001, 0.01, 0.1, 0.5, 2)
c_n_yE <- c(1, 3, 10, 25)
n_mY = 0
R2_mY = 0
BetamY = 0

##initialization of table of results

comp_method <-
  data.frame(
    Methods = vector(),
    Association_tested = vector(),
    Nb_true_predictors_of_BMI_in_M = numeric(0),
    Nb_predicted_by_BMI_in_M = numeric(0),
    Nb_predicted_by_BMI_in_E = numeric(0),
    Total_variability_of_BMI_explained_by_M = numeric(0),
    Total_variability_of_BMI_explained_by_M_measured =
      numeric(0),
    Mean_variability_of_M_explained_by_Y = numeric(0),
    Mean_variability_of_E_explained_by_Y = numeric(0),
    Mean_SD_variability_of_M_explained_by_Y = numeric(0),
    Mean_SD_variability_of_E_explained_by_Y = numeric(0),
    Number_iterations = numeric(0),

```

```

Mean_number_exp_selected_to_be_randomly_tested =
  numeric(0),
Mean_number_predictors_found = numeric(0),
Mean_sensitivity_rv = numeric(0),
Mean_specificity_rv = numeric(0),
Mean_fdp_rv = numeric(0),
Mean_R2_test_rv = numeric(0),
Mean_sensitivity_truepred = numeric(0),
Mean_specificity_truepred = numeric(0),
Mean_fdp_truepred = numeric(0),
Mean_R2_test_truepred = numeric(0),
SD_number_predictors_found = numeric(0),
SD_sensitivity_rv = numeric(0),
SD_specificity_rv = numeric(0),
SD_fdp_rv = numeric(0),
SD_R2_test_rv = numeric(0),
SD_sensitivity_truepred = numeric(0),
SD_specificity_truepred = numeric(0),
SD_fdp_truepred = numeric(0),
SD_R2_test_truepred = numeric(0),
Which_iteration = numeric(0)
)

```

$n = 1$

```

##looping on the different vectors of parameters
for (i2 in 1:length(c_BetayE)) {
  BetayE <- c_BetayE[i2]
  for (i3 in 1:length(c_n_yE)) {
    n_yE <- c_n_yE[i3]
    for (i4 in 1:length(c_BetayM)) {
      BetayM <- c_BetayM[i4]
      for (i1 in 1:length(c_n_yM)) {
        n_yM <- c_n_yM[i1]

        n_row = nrow(comp_method)
        simulated_data <- list()
        list_predBMI_E <- list()
        list_predBMI_M <- list()
        list_nl_exp_select <- list()
        start_time <- Sys.time()
        ##parallelization of f0
        cl <-
          makeClustergetOption("cl.cores", round(detectCores())))
        clusterExport(

```

```

cl,
list(
  "simulator",
  "simResponseSimple",
  "estimatedR2",
  "getresiduals_2df",
  "ewas",
  "lasso",
  "lasso_stab",
  "DSAreg",
  "sensitivity",
  "fdp",
  "specificity",
  "f0",
  "dataExp_true",
  "M1_true",
  "Y_true",
  "BetayE",
  "n_yE",
  "n_yM",
  "BetayM",
  "n_mY",
  "R2_mY",
  "BetamY"
)
)
)
clusterEvalQ(cl, list(library("boot"), library("reshape"),
  library("glmnet"), library("DSA")))
results_1_jeu <- clusterApply(cl, 1:n_iter, f0)
stopCluster(cl)
simulated_data <-
  lapply(results_1_jeu, function(x)
    x$simu)

##structure of results prioritized by methods and not anymore
##prioritized by datasets
list_predBMI_E <-
  lapply(results_1_jeu, function(x)
    x$predBMI_E)
list_predBMI_M <-
  lapply(results_1_jeu, function(x)
    x$predBMI_M)
list_nl_exp_select <-
  lapply(results_1_jeu, function(x)
    x$nl_exp_select)
remove(results_1_jeu)

```

```

####compilation of results for this scenario

##table describing the empirical characteristics of
##the simulated datasets
param_simu <-
  data.frame(
    Parameters = vector(),
    Fixed_or_measured = vector(),
    Value = numeric(0)
  )
param_simu[1, ] <-
  c(
    as.character("Nb_true predictors of BMI in M"),
    as.character("Fixed"),
    mean(unlist(
      lapply(simulated_data, function(X)
        length(X$cpg_predictors$betas)))
    )))
param_simu[2, ] <-
  c("Nb_predicted_by_BMI_in_M", "Fixed", mean(unlist(
    lapply(simulated_data, function(X)
      length(X$cpg_predicted$betas)))
  )))
param_simu[3, ] <-
  c("Nb_predicted_by_BMI_in_E", "Fixed", mean(unlist(
    lapply(simulated_data, function(X)
      length(X$exp_predicted$betas)))
  )))
param_simu[4, ] <-
  c("Total_variability_of_BMI_explained_by_M",
    "Fixed",
    R2_mY)
param_simu[5, ] <-
  c("Total_variability_of_BMI_explained_by_M",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X$R2_mY_measured)))
    )))
param_simu[6, ] <-
  c("Mean_variability_of_M_explained_by_Y",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X$R2_yM_mean)))
    )))

```

```

)))
param_simu[7, ] <-
  c("Mean_variability_of_E_explained_by_Y",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X$R2_yE_mean)))
    )))
param_simu[8, ] <-
  c("Mean_SD_variability_of_M_explained_by_Y",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X$R2_yM_SD)))
    )))
param_simu[9, ] <-
  c("Mean_SD_variability_of_E_explained_by_Y",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X$R2_yE_SD)))
    )))
param_simu[10, ] <- c("Number_iterations", "Fixed", n_iter)
param_simu[11, ] <-
  c(
    "Mean_number_exp_selected_to_be_randomly_tested",
    "Measured",
    mean(unlist(list_nl_exp_select)))
  )

##summarizing each method performance by a line in comp_method datadrame
##within this scenario
for (k1 in (1:length(list_predBMI_M[[1]]))) {
  comp_method[n_row + k1, ] <-
    c(
      names(list_predBMI_M[[1]][k1]),
      "BMI - M",
      param_simu[1, 3],
      param_simu[2, 3],
      param_simu[3, 3],
      param_simu[4, 3],
      param_simu[5, 3],
      param_simu[6, 3],
      param_simu[7, 3],
      param_simu[8, 3],
      param_simu[9, 3],
      n_iter,
      NA,

```



```

(X[[k1]][[4]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[5]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[6]]))
), na.rm = TRUE),

sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[7]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[8]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[9]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_predBMI_M, function(X)
    (X[[k1]][[10]]))
), na.rm = TRUE),
n
)
}

for (k2 in (1:length(list_predBMI_E[[1]]))) {
  comp_method[n_row + k2 + k1, ] <-
  c(
    names(list_predBMI_E[[1]][k2]),
    "BMI - E",
    param_simu[1, 3],
    param_simu[2, 3],
    param_simu[3, 3],
    param_simu[4, 3],
    param_simu[5, 3],
    param_simu[6, 3],
    param_simu[7, 3],
    param_simu[8, 3],
    param_simu[9, 3],
    n_iter,
    param_simu[11, 3],
    mean(unlist(

```

```

lapply(list_predBMI_E, function(X)
  length(X[[k2]][[1]]))
)),
mean(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[3]])))
)),
mean(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[4]])))
)),
mean(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[5]])))
)),
mean(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[6]])))
)),
NA,
NA,
NA,
NA,

```

```

sd(unlist(
  lapply(list_predBMI_E, function(X)
    length(X[[k2]][[1]])))
)),
sd(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[3]])))
)),
sd(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[4]])))
)),
sd(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[5]])))
)),
sd(unlist(
  lapply(list_predBMI_E, function(X)
    (X[[k2]][[6]])))
)),
NA,
NA,

```

```

NA,
NA,
n
)
}
print(n)

##storing generated datasets and methods results for this scenario
list_list_predBMI_E <-
  c(list_list_predBMI_E, list(list_predBMI_E))
list_list_predBMI_M <-
  c(list_list_predBMI_M, list(list_predBMI_M))
list_list_nl_exp_select <-
  c(list_list_nl_exp_select, list(list_nl_exp_select))
end_time <- Sys.time()
end_time - start_time
##saving
saveRDS(comp_method, "comp_method_rev_caus_sans_mY.Rds")
saveRDS(
  simulated_data,
  file = paste(
    n,
    '_simulated_data_scenario_iteration_rev_caus_sans_mY.Rds'
  )
)
saveRDS(list_list_predBMI_E,
        "list_list_predBMI_E_rev_caus_sans_mY.rds")
saveRDS(list_list_predBMI_M,
        "list_list_predBMI_M_rev_caus_sans_mY.rds")
saveRDS(list_list_nl_exp_select,
        "list_list_nl_exp_select_rev_caus_sans_mY.rds")

remove(simulated_data)
remove(list_predBMI_E)
remove(list_predBMI_M)
remove(list_nl_exp_select)

n = n + 1

print(n)
}
}
}
}

##save
saveRDS(comp_method, "comp_method_rev_caus_sans_mY.Rds")

```

```
#saveRDS(list_simulated_data,"list_simulated_data_mediation.Rds")
saveRDS(list_list_predBMI_E,
        "list_list_predBMI_E_rev_caus_sans_mY.Rds")
saveRDS(list_list_predBMI_M,
        "list_list_predBMI_M_rev_caus_sans_mY.Rds")
saveRDS(list_list_nl_exp_select,
        "list_list_nl_exp_select_rev_caus_sans_mY.Rds")
```

Supplementary Table V.3: List of variables selected by default LASSO and all tested stabilized LASSO for each of the 10 runs applied to relate an exposome of 173 prenatal and postnatal quantitative exposures (A) or only the smaller exposome of 74 prenatal quantitative variables (B), to zBMI in 1301 mother-child pairs of the Helix cohorts. For each run, the direction of association with zBMI in a multivariate model including all exposures selected in the run and adjusted for relevant covariates is also given.

A.

Method	Run	Exposures selected and corresponding direction of association
Default LASSO	1	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	2	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), Lead - Postnatal (+), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-)

Default LASSO	3	Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-) Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	4	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	5	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE -

		Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (-), Lead - Postnatal (+), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	6	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	7	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+),

		Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	8	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	9	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), PM2.5 (preg) - Pregnancy (-), Pressure (t1) - Pregnancy (-), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cotinine - Pregnancy (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Mercury - Postnatal (+), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (-), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), MiBP - Postnatal (-), Manganese - Postnatal (+), MnBP - Pregnancy (-), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (+), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 -

Default LASSO	10	Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), TRCS - Postnatal (-), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	11	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (-), Lead - Postnatal (+), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)

Default LASSO	12	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), PM2.5 (preg) - Pregnancy (-), Pressure (t1) - Pregnancy (-), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cotinine - Pregnancy (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Mercury - Postnatal (+), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (-), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), MiBP - Postnatal (-), Manganese - Postnatal (+), MnBP - Pregnancy (-), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (+), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), TRCS - Postnatal (-), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	13	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	14	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), PM2.5 (preg) - Pregnancy (-), Pressure (t1) - Pregnancy (-), Temperature (preg) - Pregnancy (+), Road traffic load (100 m) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt -

		Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cotinine - Pregnancy (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Mercury - Postnatal (+), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (-), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), MiBP - Postnatal (-), Manganese - Postnatal (+), MnBP - Pregnancy (-), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (-), oxo-MiNP - Pregnancy (-), Lead - Postnatal (+), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 118 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), PM2.5 (week) - Postnatal (-), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), TRCS - Postnatal (-), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Default LASSO	15	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	1	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP -

		Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	2	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	3	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+),

		Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	4	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	5	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+),

		Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	6	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	7	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)

CV ₁	8	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	9	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), MiBP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	10	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP -

		Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	11	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), MiBP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	12	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+),

		Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	13	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	14	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+),

		Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₁	15	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	1	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)

CV ₂	2	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	3	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	4	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP -

		Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	5	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	6	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+),

		Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	7	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	8	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+),

		Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	9	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	10	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)

CV ₂	11	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	12	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO ₂ (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	13	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP -

		Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	14	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+), Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
CV ₂	15	Cigarette - Pregnancy (-), Indoor PMabsorbance - Postnatal (+), Indoor PM2.5 - Postnatal (+), Temperature (preg) - Pregnancy (+), BPA - Pregnancy (+), BUPA - Pregnancy (+), Cobalt - Postnatal (-), Connectivity density (300m - school) - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DDE - Pregnancy (+), DEP - Postnatal (-), DEP - Pregnancy (+), Sleep duration - Postnatal (-), DMP - Postnatal (-), DMP - Pregnancy (-), HCB - Postnatal (-), Humidity (day) - Postnatal (+), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m) - Postnatal (+), Land use (300m - school) - Postnatal (-), Traffic noise (24h - school) - Postnatal (-), MEHP - Postnatal (-), MEOHP - Pregnancy (-), MEP - Postnatal (-), MEPA - Pregnancy (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), Molybdenum - Pregnancy (+),

		Moderate and vigorous PA - Postnatal (-), NO2 (year) - Postnatal (+), OH-MiNP - Pregnancy (-), OXBE - Postnatal (+), OXBE - Pregnancy (+), oxo-MiNP - Pregnancy (-), Lead - Postnatal (-), PBDE 153 - Postnatal (-), PBDE 47 - Postnatal (+), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFNA - Pregnancy (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density (school) - Postnatal (+), PRPA - Pregnancy (-), PCBs (sum) - Pregnancy (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+), Walkability index - Postnatal (-), Facility density (300m) - Postnatal (-), Accessibility (bus stops 300m) - Postnatal (-)
Meinshau	1	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	2	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	3	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	4	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	5	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	6	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	7	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	8	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	9	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	10	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	11	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	12	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	13	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	14	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	15	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-)
sen ₁		
Meinshau	1	HCB - Postnatal (-)
sen ₂		
Meinshau	2	HCB - Postnatal (-)
sen ₂		
Meinshau	3	
sen ₂		
Meinshau	4	HCB - Postnatal (-)
sen ₂		
Meinshau	5	HCB - Postnatal (-)
sen ₂		

Meinshau	6	
sen ₂		
Meinshau	7	DDE - Postnatal (-), HCB - Postnatal (-)
sen ₂		
Meinshau	8	DDE - Postnatal (-), HCB - Postnatal (-)
sen ₂		
Meinshau	9	
sen ₂		
Meinshau	10	HCB - Postnatal (-)
sen ₂		
Meinshau	11	DDE - Postnatal (-), HCB - Postnatal (-)
sen ₂		
Meinshau	12	
sen ₂		
Meinshau	13	HCB - Postnatal (-)
sen ₂		
Meinshau	14	DDE - Postnatal (-), HCB - Postnatal (-)
sen ₂		
Meinshau	15	HCB - Postnatal (-)
sen ₂		
Mix	1	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	2	Temperature (preg) - Pregnancy (+), Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+)
Mix	3	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m - school) - Postnatal (-), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	4	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-),

Mix	5	PFOA - Pregnancy (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+) Accessibility (bus stops 300m) - Postnatal (-), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	6	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DEP - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+)
Mix	7	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DEP - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	8	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	9	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	10	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), MEP - Postnatal (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), Population density - Postnatal (+)
Mix	11	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), DDE - Postnatal

		(-), DEP - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (+), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), Population density - Postnatal (+)
Mix	12	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), DEP - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	13	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+)
Mix	14	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (-), Manganese - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), PFUNDA - Pregnancy (+), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Population density - Postnatal (+), Road traffic load (100 m) - Postnatal (+), Inverse distance to nearest road (school) - Postnatal (+)
Mix	15	Accessibility (bus stops 300m) - Postnatal (-), BPA - Pregnancy (+), Cobalt - Postnatal (-), Cesium - Postnatal (+), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), MEP - Postnatal (+), Manganese - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PFUNDA - Postnatal (-), Population density - Postnatal (+)
LASSO	1	BPA - Pregnancy (+), Cobalt - Postnatal (-), Copper - Postnatal (+), Copper - Pregnancy (-), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), KIDMED score - Postnatal (+), Land use (300m - school) - Postnatal (-), MEP - Postnatal (-), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), PM10 (day) - Postnatal (+), Population density - Postnatal (+), Accessibility (bus stops 300m) - Postnatal (-)
LASSO	2	Cobalt - Postnatal (-), Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-)

LASSO 1SE	13	Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-)
LASSO 1SE	14	BPA - Pregnancy (+), Cobalt - Postnatal (-), Copper - Postnatal (+), DDE - Postnatal (-), Sleep duration - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), Land use (300m - school) - Postnatal (-), MEP - Postnatal (+), Molybdenum - Postnatal (-), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-), PFOA - Pregnancy (-), Population density - Postnatal (+), Accessibility (bus stops 300m) - Postnatal (-)
LASSO 1SE	15	Cobalt - Postnatal (-), Copper - Postnatal (+), DDE - Postnatal (-), HCB - Postnatal (-), Humidity (month) - Postnatal (+), PBDE 153 - Postnatal (-), PCB 170 - Postnatal (-), PCB 180 - Postnatal (+), PFOA - Postnatal (-)

B.

Method	Run	Exposures selected and corresponding direction of association
Default LASSO	1	
Default LASSO	2	
Default LASSO	3	
Default LASSO	4	
Default LASSO	5	
Default LASSO	6	
Default LASSO	7	
Default LASSO	8	
Default LASSO	9	
Default LASSO	10	
Default LASSO	11	
Default LASSO	12	
Default LASSO	13	
Default LASSO	14	
Default LASSO	15	
CV ₁	1	Cigarette - Pregnancy (+)
CV ₁	2	Cigarette - Pregnancy (+)
CV ₁	3	Cigarette - Pregnancy (+)
CV ₁	4	Cigarette - Pregnancy (+)
CV ₁	5	Cigarette - Pregnancy (+)
CV ₁	6	Cigarette - Pregnancy (+)
CV ₁	7	Cigarette - Pregnancy (+)
CV ₁	8	Cigarette - Pregnancy (+)
CV ₁	9	Cigarette - Pregnancy (+)
CV ₁	10	Cigarette - Pregnancy (+)
CV ₁	11	Cigarette - Pregnancy (+)
CV ₁	12	Cigarette - Pregnancy (+)
CV ₁	13	Cigarette - Pregnancy (+)
CV ₁	14	Cigarette - Pregnancy (+)
CV ₁	15	Cigarette - Pregnancy (+)
CV ₂	1	Cigarette - Pregnancy (+)
CV ₂	2	Cigarette - Pregnancy (+)
CV ₂	3	Cigarette - Pregnancy (+)
CV ₂	4	Cigarette - Pregnancy (+)
CV ₂	5	Cigarette - Pregnancy (+)
CV ₂	6	Cigarette - Pregnancy (+)
CV ₂	7	Cigarette - Pregnancy (+)
CV ₂	8	Cigarette - Pregnancy (+)
CV ₂	9	Cigarette - Pregnancy (+)
CV ₂	10	Cigarette - Pregnancy (+)
CV ₂	11	Cigarette - Pregnancy (+)

CV ₂	12	Cigarette - Pregnancy (+)
CV ₂	13	Cigarette - Pregnancy (+)
CV ₂	14	Cigarette - Pregnancy (+)
CV ₂	15	Cigarette - Pregnancy (+)
Meinshausen ₁	1	Facility density (300m) - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	2	DMP - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	3	OH-MiNP - Pregnancy (-)
Meinshausen ₁	4	Molybdenum - Pregnancy (+), OH-MiNP - Pregnancy (-)
Meinshausen ₁	5	
Meinshausen ₁	6	Molybdenum - Pregnancy (+), OH-MiNP - Pregnancy (-)
Meinshausen ₁	7	Facility density (300m) - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	8	DMP - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	9	DMP - Pregnancy (-)
Meinshausen ₁	10	Molybdenum - Pregnancy (+), OH-MiNP - Pregnancy (-)
Meinshausen ₁	11	Facility density (300m) - Pregnancy (-), DMP - Pregnancy (-)
Meinshausen ₁	12	Facility density (300m) - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	13	Facility density (300m) - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	14	DMP - Pregnancy (-), OH-MiNP - Pregnancy (-)
Meinshausen ₁	15	OH-MiNP - Pregnancy (-)
Meinshausen ₂	1	
Meinshausen ₂	2	
Meinshausen ₂	3	
Meinshausen ₂	4	
Meinshausen ₂	5	
Meinshausen ₂	6	
Meinshausen ₂	7	
Meinshausen ₂	8	
Meinshausen ₂	9	
Meinshausen ₂	10	
Meinshausen ₂	11	
Meinshausen ₂	12	
Meinshausen ₂	13	
Meinshausen ₂	14	
Meinshausen ₂	15	
Mix	1	
Mix	2	
Mix	3	
Mix	4	
Mix	5	
Mix	6	
Mix	7	
Mix	8	
Mix	9	
Mix	10	

Mix	11
Mix	12
Mix	13
Mix	14
Mix	15
LASSO 1SE	1
LASSO 1SE	2
LASSO 1SE	3
LASSO 1SE	4
LASSO 1SE	5
LASSO 1SE	6
LASSO 1SE	7
LASSO 1SE	8
LASSO 1SE	9
LASSO 1SE	10
LASSO 1SE	11
LASSO 1SE	12
LASSO 1SE	13
LASSO 1SE	14
LASSO 1SE	15

Supplementary Material V.1: Commented Script

This script will be available on github (<https://github.com/SoCadiou>) once the corresponding draft will be published.

```
#####
#####
##this code allows to perform a simulation to assess performance in terms of
##stability, sensitivity and specificity of prespecified statistical methods
##used to find the true predictors of an health among the exposome it contains 5
##parts:

##1. defining the functions allowing to generate a realistic dataset of exposome
##and an outcome linearly related to some variables of this exposome.

#It needs a real exposome dataset as input, as well as parameters allowing to
#define the #association between the exposome and the outcome (number of
#predictors, #variability explained, correlation..)

##2. defining the methods assessed

##3. defining some functions used to assess methods performance

##4. defining the simulation function, which, for a given scenario, generates
##the datasets, applies the methods and assess their performance. This function
##allows to parallelize the simulation.
```

```

##5. running the simulation itself with parallelization, repeating X times the
##function defined in 4. for each scenario and saving the results.
#####
#####

library(mvtnorm)
library(boot)
library(parallel)
library(reshape)
library(glmnet)
library(DSA)
library(OmicsMarkeR)
library(Rcpp)
library(stringr)

#####
#####

##### 1. Generating datasets - functions #####
#####

##Define functions used to generate datasets function which generates a
##dataset of exposures with same number of variables and individuals and similar
##correlation structure than a real exposome matrix provided and an outcome
##linearly generated

simulator_2layers <- function(E_true,
  #real exposome data
  R2_tot = 0.1 ,
  #total variability explained all predictors
  n_Ey = 5,
  #number of predictors
  BetaEy = 0.01,
  #Beta coefficient for each predictors. Can be a
  #vector of values or a unique value
  test_and_training = TRUE,
  #generate a dataset of the same size of E_true (if
  #FALSE) or double the number of individuals (if
  #TRUE)
  pos_and_neg = FALSE,
  #if FALSE, all effects are positive; if TRUE, half
  #are negative
  corr = F,
  #if TRUE, the correlation between the predictors
  #is controled
  range_corr = c(0, 1)) {
  #if corr = TRUE, range of correlation for true
  #predictors

##creating a new exposome dataset by bootstrapping
data.X <- as.data.frame(E_true)

```

```

names_row <- rownames(data.X)
data.X <- data.X[sample(1:nrow(data.X), 2 * nrow(data.X), replace = TRUE),]
rownames(data.X) <- c(names_row, sprintf('boot%0s', names_row))
dataExp <- data.X

remove(data.X)

##creating a linearly generated outcome
##defining the vector of Beta coefficients
if (length(BetaEy) == 1) {
  if (pos_and_neg == FALSE) {
    Betapred_yE <- rep(BetaEy, n_Ey)
  } else{
    Betapred_yE <- rep(BetaEy, round(n_Ey / 2))
    Betapred_yE <- c(Betapred_yE, rep(-BetaEy, n_Ey - length(Betapred_yE)))
  }
} else{
  if (length(BetaEy) != n_Ey) {
    stop(
      "error: Betas for M explaining Y not explained by E are not
      consistent with the number of predictors"
    )
  }
  Betapred_yE <- BetaEy
}
##creating the linear combination
yE <- simResponseSimple(
  met = dataExp,
  Nmet = n_Ey,
  beta = Betapred_yE,
  corr = corr,
  range_corr = range_corr
)

##adding a gaussian to Y to reach the wanted level of variability explained by
##the linear combination of predictors
Y <- yE$resp
if (!is.na(R2_tot)) {
  if ((R2_tot) != 0) {
    sigma <- var(Y) * (1 / R2_tot - 1)
  } else{
    R2 = 0.00000001
    sigma <- var(Y) * (1 / R2_tot - 1)
  }
  Y <-
    as.matrix(Y + rnorm(length(Y), mean(Y), sqrt(sigma)), ncol = 1)
  Y <- as.data.frame(Y)
}

##estimating the R2

```

```

R2 <- estimatedR2(dataExp, yE$predictors, Y)$r.squared
##results to return
resultats <- list(
  Y_train = Y[1:(nrow(dataExp) / 2), , drop = FALSE],
  ##train part of the generated Y vector
  E_train = dataExp[1:(nrow(dataExp) / 2), , drop = FALSE],
  ##train part of the generated exposome dataset
  Y_test = Y[(nrow(dataExp) / 2):nrow(dataExp), , drop =
    FALSE],
  ##test part of the generated Y vector
  E_test = dataExp[(nrow(dataExp) / 2):nrow(dataExp), , drop = FALSE],
  ##test part of the generated exposome dataset
  yE = yE,
  ##yE (output of the simResponseSimple) object containing the list of
  ##predictors and the Betas
  R2 = R2,
  ##estimated R2
  list_predictor = as.character(yE$predictors) ##vectors of predictors
)
return(resultats)
}

```

```

#####function to generate a linear response#####
simResponseSimple <- function(met,
  ##dataframe of potential predictors
  Nmet = NA,
  ##number of predictors
  beta = NULL,
  ##Betas coefficient for predictors. Can be a vector of values or a
  ##unique value
  cpg = NULL,
  ##name of forced predictors if necessary
  corr = FALSE,
  #if TRUE, the correlation between the predictors is controled
  range_corr = c(0, 1) {
  #if corr = TRUE, range of correlation for true predictors
  if (all(c(is.na(Nmet), is.null(cpg))) == TRUE) {
    ##case with no link between the response and the dataset of potential
    ##predictors
    return (list(
      resp = as.matrix(rep(0, nrow(met))), ncol = 1),
      beta = NA,
      predictors = NA
    ))
  }
  if (corr == FALSE | Nmet == 1) {
    ##case of only 1 predictors and no correlation control
    temp <- Nmet - length(cpg)
    if (temp != 0) {
      if (length(cpg) == 0) {

```

```

wh <- sample((1:ncol(met)), temp) ##drawing predictors
} else{
  ##drawing predictors while conserving those specified if some were
  ##specified as input
  wh <- sample((1:ncol(met)[-cpg]), temp)
  wh <- c(cpg, wh)
}
} else{
  wh <- cpg
}
} else{
  if (length(cpg) != 0) {
    stop("set corr to true is only possible when names of predictors
      are not provided")
  }
  ##if a specified correlation between predictors is set by the users,
  ##selecting the predictors to be in this specified range
  wh <- submatFindSimpl(Mat <- as.matrix(met), range = range_corr, Nvar = Nmet)
  wh <- which(colnames(met) %in% wh)
}
##defining a matrix of predictors
CovMat <- as.matrix(met[, wh])
colnames(CovMat) <- colnames(met)[wh]
# computing the response
mean <- CovMat %*% matrix(beta, ncol = 1)
rownames(mean) <- rownames(met)
names(beta) <- colnames(CovMat)
return (list(
  resp = mean,
  ##response vector
  beta = beta,
  ##Betas coefficient vector
  predictors = colnames(CovMat) ##vector of predictors
))
}

####function to choose a set of predictors among a set of variables with a
####constraint on the correlation range between predictors#####
submatFindSimpl <- function(Mat, range = c(0, 1), Nvar) {
  # verifying formats and values of inputs
  if (Nvar > ncol(Mat))
    stop("No matrix of the correct size meeting the range criterion")
  if (Nvar < 2)
    stop("Nvar must be at least 2")
  # computing the correlation matrix
  Mat <- abs(cor(Mat))
  diag(Mat) <- NA
  # removing rows with no correlation value in the given range
  wh <- which(apply(Mat, 1, min, na.rm = T) > range[2] |
    apply(Mat, 1, max, na.rm = T) < range[1])
  if (length(wh) > 0)

```

```

Mat <- Mat[-wh, -wh]
# iteratively selecting and testing samples for the correct correlation
Res <- NA
samp1 <- sample(1:ncol(Mat), size = ncol(Mat))
t1 <- 1
while (t1 <= ncol(Mat) & all(is.na(Res))) {
  var <- array(NA, 0)
  t2 <- samp1[t1]
  while (all(is.na(Res)) & length(t2) > 0) {
    if (length(t2) == 1) var <- c(var, t2)
    if (length(t2) > 1) var <- c(var, sample(t2, size = 1))
    t2 <- (1:ncol(Mat))[-var]
    if (length(t2) > 1)
      t2 <- t2[which(apply(as.matrix(Mat[t2, var]) <= range[2] &
                           Mat[t2, var] >= range[1] &
                           Mat[t2, var] < 1), 1, min) == 1]
  if (length(t2) == 1){
    if (min(c(Mat[t2, var] <= range[2], Mat[t2, var] >= range[1],
            Mat[t2, var] < 1)) != 1)
      t2 <- NA
  }
  if (length(var) == Nvar)
    Res <- var
}
t1 <- t1 + 1
}
if (all(is.na(Res)))
  stop("Not enough variable with the given correlation range")
return(colnames(Mat)[Res])
}

```

```

#####function which estimates R2 from a dataset of potential predictors, the list
#####of true predictors and a vector of outcome#####
estimatedR2 <- function(X, truepred, Y) {
  if ("y" %in% truepred) {
    stop("error: one of the true predictors is named y")
  }
  if (ncol(Y) != 1) {
    stop("error:Y is multidimensionnal")
  }
  if (nrow(X) != nrow(Y)) {
    stop("error: not the same number of rows")
  }
  if (isTRUE(all.equal(rownames(X), rownames(Y))) == FALSE) {
    stop("error: individuals are not ordered similarly in X and Y")
  }
  if (all(truepred %in% colnames(X))) {
    data <- X[, colnames(X) %in% truepred, drop = FALSE]
    data <- cbind(Y, data)
    colnames(data)[1] <- "y"
  }
}

```

```

mod <- lm(y ~ ., as.data.frame(data))
toselect.x <- summary(mod)$coeff[-1, 4]
r <- list(summary(mod)$r.squared,
          summary(mod)$adj.r.squared,
          names(toselect.x)[toselect.x == TRUE])
names(r) <- c("r.squared", "adj.r.squared", "pred")
return(r)
} else{
  stop("error: X does not contain all true predictors")
}

#####
#####

##an other function simulator_2layers can be used if one want to control the
##correlation of the overall dataset in this case, the simulated exposome matrix
##is no longer obtained by bootstrapping the real exposome but from a
##correlation matrix the correlation matrix must be provided by the user or
##specified as nulll uncomment the section below and adapt the input of
##functions f0 (section 4.) and clusterApply (section 5.) to use it

# simulator_2layers <-
#   function(names_rows_true,
#     cormat,
#     R2_tot = 0.1 ,
#     n_Ey = 5,
#     BetaEy = 0.01,
#     test_and_training = TRUE,
#     pos_and_neg = FALSE,
#     corr_all = "real",
#     corr_pred = TRUE,
#     range_corr = c(0, 1)) {
#   ##generation of exposome dataset DIFFERENT FROM THE OTHER FUNCTION
#   simulator_2layers
#   if (corr_all == "real") {
#     data.X <- data.frame(rmvnorm(1173 * 2, rep(0, ncol(cormat)),
#                               cormat))
#   } else{
#     if (corr_all == "null") {
#       cormat2 <- diag(ncol(cormat))
#       data.X <- data.frame(rmvnorm(1173 * 2, rep(0, ncol(cormat)),
#                                 cormat2))
#     } else{
#       stop("Correlation for the whole exposome (null or real) must
#            be specified")
#     }
#   }
#   colnames(data.X) <- colnames(cormat)
#   rownames(data.X) <-
#   c(names_rows_true, sprintf('boot%os', (names_rows_true)))

```

```

# dataExp <- data.X
# remove(data.X)
#
# ##FROM HERE, THE FUNCTION IS IDENTICAL TO THE OTHER
simulator_2layers
# function ##creating a linearly generated outcome ##defining the vector of
# Beta coefficients
#
# if (length(BetaEy) == 1) {
#   if (pos_and_neg == FALSE) {
#     Betapred_yE <- rep(BetaEy, n_Ey)
#   } else{
#     Betapred_yE <- rep(BetaEy, round(n_Ey / 2))
#     Betapred_yE <-
#       c(Betapred_yE, rep(-BetaEy, n_Ey - length(Betapred_yE)))
#   }
# } else{
#   if (length(BetaEy) != n_Ey) {
#     stop(
#       "error: Betas for M explaining Y not explained by E are not
#       consistent with the number of predictors"
#     )
#   }
#   Betapred_yE <- BetaEy
# }
# ##creating the linear combination
# yE <-
#   simResponseSimple(
#     met = dataExp,
#     Nmet = n_Ey,
#     beta = Betapred_yE,
#     corr = corr,
#     range_corr = range_corr
#   )
#
# ##adding a gaussian to Y to reach the wanted level of variability
# explained by the linear combination of predictors
# Y <- yE$resp
# if (!is.na(R2_tot)) {
#   if (((R2_tot) != 0)) {
#     sigma <- var(Y) * (1 / R2_tot - 1)
#   } else{
#     R2 = 0.00000001
#     sigma <- var(Y) * (1 / R2_tot - 1)
#   }
#   Y <-
#     as.matrix(Y + rnorm(length(Y), mean(Y), sqrt(sigma)), ncol = 1)
#   Y <- as.data.frame(Y)
# }
# ##estimating the R2

```

```

# R2 <- estimatedR2(dataExp, yE$predictors, Y)$r.squared
# ##results to return
# resultats <-
#   list(
#     Y_train = Y[1:(nrow(dataExp) / 2), , drop = FALSE],
#     ##train part of the generated Y vector
#     E_train = dataExp[1:(nrow(dataExp) / 2), , drop = FALSE],
#     ##train part of the generated exposome dataset
#     Y_test = Y[(nrow(dataExp) / 2):nrow(dataExp), , drop =
#       FALSE],
#     ##test part of the generated Y vector
#     E_test = dataExp[(nrow(dataExp) / 2):nrow(dataExp), , drop = FALSE],
#     ##test part of the generated exposome dataset
#     yE = yE,
#     ##yE (output of the simResponseSimple) object containing the list of
#     ##predictors and the Betas
#     R2 = R2,
#     ##estimated R2
#     list_predictor = as.character(yE$predictors) ##vectors of predictors
#   )
#   return(resultats)
# }

#####
#### 2. Methods to be tested - functions #####
#####

#####a function used to compute residuals from a linear model if covariates are
#####part of the inputs of any the function of the methods tested#####
getresiduals_2df <-
  function(data_Y_in, data_covar_in, name_Y, covar) {
    data_covar <-
      data_covar_in[, colnames(data_covar_in) %in% covar, drop = FALSE]
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_covar),
      colnames(data_Y_in) == name_Y, drop = FALSE]
    data_covar <-
      data_covar[rownames(data_covar) %in% rownames(data_Y), , drop = FALSE]
    data_covar <- data_covar[rownames(data_Y), , drop = FALSE]
    data_output <- data_Y
    data <- cbind(data_Y, data_covar)
    mod <- lm(data = data)
    data_output[, 1] <- as.data.frame(residuals(mod))
    return(data_output)
  }

#####

```

```

ewas <-
function(data_Xs_in,
  ##dataset of explanatory variables ("exposures")
  data_Y_in,
  ##dataset of univariate variable of interest ("outcome")
  name_Y,
  ##variable of interest name
  data_covar_in = NULL,
  ##if necessary, dataset of covariates ("confounders")
  covar = character(0),
  ##if necessary, vector of covariates name
  corr = "BY",
  ##name of multiple testing correction to be applied ("BH" or "Bon" or
  ##"BY" or "None")
  ntest = NULL) {
  ##if ntest is a numeric, correction of multiple testing will be a Bonferroni
  ##correction considering ntest as the number of tests performed
  require(parallel)
  if (length(covar) > 0) {
    ##if necessary, computing residuals of the linear model explaining the
    ##variable of interest by the covariates
    data_covar<-data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
      rownames(data_covar_in) %in% rownames(data_Xs_in),
      colnames(data_covar_in) %in% covar, drop = FALSE]
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
        rownames(data_Y_in) %in% rownames(data_Xs_in),
        colnames(data_Y_in) == name_Y, drop =
        FALSE]
    data_Xs <-
      data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
        rownames(data_Xs_in) %in% rownames(data_Y), , drop = FALSE]
    data_covar <- data_covar[rownames(data_Y), , drop = FALSE]
    data_Xs <- data_Xs[rownames(data_Y), , drop = FALSE]
    data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
  } else{
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) ==
        name_Y, drop = FALSE]
    data_Xs <-
      data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in),
      , drop = FALSE]
    data_Xs <- as.data.frame(data_Xs[rownames(data_Y), ])
    colnames(data_Xs) <- colnames(data_Xs_in)
  }
  ##checking consistency of the datasets
  if (is.null(data_Y) == TRUE |
    is.null(data_Xs) == TRUE |
    !(name_Y %in% colnames(data_Y))) {
    stop("Données incohérentes entre elles")
  }
}

```

```

}

##applying univariate regression for each exposure outcome association
p.values <- mclapply(1:ncol(data_Xs), function(x, data_Xs) {
  c(colnames(data_Xs)[x], summary(lm(Y ~ ,
    data = data.frame(
      cbind(var1 = data_Xs[, x],
        Y = data_Y[, 1])
    )))$coefficients[2, ])
},
data_Xs)
if (length(p.values) == 1) {
  p.values <-
    as.matrix(as.vector(unlist(p.values[[1]])), ncol = 5, byrow = TRUE)[-4, ]
  p.values <- t(as.data.frame(p.values))
} else{
  p.values <-
    cbind(matrix(unlist(p.values), ncol = 5, byrow = TRUE)[, -4])
}
p.values <- as.data.frame(p.values)
colnames(p.values) <- c("var", "Est", "Sd", "pVal")
p.values <- p.values[p.values$var != "Intercept", ]
p.values$pVal <- as.numeric(as.character(p.values$pVal))
p.values.adj <- p.values
pVal <- as.numeric(as.character(p.values$pVal))
##applying correction for multiple testing
if (corr == "None") {
  wh <- which(pVal <= 0.05)
  p.values.adj$pVal_adj <- pVal
}
if (corr == "Bon") {
  wh <- which(pVal <= 0.05 / nrow(p.values))
  p.values.adj$pVal_adj <- pVal * nrow(p.values)
}
if (corr == "BH") {
  wh <- which(p.adjust(pVal, "BH") <= 0.05)
  p.values.adj$pVal_adj <- p.adjust(pVal, "BH")
}
if (corr == "BY") {
  wh <- which(p.adjust(pVal, "BY") <= 0.05)
  p.values.adj$pVal_adj <- p.adjust(pVal, "BY")
}
if (!corr %in% c("Bon", "BH", "BY", "", "None"))
  stop("Please specify a known correction method for
multiple testing")
if (!is.null(ntest)) {
  p.values.adj$pVal_adj <- pVal * ntest
}
wh_num <- wh
wh <- p.values$var[wh]
a <- list(wh, wh_num, p.values.adj)

```

```

##returning selected exposures and pvalues
names(a) <- c("selected", "indices_selected", "pval")
return(a)
}

#####lasso - basic implementation##### this includes a basic 10-fold cross
##validation process as implemented in the CVglmnet package
lasso <-
  function(data_Xs_in,
    ##dataset of explanatory variables ("exposures")
    data_Y_in,
    ##dataset of univariate variable of interest ("outcome")
    name_Y,
    ##variable of interest name
    data_covar_in = NULL,
    ##if neccessary, dataset of covariates ("confounders")
    covar = character(0)) {
    ##if necessary, vector of covariates name

    if (length(covar) > 0) {
      ##if necessary, computing residuals of the linear model explaining the
      ##variable of interest by the covariates
      data_covar <-
        data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
          rownames(data_covar_in) %in% rownames(data_Xs_in),
          colnames(data_covar_in) %in%
            covar, drop = FALSE]
      data_Y <-
        data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
          rownames(data_Y_in) %in% rownames(data_Xs_in),
          colnames(data_Y_in) == name_Y, drop =
            FALSE]
      data_Xs <-
        data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
          rownames(data_Xs_in) %in% rownames(data_Y_in),
          , drop = FALSE]
      data_covar <- data_covar[rownames(data_Y), ]
      data_Xs <- data_Xs[rownames(data_Y), ]
      data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
    } else{
      data_Y <-
        data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
          colnames(data_Y_in) == name_Y, drop = FALSE]
      data_Xs <-
        data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
      data_Xs <- data_Xs[rownames(data_Y), ]
    }
    data_Y <- data.matrix(data_Y)
    data_Xs <- data.matrix(data_Xs)
    ##applying lasso (as implemented in glmnet package, a path of penalization
  }

```

```

##parameter lambda according to the MSE computed is computed by 10-fold
##cross-validation)
cvfit <- cv.glmnet(data_Xs, data_Y, family = "gaussian",
                     alpha = 1)

##Compute predicted Y "Y_predit"
Y_predit <-
  predict(cvfit, newx = data_Xs, s = "lambda.min")
##selecting the model with the penalization parameter minimizing MSE
Y_predit <- Y_predit[rownames(Y_predit),]

##dataframe of predictors selected
tmp_coeffs <-
  coef(cvfit, s = "lambda.min")
##selecting the model with the penalization parameter minimizing MSE
cg_select <-
  data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1],
             coefficient = tmp_coeffs@x)

cg_select <- cg_select$name[cg_select$name != "(Intercept)"]
a <- list()
if (length(cg_select) != 0) {
  a <- list("selected" = cg_select,
            "prediction" = Y_predit,
            "null" = "nul")
} else{
  a <-
  list(
    "selected" = character(),
    "prediction" = "pas_de_prediction",
    "null" = "nul"
  )
}
return(a)
}

#####lasso - basic implementation but using lambda.1se instead of lambda.min
##### this includes a basic 10-fold cross
##validation process as implemented in the CVglmnet package
lasso_1SE <-
  function(data_Xs_in,
          ##dataset of explanatory variables ("exposures")
          data_Y_in,
          ##dataset of univariate variable of interest ("outcome")
          name_Y,
          ##variable of interest name
          data_covar_in = NULL,
          ##if necessary, dataset of covariates ("confounders")
          covar = character(0)) {
  ##if necessary, vector of covariates name

```

```

if (length(covar) > 0) {
  ##if necessary, computing residuals of the linear model explaining the
  ##variable of interest by the covariates
  data_covar <-
    data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
      rownames(data_covar_in) %in% rownames(data_Xs_in),
      colnames(data_covar_in) %in%%
        covar, drop = FALSE]
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
      rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop =
        FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
      rownames(data_Xs_in) %in% rownames(data_Y_in),
      , drop = FALSE]
  data_covar <- data_covar[rownames(data_Y), ]
  data_Xs <- data_Xs[rownames(data_Y), ]
  data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
##applying lasso (as implemented in glmnet package, a path of penalization
##parameter lambda according to the MSE computed is computed by 10-fold
##cross-validation)
cvfit <- cv.glmnet(data_Xs, data_Y, family = "gaussian",
  alpha = 1)

##Compute predicted Y "Y_predit"
Y_predit <-
  predict(cvfit, newx = data_Xs, s = "lambda.1se")
##selecting the model within 1MSE of the model using
##the penalization parameter minimizing MSE
Y_predit <- Y_predit[rownames(Y_predit), ]

##dataframe of predictors selected
tmp_coeffs <-
  coef(cvfit, s = "lambda.min")
##selecting the model with the penalization parameter minimizing MSE
cg_select <-
  data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1],
  coefficient = tmp_coeffs@x)

```

```

cg_select <- cg_select$name[cg_select$name != "(Intercept)"]
a <- list()
if (length(cg_select) != 0) {
  a <- list("selected" = cg_select,
            "prediction" = Y_predit,
            "null" = "nul")
} else{
  a <-
  list(
    "selected" = character(),
    "prediction" = "pas_de_prediction",
    "null" = "nul"
  )
}
return(a)
}

#####
#####lasso_stab : implementation of Meinshausen 2010
##### repeating lasso on
##subsamples and performing the selection according to empirical probability of
##selection computed over the repeated runs using a threshold specified by user
lasso_stab_Meinshausen <-
  function(data_Xs_in,
    ##dataset of explanatory variables ("exposures")
    data_Y_in,
    ##dataset of univariate variable of interest ("outcome")
    name_Y,
    ##variable of interest name
    data_covar_in = NULL,
    ##if necessary, dataset of covariates ("confounders")
    covar = character(0),
    ##if necessary, vector of covariates name
    prop = 0.85) {
  #minimal threshold for an empirical probability to make the corresponding
  #variable selected
  if (length(covar) > 0) {
    ##if necessary, computing residuals of the linear model explaining the
    ##variable of interest by the covariates
    data_covar <-
      data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
                    rownames(data_covar_in) %in% rownames(data_Xs_in),
                    colnames(data_covar_in) %in% 
                    covar, drop = FALSE]
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
              rownames(data_Y_in) %in% rownames(data_Xs_in),
              colnames(data_Y_in) == name_Y, drop =
              FALSE]
}

```

```

data_Xs <-
  data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
  rownames(data_Xs_in) %in% rownames(data_Y_in),
  , drop = FALSE]
data_covar <- data_covar[rownames(data_Y), ]
data_Xs <- data_Xs[rownames(data_Y), ]
data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
    colnames(data_Y_in) ==
      name_Y, drop = FALSE]
data_Xs <-
  data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y), , drop = FALSE]
data_Xs <- data_Xs[rownames(data_Y), ]
}
if (is.null(data_Y)) {
  stop("Données incohérentes entre elles _ Y")
}
if (is.null(data_Xs)) {
  stop("Données incohérentes entre elles _ Xs")
}
if (!(name_Y %in% colnames(data_Y))) {
  stop("Données incohérentes entre elles _ nom Y")
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
list_iter <- list()

length_lambdas <- numeric()
n_iter_stab <- 100
##setting the number of repetitions from which the empirical probabilities
##will be computed
for (k2 in (1:n_iter_stab)) {
  #randomly selecting a subsample containing 50% of individuals
  selec <-
    sample(rownames(data_Xs), round(1 * nrow(data_Xs) / 2)) #
  i_df <- data.matrix(data_Xs[rownames(data_Xs) %in% selec, ])
  block_pheno <-
    data_Y[rownames(data_Y) %in% rownames(i_df), , drop = FALSE]
  block_pheno <- block_pheno[rownames(i_df), , drop = FALSE]
  ##applying lasso to the subsample
  model.lasso <-
    glmnet(
      x = i_df,
      y = data.matrix(block_pheno),
      family = "gaussian",
      alpha = 1
    )
  #saving the selection for each lambda in a dataframe
  selection_pour_une_iter <-
}

```

```

as.data.frame(cbind(rownames(model.lasso$beta), as.numeric(tabulate(
  model.lasso$beta@i + 1
))))
colnames(selection_pour_une_iter) <-
  c("variables", paste("iter", k2))
##adding this datafram to the list of dataframes computed for all
##precedent subsamples
list_iter <- c(list_iter, list(selection_pour_une_iter))
##saving the length of penalization path
length_lambdas <-
  c(length_lambdas, length(model.lasso$lambda))
}
##merging all dataframes of selection by variable
M <-
  as.data.frame(Reduce(function(x, y)
    merge(x, y, by = "variables"), list_iter))
M[, 2:ncol(M)] <-
  lapply(M[, 2:ncol(M)], function(x)
    as.numeric(as.character(x)))
##summing the frequencies of selection by variables
M$sum <- rowSums(M[, 2:ncol(M)])
M1 <- M[, c(1, ncol(M))]
##computing the empirical probabilities from the sum of frequencies of
##selection by variables and the number of possible selections
M1$proba_estimee <- M1$sum / (sum(length_lambdas))
##selecting variables from the empirical probabilities and the threshold
cg_select <- M1$variables[M1$proba_estimee >= prop]
##returning the selection
a <- list()
if (length(cg_select) != 0) {
  a <-
    list(
      "selected" = cg_select,
      "selection_iter" = list(M, M1),
      "iteration" = n_iter_stab
    )
} else{
  a <-
    list(
      "selected" = character(0),
      "selection_iter" = list(M, M1),
      "iteration" = n_iter_stab
    )
}
return(a)
}

#####
# lasso_stab : second implementation of Meinshausen 2010
##### selection of a
##set of lambda parameters on a subsample then repeating lasso on subsamples

```

```

##with those lambdas and performing the selection according to empirical
##probability of selection computed over the repeated runs using a threshold
##specified by user
lasso_stab_Meinshausen2 <-
  function(data_Xs_in,
    ##dataset of explanatory variables ("exposures")
    data_Y_in,
    ##dataset of univariate variable of interest ("outcome")
    name_Y,
    ##variable of interest name
    data_covar_in = NULL,
    ##if necessary, dataset of covariates ("confounders")
    covar = character(0),
    ##if necessary, vector of covariates names
    prop = 0.95) {
  #minimal threshold for an empirical probability to make the corresponding
  #variable selected
  if (length(covar) > 0) {
    ##if necessary, computing residuals of the linear model explaining the
    ##variable of interest by the covariates
    data_covar <-
      data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
        rownames(data_covar_in) %in% rownames(data_Xs_in),
        colnames(data_covar_in) %in%
          covar, drop = FALSE]
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
        rownames(data_Y_in) %in% rownames(data_Xs_in),
        colnames(data_Y_in) == name_Y, drop =
          FALSE]
    data_Xs <-
      data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
        rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
    data_covar <- data_covar[rownames(data_Y), ]
    data_Xs <- data_Xs[rownames(data_Y), ]
    data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
  } else{
    data_Y <-
      data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
        colnames(data_Y_in) ==
          name_Y, drop = FALSE]
    data_Xs <-
      data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y), , drop = FALSE]
    data_Xs <- data_Xs[rownames(data_Y), ]
  }
  if (is.null(data_Y)) {
    stop("Données incohérentes entre elles _ Y")
  }
  if (is.null(data_Xs)) {
    stop("Données incohérentes entre elles _ Xs")
  }
}

```

```

if (!(name_Y %in% colnames(data_Y))) {
  stop("Données incohérentes entre elles _ nom Y")
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)

list_iter <- list()
####defining a vector of penalized parameters lambda on a subsample
selec <- sample(rownames(data_Xs), round(1 * nrow(data_Xs) / 2))
i_df <- data.matrix(data_Xs[rownames(data_Xs) %in% selec, ])
block_pheno <-
  data_Y[rownames(data_Y) %in% rownames(i_df), , drop = FALSE]
block_pheno <- block_pheno[rownames(i_df), , drop = FALSE]
temp <-
  glmnet(
    x = i_df,
    y = block_pheno,
    family = "gaussian",
    alpha = 1
  )
##lasso is applied; the path of lambdas will be used for all other
##subsamples
lambdas <- temp$lambda
####repeating lasso fo all subsamples
n_iter_stab <- 100
##setting the number of repetitions from which the empirical probabilities
##will be computed
for (k2 in (1:n_iter_stab)) {
  selec <-
    sample(rownames(data_Xs), round(1 * nrow(data_Xs) / 2))
  #randomly selecting a subsample containing 50% of individuals
  i_df <- data.matrix(data_Xs[rownames(data_Xs) %in% selec, ])
  block_pheno <-
    data_Y[rownames(data_Y) %in% rownames(i_df), , drop = FALSE]
  block_pheno <- block_pheno[rownames(i_df), , drop = FALSE]
  ##applying lasso to the subsample
  model.lasso <-
    glmnet(
      x = i_df,
      y = data.matrix(block_pheno),
      family = "gaussian",
      alpha = 1,
      lambda = as.numeric(lambdas)
    )
  #saving the selection for each lambda in a dataframe
  selection_pour_une_iter <-
    as.data.frame(cbind(rownames(model.lasso$beta), as.numeric(tabulate(
      model.lasso$beta@i + 1
    ))))
  colnames(selection_pour_une_iter) <-

```

```

c("variables", paste("iter ", k2))
##adding this dataframe to the list of dataframes computed for all
##precedent subsamples
list_iter <- c(list_iter, list(selection_pour_une_iter))
}
##merging all dataframe of selection by variable
M <-
  as.data.frame(Reduce(function(x, y)
    merge(x, y, by = "variables"), list_iter))
M[, 2:ncol(M)] <-
  lapply(M[, 2:ncol(M)], function(x)
    as.numeric(as.character(x)))
##summing the frequencies of selection by variables
M$sum <- rowSums(M[, 2:ncol(M)])
M1 <- M[, c(1, ncol(M))]
##computing the empirical probabilities from the sum of frequencies of
##selection by variables and the number of possible selections
M1$proba_estimee <- M1$sum / (length(lambdas) * n_iter_stab)
M1$proba_estimee <- M1$sum / (sum(length_lambdas))
##selecting variables from the empirical probabilities and the threshold
cg_select <- M1$variables[M1$proba_estimee >= prop]
##returning the selection
a <- list()
if (length(cg_select) != 0) {
  a <-
    list(
      "selected" = cg_select,
      "selection_iter" = list(M, M1),
      "iteration" = n_iter_stab
    )
} else{
  a <-
    list(
      "selected" = character(0),
      "selection_iter" = list(M, M1),
      "iteration" = n_iter_stab
    )
}
return(a)
}

#####LASSO_CV2#####

#### loop applying 100 times lasso with the 10-fold validations procedures on the
#### whole dataset the average of the mean error curves (MSE as a function of the
#### MSE) gives the penalization parameter which minimizes this averaged MSE and
#### which will be used in the final model
lasso_moy_MSE <-
  function(data_Xs_in,
    ##dataset of explanatory variables ("exposures")

```

```

data_Y_in,
##dataset of univariate variable of interest ("outcome")
name_Y,
##variable of interest name
data_covar_in = NULL,
##if necessary, dataset of covariates ("confounders")
covar = character(0)) {
##if necessary, vector of covariates name
if (length(covar) > 0) {
##if necessary, computing residuals of the linear model explaining the
##variable of interest by the covariates
data_covar <-
  data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
    rownames(data_covar_in) %in% rownames(data_Xs_in),
    colnames(data_covar_in) %in%
      covar, drop = FALSE]
data_Y <-
  data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
    rownames(data_Y_in) %in% rownames(data_Xs_in),
    colnames(data_Y_in) == name_Y, drop =
      FALSE]
data_Xs <-
  data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
    rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
data_covar <- data_covar[rownames(data_Y), ]
data_Xs <- data_Xs[rownames(data_Y), ]
data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) ==
        name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in),
      , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
lambdas <- numeric(0)
MSEs <- data.frame(matrix(NA, nrow = 100, ncol = 100))
##repeating 100 times the cross-validation process
for (i in 1:100) {
  cv <- cv.glmnet(x = data_Xs, y = data_Y, alpha = 1) ##applying lasso
  MSEs[1:length(cv$lambda), i] <-
    cv$cvm
  ##saving the cross-validation path (ie MSE values for each lambdas) as a
  ##colum of MSEs dataframe
  if (length(cv$lambda) > length(lambdas)) {
    lambdas <-
      cv$lambda ##saving the vector of lambdas of maximum length
}

```

```

}

print(cv$lambda[1:10])

}

##restricting MSEs to non empty rows
MSEs <- MSEs[1:length(lambdas), ]
rownames(MSEs) <- lambdas
##choosing the lambda minimizing the mean of MSE computed across all
##lambdas
lambda.min <-
  as.numeric(names(which.min(rowMeans(MSEs, na.rm = TRUE)))))

##applying lasso with this lambda as forced penalization parameter
model.enet <- glmnet(
  data_Xs,
  data_Y,
  family = "gaussian",
  alpha = 1,
  lambda = lambda.min
)

##selecting variables
tmp_coeffs <- coef(model.enet)
cg_select <-
  data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1],
             coefficient = tmp_coeffs@x)
cg_select <- cg_select$name[cg_select$name != "(Intercept)"]
##returning selected variables
a <- list()
if (length(cg_select) != 0) {
  a <-
    list("selected" = cg_select,
         "prediction" = "prediction",
         "null" = "nul")
} else{
  a <-
    list(
      "selected" = character(),
      "prediction" = "pas_de_prediction",
      "null" = "nul"
    )
}
return(a)
}

####LASSO_CV1#####
### loop applying 100 times lasso with the 10-fold validations procedures on the
### whole dataset the average of the penalization parameters minimizing the MSE
### for each run gives the penalization parameter which will be used in the

```

```

#### final model

lasso_moy_lambda <-
  function(data_Xs_in,
    ##dataset of explanatory variables ("exposures")
    data_Y_in,
    ##dataset of univariate variable of interest ("outcome")
    name_Y,
    ##variable of interest name
    data_covar_in = NULL,
    ##if necessary, dataset of covariates ("confounders")
    covar = character(0)) {
  ##if necessary, vector of covariates name
  if (length(covar) > 0) {
    ##if necessary, computing residuals of the linear model explaining the
    ##variable of interest by the covariates
    data_covar <-
      data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
        rownames(data_covar_in) %in% rownames(data_Xs_in),
        colnames(data_covar_in) %in%
          covar, drop = FALSE]
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
      rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop =
        FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
      rownames(data_Xs_in) %in% rownames(data_Y_in),
      , drop = FALSE]
  data_covar <- data_covar[rownames(data_Y), ]
  data_Xs <- data_Xs[rownames(data_Y), ]
  data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)

##initialization
lambdas <- NULL
##repeating 100 times the cross-validation process
for (i in 1:100) {
  cv <- cv.glmnet(x = data_Xs, y = data_Y, alpha = 1)
  lambdas <- c(lambdas, cv$lambda.min)
}

```

```

##averaging the optimal parameters across repetitions
lambda_min <- mean(lambdas)
##applying lasso with this lambda as forced penalization parameter
model.enet <- glmnet(
  data_Xs,
  data_Y,
  family = "gaussian",
  alpha = 1,
  lambda = lambda_min
)

##selecting variables
tmp_coeffs <- coef(model.enet)
cg_select <-
  data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1],
             coefficient = tmp_coeffs@x)

cg_select <- cg_select$name[cg_select$name != "(Intercept)"]
##returning selected variables
a <- list()
if (length(cg_select) != 0) {
  a <-
    list("selected" = cg_select,
         "prediction" = "prediction",
         "null" = "nul")
} else{
  a <-
    list(
      "selected" = character(),
      "prediction" = "pas_de_prediction",
      "null" = "nul"
    )
}
return(a)
}

#####Mix Method#####
##Computing empirical probabilities derived from selection frequencies when
##running cross-validated lasso on subsamples
lasso_moy_Meinshausen <-
  function(data_Xs_in,
    ##dataset of explanatory variables ("exposures")
    data_Y_in,
    ##dataset of univariate variable of interest ("outcome")
    name_Y,
    ##variable of interest name
    data_covar_in = NULL,
    ##if necessary, dataset of covariates ("confounders")

```

```

covar = character(0),
##if necessary, vector of covariates name
prop = 0.5)
##threshold to select variables from their empirical probabilities
if (length(covar) > 0) {
  ##if necessary, computing residuals of the linear model explaining the
  ##variable of interest by the covariates
  data_covar <-
    data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
      rownames(data_covar_in) %in% rownames(data_Xs_in),
      colnames(data_covar_in) %in%%
        covar, drop = FALSE]
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
      rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop =
        FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
      rownames(data_Xs_in) %in% rownames(data_Y_in),
      , drop = FALSE]
  data_covar <- data_covar[rownames(data_Y), ]
  data_Xs <- data_Xs[rownames(data_Y), ]
  data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in),
      colnames(data_Y_in) == name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
if (is.null(data_Y)) {
  stop("Données incohérentes entre elles _ Y")
}
if (is.null(data_Xs)) {
  stop("Données incohérentes entre elles _ Xs")
}
if (!(name_Y %in% colnames(data_Y))) {
  stop("Données incohérentes entre elles _ nom Y")
}
##initialization
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
vector_selected <- character()
n_iter_stab <- 100
#setting the number of repetitions from which the empirical probabilities
#will be computed
for (k2 in (1:n_iter_stab)) {
  #randomly selecting a subsample containing 50% of individuals
  selec <-

```

```

sample(rownames(data_Xs), round(1 * nrow(data_Xs) / 2))
i_df <- data.matrix(data_Xs[rownames(data_Xs) %in% selec, ])
block_pheno <-
  data_Y[rownames(data_Y) %in% rownames(i_df), , drop = FALSE]
block_pheno <- block_pheno[rownames(i_df), , drop = FALSE]
##applying cross_validated lasso to the subsample
model.lasso <-
  cv.glmnet(
    x = i_df,
    y = data.matrix(block_pheno),
    family = "gaussian",
    alpha = 1
  )
##selecting the model minimizing the MSE
tmp_coeffs <- coef(model.lasso, s = "lambda.min")
cg_select <-
  data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1], coefficient =
tmp_coeffs@x)
cg_select <- cg_select$name[cg_select$name != "(Intercept)"]
#saving the variables selected for this iteration
vector_selected <- c(vector_selected, as.character(cg_select))
}
##computing frequencies of selection
tab <- as.data.frame(table(vector_selected))
colnames(tab) <- c("exp", "proba")
##computing empirical probabilities
tab$proba <- tab$proba / n_iter_stab
###selecting variables from the empirical probabilities and the threshold
cg_select_tot <- tab$exp[tab$proba >= prop]
##returning the selection
a <- list()
if (length(cg_select_tot) != 0) {
  a <-
    list("selected" = cg_select_tot,
         "freq" = tab,
         "iteration" = n_iter_stab)
} else{
  a <- list(
    "selected" = character(0),
    "freq" = tab,
    "iteration" = n_iter_stab
  )
}
return(a)
}

####Elastic-Net#####
Enet <-
  function(data_Xs_in,

```

```

##dataset of explanatory variables ("exposures")
data_Y_in,
##dataset of univariate variable of interest ("outcome")
name_Y,
##variable of interest name
data_covar_in = NULL,
##if necessary, dataset of covariates ("confounders")
covar = character(0) {
##if necessary, vector of covariates name
if (length(covar) > 0) {
##if necessary, computing residuals of the linear model explaining the
##variable of interest by the covariates
data_covar <-
  data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
    rownames(data_covar_in) %in% rownames(data_Xs_in),
  colnames(data_covar_in) %in%
    covar, drop = FALSE]
data_Y <-
  data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
    rownames(data_Y_in) %in% rownames(data_Xs_in), colnames(data_Y_in) ==
  name_Y, drop =
    FALSE]
data_Xs <-
  data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
    rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
data_covar <- data_covar[rownames(data_Y), ]
data_Xs <- data_Xs[rownames(data_Y), ]
data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in), colnames(data_Y_in) ==
      name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)
##preparation of CV folds (which must be the same for alpha and lambda CV)
nfolds=10

n <- nrow(data_Xs)
folds <- rep(1:nfolds,length.out=n)[sample(n,n)]
#step 1: do all crossvalidations for each alpha in range 0.1 - 1.0
alphasOfInterest <- seq(1,0.1,-0.1)
cvs <- lapply(alphasOfInterest,
  function(curAlpha) {cv.glmnet(data_Xs,data_Y,alpha=curAlpha,
    family="gaussian",nfolds=nfolds,
    foldid=folds,standardize=FALSE)})

#step 2: collect the optimum lambda for each alpha
optimumPerAlpha <- sapply(seq_along(alphasOfInterest), function(cur) {

```

```

curcvs <- cvs[[curi]]
curAlpha <- alphasOfInterest[curi]
indOfMin <- match(curcvs$lambda.min, curcvs$lambda)
return(c(lam=curcvs$lambda.min, alph=curAlpha, cvm=curcvs$cvm[indOfMin],
        cvup=curcvs$cvup[indOfMin])) })
#step 3: find the overall optimum
posOfOptimum <- which.min(optimumPerAlpha["cvm",])
overall.alpha.min <- optimumPerAlpha["alph",posOfOptimum]
overall.lambda.min <- optimumPerAlpha["lam",posOfOptimum]
overall.criterionthreshold <- optimumPerAlpha["cvup",posOfOptimum]

##final model
model <-
  glmnet(x=data_Xs,y=data_Y,alpha=overall.alpha.min,
         lambda=overall.lambda.min,family="gaussian",standardize=FALSE)
##selecting variables in the model minimizig the MSE
tmp_coeffs <- coef(model, s =as.numeric(overall.lambda.min))
cg_select<-data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1],
                        coefficient = tmp_coeffs@x)
cg_select<-
cg_select$name[cg_select$name!="(Intercept)"&!is.na(cg_select$name)&cg_select$name!="<NA
>"]
##returning selection
a<-list()
if (length(cg_select)!=0){
  a<-list("selected"=cg_select,"alpha_final"=overall.alpha.min,
          "lambda_final"=overall.lambda.min)
} else{
  a<-list("selected"=character(0),"alpha_final"=overall.alpha.min,
          "lambda_final"=overall.lambda.min)
}
return(a)
}

#####Elastic-Net stabilized by repeating the CV process#####
Enet_CV <-
  function(data_Xs_in,
          ##dataset of explanatory variables ("exposures")
          data_Y_in,
          ##dataset of univariate variable of interest ("outcome")
          name_Y,
          ##variable of interest name
          data_covar_in = NULL,
          ##if neccessary, dataset of covariates ("confounders")
          covar = character(0)) {
  ##if necessary, vector of covariates name
  if (length(covar) > 0) {
    ##if necessary, computing residuals of the linear model explaining the

```

```

##variable of interest by the covariates
data_covar <-
  data_covar_in[rownames(data_covar_in) %in% rownames(data_Y_in) &
    rownames(data_covar_in) %in% rownames(data_Xs_in),
  colnames(data_covar_in) %in%
    covar, drop = FALSE]
data_Y <-
  data_Y_in[rownames(data_Y_in) %in% rownames(data_covar) &
    rownames(data_Y_in) %in% rownames(data_Xs_in), colnames(data_Y_in) ==
  name_Y, drop =
    FALSE]
data_Xs <-
  data_Xs_in[rownames(data_Xs_in) %in% rownames(data_covar) &
    rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
data_covar <- data_covar[rownames(data_Y), ]
data_Xs <- data_Xs[rownames(data_Y), ]
data_Y <- getresiduals_2df(data_Y, data_covar, name_Y, covar)
} else{
  data_Y <-
    data_Y_in[rownames(data_Y_in) %in% rownames(data_Xs_in), colnames(data_Y_in) ==
      name_Y, drop = FALSE]
  data_Xs <-
    data_Xs_in[rownames(data_Xs_in) %in% rownames(data_Y_in), , drop = FALSE]
  data_Xs <- data_Xs[rownames(data_Y), ]
}
data_Y <- data.matrix(data_Y)
data_Xs <- data.matrix(data_Xs)

```

nfolds=10

```

##the cross-validation process is repeated 100 times for alpha
all_alpha_min<-numeric(0)
n <- nrow(data_Xs)
optimumAlpha<-NULL
for (i in 1:100){
  print(i)
  ##preparation of CV folds (which must be the same for alpha and lambda CV)
  folds <- rep(1:nfolds,length.out=n)[sample(n,n)]
  #step 1: do all crossvalidations for each alpha in range 0.1 - 1.0
  alphasOfInterest <- seq(1,0.1,-0.1)
  cvs <- lapply(alphasOfInterest,
    function(curAlpha) {cv.glmnet(data_Xs,data_Y,alpha=curAlpha,
      family="gaussian",nfolds=nfolds,
      foldid=folds,standardize=FALSE)})
  #step 2: collect the optimum lambda for each alpha
  optimumPerAlpha <- sapply(seq_along(alphasOfInterest), function(curi){
    curcvs <- cvs[[curi]]
    curAlpha <- alphasOfInterest[curi]
    indOfMin <- match(curcvs$lambda.min, curcvs$lambda)
    return(c(lam=curcvs$lambda.min,alph=curAlpha,cvm=curcvs$cvm[indOfMin],
  
```

```

cvup=curcvs$cvup[indOfMin])) })
posOfOptimum <- which.min(optimumPerAlpha["cvm",])
overall.alpha.min <- optimumPerAlpha["alph",posOfOptimum]
#list_optimumPerAlpha<-cbind(list_optimumPerAlpha,list(optimumPerAlpha))
optimumAlpha<-c(optimumAlpha,overall.alpha.min)
}
##the final alpha value is computed
alpha_final<-mean(optimumAlpha)
##comment : it is not possible to obtain simultaneously the average values of
##lambda and alpha, as lambda must be computed according to alpha

##the cross-validation process is repeated 100 times for lambda
all_lambda_min<-numeric(0)
for (i in 1:100){
  overall.lambda.min<-cv.glmnet(x=data_Xs,y=data_Y,alpha=alpha_final,family="gaussian")
  plot(overall.lambda.min)

  all_lambda_min<-c(all_lambda_min,overall.lambda.min)
}
lambda_final<-mean(all_lambda_min)

##final model
model <-

glmnet(x=data_Xs,y=data_Y,alpha=alpha_final,lambda=lambda_final,family="gaussian",standar
dize=FALSE)
##selecting variables in the model minimizig the MSE
tmp_coeffs <- coef(model, s = as.numeric(overall.lambda.min))
cg_select<-data.frame(name = tmp_coeffs@Dimnames[[1]][tmp_coeffs@i + 1], coefficient =
tmp_coeffs@x)
cg_select<-
cg_select$name[cg_select$name!="(Intercept)"&!is.na(cg_select$name)&cg_select$name!="<NA
>"]
##returning selection
a<-list()
if (length(cg_select)!=0){
  a<-
  list("selected"=cg_select,"alpha_final"=overall.alpha.min,"lambda_final"=overall.lambda.min)
}else{
  a<-
  list("selected"=character(0),"alpha_final"=overall.alpha.min,"lambda_final"=overall.lambda.min)
}
return(a)
}

####DSA#####
##DSA is an iterative linear regression model search algorithm (Sinisi and van
##der Laan 2004) following three constraints: maximum order of interaction
##amongst predictors, the maximum power for a given predictor, and the maximum

```

```

##model size.
DSAreg <-
  function(Exp,
    #####dataset of explanatory variables ("exposures")
    resp,
    ##dataset of univariate variable of interest ("outcome")
    family = gaussian,
    ##family of the outcome
    maxsize = 15,
    ##maximum size of the model
    maxsumofpow = 2,
    ##maximum power for a given predictor
    maxorderint = 2) {
  ##maximum order of interaction
  Exp <-
    data.frame(cbind(resp = resp, data.frame(Exp)))
  ##merging exp and resp in a unique dataframe
  colnames(Exp)[1] <- "resp"
  ##applying DSA function with 5 fold split and 1 cross-validation process
  res <-
    DSA(
      resp ~ 1,
      data = Exp,
      family = family,
      maxsize = maxsize,
      maxsumofpow
      = maxsumofpow,
      maxorderint = maxorderint ,
      nsplits = 1,
      usersplits = NULL
    )
  ##extracting the selected variables in case there are power or interaction
  form <- gsub("I[()", "", colnames(coefficients(res)))
  form <-
    gsub("[*]", ":" , gsub("[]]", "", gsub("[^:]1", "", form)))
  if (length(grep(":", form)) > 0) {
    nam <- strsplit(form[grep(":", form)], ":")
    for (j in 1:length(nam)) {
      nam[[j]] <- gsub("[[:space:]]", "", nam[[j]])
      name <- nam[[j]][1]
      for (k in 2:length(nam[[j]]))
        name <- paste(name, ":", nam[[j]][k], sep = "")
      Exp <- cbind(Exp, name = apply(Exp[, nam[[j]]], 1, prod))
    }
  }
  form2 <- "resp~1"
  if (length(form) > 1)
    for (i in 2:length(form))
      form2 <- paste(form2, "+", form[i])
  ##putting the selected variables in a linear model
  res2 <- lm(form2, data = data.frame(Exp))

```

```

#pred <- predict(res2,Exp)
##obtaining beta coefficients
coef <- summary(res2)$coefficients
coef <-
  as.character(rownames(coef)[rownames(coef) != "Intercept"])
##returning selecting
return(list(
  selected = coef[coef != "(Intercept)"],
  pred = "prediction",
  null = "null"
))
}

##multivariate regression
multi <- function(Exp, resp) {
  ##Exp is the dataset of potential predictors ("the exposome")
  ##resp is the variable to explain ("the outcome")
  var <- colnames(Exp)
  Exp <- data.frame(cbind(resp = resp, data.frame(Exp)))
  colnames(Exp)[1] <- "resp"
  formula <-
    as.formula(paste("resp ~", paste(var, collapse = " + ")))
  model <- lm(formula = formula, data = Exp)
  selection <- as.data.frame(summary(model)$coeff[-1, 4])
  colnames(selection) <- "pVal"
  selection$pVal_adj <- p.adjust(selection$pVal, "BH")
  selected <-
    as.character(unique(rownames(selection)[selection$pVal_adj <= 0.05]))
  return(list(
    selected = selected,
    pred = "prediction",
    null = "null"
))
}

#####
#
##### 3. Methods assessment - functions #####
#####

#####computing sensitivity#####
sensitivity <- function(truepred, predfound) {
  return(length(truepred[truepred %in% predfound]) / length(truepred))
}

#####computing false discovery proportion#####
fdp <- function(truepred, predfound) {
  if (length(predfound) == 0) {
    return(0)
  } else{
    return(length(predfound[!predfound %in% truepred]) / length(predfound))
  }
}

```

```

}

####computing specificity#####
specificity <- function(truepred, predfound, n_base) {
  return((n_base - length(truepred) - length(predfound[!predfound %in% truepred]))) /
    (n_base - length(truepred)))
}

#####
##### 4. Function to parallelize - performing simulation and assessment #####
#####

##take only the iteration number as input and use it as a seed
f0 <- function(x) {
  ##setting seed
  set.seed(x)
  ##generating datasets
  simu <-
    simulator_2layers(
      E_true = dataExp_true,
      R2_tot = R2_fixed,
      n_Ey = n_Ey,
      BetaEy = 0.1,
      test_and_training = TRUE,
      pos_and_neg = pos_and_neg,
      corr = corr,
      range_corr = corr_range
    )
  simu$Y_train <- scale(simu$Y_train)
  simu$Y_test <- scale(simu$Y_test)
  list_predBMI_E <- list()
  ##creating a list where to save methods results
  predBMI_E <-
    lapply(1:n_method, function(i)
      lapply(1:7, function(x)
        list())))
  ####repeating methods on the datasets
  for (j_stab in (1:n_iter_stab)) {
    ##setting seed
    set.seed(x + j_stab)
    ##measuring computation time
    start_time_meth <- Sys.time()
    ##ExWAS
    pred_iter <-
      list(ewas_BH = ewas(
        as.data.frame(simu$E_train),
        as.data.frame(simu$Y_train),
        colnames(as.data.frame(simu$Y_train)),
        corr = "BH"))
  }
}

```

```

))
end_time_meth <- Sys.time()
pred_iter$ewas_BH[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))
print("ewas")
###LASSO
start_time_meth <- Sys.time()
predlasso <-
  lasso(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train))
  )
end_time_meth <- Sys.time()
predlasso[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))
print("lasso")
###LASSO Meinshausen 1
start_time_meth <- Sys.time()
predlasso_stab_Meinshausen <-
  lasso_stab_Meinshausen(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    prop = 0.85
  )
end_time_meth <- Sys.time()
###LASSO Mix
predlasso_stab_Meinshausen[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))
print("lasso_stab")
start_time_meth <- Sys.time()
predlasso_moy_Meinshausen <-
  lasso_moy_Meinshausen(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    prop = 0.5
  )
end_time_meth <- Sys.time()
predlasso_moy_Meinshausen[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))

start_time_meth <- Sys.time()
###LASSO Meinshausen 2
predlasso_stab_Meinshausen2 <-
  lasso_stab_Meinshausen2(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train)),
    prop = 0.95
  )

```

```

)
end_time_meth <- Sys.time()
predlasso_stab_Meinshausen2[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))

start_time_meth <- Sys.time()
###LASSO CV2
predlasso_moy_MSE <-
  lasso_moy_MSE(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train))
  )
end_time_meth <- Sys.time()
predlasso_moy_MSE[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))

###LASSO CV1
start_time_meth <- Sys.time()
predlasso_moy_lambda <-
  lasso_moy_lambda(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train))
  )
end_time_meth <- Sys.time()
predlasso_moy_lambda[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))

##ElasticNet
start_time_meth <- Sys.time()
predEnet <-
  Enet(
    data_Xs_in = as.data.frame(simu$E_train),
    data_Y_in = as.data.frame(simu$Y_train),
    colnames(as.data.frame(simu$Y_train))
  )
end_time_meth <- Sys.time()
predEnet[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))
print("enet")

start_time_meth <- Sys.time()

predDSA <-
  DSAreg(
    Exp = simu$E_train,
    resp = simu$Y_train,
    maxsize = floor(ncol(simu$E_train) / 10),
    maxsumofpow = 1,

```

```

maxorderint = 1
)
end_time_meth <- Sys.time()
predDSA[[2]] <-
  as.numeric(difftime(end_time_meth, start_time_meth, units = c("secs")))
print("DSA effectué")
##combining results of the different methods for this iteration
pred_iter <-
  c(
    pred_iter,
    lasso = list(predlasso),
    lasso_stab_Meinshausen = list(predlasso_stab_Meinshausen),
    lasso_stab_Meinshausen2 = list(predlasso_stab_Meinshausen2),
    lasso_moy_Meinshausen = list(predlasso_moy_Meinshausen),
    lasso_moy_MSE = list(predlasso_moy_MSE),
    lasso_moy_lambda = list(predlasso_moy_lambda),
    Enet = list(predEnet),
    DSA = list(predDSA))
)
#pred_iter<-c(pred_iter,lasso=list(predlasso),
#lasso_stab=list(predlasso_stab),
#lasso_stab_plus_spec=list(predlasso_stab_ps),
#Enet=list(predEnet),DSA=list(predDSA))
#pred_iter<-c(pred_iter,lasso=list(predlasso)) #assessing performance of
#each method
truepred <- simu$yE$predictors
for (k1 in (1:length(pred_iter))) {
  predfound <- pred_iter[[k1]]$selected
  if (exists("predfound") & exists("truepred")) {
    if (length(predfound) == 0) {
      print("no predictors found")
    }
    a <- sensitivity(truepred, predfound)
    b <- specificity(truepred, predfound, ncol(simu$E_train))
    c <- fdp(truepred, predfound)
    d <-
      estimatedR2(simu$E_test, predfound, simu$Y_test)$r.squared
    # print(a)
    # print(b)
    # print(c)
    # print(d)
    pred_iter[[k1]] <-
      c(
        pred_iter[[k1]],
        sens = a,
        spec = b,
        fdp = c,
        R2_test = d
      )
    remove(a)
  }
}

```

```

remove(b)
remove
remove(d)
remove(predfound)
}
k1 <- k1 + 1
}
##reshaping results saves
for (k2 in (1:length(pred_iter))) {
  for (k3 in (1:length(pred_iter[[k2]]))) {
    predBMI_E[[k2]][[k3]] <-
      c(predBMI_E[[k2]][[k3]], list(pred_iter[[k2]][[k3]]))
    names(predBMI_E[[k2]]) <- names(pred_iter[[k2]])
  }
  names(predBMI_E) <- names(pred_iter)
}
print(j_stab)
j_stab <- j_stab + 1

}
list_predBMI_E <- c(list_predBMI_E, list(predBMI_E))
performance <- list()
##Assessing stability
####Assessing the frequency of selection of each hits
for (k4 in (1:length(predBMI_E))) {
  tab_freq <-
    as.data.frame(table(table(as.character(
      unlist(predBMI_E[[k4]][[1]])
    ))))
  tab_freq$Var1 <-
    as.numeric(as.character(tab_freq$Var1)) / n_iter_stab
  colnames(tab_freq) <- c("freq", "nb_exp")
  tab_freq <- tab_freq[order(tab_freq$freq), ]
  ####Computing average Sorensen index as a measure of stability
  c_sor <- numeric(0)
  for (i1 in 1:(length(predBMI_E[[k4]][[1]]) - 1)) {
    for (i2 in (i1 + 1):length(predBMI_E[[k4]][[1]])) {
      if ((length(predBMI_E[[k4]][[1]][[i1]]) == 0) &
          length(predBMI_E[[k4]][[1]][[i2]]) == 0) {
        sor <- 1
      } else{
        sor <-
          sorensen(as.character(predBMI_E[[k4]][[1]][[i1]]),
                   as.character((predBMI_E[[k4]][[1]][[i2]])))
      }
      c_sor <- c(c_sor, sor)
    }
  }
  sor_moy <- mean(c_sor)
  ##Saving performance measurement
}

```

```

performance[[k4]] <-
list(
  mean_nb_selec = mean(unlist(
    lapply(predBMI_E[[k4]][[1]], function(X)
      length(X))
  )),
  mean_sens = mean(unlist(predBMI_E[[k4]][[4]]), na.rm =
    TRUE),
  mean_spec = mean(unlist(predBMI_E[[k4]][[5]]), na.rm =
    TRUE),
  mean_fdp = mean(unlist(predBMI_E[[k4]][[6]]), na.rm =
    TRUE),
  mean_R2_test = mean(unlist(predBMI_E[[k4]][[7]]), na.rm =
    TRUE),
  nb_sup_20percent = sum(tab_freq$nb_exp[(which(tab_freq$freq >=
    0.2))]),
  nb_sup_60percent = sum(tab_freq$nb_exp[(which(tab_freq$freq >=
    0.6))]),
  sorensen = sor_moy,
  sd_nb_selec = sd(unlist(
    lapply(predBMI_E[[k4]][[1]], function(X)
      length(X))
  ), na.rm = TRUE),
  sd_sens = sd(unlist(predBMI_E[[k4]][[4]]), na.rm =
    TRUE),
  sd_spec = sd(unlist(predBMI_E[[k4]][[5]]), na.rm =
    TRUE),
  sd_fdp = sd(unlist(predBMI_E[[k4]][[6]]), na.rm =
    TRUE),
  sd_R2_test = sd(unlist(predBMI_E[[k4]][[7]]), na.rm =
    TRUE),
  run_time <-
    mean(unlist(predBMI_E[[k4]][[2]]), na.rm = TRUE)
)
}

remove(sor_moy)
remove(tab_freq)
names(performance) = names(predBMI_E)
##returning an object containing the simulated datasets with their
##characteristics, the results of each method and the performance measurements
A <-
list(simu = simu,
  performance = performance,
  list_predBMI_E = list_predBMI_E)
remove(simu)
remove(performance)
remove(list_predBMI_E)
gc()
return(A)

```

```

}

#####
##### 5. SIMULATIONS #####
#####

#####loading input needed to generate datasets#####
dataExp_true <- readRDS("20190205 Exposome simu borne.rds")
##initialization
list_list_list_predBMI_E <- list()
list_list_performance <- list()
n_iter <- 30 ##number of iterations for each scenarios
n_iter_stab <- 15 ##number of iterations for each dataset
n_method <- 8 ##number of methods tested
c_n_R2_fixed <-
  c(0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8)
##values tested for R2 (variability of the outcome explained)
c_n_Ey <-
  c(1, 3, 10, 25, 50, 100) ##values tested for the number of predictors

#list_corr<-list(all=c(0,1),low=c(0,0.2),middle=c(0.2,0.4),high=c(0.5,1))
list_corr <- list(all = c(0, 1))
neg <- FALSE
#neg<-c(FALSE,TRUE)

seed = 22
##initializing the dataset in which the performance of each method will be
##summarized by scenarios
comp_stability_method <- data.frame(
  Methods = vector(),
  Nb_true_predictors_of_BMI_in_E = numeric(0),
  Total_variability_of_BMI_explained_by_E =
    numeric(0),
  Mean_measured_variability_of_BMI_explained_by_E =
    numeric(0),
  Number_iterations = numeric(0),
  Number_iterations_stab = numeric(0),
  Mean_mean_number_predictors_found = numeric(0),
  Mean_mean_sensitivity = numeric(0),
  Mean_mean_specificity = numeric(0),
  Mean_mean_fdp = numeric(0),
  Mean_mean_R2_test = numeric(0),
  Mean_nb_sup_20percent = numeric(0),
  Mean_nb_sup_60percent = numeric(0),
  Mean_Sorensen = numeric(0),
  Mean_sd_number_predictors_found = numeric(0),
  Mean_sd_sensitivity = numeric(0),
  Mean_sd_specificity = numeric(0),
  Mean_sd_fdp = numeric(0),

```

```

Mean_sd_R2_test = numeric(0),
SD_mean_number_predictors_found = numeric(0),
SD_mean_sensitivity = numeric(0),
SD_mean_specificity = numeric(0),
SD_mean_fdp = numeric(0),
SD_mean_R2_test = numeric(0),
SD_nb_sup_20percent = numeric(0),
SD_nb_sup_60percent = numeric(0),
SD_sd_number_predictors_found = numeric(0),
SD_sd_sensitivity = numeric(0),
SD_sd_specificity = numeric(0),
SD_sd_fdp = numeric(0),
SD_sd_R2_test = numeric(0),
SD_Sorensen = numeric(0),
which_scenario = numeric(0),
Correlation = vector(),
Negative_coefficient = vector(),
Run_time = numeric(0)
)

#####
#####Run of the simulation#####
#####
n = 1
##looping on each scenario ie looping on each list of parameters which define
##scenarios
for (i4 in 1:length(neg)) {
  pos_and_neg <- neg[i4]
  for (i3 in 1:length(list_corr)) {
    corr_range <- list_corr[[i3]]
    corr = TRUE
    if (corr_range[1] == 0 & corr_range[2] == 1) {
      corr = FALSE
    }
    for (i1 in 1:length(c_n_Ey)) {
      n_Ey <- c_n_Ey[i1]
      for (i2 in 1:length(c_n_R2_fixed)) {
        R2_fixed <- c_n_R2_fixed[i2]

      ##inside the loops a scenario is set

      test_correl <- tryCatch({
        simu <-
        simulator_2layers(
          E_true = dataExp_true,
          R2_tot = R2_fixed,
          n_Ey = n_Ey,
          BetaEy = 0.1,
          test_and_training = TRUE,
          pos_and_neg = pos_and_neg,
          corr = corr,

```

```

    range_corr = corr_range
  )
}, silent = FALSE)
if (class(test_correl) == "try-error") {
  print(n)

} else{
  n_row = nrow(comp_stability_method)
  list_list_predBMI_E <- list()
  list_performance <- list()
  print("cluster")

##parallelization
start_time <- Sys.time()
cl <-
  makeClustergetOption("cl.cores", round(detectCores())))
clusterExport(
  cl,
  list(
    "dataExp_true",
    "simulator_2layers",
    "simResponseSimple",
    "estimatedR2",
    "getresiduals_2df",
    "ewas",
    "lasso",
    "lasso_stab_Meinshausen2",
    "lasso_moy_lambda",
    "lasso_moy_MSE",
    "lasso_moy_Meinshausen",
    "lasso_stab_Meinshausen",
    "Enet",
    "wqs",
    "sensitivity",
    "fdp",
    "specificity",
    "f0",
    "R2_fixed",
    "n_Ey",
    "n_iter_stab",
    "submatFindSimpl",
    "DSAreg",
    "corr",
    "corr_range",
    "pos_and_neg",
    "n_method"
  )
)
clusterEvalQ(
  cl,
  list(

```

```

library("boot"),
library("reshape"),
library("glmnet"),
library("DSA"),
library("mvtnorm"),
library("gWQS"),
library("OmicsMarkeR"),
library("Rcpp")
)
)
##applying f0 in parallel
results_1_jeu <- clusterApply(cl, 1:n_iter, f0)
stopCluster(cl)
##getting results for this scenario
simulated_data <- lapply(results_1_jeu, function(x)
  x$simu)
##formatting results for this scenario
list_list_predBMI_E <-
  lapply(results_1_jeu, function(x)
    x$list_predBMI_E)
list_performance <- lapply(results_1_jeu, function(x)
  x$performance)
remove(results_1_jeu)

##saving simulation parameters for this scenario
param_simu <-
  data.frame(
    Parameters = vector(),
    Fixed_or_measured = vector(),
    Value = numeric(0)
  )
param_simu[1, ] <-
  c("Nb_true predictors of BMI in E",
    "Fixed",
    mean(unlist(
      lapply(simulated_data, function(X)
        length(X$yE$beta)))
  )))
param_simu[2, ] <-
  c("Total variability of BMI explained by E",
    "Fixed",
    R2_fixed)
param_simu[3, ] <-
  c("Mean variability of BMI explained by E",
    "Measured",
    mean(unlist(
      lapply(simulated_data, function(X)
        (X$R2)))
  )))
param_simu[4, ] <- c("Number_iterations", "Fixed", n_iter)
param_simu[5, ] <-

```



```

(X[[i1]][[7]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[8]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[9]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[10]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[11]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[12]]))
), na.rm = TRUE),
mean(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[13]]))
), na.rm = TRUE),

```

```

sd(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[1]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[2]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[3]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[4]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[5]]))
), na.rm = TRUE),
sd(unlist(
  lapply(list_performance, function(X
    (X[[i1]][[6]]))

```

```

  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[7]]))
  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[8]]))
  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[9]]))
  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[10]]))
  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[11]]))
  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[12]]))
  ), na.rm = TRUE),
  sd(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[13]]))
  ), na.rm = TRUE),
  n,
  param_simu[6, 3],
  ifelse(pos_and_neg == TRUE, "yes", "no"),
  mean(unlist(
    lapply(list_performance, function(X)
      (X[[i1]][[14]]))
  ), na.rm = TRUE)

)
i1 <- i1 + 1
}

print(n)

list_list_list_predBMI_E <-
  c(list_list_list_predBMI_E,
    list(list_list_predBMI_E))
end_time <- Sys.time()
end_time - start_time
##saving results externally
saveRDS(comp_stability_method,

```

```
"comp_stability_method_n30_15.Rds")
saveRDS(
  simulated_data,
  file = paste(n,
    '_simulated_data_scenario_iteration_stability_n30_15.Rds')
)
saveRDS(
  list_list_list_predBMI_E,
  "list_list_list_predBMI_E_stability_n30_15.Rds"
)
saveRDS(list_list_performance,
  "list_list_performance_stability_n30_15.Rds")
remove(simulated_data)
remove(list_list_predBMI_E)
remove(list_performance)

}
n = n + 1

}
}
```