Amélioration de Scores de Risque Environmental par Machine Learning Informé et AI Explicable

Jean-Baptiste Guimbaud

Université Claude Bernard Lyon 1

Universitat Pompeu Fabra Barcelona

PhD THESIS

Conducted in cotutella at :

Claude Bernard Lyon 1 University, Department of Informatics, and
Pompeu Fabra University, Department of Medicine and Life Sciences

French Doctoral School N° 512
Informatique et Mathématiques

Specialty : Computer Science, Epidemiology

Publicly defended on 11/10/2024, by :
Jean-Baptiste Guimbaud

Enhancing Environmental Risk Scores
with Informed Machine Learning and Explainable AI

Jury :

Dr. Mohand-Saïd HACID Full professor at Claude Bernard Lyon 1 university, Lyon, France	**Chairman**
Dr. Sandra BRINGAY Full professor at Paul Valéry university, Montpellier, France	**Reviewer**
Dr. Sabina TANGARO Associate professor at university of Bari Aldo Moro, Bari, Italy	**Reviewer**
Dr. Valérie SIROUX Full professor at Grenoble Alpes university, Grenoble, France	**Examiner**
Dr. Rémy CAZABET Associate professor at Claude Bernard Lyon 1 university, Lyon, France	**Thesis supervisor**
Dr. Léa MAITRE Assistant professor at Pompeu Fabra university, Barcelona, Spain	**Thesis supervisor**

Dedication

To my beloved and our little one on the way.

Une aventure s'achève et une autre commence.
Apollon fait place à Dionysos.
Je t'aime.

Acknowledgement

First and foremost, I want to express my deepest gratitude to my PhD supervisors, Dr. Rémy Cazabet and Dr. Léa Maître, for their trust in me, their support, and their valuable guidance throughout this thesis.

Dr. Rémy Cazabet, votre expertise, votre patience, et votre soutien ont été d'une grande aide pour moi durant ces trois ans. Votre capacité à rendre limpide des concepts complexes, témoignant à la fois d'une grande maitrise technique et d'un véritable souci de l'autre, est une source d'inspiration pour moi.

Dr. Léa Maître, votre soutien, votre expertise et votre sens du détail tout au long de ces trois années, et plus encore durant la première année de ma thèse, ont été d'une valeur inestimable. Vos conseils, à la manière d'un phare, m'ont permis de naviguer avec confiance dans les eaux inconnues d'un domaine initialement nouveau pour moi.

Merci à tous les deux pour votre confiance, pour cette opportunité, et pour le privilège d'avoir pu bénéficier de votre supervision. Merci pour les innombrables heures que vous avez passées à revoir mes travaux. Vos retours ont permis d'en améliorer grandement la qualité. Je n'oublierai jamais ce que vous avez fait pour moi et j'espère pouvoir encore avoir la chance de travailler avec vous dans le futur.

I'm also deeply grateful to my corporate mentor at Meersens, Louis Stockreisser, for his friendliness and his understanding of the time pressure inherent to the doctoral work. I'm grateful to my employer, Morane Rey Huet, for the opportunity he gave me to begin this journey. I'm also not forgetting the team at Meersens, a small team of young and talented engineers, for which I'm grateful for the fun, good vibes and (endless) jokes.

I reserve the final words to my family, offering them my sincere thanks for their unwavering love and support. I'm especially grateful to my parents, who not only gave me the gift of life but also transmitted to me their moral framework. As I embark on the next chapter of my life, I aspire to provide an equally valuable nurturing environment to our child in the hopes of shaping

for him/her a future filled with love and possibilities.

Abstract

From conception onward, environmental factors such as air quality or dietary habits can significantly impact the risk of developing various chronic diseases. Within the epidemiological literature, indicators known as Environmental Risk Scores (ERSs) are used not only to identify individuals at risk but also to study the relationships between environmental factors and health. A limit of most ERSs is that they are expressed as linear combinations of a limited number of factors. This doctoral thesis aims to develop ERS indicators able to investigate nonlinear relationships and interactions across a broad range of exposures while discovering actionable factors to guide preventive measures and interventions, both in adults and children.

To achieve this aim, we leverage the predictive abilities of non-parametric machine learning methods, combined with recent Explainable AI tools and existing domain knowledge. In the first part of this thesis, we compute machine learning-based environmental risk scores for mental, cardiometabolic, and respiratory general health for children. On top of identifying nonlinear relationships and exposure-exposure interactions, we identified new predictors of disease in childhood. The scores could explain a significant proportion of variance and their performances were stable across different cohorts.

In the second part, we propose SEANN, a new approach integrating expert knowledge in the form of Pooled Effect Sizes (PESs) into the training of deep neural networks for the computation of *informed environmental risk scores*. SEANN aims to compute more robust ERSs, generalizable to a broader population, and able to capture exposure relationships that are closer to evidence known from the literature. We experimentally illustrate the approach's benefits using synthetic data, showing improved prediction generalizability in noisy contexts (i.e., observational settings) and improved reliability of interpretation using Explainable Artificial Intelligence (XAI) methods compared to an agnostic neural network.

In the last part of this thesis, we propose a concrete application for SEANN using data from a cohort of Spanish adults. Compared to an agnos-

tic neural network-based ERS, the score obtained with SEANN effectively captures relationships more in line with the literature-based associations without deteriorating the predictive performances. Moreover, exposures with poor literature coverage significantly differ from those obtained with the agnostic baseline method with more plausible directions of associations.

In conclusion, our risk scores demonstrate substantial potential for the data-driven discovery of unknown nonlinear environmental health relationships by leveraging existing knowledge about well-known relationships. Beyond their utility in epidemiological research, our risk indicators are able to capture holistic individual-level non-hereditary risk associations that can inform practitioners about actionable factors in high-risk individuals. As in the post-genetic era, personalized medicine prevention will focus more and more on modifiable factors, we believe that such approaches will be instrumental in shaping future healthcare paradigms.

Keywords: Machine Learning, Informed Machine Learning, Exposome, Environmental Risk Scores, Deep Neural Networks.

Dès la conception, des facteurs environnementaux tels que la qualité de l'air ou les habitudes alimentaires peuvent significativement influencer le risque de développer diverses maladies chroniques. Dans la littérature épidémiologique, des indicateurs connus sous le nom de Scores de Risque Environnemental (*Environmental Risk Score*, ERS) sont utilisés non seulement pour identifier les individus à risque, mais aussi pour étudier les relations entre les facteurs environnementaux et la santé. Une limite de la plupart des ERSs est qu'ils sont exprimés sous forme de combinaisons linéaires d'un nombre limité de facteurs. Cette thèse de doctorat vise à développer des indicateurs ERSs capables d'investiguer des relations non linéaires et des interactions à travers un large éventail d'expositions tout en découvrant des facteurs actionnables pour guider des mesures et interventions préventives, tant chez les adultes que chez les enfants.

Pour atteindre cet objectif, nous exploitons les capacités prédictives des méthodes d'apprentissage automatique non paramétriques, combinées avec des outils récents d'IA explicable et des connaissances existantes du domaine. Dans la première partie de cette thèse, nous calculons des scores de risque environnemental basés sur l'apprentissage automatique pour la santé mentale, cardiométabolique et respiratoire de l'enfant. En plus d'identifier des relations non linéaires et des interactions entre expositions, nous avons identifié de nouveaux prédicteurs de maladies chez les enfants. Les scores peuvent expliquer une proportion significative de la variance des données et leurs performances sont stables à travers différentes cohortes.

Dans la deuxième partie, nous proposons SEANN, une nouvelle approche intégrant des connaissances expertes sous forme d'Effet Agrégées (*Pooled Effect Size*, PES) dans l'entraînement de réseaux neuronaux profonds pour le calcul de scores de risque environnemental informés (*Informed ERS*). SEANN vise à calculer des ERSs plus robustes, généralisables à une population plus large, et capables de capturer des relations d'exposition plus proches de celles connues dans la littérature. Nous illustrons expérimentale-

ment les avantages de cette approche en utilisant des données synthétiques. Par rapport à un réseau neuronal agnostique, nous obtenons une meilleure généralisation des prédictions dans des contextes de données bruitées et une fiabilité améliorée des interprétations obtenues en utilisant des méthodes d'Intelligence Artificielle Explicable (*Explainable AI* - XAI).

Dans la dernière partie de cette thèse, nous proposons une application concrète de SEANN en utilisant les données d'une cohorte espagnole composée d'adultes. Comparé à un score de risque environnemental basé sur un réseau neuronal agnostique, le score obtenu avec SEANN capture des relations mieux alignées avec les associations de la littérature sans détériorer les performances prédictives. De plus, les expositions ayant une couverture littéraire limitée diffèrent significativement de celles obtenues avec la méthode agnostique de référence en bénéficiant de directions d'associations plus plausibles.

En conclusion, nos scores de risque démontrent un indubitable potentiel pour la découverte informée de relation environnement-santé non linéaires peu connues, tirant parti des connaissances existantes sur les relations bien connues. Au-delà de leur utilité dans la recherche épidémiologique, nos indicateurs de risque sont capables de capturer, de manière holistique, des relations de risque au niveau individuel et d'informer les praticiens sur des facteurs de risque actionnables identifiés. Alors que dans l'ère post-génétique, la prévention en médecine personnalisée se concentrera de plus en plus sur les facteurs non héréditaires et actionnables, nous pensons que ces approches seront déterminantes pour façonner les futurs paradigmes de la santé.

Mots-clés: Machine Learning, Informed Machine Learning, Exposome, Scores de Risque Environmental , Réseaux de Neurones Profonds.

Desde la concepción, factores ambientales como la calidad del aire o los hábitos alimentarios pueden influir significativamente en el riesgo de desarrollar diversas enfermedades crónicas. En la literatura epidemiológica, se utilizan indicadores conocidos como Puntuaciones de Riesgo Ambiental (*Environmental Risk Score*, ERS) no solo para identificar a individuos en riesgo, sino también para estudiar las relaciones entre los factores ambientales y la salud. Una limitación de la mayoría de los ERS es que se expresan en forma de combinaciones lineales de un número limitado de factores. Esta tesis doctoral tiene como objetivo desarrollar indicadores ERS capaces de investigar relaciones no lineales e interacciones a través de una amplia gama de exposiciones, descubriendo al mismo tiempo factores accionables para guiar medidas e intervenciones preventivas, tanto en adultos como en niños.

Para alcanzar este objetivo, aprovechamos las capacidades predictivas de métodos de aprendizaje automático no paramétrico, combinados con herramientas recientes de IA explicable y conocimientos existentes del campo. En la primera parte de esta tesis, calculamos puntuaciones de riesgo ambiental basadas en aprendizaje automático para la salud mental, cardiometabólica y respiratoria infantil. Además de identificar relaciones no lineales e interacciones entre exposiciones, identificamos nuevos predictores de enfermedades en la infancia. Las puntuaciones pueden explicar una proporción significativa de la varianza de los datos y sus rendimientos son estables en diferentes cohortes.

En la segunda parte, proponemos SEANN, un nuevo enfoque que integra conocimientos expertos en forma de Tamaños de Efecto Agrupados (*Pooled Effect Size*, PES) en el entrenamiento de redes neuronales profundas para el cálculo de puntuaciones de riesgo ambiental informadas (*Informed ERS*). SEANN tiene como objetivo calcular ERS más robustos, generalizables a una población más amplia y capaces de capturar relaciones de exposición más cercanas a las conocidas en la literatura. Ilustramos experimentalmente las ventajas de este enfoque utilizando datos sintéticos. En comparación

con una red neuronal agnóstica, obtenemos una mejor generalización de las predicciones en contextos de datos ruidosos y una mayor fiabilidad de las interpretaciones obtenidas utilizando métodos de Inteligencia Artificial Explicable (*Explainable AI* - XAI).

En la última parte de esta tesis, proponemos una aplicación concreta de SEANN utilizando los datos de una cohorte española compuesta por adultos. En comparación con una puntuación de riesgo ambiental basada en una red neuronal agnóstica, la puntuación obtenida con SEANN captura relaciones mejor alineadas con las asociaciones de la literatura sin deteriorar el rendimiento predictivo. Además, las exposiciones con una cobertura limitada en la literatura difieren significativamente de las obtenidas con el método agnóstico de referencia, beneficiándose de direcciones de asociación más plausibles.

En conclusión, nuestras puntuaciones de riesgo demuestran un indudable potencial para el descubrimiento informado de relaciones ambientales-salud no lineales poco conocidas, aprovechando los conocimientos existentes sobre las relaciones bien conocidas. Más allá de su utilidad en la investigación epidemiológica, nuestros indicadores de riesgo son capaces de capturar, de manera holística, relaciones de riesgo a nivel individual e informar a los profesionales sobre los factores de riesgo accionables identificados. Mientras que en la era post-genética, la prevención en medicina personalizada se centrará cada vez más en los factores no hereditarios y accionables, creemos que estos enfoques serán determinantes para dar forma a los futuros paradigmas de la salud.

Palabras clave: Aprendizaje Automático, Aprendizaje Automático Informado, Exposome, Puntuaciones de Riesgo Ambiental, Redes Neuronales Profundas.

Preface

The present Thesis has been funded by Meersens, Lyon, France, an industrial company, as part of the *Conventions Industrielles de Formation par la REcherche* (Cifre) program launched by the French Ministry of Higher Education, Research and Innovation (MESRI). It was carried out under the cotutelle of University Pompeu Fabra (UPF), Barcelona, Spain, and University Claude Bernard (UCBL), Lyon, France and complies with the joint procedures and regulations of the Biomedicine PhD program of the Department of Medicine and Life Sciences of the UPF and doctoral school Infomaths (ED512), Lyon, France. It is an interdisciplinary thesis in two specialties, environmental epidemiology and computer science.

The research described in this work was first carried out at the Barcelona Institute of Global Health (ISGlobal) between September 2021 and June 2022 and then, in alternation between the *Laboratoire d'InfoRmatique en Image et Systèmes d'information* (LIRIS) and Meersens between June 2022 and May 2024. It was conducted under the joint supervision of Dr. Léa Maitre and Dr. Rémy Cazabet.

The main aim of this thesis was to study environmental-health relationships and derive actionable measures for prevention and intervention through the computation of machine learning-based environmental risk scores. The Thesis contains three original research papers first authored by the PhD candidate (1 published, 2 submitted). For all the scientific papers, the PhD candidate, in collaboration with his supervisors, formulated the research objective, conceptualized the study design, performed the data management and statistical analyses, interpreted the findings, and wrote the scientific articles. Besides the three manuscripts enclosed in this Thesis, the PhD candidate has co-authored, as second author, an additional research paper describing state-of-the-art methods used during a data challenge event held in ISGlobal in April 2021 in which he competed.

During his doctoral years, the PhD candidate attended and presented his work at several international conferences, attended additional epidemio-

logical and statistical lectures, peer-reviewed scientific articles, and provided technical guidance and support for Meersens.

Few lines about Meersens Meersens is a French startup company founded in 2017 by Morane Rey-Huet and Louis Stockreisser. Their mission is to measure, analyze, and model environmental exposures geographically and provide preventive recommendations to promote health and well-being. An important part of their business lies in the development of a Software as a Service (SaaS) application that enables a stakeholder (e.g., a local authority, company, or hospital) to assess the exposure of the people for whom it is responsible and to determine the potential resulting health impacts in order to set up adapted prevention policies. The objective of Meersens through this thesis is to improve its risk assessment solution by using performant and adapted machine learning models. Details about how this work is or will be integrated into its industrial solution are not discussed in this document.

AENET-I Adaptive Elastic-Net. 18

ARMA auto regressive moving average. 25

BiB Born in Bradford. 35, 38, 45, 55

BKMR Bayesian Kernel Model Regression. 19

BPA bisphenol A. 38

BUPA n-Butyl paraben. 38

CBCL Child Behavior CheckList. 44, 56

CDC US Center for Disease Control. 25

CDSS clinical decision support system. 25

CI confidence interval. 95

CNN convolutional neural network. 25

DDE dichlorodiphenyldichloroethylene. 38

DDT dichlorodiphenyltrichloroethane. 38

DEP diethyl phosphate. 38

DETP diethyl thiophosphate. 38

DMDTP dimethyl dithiophosphate. 38

DMP dimethyl phosphate. 38

DMTP dimethyl thiophosphate. 38

DNN deep neural network. 53, 74, 75, 80, 87

ECDC European Centre for Disease Prevention and Control. 25

EDEN Etude des Déterminants pré- et postnatals du Développement et de la santé de l'ENfant. 35, 38, 45, 55

EEA European Environment Agency. 3

EHR electronic health records. 24, 35

EPA United States Environmental Protection Agency. 3

ERS environmental risk score. 9, 12, 21, 22, 44, 65, 74, 90, 108, 109

ETPA ethyl paraben. 38

EWAS Environment-Wide Association Study. 17, 21

ExE Environment x Environment. 8, 11, 16

ExWAS Exposome-Wide Association Study. 17

FAS family affluence score. 39

FDA federated data analysis. 111

FEV_1 Forced Expiratory air Volume in one second. 37, 46, 47, 52, 57

GAM generalized additive model. 19

GIS geographic information system. 41, 47

GLINTERNET Group-Lasso INTERaction-NET. 18

GWAS Genome-Wide Association Study. 17

GxE Genetics x Environment. 8, 16

HCB hexachlorobenzene. 38

HGP human genome project. 8

INMA Infancia y Medio Ambiente. 34, 35, 38, 45, 55

KANC Kaunas Cohort. 34, 35, 45, 55, 64

LASSO Least Absolute Shrinkage and Selection Operator. 18, 20, 53, 56, 57

MAE mean absolute error. 82

MBzP mono benzyl phthalate. 38

MECPP mono-2-ethyl 5-carboxypentyl phthalate. 38

MEHHP mono-2-ethyl-5hydroxyhexyl phthalate. 38

MEHP mono-2-ethylhexyl phthalate. 38

MEOHP mono-2-ethylhexyl phthalate. 39

MEP mono-2-ethylhexyl phthalate. 39

MEPA methyl paraben. 38

MET metabolic equivalent of task. 40

MetS Metabolic Syndrome. 37, 44, 51, 52, 57

MiBP mono-2-ethylhexyl phthalate. 39

ML machine learning. 13, 24, 27

MnBP mono-n-butyl phthalate. 39

MoBa Norwegian Mother, Father and Child Cohort Study. 34, 35, 38, 45, 55

NDVI Normalized Difference Vegetation Index. 41, 95

NO_2 nitrogen dioxide. 41, 47, 95

O_3 ozone. 41, 95

OC organochlorine compound. 38

oh-MiNP mono-4-methyl-7-hydroxyoctyl phthalate. 39

OP organophosphate pesticide. 38

OR odd ratio. 10, 74, 75, 99, 109

OXBE oxybenzone. 38

oxo.MiNP mono-4-methyl-7-oxooctyl phthalate. 39

PAQ Physical Activity Questionnaire. 40

PBDE polybrominated diphenyl ether. 38

PCA principal component analysis. 21

PCB polychlorinated biphenyl. 38

PES pooled effect size. 5, 10, 12, 30, 74–76, 80, 86, 87, 90, 109

PFAS per- and polyfluoroalkyl substance. 39

PFHxS perfluorohexane sulfonate. 39

PFNA perfluorononanoate. 39

PFOA perfluorooctanoate. 39

PFOS perfluorooctane sulfonate. 39

PFUnDA perfluoroundecanoate. 39

PM_{10} particulate matter with an aerodynamic diameter of less than 10 μm. 41

$PM_{2.5}$ particulate matter with an aerodynamic diameter of less than 2.5 μm. 41, 95

PRPA propyl paraben. 38

PRS polygenic risk score. 9, 22, 65, 74

RCT randomized controlled trial. 5, 7

RHEA Mother-Child Cohort in Crete. 34, 35, 38, 45, 55

RNN recurrent neural network. 25

RR risk ratio. 10, 74, 75, 99

SARIMA seasonal autoregressive integrated moving average. 25

SEANN Summary Effects Adapted Neural Network. 11, 12, 14, 74, 76, 78, 97, 101, 106, 109

SHAP SHapley Additive exPlanations. 27, 31, 53, 57, 61, 111

SNP single nucleotide polymorphism. 22

TCS triclosan. 38

THM trihalomethane. 39

WHO World Health Organization. 3, 56

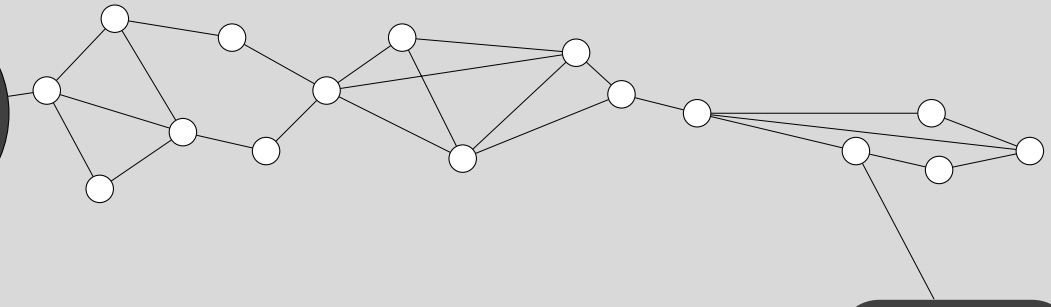
WQS Weighted Quantile Sum. 20

Contents

List of Figures

List of Tables

1



Introduction

1.1 Context and motivation

1.1.1 The Epidemiological Transition and the Importance of Environmental Health Initiatives

Over the past century, technological progress has deeply reshaped our societies. One of the most important changes we can observe is the massive increase in life expectancy (multiplied by 2.5 in developed countries) over the past 150 years. Not only do we live longer, but we are also in better health than our ancestors [Riley, 2001]. In 1971, Abdul Omran explained this phenomenon with the *epidemiological transition* theory [Omran, 2005]. In essence, the idea is that the increase in life expectancy is largely due to a shift from deaths due to infectious diseases such as pneumonia or influenza, to deaths attributed to chronic diseases like cancer or cardiovascular diseases from which we mainly die at an older age. While pandemics like Covid-19 are still major threats to public health, the reduction of mortality from infectious diseases, once the leading cause of death worldwide, combined with the augmentation of the chronic disease burden marks a profound shift in modern public health challenges. For instance, in the US, infectious diseases were responsible for 47% of all deaths at the beginning of the 20th century [Armstrong, 1999] compared to only 5.4% from 1980 to 2014 [Hansen et al., 2016]. In contrast, chronic diseases are now responsible for 74% of deaths worldwide [Thomas et al., 2023].

This shift was facilitated by significant advances in medical science, notably including the development of vaccines and antibiotics that effectively combat infectious diseases. Simultaneously, the scale of the human impact on the environment has escalated to the point that scientists such as Crutzen relate to a new geological era, the Anthropocene, defined by human activity's profound effect on the Earth to designate our epoch [Crutzen, 2006]. While the term itself can be controversial, we know that the recent history of mankind is marked by a more pronounced release of pollutants into the environment, a consequence of industrialization, urbanization, and other human endeavors, which have proven to impact human health [Fuller et al., 2022].

Unlike infectious diseases, for which the cause is simple and direct, i.e., exposure to an infectious agent, chronic diseases often result from the ac-

cumulation of multiple low-dose exposures to various environmental factors combined with a genetic liability. A recurring phrase in environmental science put it this way: "*Genetics load the gun, environment pulls the trigger*". The first statement of it should be accredited to MD Elliott Joslin in the early 20s'. While it is an excellent punchline, a sentence like "*Genetics set the board, environment makes the moves*" would be less fatalistic—the environment can also have positive effects on health—and would better highlight the actionable property of the environment in opposition to the genome. Moving forward, as the growing understanding of the impact of long-term exposure to environmental factors spreads outside the scientific community, there has been an increase in public interest in this research area.
In such context, environmental health assessment has emerged as a cornerstone for public health initiatives, aiming to quantify and mitigate the adverse effects of environmental hazards on human health. Over the past decades, several governmental agencies worldwide, such as the World Health Organization (WHO), the United States Environmental Protection Agency (EPA), or the European Environment Agency (EEA), have been tasked with developing and enforcing environmental health policies and regulations. The significance of chronic diseases was, for instance, recognized in the United Nations 2030 Agenda for Sustainable Development, which set targets to "*reduce by one-third premature mortality from noncommunicable diseases through prevention and treatment*" [Cf, 2015].

1.1.2 About environmental health

Environmental health is a field of research aiming at studying the effect of exposure to environmental factors on health. Those factors refer to any external substance or condition that an individual comes into contact with, which could potentially affect their health. Exposures to such factors can include a wide range of physical, chemical, biological, and social elements, such as air and water pollutants, radiation, infectious agents, dietary components, and socioeconomic conditions. Their effects can be direct and obvious (e.g., car accidents, gunshots, lethal dose injection of poisonous substances) but also very subtle and not visible without adequate tools, for instance in the case of long-term and repetitive exposures to low concentrations of air pollution. While typical approaches in modern medicine are able to identify the biological mechanisms that caused the illness, they are unable to iden-

tify their cause. Environmental health science aims to identify such distal causes (i.e., the cause of the cause).
The field is traditionally studied through three distinct types of approaches. Environmental epidemiologists are looking for associations at the population level, exposure scientists at the individual level, and toxicologists at the molecular-cellular level [Miller and Jones, 2013]. All three disciplines are naturally interconnected and rely on each other. For instance, toxicologists rely on epidemiological exploratory studies to target relevant exposure to assess, while epidemiologists rely on toxicological studies to verify if their observed associations are in line with the known biological pathways. In this thesis, we were interested in environmental-health relationships at the epidemiological level.

1.1.3 About environmental epidemiology

In their population studies, environmental epidemiologists generate statistical estimates about the effect on health of some environmental factors. Those are generally in the form of associations, i.e., statistical relationships, between a variable of interest (called the explanatory or the independent variable) and a health outcome (called the response or the dependent variable). Such associations are qualified as positive when an increase in one variable (such as smoking) is associated with an increase in another variable (such as the prevalence of lung cancer) or negative when the increase in one variable is associated with a decrease in the other. In practice, an association is quantified as a single value estimate, such as a beta coefficient in a linear regression model or an odds ratio in a logistic model, which represents the effect of the predictor on the outcome. Additionally, these estimates are commonly accompanied by a measure of statistical significance (i.e., the p-value), which quantifies the likelihood that the observed relationship could have occurred by chance.
Those associations can be obtained from different study designs that yield different levels of evidence in terms of strength and reliability. Those different levels are traditionally ranked from bottom to top in a pyramid (Fig.1.1) [Wallace et al., 2022]. The first and stronger category of evidence in this pyramid is *Filtered Information*, where evidence has been critically evaluated and synthesized from multiple studies. Those include meta-analyses, systematic reviews, and guidelines developed by panels of experts based on

previous studies. **Understanding the concept of meta-analysis is important in this thesis**. Meta-analyses are studies that aim to provide a more precise estimate of the effect size, i.e., statistically stronger than the results of any single study, by pooling association estimates between a given predictor on a given outcome from several studies [Borenstein et al., 2021]. The pooled effect size (PES) estimate is obtained by weighting the estimate in each study by its statistical significance. In practice, this is generally performed by using each study sample size or the inverse of the estimate's variance.

Considered less robust in general, *Unfiltered Information* represents evidence obtained from a single analysis. This is obviously the main source of knowledge in the field in terms of quantity. The different study designs encompassed in this category have been ranked according to their evidence quality [Wallace et al., 2022] from the highest to the lowest:

- **Randomized controlled trials (RCTs)**. An experiment involving one or a few variables in which participants are randomly assigned to either a treatment or control group to rigorously test the effects of an intervention. While being the gold standard in this category, they are costly to implement and can be impracticable (when the exposure is hard to administrate) and/or unethical (for instance when testing prenatal potential harmful effects).

- **Cohort studies** consist of following a group of people (a cohort) over a period to see how different exposures affect their outcomes. Participants are not assigned by the researcher to exposed or non-exposed groups; rather, they are observed based on their real-life exposure statuses. While they can track changes over time and study exposures that would be unethical to assign deliberately, they are also more susceptible to confounding factors and bias as the assignment to exposure is not controlled by the researcher. In this thesis, we used this design in Paper 3.

- In **cross-sectional studies**, data are collected from a population at a single point in time. The focus is on assessing the prevalence of an outcome or a particular set of variables within the study population at that specific time. Compared with cohort studies, since the data are collected at one point in time, it is difficult to ascertain temporal

sequences or causality between variables. In this thesis, we used this design in Paper 1.

- Finally, the lower level of evidence is **case-control studies**, where researchers start with an outcome (i.e., the presence or absence of a disease) and then look backward to find exposures or risk factors. They select a group of individuals who have the disease (cases) and a group without the disease (controls), and then compare their past exposures. While being less costly than cohort studies and requiring less time (the outcome has already occurred), they are prone to bias, especially when participants do not recall precisely their past exposures.

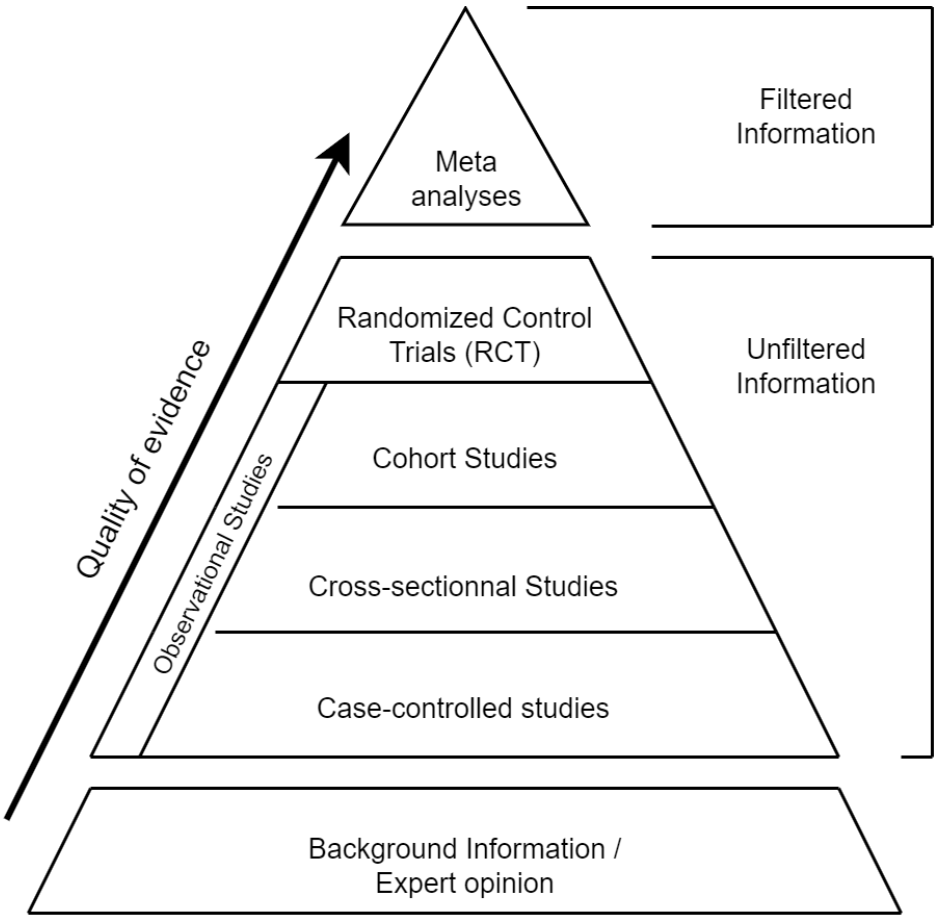


Figure 1.1: Evidence pyramid

1.1.4 Confounding in epidemiology

As mentioned in the previous section, this thesis leverages observational data. An important source of bias in observational studies is called *confounding*. It occurs when a variable (the confounder) influences both an exposure under study and the outcome, leading to the measure of a spurious association. This phenomenon can be difficult to discern when no data about the confounder is available.

Let us illustrate this phenomenon with an example. The consumption of fish (moderate) is known to influence health positively in general as it is filled with omega-3 fatty acids and vitamins such as D and B [Stratakis et al., 2020]. However, it is also a major source of exposure to heavy metals such as mercury, which are notably detrimental to the nervous system. Hence, if not accounted for correctly, the effect of mercury on health could be confounded by healthy fish consumption and appear as a protective factor.

Several approaches are possible to mitigate the impact of confounding, such as randomly assigning individuals to exposure and control groups in RCTs. In observational studies, common approaches imply stratifying study populations according to known confounders or using multivariate models (such as multivariate regression) that incorporate (i.e., adjust for) the confounders' effects.

1.1.5 The exposome paradigm

Historically in epidemiological studies, the impact of environmental health associations was largely studied using a 'one-exposure-one-health-effect' approach [Vrijheid, 2014]. While such targeted approaches are still useful, scaling them to the diversity of existing environmental factors is expensive. Additionally, they can be subjected to unaccounted confounding and they may miss subtle, unattended effects (e.g., potential interactions with other exposures or genetic liabilities). To address those challenges, a new research paradigm was needed.

Concerned about a lack of measures and adequate tools for researchers to use in order to explore and identify new environmental exposures, Dr. Christopher Wild, an epidemiologist, coined the term 'exposome' in 2005 as an environmental equivalent to the human genome and an indispensable complement regarding its impact on health [Wild, 2005]. Dr. Wild wanted

to draw attention to the need for better and more complete environmental exposure data, in order to balance the investment, tools and knowledge in genetics. At the time, significant advancements were made regarding the mapping and understanding of the human genome. The human genome project (HGP), a massive international project aiming at mapping the complete set of nucleotides contained in the human reference genome, had been completed just two years earlier. This project is still, to that day, the largest collaborative project in biological science.

Wild defined the exposome as a holistic view of *"all exposures from conception onward, including those from lifestyle, diet and environment"* that complements the genome. In a subsequent publication [Wild, 2012], he further divided the exposome into three main complementary yet distinct domains: the specific external exposome (encoding the immediate local environment such as air quality, radiation, infectious agents, professional occupation and lifestyle), the general external exposome (the socio-economic setting) and the internal exposome (encoding measurements of the body response to exposures). This categorization aimed to better federate a variety of research fields, ranging from toxicology to epidemiology, ecology, and the social sciences.

The exposome as a research paradigm recognizes that individuals are simultaneously exposed to a multitude of environmental factors and takes a holistic approach to the discovery of etiological factors for disease. Its main advantage over traditional "one-exposure-one-disease" approaches is that it provides a conceptual framework for investigating multiple environmental hazards (e.g., urban, chemical, lifestyle, social) and their combined effects. Classical single exposure analyses may be limited as the studied exposure association could arise from another correlated factor not taken into account and are, moreover, unable to capture interactions or cumulative effects from the exposure mixture. Exposures are not isolated, can be correlated with one another, and are likely to interact both among themselves, i.e., Environment x Environment (ExE) and with genetics, i.e., Genetics x Environment (GxE) to drive health and non-communicable diseases [Jaffee and Price, 2008, Johns et al., 2012].

To address the exposome's inherent challenges, a wide range of environmental data coupled with clinical biomarkers is needed to comprehensively capture its main domains. Consequently, advanced modeling approaches, suit-

able for the analysis of complex mixtures of exposures in observational and clinical studies, are needed to process such data. In addition to traditional biostatistical methods, recent advances in explainable AI allow extracting and making intelligible the relationships captured by powerful predictive algorithms (e.g., [Guidotti et al., 2018]), whose use was previously limited to prediction and forecast only.

1.1.6 Environmental risk scores

Inspired by risk prediction models, such as the Framingham risk score for coronary heart disease [D'Agostino et al., 2008] or polygenic risk scores (PRSs) [Khera et al., 2018] from genetic research, ERS are summary measures of the effects of multiple exposures used to estimate the environmental liability for a particular health outcome at an individual level [Park et al., 2014]. Those scores are useful tools for screening individuals to select for more expensive testing, and, more importantly, they can be used to study the effect of environmental exposure on health [Park et al., 2014].
ERSs are usually built as a weighted sum of the individual exposure estimates, obtained either from previous literature studies (e.g., from meta-analyses) or derived through linear regression models (from either single multivariate models or several single exposures models, adjusted for multiple testing, c.f. **Section 2.2.1**) [Pries et al., 2021]. This scheme, however, assumes that each environmental stressor individually acts in a linear dose-response relationship, while previous research has shown that their combined effect does not necessarily follow this rule [Le Magueresse-Battistoni et al., 2018].
Unlike genetics, some environmental factors are actionable, which gives ERSs a broader potential for informing public health policies by identifying actionable key factors that facilitate the implementation of preventive measures. Combined with PRSs, these scores could also serve as an initial step in identifying at-risk populations, who can then be directed to more specific clinical diagnostic tests, thereby serving as a complementary tool in the healthcare decision-making process [Murray et al., 2021, Wray et al., 2021]. By giving recommendations at an individual level, ERSs combined with PRSs are a first step towards personalized medicine [1].

[1] Considered the future of healthcare, personalized medicine is a shift from the "one-size-fits-all" approach to a more precise and patient-centered one, performed by taking

1.1.7 Incorporating Domain Knowledge in Exposome Studies

As recent research on the exposome involves more and more observational studies collecting large varieties of environmental factors, there is a recognized need for data-driven methods able to handle large amounts of variables for analysis [Miller and Jones, 2013, Haddad et al., 2019]. However, purely data-driven approaches can lead to unsatisfactory results, such as capturing spurious associations or poorly generalizable performances. Such results are more likely to occur when facing a small available sample size for the training data, high measurement errors, or confounding effects. Additionally, purely data-driven methods do not necessarily follow important guidelines about security or fairness nor necessarily satisfy known natural laws (e.g., biological pathways). Incorporating additional knowledge to complement the training data is a way to address those issues that we explored in this doctoral work.

The information gathered in previous studies about exposure-health relationships is a particularly useful form of knowledge for studying the exposome concerns. Incorporating the known relationships into the machine learning models can help better capture less studied associations in case of confounding, for instance, using the firsts to adjust the seconds. Considered one of the most reliable forms of knowledge in epidemiological studies [Rosner, 2012], PESs represent well-known exposure effects aggregated across several studies in meta-analyses [Pathak et al., 2020], as previously mentioned in section 1.1.3. In practice, these estimates are generally encoded as odd ratios (ORs), risk ratios (RRs), or simple linear estimates (cf., section 2.2.1).

1.2 Objectives

Our overarching goal in this work was, in line with the exposome concept, to study the combined effects of environmental exposures on human health. More specifically, our work focused on the exploratory analysis of observational cohort data, examining a broad range of exposures. It was not specific to any health outcome in particular but rather on proposing new ap-

into account genetic, environmental, and lifestyle factors.

proaches for the untargeted discovery of environmental health relationships. To achieve this, we leveraged the predictive abilities of non-parametric machine learning methods, which are still uncommon in the field, combined with the recent advancements in Explainable AI and the availability of existing domain knowledge to derive informed summary risk scores aiming to address some of the exposome's challenges, including:

- The extraction of complex non-linear exposure-health relationships.
- The disentanglement of the marginal effects on health of exposures within intricate mixtures.
- The extraction of ExE interactions (i.e., synergies or "cocktail-effects" within exposures).
- The learning and the extraction of plausible effects according to the literature's established knowledge.
- The identification of individuals at risk for further targeted testing and prevention.
- The identification of actionable factors of disease for preventive actions.

This thesis being co-supervised by two universities specialized in different fields, we separate our research objectives into two main types, methodological and applicative. Our methodological objectives were:

1) To demonstrate the benefits of using highly expressive but complex models combined with adequate explainability tools to study exposome-health relationships *(Paper 1)*.

2) To develop a new approach, i.e., Summary Effects Adapted Neural Network (SEANN), that incorporates relevant scientific knowledge into highly expressive predictive models to provide informed environmental risk scores and to demonstrate the benefits of this approach in a controlled environment by using generated data *(Paper 2)*.

3) To demonstrate the benefits of this approach compared with knowledge-agnostic risk scores (similar to those mentioned in point 1) on real observational data *(Paper 3)*.

Our applicative objectives were to explore complex relationships and interactions on a wide range of exposures while being able to identify individuals at risk and actionable factors to guide preventive measures and intervention in both adults (*Paper 3*) and children (*Paper 1*).

1.3 Contributions

In more detail, the contributions presented in this work are:

- **New machine learning-based early life environmental risk scores for the European population:** In this work (*Paper 1*), we computed environmental risk scores for mental, cardiometabolic, and respiratory general health using machine learning and explaining AI for mother and child pairs in the HELIX cohorts. Compared with the previous HELIX cohort studies, our approach identified new important predictors of disease in childhood on top of identifying nonlinear relationships and exposure-exposure interactions. The scores could also explain a significant proportion of variance (meaning that they had some predictive value), and their performances were stable across all six cohorts, meaning that they were generalizable across different European countries. Raw performances however were comparable with a traditional method.

- **SEANN, a novel informed machine learning approach for the computation of environmental risk scores:** In this work (*Paper 2*), we proposed SEANN, a new approach integrating external knowledge into the training of deep neural networks for the computation of informed ERSs. This approach integrates literature-based PESs to the training of DNNs, which are estimates considered to be one of the best levels of evidence in epidemiological studies. SEANN aims to compute more robust ERSs, generalizable to a broader population, and able to capture exposure relationships that are closer to the known evidence. Using the available knowledge about well-known relationships, SEANN can better capture those that are still poorly studied. In this work, we experimentally illustrated the approach's benefits using synthetic data only.

- **An informed machine learning risk score for hypertension in adults:** In this work (*Paper 3*), we proposed a concrete application for SEANN using data from GCAT, a cohort of Spanish adults. We also refined our approach by proposing another way to determine the relative weights of the literature knowledge with regard to the available data. Compared to an agnostic neural network-based ERS, the

score obtained with SEANN effectively captured relationships that were more in line with the known meta-estimates without deteriorating the predictive performances. Additionally, the exposures with poor literature coverage significantly differed from those obtained with the agnostic NN with more plausible directions of associations. Similarly to our previous work on the HELIX cohorts, we identified important actionable environmental factors of diseases and nonlinear associations between a wide range of environmental factors and hypertension.

1.4 Thesis structure

This document is organized according to the three main papers produced during the three years of the PhD. Before presenting the contributions mentioned in the previous section, we provide an overview of related work previously performed in the literature to help the reader understand where our work stands in the exposome literature.

- **Chapter 2 - Related work** A literature overview of previous work is described in this chapter, which first presents the methods developed to study the exposome and compute environmental risk scores. Additionally, besides the traditional methods of biostatistics, we provide an overview of the use of machine learning models in healthcare and, more particularly, in studies of the exposome. We then discuss previous works that have been proposed to integrate domain knowledge in machine learning methods, in particular in healthcare.

- **Chapter 3 - Methods.** In this chapter, we propose a quick presentation of the data sources, study designs, study populations, health outcomes and exposures used in this thesis.

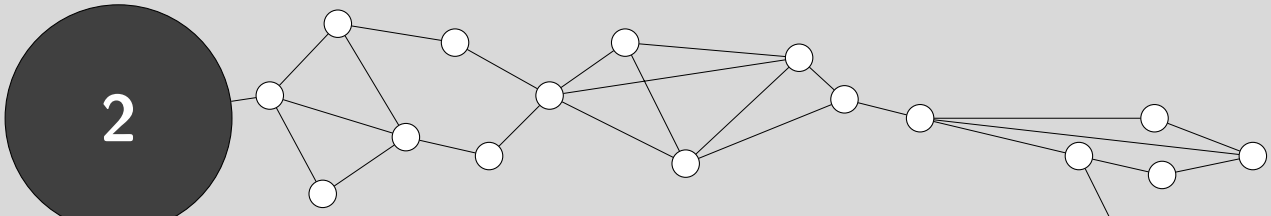
- **Chapter 4 - Machine Learning Based Environmental-Clinical Risk Scores in Children.** In this chapter, we relate the work performed in (*Paper 1*), where we proposed early-life environmental risk scores for various health outcomes in European populations computed using a combination of existing non-parametric machine learning (ML) methods and Explainable AI tools. This approach, still novel in the field, allowed the identification of new exposure-health relationships missed in previous studies on the HELIX cohorts (i.e., European

mother-child cohorts). However, challenges such as the limited sample sizes typically provided in such observational studies on a wide variety of factors prevented leveraging its full potential.

- **Chapter 5 - A Deep Learning Approach for Informed Environmental Risk Scores.** In this chapter, we relate the work performed in *(Paper 2)*, where we presented a novel informed machine learning approach, namely SEANN, that incorporated literature-extracted effect size estimates for computing environmental risk scores. Using synthetic data, we provided an experimental illustration of the approach's benefits in a controlled environment.

- **Chapter 6 - An Informed Environmental Risk Score for Adult Hypertension.** In this chapter, we relate the work performed in (*Paper 3*), where we leveraged SEANN for the computation of an environmental risk score of hypertension in adults based on the GCAT Spanish cohort. The paper further demonstrates the approach's benefits in a real setting, notably identifying exposure to health relationships more aligned with literature knowledge.

- **Chapter 7 - Conclusion and future directions**



2 Related work

2.1 Introduction

As research in environmental epidemiology increasingly focuses on complex mixtures—that is, multiple exposures analyzed simultaneously—there has been a corresponding adaptation in analytical tools to meet the challenges introduced by this paradigm shift. This chapter provides an overview of those tools and methods and explores promising new directions. While this is not an exhaustive review of all the methods ever used in the field, we aim to provide the necessary background to understand the types of methods typically used during recent years, their limitations, and the pertinence of the contributions proposed in this doctoral work.

We organize this overview of methods into three main parts: First, we focus on the biostatistical methods used to study the exposome, which are the most commonly used but have limitations. Second, we discuss the incorporation of non-statistical machine learning methods originating from various computer science disciplines in healthcare and, more specifically, in exposome studies. Those have become more prominently used over the past few years. Finally, we discuss the integration of domain knowledge in healthcare and its potential to address various challenges of the exposome.

2.2 Statistical Methods for Studying the Exposome

Statistical methods have been widely used over the past few years to study the effects of environmental exposure on health. Initially tailored to study the effects of a limited number of pre-selected exposures, some of those methods had to be adapted for exposome-wide approaches, cf. **Section 1.1.5**. Extensive literature describes those methods, e.g., [Billionnet et al., 2012, Sun et al., 2013, Stafoggia et al., 2017, Barrera-Gómez et al., 2017, Oskar and Stingone, 2020, Maitre et al., 2022b]. In this section, we organize these into two research axes, namely: 1) Methods measuring the individual effects on health and 2) methods measuring Environment x Environment (ExE) and/or Genetics x Environment (GxE) interactions.

2.2.1 Methods for studying marginal effects of exposure on health

In the following subsections, we present an overview of the methods used to study health relationships from a group of exposures. These relationships can be analyzed in two ways: independently cf. **subsection 2.2.1.1**, disregarding the influence of other factors, or simultaneously cf. **subsection 2.2.1.2**, adjusting for the effects of each exposure. When the number of exposures is large, specialized techniques may be required to filter the relevant information cf. **subsection 2.2.1.3**. Additionally, predictive models designed to identify individuals at risk based on their exposure profiles can be used to explore these relationships cf. **subsection 2.2.1.4**.

2.2.1.1 Methods for estimating effects independently: EWAS and ExWAS

Environment-Wide Association Studies (EWASs) [Patel et al., 2010] is an approach designed to estimate the impact of a diverse array of environmental factors in a high-dimensional setting. It is adapted from Genome-Wide Association Studies (GWASs) [Chang et al., 2018], an approach widely used for analyzing high dimensional genomic data. Subsequently, [Rappaport, 2012] proposed Exposome-Wide Association Studies (ExWASs), a similar approach that includes biomarkers of exposures and diseases in order to encompass all domains of the exposome. Both methods agnostically explore environmental health associations within wide mixtures of exposures. Their primary aim is to discover untargeted associations at the population level that could subsequently be further validated at the toxicological level.
EWASs and ExWASs typically designate studies where each environmental exposure is individually tested to determine its association with a health outcome [Zheng et al., 2020]. This is usually done with a logistic or linear regression model. As many tests are conducted simultaneously, which increases the probability of type I error, those approaches subsequently apply a correction for multiple testing (i.e., a Bonferroni or Benjamini-Hochberg correction) on the significance thresholds (i.e., the p-values).
EWAS and ExWAS approaches have limitations due to their reliance on single-pollutant models. In particular, they are unable to properly adjust each exposure effect relative to the others and disentangle each effective

contribution. In addition, EWAS and ExWAS face significant challenges compared to their genetic analogs. While genetic factors are fully measurable with recent genome sequencing approaches [Petersen et al., 2017], environmental factors are diverse in nature and must be measured from different sources and methods. Consequently, those approaches might be facing issues such as limited exploitation of data sources, high heterogeneity in analytical approaches, and lack of replication across the studies [Zheng et al., 2020].

2.2.1.2 Methods for estimating effects simultaneously

Linear methods have the advantage of being easily understandable by the human mind. They measure simple estimates encoding the effect of a factor on health in a single coefficient and assume that each effect is additive. Often based on linear or logistic regression, such methods usually incorporate a variable selection mechanism. Prominent examples are the Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani, 1996] and Elastic-Net [Zou and Hastie, 2005] methods, which add a penalty term to the optimization function. This penalty leads to a selection of the most impactful effects via a shrinkage of weakly associated coefficients toward zero. Some studies leverage more elaborated methods such as Group-Lasso INTERaction-NET (GLINTERNET) [Lim and Hastie, 2015] that captures pairwise interactions or Adaptive Elastic-Net (AENET-I) [Zou and Zhang, 2009] that provides linear estimates with desirable properties. Penalized methods perform well in high-dimensional settings and are an attempt to handle the problem of multi-collinearities, where a model cannot disentangle the effects among highly correlated factors. However, as they tend to provide biased estimates [Tibshirani, 1996], some researchers adopt a two-step procedure where variable selection is first performed via weight shrinkage (i.e., using Lasso or Elastic-Net), and then exposures with non-null estimates are studied in a standard regression procedure [Gibson et al., 2019].
Obviously, assuming that exposure-health relationships are linear is a strong assumption. It means that each unit increase in exposure is associated with a constant change in the health outcome, regardless of the exposure level. For instance, if we assume a linear relationship between air pollution and respiratory health, we would predict the same incremental impact of pollution on health, whether the air quality is slightly below or far above

the safety threshold. This can be a severe limitation because, in reality, health effects may increase disproportionately at higher levels of exposure. Consequently, other methods have been developed to capture more complex relationships.

Nonlinear methods are less commonly used in exposome studies due to their increased complexity, higher computational demands, larger data requirements, and the historical precedent and interpretability advantages of linear methods. Two prominent approaches from biostatistics are Bayesian Kernel Model Regression (BKMR) and generalized additive models (GAMs) [Stafoggia et al., 2017]. Similar to Gaussian processes [Quadrianto et al., 2011], BKMR uses a Gaussian kernel to estimate the exposure-outcome relationship as a nonlinear function. BKMR also includes a variable selection process, directly embedded into its Bayesian framework, by estimating the posterior probabilities that each variable has an effect on the outcome. This variable selection can help with groups of highly correlated exposures by selecting only one representative among them (not necessarily one with a causal effect). GAMs express the overall effect on health as a weighted sum of the individual effects, similar to linear regression, but using nonlinear functions (generally splines or a set of basic functions). Both approaches, however, can be computationally demanding when the number of predictors increases. BKMR can have convergence issues and be sensible to prior choices, while GAMs assume that effects are additives.

2.2.1.3 Methods for addressing the curse of dimensionality

A major challenge associated with analyzing exposure-health associations across a large number of exposures is commonly referred to as the "curse of dimensionality". This "curse" refers to poor performances due to a low signal-to-noise ratio, where the amount of useful information for the task is diluted into the number of factors that are often unrelated. Hence, as the attentive reader would have already noticed in the previous subsections, a lot of the methods used in exposome studies aim to reduce this dimensionality by using different approaches.
While some already presented methods use a variable selection process embedded within a modeling procedure, other ones perform only data selection, and can thus be used in conjunction with any type of model. The Deletion-

Substitution-Addition (DSA) algorithm [Sinisi and van der Laan, 2004] is a popular iterative feature selection procedure originally developed for omics data analysis that has been adapted for environmental observational data. DSA iteratively selects the most significant exposures through deletion, substitution, or addition steps based on the predictive performance of predictive models (originally generalized linear models). Such iterative procedures can be very computationally demanding as the number of variables from which performing the selection increases and, similar to penalized regression approaches, they provide no guarantee of selecting predictors with an actual effect. A recent approach tries to address this limitation by incorporating external knowledge on existing causal relationships into a LASSO-performed feature selection [Frndak et al., 2023].

In the same vein as variable selection, another approach consists in grouping similar exposures together and computing the effects on health of those groups/families. Weighted Quantile Sum (WQS) regression is an index-based regression designed for environmental exposure analysis [Carrico et al., 2014]. An index is a single numerical measure summarizing the effects of a group of exposures. WQS builds this index as weighted sums using a linear regression scheme. More recently, [Masselot et al., 2022] proposed the constrained groupwise additive index model (CGAIM) using the same underlying principle while being able to capture nonlinear relationships. A similar approach is the group Lasso, an extension of the Lasso technique where nonoverlapping variables are grouped based on prior knowledge [Yuan and Lin, 2005]. Group Lasso performs variable selection and regularization across these groups rather than individual variables.

Instead of grouping exposures, some methods aim to identify distinct exposure profiles among individuals. They categorize them into groups where members within the same group have similar exposures while ensuring that these exposures differ significantly from those in other groups. Once the grouping of individuals has been done, usually using unsupervised algorithms such as k-means [Lloyd, 1982] or hierarchical clustering [Nielsen, 2016], these methods typically consist of leveraging the indicators of group membership as predictors in a regression procedure for a health outcome. Another approach consists of selecting one observation—called a prototype—to represent each cluster [Reid and Tibshirani, 2015]. These prototypes are then used in a classical statistical analysis, often employing LASSO

or EWAS, using post-selection inference theory [Tibshirani et al., 2016] to compute exact p-values and confidence intervals.
Lastly, a range of methods has been used to summarize the exposure matrix in a lower dimensional space. Most commonly used methods within this scheme are principal component analysis (PCA) and its supervised variant [Bair et al., 2006]. PCA computes a new set of variables, called principal components, which are ordered so that the first few retain most of the variance present in the original variables. Supervised PCA is an extension of PCA incorporating health outcomes into the dimensionality reduction process to better capture the variance relevant to the predictive task. Positive Matrix Factorization (PMF) [Paatero and Tapper, 1994] is a similar method commonly used to estimate source apportionment of air pollutants, more specifically the source profiles (i.e., the type of emission) and the source contributions (i.e., the amount of emitted pollution), constrained to be positives. Such source estimates are used in several epidemiological studies [Krall and Strickland, 2017, Zhang et al., 2024]. PMF, similarly to PCA (and SPCA), assumes linear relationships between variables.

2.2.1.4 Methods for computing Environmental Risk Scores

Using similar methods to those presented in **Section 2.2.1.2**, a subset of studies, also interested in identifying individuals at risk for a given health condition, propose building predictive models to estimate risk indicators while exploring exposure-health relationships. Such indicators are referred to as Environmental Risk scores (ERSs). Understanding what those scores are and how they are typically computed is important in this thesis, as our main contribution is a new approach for estimating them.
Historically, predictive models have been used to predict health liabilities according to various health factors for years. The first proposed health risk scores were developed as cost-effective tools for population screening applicable in clinical settings using few variables. For instance, in 2008, the Framingham risk score for coronary heart disease was developed as a simple tool predicting a person's liability from a restricted set of factors that are easily measurable in clinical settings, such as age, blood pressure, or smoking status [D'Agostino et al., 2008].
Health risk scores encompassing a wide variety of environmental factors are more recent. Inspired by the progress in genetic epidemiology, in which high-

dimensional polygenic risk scores (PRSs) were developed to assess genetic liabilities [Khera et al., 2018], [Park et al., 2014] proposed high-dimensional environmental risk score as a new tool for studying multi-pollutant effects [1]. According to their definition, ERSs in epidemiology should not only discriminate the individuals at risk of developing a disease but also point out potential unexplored environmental effects on health.

Similarly to PRSs, ERSs are typically built as simple weighted sums of linear individual estimates of exposure liability [Pries et al., 2021]. While those scores can be directly built from data using regression procedures such as logistic regression, LASSO, or ExWAS (e.g., [He et al., 2023]), this scheme also allows ERSs to be built based on the estimates of previous studies [Pries et al., 2021, Vassos et al., 2019, Padmanabhan et al., 2017]. The underlying idea behind literature-based ERS is to leverage large populations used to derive estimates of environmental health associations from previous studies to establish a robust score that can be validated on a smaller population. Examples of such literature-based ERSs include [Padmanabhan et al., 2017, Vassos et al., 2019, Mas et al., 2020].

In this thesis, we propose an approach that allows the computation of more powerful ERSs. Derived from machine learning methods, those ERSs can more precisely discriminate the individuals at risk. They can also be used to explore more complex exposome-health relationships (e.g., nonlinearities) using explainable tools. Finally, they can incorporate knowledge from previous large-scale studies by using our informed machine-learning approach.

2.2.2 Methods for estimating interaction effects

In this section, we focus on methods that study the interactions, i.e., the synergies or "cocktail effects" that could arise between the effects on health of several combined exposures, or even between environmental exposures and omics factors.

A straightforward method for assessing those is to add additional multiplicative terms in linear regression procedures. However, even when considering

[1] Their proposed ERSs, computed for a north american population, encompassed hundreds of exposures. Compared to the number of single nucleotide polymorphisms (SNPs) commonly included in PRSs (hundreds of thousands), the number of exposures included in this score was small but still represented a substantial gap in scales compared to previous risk scores.

only 2-ways interactions[2] (i.e., pairwise interactions), the number of parameters to be estimated can grow significantly depending on the number of exposures considered ($n = 2p + \binom{p}{2}$). Several methods have been proposed to address this issue.

Some of these (e.g., [Yuan et al., 2009, Choi et al., 2010, Bien et al., 2013]), propose to integrate two types of assumptions, a strong and weak heredity assumption [Chipman, 1996]. A model with a strong heredity assumption refers to a model in which pairwise interactions are included only if both main effects are considered significant, while a weak hereditary assumption refers to a model in which pairwise interactions are included if at least one main effect is. The underlying concept, that an interaction term can be selected only if their parents are in the model, is also referred to as the marginality principle [Nelder, 1977]. These methods are feasible when the number of exposures is a few hundred or less [Choi et al., 2010]. For higher numbers, an alternative procedure is to first perform the selection using a penalized regression on the marginal effects only and then perform a selection on the remaining derived pairwise interactions [Hao et al., 2018].

Bayesian procedures have been proposed to extend linear regression-based methods by including prior information about the data and providing measures of uncertainty for the estimates, e.g., [Bondell and Reich, 2012, Nishimura and Suchard, 2018]. Bayesian procedures can be computationally demanding as the size of the input data increases. However, recent approaches designed for exposome studies such as [Ferrari and Dunson, 2020] are able to handle hundreds of predictors and thousands of rows (i.e., individuals) to derive linear estimates.

Some methods have also been proposed for the estimation of nonlinear interactions. Both [Radchenko and James, 2010] and [Ma et al., 2015] directly extend the quadratic regression procedures by introducing nonlinear functions in the regression equation. Similar to their linear counterparts, they also rely on sparsity assumptions, obtained with different forms of penalization, to be computationally tractable. Other types of variable selection for pairwise interactions have been proposed, such as a forward stepwise feature selection procedure [Narisetty et al., 2018], where most significant marginal effects and interaction terms are progressively added to the model, continuing until a maximum number of iteration k is achieved.

[2]This is commonly referred to as quadratic regression.

While estimating interactions was not our primary aim, the first approach used in this thesis (c.f., Paper 1) can extract nonlinear pairwise interactions learned by tree-based models, using already existing tools, without making sparsity assumptions or assuming known basic functions. However, deriving those from our informed machine learning method (i.e., neural networks) is computationally untractable with our approach.

2.3 Machine Learning Methods in Healthcare

The term "machine learning" (ML) was coined in 1959 by Arthur Samuel to describe a pioneering computer program that he invented, which could learn to play the game of checkers more proficiently than an average human player [Samuel, 1959]. Today, machine learning can be viewed as a subfield of Artificial Intelligence that encompasses a wide variety of supervised and unsupervised algorithmic procedures, where a computer is able to acquire its own knowledge by extracting patterns from raw data [Goodfellow et al., 2016]. In that regard, many of the biostatistical methods presented earlier, such as linear regression, could be qualified as machine learning methods. However, traditional epidemiological methods are typically frequentists, sometimes Bayesians, and yield statistical measures, such as effect size estimates and p-values, that are tailored for hypothesis testing. In contrast, "machine learning methods" typically refer to algorithmic procedures primarily designed for raw predictive performances rather than deriving specific statistical measures. Those methods can generally capture complex patterns and relationships within the data.
In this section, we will discuss the use of machine learning methods in healthcare in opposition to traditional biostatistics. We will discuss how, although machine learning has been employed for a variety of medical applications, its use in studying the health effects of exposure to environmental factors, i.e., the main interest of this thesis, has been limited due to their lack of interpretability.

2.3.1 Machine learning in clinical setting

The availability of large volumes of data in clinical settings through the collection of large electronic health records (EHR) databases of diverse natures (e.g., texts, images, tabular) has lead to a surge of ML models us-

age for the implementation of clinical decision support systems (CDSSs) [Alanazi, 2022]. CDSSs are typically used for diagnosing various health conditions (e.g., cancer [Teramoto et al., 2020], cardiovascular risk [Kennedy et al., 2013]) but also include other use cases such as identifying candidate molecules for research more likely to pass through regulatory processes [Onay and Onay, 2020]. They have also been widely applied for image data processing [Shailaja et al., 2018, Rahman et al., 2023], including identifying body organs from medical images [Yan et al., 2016], automated tissue characterization [Anthimopoulos et al., 2016], reconstructing medical images [Schlemper et al., 2017], segmenting brain tumors [Mehta and Majumdar, 2017], or other specific sources such as voice waveform data to predict the onset of dementia [Xue et al., 2021]. Overall, these tools provide guidance for decision-making, increase medical experts' efficiency and reduce costs by highlighting particular areas worth investigating in large volumes of data.
In contrast, epidemiological studies aim to explore the relationships between various factors and health. Additionally, the data collected through different processes (e.g., observational studies) are generally less abundant in terms of sample sizes. Hence, the use of such methods has been scarcer compared with the typical parametric regression methods.

2.3.2 Machine learning in epidemiology

2.3.2.1 Predictive machine learning for public health

Population-level forecasting of disease is a critical public health issue. For instance, several public health agencies such as the European Centre for Disease Prevention and Control (ECDC) or the US Center for Disease Control (CDC), mainly specialized in infectious disease surveillance, have been focused on epidemic forecasting (e.g., influenza, COVID19) in order to estimate future demands in medical resources. Beyond traditional statistical approaches for time series forecasting (e.g., ARMA, ARIMA, SARIMA, exponential smoothing), machine learning methods (e.g., neural networks, meta-learners, support vector machines) have been widely used for the task [Volkova et al., 2017, Lee et al., 2021]. Given sufficient training data, neural network architectures (e.g., transformers [Wu et al., 2020], recurrent (RNNs) [Kondo et al., 2019], and convolutional networks (CNNs) [Lee et al., 2021]) have proven to perform far better than their statistical counterparts [Lee

et al., 2021].

For non-communicable diseases, predictive ML methods have been used for modeling and forecasting known disease factors such as air pollution [Bellinger et al., 2017], road traffic [Alsolami et al., 2019], noise sources [Bravo-Moncayo et al., 2019] or allergens [Zewdie et al., 2019]. As mentioned in **Section 2.3.1**, ML methods have also been used in clinical settings to directly forecast the onset of non-communicable diseases such as Cancer.

2.3.2.2 Machine learning for studying environmental effects on health

While unable to directly derive exposure-outcome relationships like a linear regression would, tree-based methods such as Random Forest [Breiman, 2001], Bayesian Additive Regression Trees (BART) [Chipman et al., 2010], Classification and Regression Trees (CART) [Breiman, 2001], or gradient-boosted decision trees (GBDT) [Friedman, 2001] are non-parametric approaches able to capture complex relationships such as nonlinearities or interactions in data while allowing a quick computation of a measure of feature importance. While mostly used in clinical settings, as they can guide medical researchers to isolate potential sites of intervention, they have also been used in various epidemiological studies [Stafoggia et al., 2017].

Besides tree-based procedures, the use of ML methods for studying the effects on health of exposure to environmental factors has been limited due to their lack of interpretability [Cheng et al., 2020]. However, recent explainable AI (XAI) methods are a promising solution to address this challenge. XAI techniques aim to make the decision-making processes of complex models, such as deep neural networks and ensemble methods, more transparent and understandable to humans. These methods provide insights into how and why certain predictions are made, which is crucial for applications in epidemiology to understand the relationships between various factors and health outcomes. For instance, methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) help elucidate the contribution of individual features to the model's predictions, thereby enhancing trust and facilitating better decision-making in public health contexts [Ribeiro et al., 2016, Lundberg and Lee, 2017]. XAI not only improves model transparency but also aids in identifying potential biases and ensuring the ethical use of AI in healthcare [Guidotti et al., 2018].

Recently, with the development of explainable AI tools, complex models such as deep neural networks or ensemble learning methods, previously commonly referred to as 'black boxes', can now be used to study the relationships between various factors and health. This approach leverages the raw predictive power of these models, able to capture relevant information from the data with its potential complexities, and then extract and make it intelligible using appropriate tools. Similar to the approach we proposed in the first publication of this doctoral work (c.f., chapter 4), a few studies have been proposed in the last year to study the environmental impact of exposure and health. [Romano et al., 2024] studied the effect of air pollution and socioeconomic factors on respiratory cancer mortality, and [Atehortúa et al., 2023] studied the effects of a wide range of exposures on cardiometabolic risk. Both of those approaches leveraged Shapley values [Hart, 1989] approximated with the SHAP package [Lundberg and Lee, 2017] to explain their ML models.

Researchers in the field have recognized that there is a lack of effective integration of skills and knowledge between the disciplines of data science and epidemiology, which represents a challenge for the broader adoption of ML methods in epidemiological studies [Alanazi, 2022, Bi et al., 2019]. For instance, different words are used to designate the same concepts across the two domains (e.g., recall is sensitivity, label is dependant variable [Wiemken and Kelley, 2020]). [Kolachalama and Garg, 2018] stated that the integration of more data science and ML-related concepts into the curriculum of biomedical researchers is needed to facilitate broader adoption of the ML paradigm in their studies.

2.4 Incorporating Domain Knowledge in Machine Learning Models

Recent advances in machine learning have led to significant improvements in multiple fields, including natural language processing, data generation, computer vision, and many others. However, most machine learning algorithms rely on both the quantity and the quality of the available training data. In multiple applications, securing vast and representative datasets poses significant challenges (e.g., in healthcare [Mandreoli et al., 2022]) that would consequently impact the learning process and, thus, the reliability of ob-

tained predictions. Beyond predictive performances, most machine learning procedures do not consider the underlying mechanisms at play (e.g., biological pathways, physical rules, etc) when they learn patterns within the data. As such, they may learn and amplify potential biases (also known as shortcut learning [Geirhos et al., 2020]), particularly when the data is noisy and incomplete. An approach to tackle those problems consists of integrating domain knowledge into the machine learning procedure, which is known as Informed Machine Learning (IML) [Von Rueden et al., 2021].
Domain knowledge or expert knowledge refers to a form of knowledge specific to a field of application that is not contained within the input data. In the medical field, it could be, for instance, knowledge about known interactions that would occur at the molecular level, while the input data contains relationships at the population level. A taxonomy of the different representations for this knowledge, encountered across various fields, has been proposed by [Von Rueden et al., 2021]. Among them, the most relevant to our application case are:

- Probabilistic relationships between variables. Typically, they can encode assumptions on the conditional independence of random variables, such as health risk scores relative to individual input factors (e.g., [Kumar et al., 2021]).

- Algebraic Equations. Knowledge is encoded as equality or inequality relations and can generally be seen as constraints on the variables present in the input data. For instance, such equations can encode linear relationships between a predictor and an outcome, which has been incorporated into nonlinear methods (e.g., kernel methods to predict breast cancer [Mangasarian and Wild, 2008]).

- Knowledge graph. Nodes generally describe concepts, and edges relationships between them. For instance, those graph are use to represent known metabolic and regulatory pathways in databases such as [Ogata et al., 1999].

- Human Feedback. This refers to *human-in-the-loop machine learning* [Mosqueira-Rey et al., 2022], a ML design where human feedback is incorporated into the learning of a ML procedure, generally through a direct interface. For instance, expert feedback was used to enhance the

selection of features used to train a DNN for cancer survival prediction [Marschner et al., 2021].

Four different ways have been identified by [Von Rueden et al., 2021] to incorporate knowledge into the Informed ML procedure, namely:

- by directly encoding it into the training data. For instance, by generating meta features [Bergman, 2020], based on physic rules (e.g., [Wu et al., 2018]).
- by encoding it into the model's structure. For instance, by designing neurons to enforce specific rules in a neural network.
- by comparing it with the output of the ML procedure. For instance, the outputs of a model can be compared with known constraints, and noncompliant results can be discarded or labeled.
- by incorporating it into the loss function. For instance, adding penalty terms to relationships that do not comply with external-knowledge-based constraints as we did in this thesis. This approach induces the existence of a tradeoff between constraints compliance and learning patterns from the data if the two tasks diverge.

2.4.1 Informed machine learning in healthcare

A first reason to use Informed ML in healthcare research is to mitigate the problem of data accessibility. Gathering the data needed to train Ml models can be difficult and expensive. Medical data can be sensitive and subject to legal and ethical restrictions. It can be fragmented across various institutions and encoded into diverse forms. Integrating domain knowledge can thus be a solution to supplement potential data deficiencies.
Another major factor hindering the use of ML methods in healthcare is the need for interpretability and trustworthiness. Integrating domain knowledge can also help address this issue. While it is still an emergent stream of research, the number of papers published per year has approximately doubled every year, starting from 10 published studies in 2018 to 58 in 2021, [Leiser et al., 2023].
In a recent literature review (2023) on Informed ML in healthcare, [Leiser et al., 2023] reported that most approaches focused on image data and consequently, most of these models use CNNs. More generally, neural networks

are largely the most frequently used Informed ML models [Leiser et al., 2023]. Other methods used included Bayesian networks and biostatistical methods such as logistic regression (e.g., [Radovanović et al., 2019]). The five most prominent forms of domain knowledge incorporated in medical Informed ML [Leiser et al., 2023], ranked by decreasing order of importance, includes: 1) spatial invariances (e.g., [Chawla et al., 2009]) widely used for image processing, 2) probabilistic relations (e.g., [Rahaman and Hossain, 2013]), 3) knowledge graphs (e.g., [Chen et al., 2019]), 4) algebraic equations (e.g., [Demirel et al., 2021]) and 5) human feedback (e.g., [Sampedro et al., 2014]).

In this thesis, we propose a novel method to integrate pooled effect estimates (PESs) from meta-analyses into machine learning procedures, specifically neural networks. In epidemiological studies, PESs are considered to be among the strongest levels of confidence for a factor's relationship with health at a population level [Rosner, 2012]. Despite their significance, there has been limited research on incorporating them into informed machine learning procedures, particularly using deep neural networks (DNNs). To our knowledge, our approach is the first to achieve this. In recent work, [Neri et al., 2022] proposed to integrate them into a naive Bayes model for the computation of health risk scores, the CArdiovascular LIterature-Based Risk Algorithm (CALIBRA). While their approach can combine input data with literature estimates to learn health relationships, it is limited to naive Bayes models and cannot be directly applied to other types of machine learning procedures.

The use of DNNs in epidemiology has been limited due to their black-box nature and the requirement for large training datasets. However, by incorporating domain knowledge such as PESs, these challenges can be mitigated. Our work demonstrates that integrating PESs into DNNs can enhance their effectiveness and trustworthiness, thereby addressing some of the key issues associated with their application in this field.

2.5 Conclusion

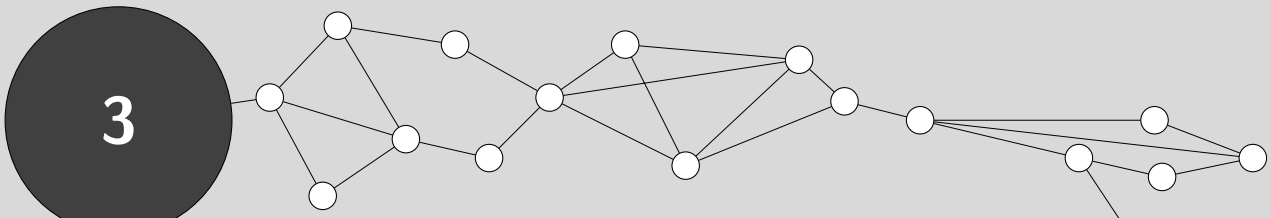
To summarize, most epidemiological studies applying the exposome paradigm use simple biostatistical methods, making various assumptions about the nature of exposure-health associations, such as linearity, additivity, and spar-

sity. Advanced statistical methods have been developed to address those limitations. However, depending on their assumptions, they can be computationally expensive when applied to large amounts of data or may miss intricacies within the exposure relationships.

A promising approach for exposome studies departs from traditional statistical methods to leverage the predictive capabilities of non-parametric models, such as deep neural networks or ensembles of trees, to efficiently capture complex patterns of associations and interactions. While nonparametric machine learning methods are widely used in various domains of healthcare, particularly for automated diagnosis in clinical settings, their application in studying health-exposure relationships remains limited. This is primarily because these models require recently developed tools and techniques to extract and interpret the captured information.

Purely data-driven approaches, however, are heavily dependent on the quality and quantity of the input data. Complex models, in particular Deep Neural Networks, demand large amounts of data to be trained efficiently. However, large sample sizes in observational studies covering a wide range of exposures are rare, and accurately measuring exposures such as air pollution remains challenging. Additionally, access to data can be subject to legal restrictions. Integrating domain knowledge has proven to be beneficial in many settings to improve the plausibility, trustworthiness, and generalizability of Informed ML procedures, but their use within the exposome paradigm has been limited.

In this thesis, we address those gaps by first proposing ML-based ERSs for European Children, leveraging nonparametric ensembles of trees and individual-level models' explanations derived with SHAP. Those scores uncovered new relationships compared with previous studies on the same observational data (the HELIX cohort) but also reported some of the same spurious associations. To address this, we propose a new Informed ML method designed to incorporate one of the most reliable knowledge in epidemiological studies, pooled effect estimates, and enhance the plausibility of captured associations.



3 Methods

This chapter gives a general overview of the data sources, study designs, study populations, health outcomes and exposures used in this thesis. Each of the papers provides a more detailed description of the methods and analyses, referring to chapters 4, 5, and 6 of this manuscript.

3.1 Data sources and study populations

In this thesis, we use two different data sources from two distinct populations: The Human Early Life Exposome (HELIX) project [Maitre et al., 2018] and the Genomes for Life (GCAT) [Obón-Santacana et al., 2018].
The HELIX project includes pregnancy and childhood data—used in paper 1—on a wide range of variables, including urban exposures, lifestyle, and clinical biomarkers. This exposome dataset was adapted for the untargeted discovery of environmental health associations with our machine learning-based approach as it is one of the richest in terms of the number and quality of early life exposure assessments and it is measured across several European countries.
The GCAT project includes longitudinal data—used in paper 3—on a more restricted selection of exposure, but benefits from a larger sample size. We needed a bigger sample size to test our novel informed machine learning approach and it was not necessary to have many exposures to prove its utility.

3.1.1 The HELIX dataset

The HELIX project is a collaborative project that comprises six established ongoing longitudinal population-based birth cohort studies from six different European countries:

- from **Greece**, the Mother-Child Cohort in Crete (RHEA).
- from **Lithuania**, the Kaunas Cohort (KANC).
- from **Norway**, the Norwegian Mother, Father and Child Cohort Study (MoBa).
- from **Spain**, the *Infancia y Medio Ambiente* (INMA) cohort.

- from the **United Kingdom**, the Born in Bradford (BiB) cohort.

- and from **France**, the *Etude des Déterminants pré- et postnatals du Développement et de la santé de l'ENfant* (EDEN) cohort.

The aim of the project was to assess and describe multiple environmental exposures during pregnancy and the first years of life and relate them with different molecular omics signatures and health outcomes. The recruitment of pregnant women was conducted between 1999 and 2010. Specifically, INMA, KANC, and RHEA recruited pregnant women during the first trimester of pregnancy between 2003 and 2008, EDEN and MoBa through the first and second trimesters from 1999 to 2008, and BiB between weeks 26 and 28 of gestation between 2007 and 2010.

The project used a multilevel study design, with an entire study population of 31,472 mother-child pairs recruited during pregnancy, a subcohort of 1,301 mother-child pairs in which the measurement of biomarkers, omics signatures and health outcomes was obtained at age 6-11 years and repeat-sampling panel studies with around 150 children and pregnant women with personal exposure data. This thesis (paper 1) leverage the wide variety of exposures available in the subcohort (N=1600) at two time points, pregnancy and childhood.

3.1.2 The GCAT dataset

The Genomes for Life (GCAT) project is one of the largest prospective cohort in Spain. It aims to assess the role of environmental and omic factors (i.e., genomic, metabolomic, proteomic, and epigenomic) in the development of chronic diseases in adults from Catalonia. The project baseline population recruitment was performed between 2014 and 2017 and covers 19 209 mid-term adults aged 40-65. Several follow-up assessments were performed through online and telephonic questionnaires in the following years, as well as the collection of electronic health records. This thesis (paper 3) leveraged both baseline exposure data and follow-up EHR data collected until the year 2022.

DATASET	The HELIX cohorts	The GCAT cohort
Type	Cross-sectional and longitudinal	Longitudinal
Sample size	1600	19000
Geographical coverage	European population	Spanish population (Catalonia)
Temporality	Early life (pregnancy and childhood)	Adulthood (40-65)
Exposome data		
Chemical	Water disinfection Byproducts Organochlorine and brominated compounds Perfluorinated alkylated substances (PFAS) Metals and elements Organophosphate pesticide Phenols and phthalates	
Psycho-social	Perceived stress Socio-economical capital	Socio-economical capital
Lifestyle	Dietary habits Physical activity Smoking, alcohol	Dietary habits Physical activity Smoking, alcohol Time to sleep
Occupational		Occupational mobility Occupational physical activity Work schedule, status, category
Outdoor and urban	Built environment (e.g., factory proximity) Traffic Natural spaces (green/blue) Air pollution Noise disturbance (traffic, neighbors, etc) Temperature UV Artificial light	Built environment Traffic Natural spaces (green) Air pollution Noise disturbance (traffic only)
Omics		
	Blood/urine metabolites, proteins	
Clinical markers		
	BMI Blood pressure Neurodevelopment Lung function test	Blood pressure Hypertenssion medication

Table 3.1: Description of the data used in this thesis.
Synthetic data used in Paper 2 are not discussed.

3.2 Outcome assessment

Unlike most scientific studies on such data, this thesis was not focused on any health outcome specifically but rather on proposing new approaches for the

untargeted discovery of environmental health relationships. The first paper proposed risk scores for three general health areas: mental, cardiometabolic, and respiratory. The mental health risk score was based on the P-factor [Cervin et al., 2021], a well-known composite measure of psychopathology in the young population. It was computed using responses from a standardized questionnaire, the 99-item Child Behavior Checklist [Achenbach, 1991] using confirmatory factor analysis [Harrington, 2009]. The cardiometabolic risk score was based on the Metabolic Syndrome (MetS) [Cornier et al., 2008], another well-known composite measure summarizing an individual liability to develop cardiometabolic conditions such as heart disease, stroke, and type II diabetes. It was obtained using a parametric model on waist circumference, HDL cholesterol, triglycerides, and blood pressure. The respiratory score was obtained based on the child's Forced Expiratory air Volume in one second (FEV_1) measured from a standardized spirometry test. The Global Lung Initiative reference equations [Quanjer et al., 2012] were used to compute standardized values (i.e., by age, height, sex, and ethnicity of the patient). The third paper proposed a more specific risk score focusing on hypertension. Individuals were considered hypertensive if they had at least one diagnosis of hypertension or took medication related to this condition (anatomical therapeutic chemical codes C02, C03, C07, C08 and C09) at baseline or during the following years.

For children, we use continuous symptom scores as the onset of disease typically comes later in life. For adults however, we used a clinically diagnosed outcome.

3.3 Exposure assessment

In Paper 1, we evaluated a broad range of environmental exposures and preclinical markers, including 63 during pregnancy and 240 during childhood. While Paper 3 assessed a more restricted selection (53 exposures) in the GCAT cohort, it benefited from a larger sample size (19 000 vs. 1600). We categorized environmental exposures into four groups: (1) chemical biomarkers of exposures, (2) Psycho-social exposures, (3) lifestyle exposures and, (4) outdoor and urban exposures. Below, we briefly describe the exposure assessment of these families, but an extensive explanation can be found in Annexe A.1 for HELIX and in the following publication [Obón-Santacana

et al., 2018] for GCAT.

3.3.1 Chemical exposures

In HELIX, different chemical contaminants were evaluated within the framework of the early life exposome investigated in paper 1. For the pregnancy period, several chemicals were already measured in some cohorts before the HELIX project was created, and their results were used. More information can be found in **Supplementary Table 6** (Paper 1). During the childhood period, the sample collection was harmonized in all six cohorts and analyzed at the Norwegian Institute of Public Health. The biological sample collection consisted of two urine and one blood sample. Below is a quick summary of the methods used for measuring those markers.

- **Metals and essential minerals** including arsenic, cadmium, cesium, cobalt, mercury, selenium, thallium, zinc, lead, manganese, molybdenum, potassium, magnesium, and sodium were measured in the whole blood according to [Rodushkin and Axelsson, 2000].
- **Organochlorine compounds (OCs) and polybrominated diphenyl ethers (PBDEs)** including dichlorodiphenyldichloroethylenes (DDEs), dichlorodiphenyltrichloroethanes (DDTs), hexachlorobenzenes (HCBs), polychlorinated biphenyls (PCBs) 118, 138, 153, 170 and 180 were measured in blood serum (maternal samples in EDEN, INMA, RHEA and BiB, and children's samples) or plasma (BiB and MoBa maternal samples) according to [Caspersen et al., 2016] and adjusted for lipids.
- **Organophosphate pesticides (OPs)** including diethyl phosphates (DEPs), diethyl thiophosphates (DETPs), dimethyl phosphates (DMPs), dimethyl thiophosphates (DMTPs) and dimethyl dithiophosphates (DMDTPs) were measured in urine according to [Cequier et al., 2016] and adjusted for creatinine.
- **Phenols** including Bisphenol A (BPA), n-Butyl paraben (BUPA), ethyl paraben (ETPA), methyl paraben (MEPA), oxybenzone (OXBE), propyl parabens (PRPAs), and triclosan (TCS) were measured in urine according to [Sakhi et al., 2018] and adjusted for creatinine.
- **Phthalates** including Mono benzyl phthalates (MBzPs), mono-2-ethyl 5-carboxypentyl phthalates (MECPPs) and others (MEHHP,MEHP,

MEOHP,MEP,MiBP,MnBP,oh-MiNP and oxo.MiNP) were measured in urine according to [Sakhi et al., 2018] and adjusted for creatinine.

- **Per- and polyfluoroalkyl substances (PFASs)** including perfluorohexane sulfonate (PFHxS), perfluorononanoate (PFNA), perfluorooctanoate (PFOA), perfluorooctane sulfonate (PFOS) and perfluoroundecanoate (PFUnDA) were measured in the whole blood using [Haug et al., 2009] method for the pregnancy period and in plasma using [Poothong et al., 2017] method for the childhood period.
- **Water disinfectant byproducts**, including trihalomethanes (THMs), brominated THMs and chloroform, were estimated during the pregnancy period following the protocol developed for the HiWate project [Jeong et al., 2012] using data from the water companies.

No chemical exposures were collected in GCAT to investigate the adult exposome (Paper 3).

3.3.2 Psycho-social exposures

In HELIX, variables related to socioeconomic positions (e.g., maternal education) were collected during pregnancy in all cohorts through a questionnaire and harmonized. During childhood, more psychosocial exposures were assessed using a follow-up questionnaire. Those included maternal stress, family Affluence Score family affluence score (FAS), house crowding, contact with friends and family, and social participation (membership in an organization). GCAT collected socioeconomic information (e.g., marital status, social network, household incomes, type of healthcare access, and education levels) through the baseline survey.

3.3.3 Lifestyle exposures

3.3.3.1 Diet

Food frequency questionnaires were used in both HELIX and GCAT for dietary assessment. In Helix, diet during pregnancy was assessed by each cohort and harmonized a posteriori for the HELIX project. During childhood, information on the child's diet was collected through the standardized HELIX subcohort questionnaire and then summarized in 15 food groups. Additionally, a dietary score representative of Mediterranean dietary patterns,

the KIDMED index, was estimated based on this questionnaire's data. In GCAT, the 14-item Mediterranean Diet Adherence Screener [Schröder et al., 2011] was used to estimate adherence to the Mediterranean diet based on the responses from the baseline food frequency questionnaire.

3.3.3.2 Physical activity

In HELIX, physical activity during pregnancy was estimated based on the harmonization of the respective cohort questionnaire data. Two variables were created: moderate (corresponding to walking) and vigorous activity (corresponding to sport). During childhood, the time spent doing moderate-to-vigorous physical activity variable (>3 MET) and sedentary behavior duration (e.g., TV, computer) were created based on questionnaire data. In GCAT, a short version of the European Prospective Investigation into Cancer and Nutrition (EPIC) Physical Activity Questionnaire (PAQ) [Consortium, 2012] is used to assess physical activity during the past year (before the baseline assessment).

3.3.3.3 Smoking and alcohol

In HELIX, tobacco smoking, and alcohol consumption were assessed during pregnancy using questionnaire data. During childhood, passive child smoking exposure was assessed using a follow-up questionnaire and through cotinine measurements in urine. In GCAT, detailed information about smoking (including electronic cigarettes and others), passive smoking, and smoking history were assessed in the baseline questionnaire. Alcohol consumption (i.e., reported number of standard glasses per day/week) was also assessed in this questionnaire.

3.3.4 Occupational exposures

No occupational exposures were collected in HELIX. In GCAT, the baseline questionnaire assessed occupational exposures, including work travel (duration, vehicle), work schedules, and job positions (further classified using CNO-11, the Spanish classification of occupations).

3.3.5 Outdoor and urban exposures

3.3.5.1 Atmospheric pollutant

In HELIX, nitrogen dioxide (NO_2), particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$), less than 10 μm PM_{10}, and absorbance of $PM_{2.5}$ filters were assessed during childhood in the HELIX subcohort. Briefly, outdoor air pollution exposures were assessed using estimates mostly based on Land Use Regression (LUR) modeling approach developed within the framework of the ESCAPE project [Beelen et al., 2009]. Estimates on air pollutants were assigned to each individual based on their residential and school geocoded addresses, which were collected through the last available follow-up survey for each cohort. Different time windows were calculated for the evaluated air pollutants by averaging them over one day, one week, and one year before the clinical and molecular assessment. In GCAT, similar to HELIX, the residences of participants were geolocated and a GIS approach was applied to estimate an annual average concentration for different air pollutants, namely NO_2, $PM_{2.5}$ and ozone (O_3).

3.3.5.2 Natural spaces

For HELIX, the amount of surrounding greenness (i.e., from trees, shrubs, and parkland) was summarized in a single numerical value, i.e., the Normalized Difference Vegetation Index (NDVI), and was calculated within 100, 300, and 500-meter buffers around residential and school geocoded addresses. NDVI was calculated using satellite imaging following the PHENOTYPE protocol [Nieuwenhuijsen et al., 2014]. Major green spaces (parks or countryside) and blue spaces (bodies of water), i.e., with an area greater than 5000 m^2, were localized using topographical maps or local sources. The straight line distance from the home or school to those spaces was measured. For GCAT, NDVI was computed around participants' residences using [Didan, 2015]. Additionally, a percentage of green spaces within a 1000m buffer around the residential addresses was computed using CORINE Land Cover.

3.3.5.3 Built environment

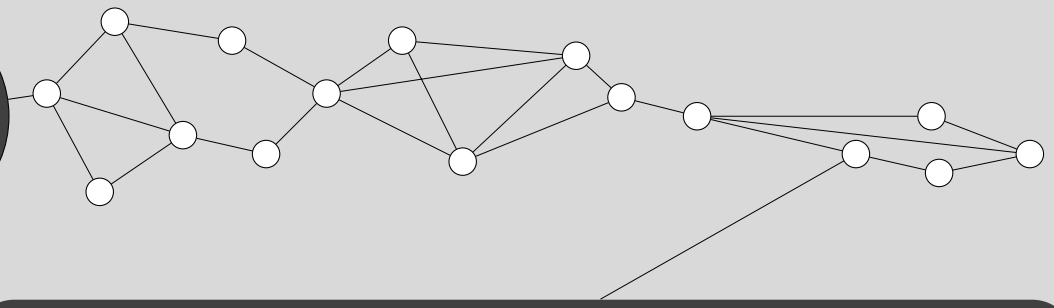
For both HELIX and GCAT, indicators of the built environment were estimated using topological maps obtained from local authorities or Europe-

wide sources. While both projects included building and population density, the HELIX project provided more information on street connectivity, walkability, facility diversity, and facility density.

3.3.5.4 Road Traffic

In both Helix and GCAT, annual average traffic noise pressure levels during the day and night were derived from noise maps produced in each local municipality under the European Noise Directive (directive 2002/49/EC), or for GCAT and INMA (the Spanish cohort within HELIX) under the Spanish Law 37/2003.

4



Machine Learning Based Environmental-Clinical Risk Scores in Children

This chapter relates the work performed for the first publication of this doctoral work [Guimbaud et al., 2024]. In this work, we computed ERSs for three general health outcomes in children: a P-factor score derived from the Child Behavior CheckList (CBCL) for mental health, a MetS severity score for cardiometabolic health, and a lung function score (spirometry test) for respiratory health. Those ERSs were computed using non-parametric machine learning models (namely, tree ensemble methods), able to capture complex exposome-health relationships and interactions. Compared with previous studies on the same observational data—i.e.,the HELIX cohorts—our approach identified new important predictors of disease in childhood on top of identifying nonlinear relationships and exposure-exposure interactions. The scores could also explain a significant proportion of variance (meaning that they had some predictive value), and their performances were stable across all six cohorts, meaning that they were generalizable across different European countries.

4.1 Introduction

The availability of rich data on multiple levels of environmental exposures in birth cohorts presents an opportunity to address the gap in large-scale studies on the association between the exposome and child and adolescent development. Previous ERS studies have been limited by the number of exposure variables or domains included [D'Agostino et al., 2008, Vassos et al., 2019, Padmanabhan et al., 2017]. In contrast, our study aims to identify a predictive environmental-clinical risk score (ECRS) based on a wide array of pregnancy and childhood environmental exposures related to both external (e.g., air quality, lifestyle, psychosocial) and internal (e.g., blood metals, pesticides) exposures, (pre)clinical factors (metabolites, proteins, co-morbidities), and link these to a range of physical and mental symptoms in the large European Human Early-Life Exposome (HELIX) cohort [Maitre et al., 2018, Vrijheid, 2014]. In the context of this study, the term *prediction* refers to the inference of diagnostical risk scores from pregnancy and childhood cross-sectional epidemiological factors to predict childhood liabilities at a single point in time.

ECRSs obtained in this study explained a substantial portion of the variance, in particular for mental and cardiometabolic ECRSs, and their performances generalized well across all six cohorts in the HELIX project. We identified predictors with an overall high impact on the predicted risk, such as maternal stress, child BMI, and noise exposure for mental health. We also extracted non-linear dose-response relationships. Our approach's main benefit lies in its ability to capture complex associations and extract insights at both a global and personal level for each exposure or group of exposures. Overall, this study highlight the potential of such approaches to compute risk scores able to inform practitioners about actionable factors in high-risk children.

4.2 Methods

4.2.1 Study participants

This study uses data from the HELIX project. This project includes data from six different European longitudinal birth cohorts, namely: 1) Born in Bradford alias BiB; UK [Wright et al., 2012], 2) Etude des Déterminants pré et postnatals du Développement et de la santé de l'Enfant, alias EDEN; France [Heude et al., 2015], 3) Infancia y Medio Ambiente alias INMA, Spain [Gascon et al., 2017], 4) Kaunas Cohort alias KANC; Lithuania [Grazuleviciene et al., 2015], 5) Norwegian Mother, Father and Child Cohort Study alias MoBa; Norway [Magnus et al., 2016, Rønningen et al., 2006], and 6) Mother-Child Cohort in Crete alias RHEA; Greece [Chatzi et al., 2017]. Children were born at different periods depending on the cohort (RHEA 2007-2008, EDEN 2003-2005, INMA and MoBa 2005-2007 and KANC 2007-2009). In total, nearly 32 000 mother-child pairs were initially followed during pregnancy and a subset into childhood from 6 to 12 years old, depending on the cohort. From these, we used data from 1622 pairs for which biological samples, environmental exposures, clinical biomarkers and health outcomes were assessed with common standardized protocols. All six cohorts in which HELIX is based had undergone the required evaluation by national ethics committees (prior to the start of the project) and confirmed that relevant, informed consent was given for secondary use of the data [Maitre et al., 2018].

4.2.2 Data

4.2.2.1 Health outcomes

Health outcomes during childhood were either directly measured or derived from variables collected between December 2013 and 2016 in the Helix subcohort follow-up visit [Maitre et al., 2018]. Hence, health outcomes and childhood environmental factors are cross-sectional.

- *Mental health.* We modeled mental health using the P-factor, a reliable measure of psychopathology in youth populations [Constantinou et al., 2019, Haltigan et al., 2018] that represents life course vulnerability to psychiatric disorders [Caspi et al., 2020] and is predictive of long-term psychiatric and functional outcomes [Cervin et al., 2021]. P-factor in childhood has been found to predict the course and severity of a multitude of psychiatric outcomes in adolescence [Rijlaarsdam et al., 2021]. It was computed using confirmatory factor analysis (CFA) to fit a hierarchical general psychopathology model with the Lavaan statistical package [Rosseel, 2012] with data from the 99-item Child Behavior Checklist [Achenbach, 1991], a questionnaire filled by the parents.

- *Cardiometabolic health.* We used an aggregated metabolic syndrome score as a summary score for cardiometabolic health [Stratakis et al., 2020]. It was calculated using the z scores of waist circumference, systolic and diastolic blood pressures, levels of triglyceride, high-density lipoprotein cholesterol, and insulin with the following formula: metabolic syndrome= z waist circumference +(–z HDL cholesterol level + z triglyceride level)/2 + z insulin + (z systolic blood pressure + z diastolic blood pressure)/2. A higher metabolic syndrome score indicated a poorer metabolic profile.

- *Respiratory health.* Finally, we assessed the lung function using the child-forced expiratory (air) volume in one second (FEV_1) percent predicted value (PPV) (i.e., values standardized by age, height, sex, and ethnicity of the patient) as in a previous HELIX study [Agier et al., 2019]. FEV_1 was measured with a spirometry test (EasyOne spirometer; NDD [New Diagnostic Design], Zurich, Switzerland) using a standard standardized protocol. Then, the Global Lung Initiative

reference equations [Quanjer et al., 2012] were used to compute FEV_1 percent predicted values.

4.2.2.2 Environmental data

We used a wide variety of environmental exposures from both mothers (during pregnancy) and children (between ages 6 to 12, depending on the cohort of inclusion) that participated in the HELIX follow-up visit organized between December 2013 and 2016 [Maitre et al., 2018]. Measurements on pregnant mothers were collected between 1999 and 2010. Information about the methods used to estimate those exposures is available in **Supplementary Notes, Supplementary Tables 1-9**.
In previous HELIX exposome-wide studies (including chemical, outdoor, and psychosocial exposures), a sub-selection of variables was made among the questionnaire data, and they did not include together external and internal exposome. Due to the exploratory nature of our study, we included new variables previously unexplored in the HELIX studies. In total, we selected 63 prenatal and 240 postnatal exposures grouped in 18 exposure families. An overview of those families is given in **Figure 4.1**. Most of the exposures used were already described in previous publications, specifically measurement available and baseline data [Maitre et al., 2018]. New exposures, previously not described in detail, were extracted from the HELIX subcohort main questionnaire, more precisely about children's time spent outside (during weekends and holidays), noise disturbance, and house cleaning products. A full list and a description of the selected variables are available in **Supplementary Data 1**.

- *Outdoor and indoor exposures.* Outdoor exposures were estimated using geographic information system (GIS), remote sensing, and spatiotemporal modeling [Robinson et al., 2018]. Considered exposures include air pollutants (e.g., particulate matter), meteorological factors (temperature, humidity, UV exposure), traffic noise, traffic indicators, natural space (green spaces, blue spaces), and built environment (e.g., building density, public transport, facilities, etc.). More details are provided in **Supplementary Notes, Supplementary Tables 1-4**. Indoor air pollution exposure to NO_2 and to volatile organic compounds, benzene, toluene, ethylbenzene, meta-xylene, para-xylene,

and ortho-xylene was measured through passive samplers installed in the homes of 150 individuals in the panel studies and extrapolated for the whole HELIX subcohort using prediction models [Maitre et al., 2018].

- Lifestyle. Lifestyle exposures were collected using a standardized questionnaire developed for HELIX and included the child's diet, physical activity, sleeping patterns, socioeconomic variables (e.g., subjective wealth, social capital of the family), exposure to environmental tobacco smoke, water consumption habits, cleaning products, noise perception and time outdoors. Water disinfection by-product measurements were collected from water companies for the entire cohorts in each HELIX center. More details are provided in **Supplementary Notes, Supplementary Table 5**.

- Biomonitored Chemical pollutants. Pollutant biomarkers were assessed during pregnancy and childhood using blood and urine samples. They include organophosphate pesticides, phenols, phthalates, metals, perfluoroalkyl (PFAS) substances, polybrominated diphenyl ethers (PBDEs), organochlorines and creatine. These measurements were adjusted for lipids and creatinine when appropriate. More details are provided in **Supplementary Notes, Supplementary Tables 6-9**.

4.2.2.3 Metabolites and proteins

We included 122 protein and metabolite measurements in the study. More specifically, we included 36 proteins that were assessed from plasma using Luminex immunoassay kits (cytokines 30-plex, apoliprotein 5-plex, and adipokine 15-plex). Forty-two blood serum metabolite indicators were assessed using the targeted Biocrates' AbsoluteIDQ p180 kit and the MetIDQTM RatioExplorer software that calculated sums and ratios of metabolites, termed metabolism indicators, to improve biological interpretation. Forty-four urine metabolites were assessed using proton nuclear magnetic resonance (1H NMR) spectroscopy [Lau et al., 2018]. Urine metabolites were normalized using the median fold change normalization method [Dieterle et al., 2006], which takes into account the distribution of relative levels of all metabolites

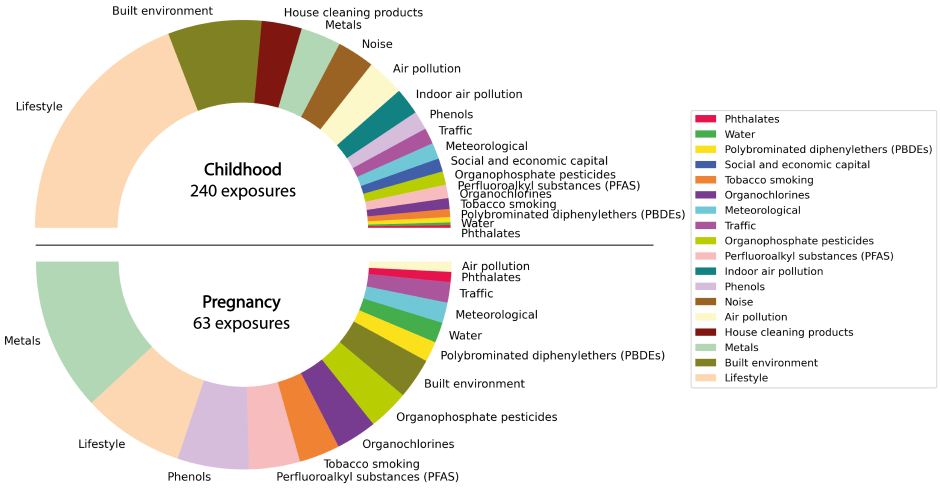


Figure 4.1: Proportions of exposures grouped into families.
Pie chart displaying the 18 different families of exposures considered in the study and their relative sizes in terms of the number of exposures considered within each family. The top half of the chart displays exposures measured during childhood (n=240), while the bottom part displays those measured during pregnancy (n=63).

compared to the reference sample in determining the most probable dilution factor. The full list and description of selected metabolites and proteins is available in **Supplementary Data 1**.

4.2.2.4 Parental and child Clinical factors

Clinical factors were collected during childhood or pregnancy from the HELIX subcohort follow-up clinical examination (between December 2013 and 2016) [Maitre et al., 2018] or initial cohort assessments on pregnant mothers (between 1999 and 2010). Childhood clinical factors included maternal mental and cognitive states (e.g., maternal perceived stress (short form version) [Cohen, 1988], maternal working memory [Sweet, 2011]), child respiratory factors (e.g. diagnosed asthma, self-reported rhinitis) and cardiometabolic factors (e.g., systolic and diastolic blood pressure, blood lipids). Pregnancy clinical factors only include maternal blood lipids collected during the initial cohorts' assessments. The complete list and description of included clinical variables is available in **Supplementary Data 1**.

4.2.2.5 Covariates

Covariates were used as predictors in the ECRSs. We used both children's characteristics (e.g., age at examination, sex, asthma medication, season of birth) and parents' characteristics (e.g., parents' nativity, paternal and maternal education, mother's age at birth, mother's parity) as covariates. A full list and description are available in **Supplementary Data 1**.

4.2.3 Statistical analysis

All data processing was performed in Python 3.9.7.

4.2.3.1 Data preparation

Figure 4.2 provides a brief description of the data selection process and the study workflow. This study aims to agnostically discover exposure-health associations while minimizing the likelihood of overfitting and maximizing ECRS models' interpretability. Hence, we performed several minimal data selection steps, from the initial selection of data to the filtering of strongly correlated and noisy data.

First, as detailed in Section 4.2.2, we used a wide selection of previously described variables enriched with new exposures that were not assessed in previous HELIX studies. In this step, similarly to previous studies on the HELIX sub-cohort data [Agier et al., 2019, Maitre et al., 2022a], we selected single representatives from groups of related and correlated variables to reduce multicollinearity and increase interpretability. For instance, we only kept the 300-meter buffers for the number of road intersections per km^2 and removed the 100m buffers, or we considered only home-based air pollutants measurements and discarded those from schools or other places. The full description of the preselection steps is available in the annex (**Supplementary Methods** – part 1).

Then we further refined this selection by filtering among strongly correlated variables ($r>0.9$), discarding 28 variables in total. A description of the applied rules for this step is also described in the annex (**Supplementary Methods** – part 2).

To reduce the amount of missingness in the data, we discarded records of individuals with more than 50% (102 records discarded) and variables with more than 60% missing data (3 variables discarded). More informa-

tion about each selection step is provided in **Supplementary Table 10**. The mean percentage of missing values per exposure was 14% (first quartile 0.66% and third quartile 20.59%). Percentages of imputed missing values for each selected variable are available in **Supplementary Data 1** and computed for each cohort in **Supplementary Data 2**. Missing values were imputed once using MissForest [Stekhoven and Bühlmann, 2011], a single iterative imputation algorithm that can handle both categorical and continuous variables and capture nonlinear relationships. We compared performance of this method with two classical single imputation methods, namely KNN and mean imputation on manually generated missing values. The number of missing values to add on top of the originals was settled to be proportional to 15% of the total sample size for each variable (i.e., 243). MissForest considerably outperformed KNN and mean imputation with a mean squared error of 0.56, 1.00 and 1.25 respectively. From this imputed dataset, individuals with non-missing outcomes were selected for the mental, cardiometabolic, and respiratory risk scores resulting in three distinct datasets.

Finally, depending on the outcome for each dataset, we excluded clinical factors closely related to the outcome to prevent data leakage, which could over-inflate the model's predictive power (e.g., blood pressure for MetS). The selection made on the basis of the outcome is available in **Supplementary Data 1**.

After selection, the data were prepared for the analysis. Depending on their nature, categorical data were one hot encoded (i.e., changed into dummy variables), labeled, or encoded with floating values (for frequencies, binned continuous variables, etc.). Out of the 478 total selected variables, 75 were categorical (11 one-hot encoded, 43 labeled). Then each phenotype was standardized into z-scores.

4.2.3.2 Modeling

We computed ECRSs with supervised machine learning methods, predicting simple scalar measures as outcomes (P-factor, MetS and lung function) and using multiple environmental and clinical variables, including metabolites and proteins as predictors. Training is first performed in a 10-fold cross-validation (CV) procedure, where hyperparameters of the methods are optimized and performances measured. 10-fold CV was chosen over, for instance,

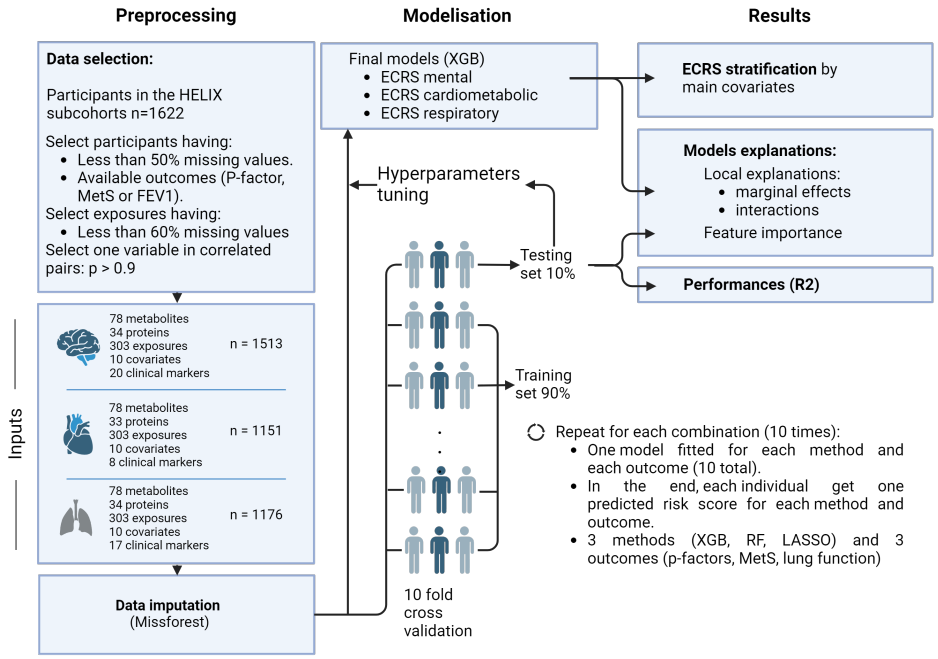


Figure 4.2: Analysis workflow.
This figure provides a concise overview of the steps performed in the study analysis, organized sequentially within three main stages corresponding to three columns. These stages are the preprocessing of input data, the modeling of ECRS, and the reporting of results.
Abbreviations: Random Forest (RF), XGBoost (XGB), Environmental-Clinical Risk Score (ECRS), metabolic syndrome (MetS), Forced expiratory volume in 1 second (FEV_1).

leave-one-out or 5-fold CV, to balance the computational efficiency and the robustness of our performance estimates. Hyperparameters optimizations of each method were achieved using the Tree-structured Parzen Estimator [Bergstra et al., 2011] from the Optuna library, which optimizes the path through the hyperparameters space. **Supplementary Tables 11** and **12** show the selected hyperparameters for each model.

We compared predictive performances of ECRSs computed with three methods: LASSO, Random Forests, and XGBoost. LASSO is a penalized linear regression method widely used in the field. Note that technically, while the model itself is linear (it models the relationships between the input features and the output using a linear equation), the optimization process due to the

L1 penalty is non-linear. One of its main advantages is its ability to handle high-dimensional data through regularization. For this method, data was standardized before training. The other two methods are nonlinear and non-parametric ensembles of trees: Random Forests and eXtreme Gradient Boosting, aka XGBoost [Chen and Guestrin, 2016]). Ensemble methods are able to handle small datasets and high dimensionality [Yang et al., 2010] while interaction effects can be captured and extracted from tree models [Lundberg et al., 2020, Lundberg et al., 2018]. In terms of prediction power, tree-based methods are still competitive with deep neural networks (DNNs) on tabular data [Grinsztajn et al., 2022] due to 1) their robustness to uninformative features, 2) their ability to preserve the data orientation, and 3) their ability to easily learn irregular functions. Machine learning, with systematic out-of-sample testing, employs a rigorous approach to test how well relationships between targets and variables are captured. The better a model performs, the more accurate associations it captures, and very poor performances may indicate unreliably captured information.
One of the HELIX project particularities is that it aggregates data from six different cohorts, with some variables having very distinct distributions across them [Tamayo-Uria et al., 2019]. Those variables are likely to be biased by cohort-related effects. Thus, we chose to penalize contributions of features strongly associated with the cohorts proportionally to their importance in the prediction. Each ECRS was computed using the following modeling (including those in the CV procedures):

$$Y = f_o(X) + g_0(Z) + R, E[R|Z, X] = 0$$

Where Y is the phenotype, X the environmental/clinical factors and usual covariates, Z the one hot encoded cohorts of inclusion, and R the residuals. f0 and g0 were estimated using our regressive methods (LASSO, RF, XGB) in two sequential steps with separate models.

Step 1: $Y = g_0(Z) + U, E[U|Z] = 0$
Step 2: $U = f_0(X) + V, E[V|X] = 0$

4.2.3.3 Model's explanations

Unlike LASSO, tree-based approaches can capture complex relationships that are not limited to a single coefficient per feature. We used SHAP [Lund-

berg and Lee, 2017], a local explanation method that uses Shapley values to extract different contribution coefficients for each individual. This allowed us to keep nonlinear relationships in the explanations. Shapley values were initially used in cooperative game theory to estimate the contribution of each player to the overall cooperation with desirable properties [Hart, 1989]. Adapted to machine learning, it gives the contribution of each feature to the overall prediction at the local (individual) level and can be aggregated (taking the average absolute Shapley values) to give a global measure of feature importance. Concretely, a Shapley value gives, for a given individual, how much the given associated variable impacted the model prediction from the mean predicted value (negatively or positively). They are additives and sum up to the mean predicted value. We used them to explore captured associations (e.g., their directions, nonlinearities) and to compute measures of the global importance of a feature (or a family/group of features) obtained by averaging its absolute (Shapley) values across individuals. Because Shapley values are additives, the contribution of a group of features is easily computed (taking the sum of Shapley values at the individual level). We also computed SHAP interactions [Lundberg et al., 2018] to explore potential (pairwise) cocktail effects.

Measures of global feature importance were computed in the 10-fold CV loop to estimate confidence intervals. Features importances were computed using XGBoost, the best-performing nonlinear tree-based method. Local explanations and stratification were conducted on the final ECRS obtained by training the XGBoost models on the whole dataset, with the hyperparameters selected by the 10-fold CV procedure. We also computed ECRS with Lasso on the whole dataset to compare extracted Shapley values.

4.2.4 Sensitivity analysis

Finally, for each ECRS, we tested the robustness of our method when applied to different populations. Leveraging the data available in our study from six different cohorts, we applied a leave-one-cohort-out cross-validation procedure, recursively training our XGBoost models on five cohorts to predict the sixth. Before training, we standardized both features and targets (e.g., P-factor, MetS, and lung function) across each cohort. The hyperparameters used were the same as before and were not optimized for this task.

4.2.5 Ethics approval

Local ethical committees approved the studies that were conducted according to the guidelines laid down in the Declaration of Helsinki. The ethical committees for each cohort were the following: BiB, Bradford Teaching Hospitals NHS Foundation Trust; EDEN, Agence nationale de sécurité du médicament et des produits de santé; INMA, Comité Ético de Inverticación Clínica Parc de Salut MAR; KANC, LIETUVOS BIOETIKOS KOMITETAS; MoBa, Regional komité for medisinsk og helsefaglig forskningsetikk; RHEA, Ethical committee of the general university hospital of Heraklion, Crete. Informed consent was obtained from a parent and/or legal guardian of all participants in the study.

4.2.6 Data availability

The raw data supporting the current study are available from the corresponding author on request subject to ethical and legislative review. The "HELIX Data External Data Request Procedures" are available with the data inventory at http://www.projecthelix.eu/data-inventory. The document describes who can apply to the data and how, the approval timings, and the conditions for data access and publication.

4.2.7 Code availability

Python code is publicly available on GitHub [Guimbaud, 2024].

4.3 Results

4.3.1 Population characteristics

To investigate the predictive potential of early-life external and internal exposome associated with mental, cardiometabolic, and lung health in children, we selected a study cohort of 1622 mother-child pairs who participated in the HELIX study. This cohort was composed of approximately half females (46.1%), mainly of European ancestry (82.9%), from highly educated families (40.1% with high maternal education), and the majority residing in urban areas (75.3% in areas with a density of population >1500 inhabitants/km^2) (see **Supplementary Figure 1**). At the time of the health

assessment, children were on average 8 years old (range: 5.5 to 12 years), 3.9% regularly visited the psychologist, and 7.3% had a neuropsychiatric diagnosis at the time of the visit (according to parent's reports, besides the CBCL screening). Based on the World Health Organization (WHO) international standards for BMI cut-offs (normal: 18.5–25 kg/m², overweight: 25–30 kg/m², obese: ≥30 kg/m²), while 69.2% of participants fell within the normal category (n=1122), 10.6% of participants were categorized as overweight (n=172), and 20.2% of participants were identified as obese (n=328). Additionally, 10.2% of children were reported to have asthma (ever diagnosed).

Leveraging data from the pregnancy (p=63) and childhood (p=240) exposome, preclinical biomarkers (p=112), and clinical factors (p=18) along with covariates (p=14) (full list available in **Supplementary Data 1**), we trained machine learning models to predict the P-factor, MetS, and lung function. We used 445 features from 1513 individuals for mental health, 432 from 1151 for cardiometabolic health, and 442 from 1176 for respiratory health (**Figure 4.2**). P-factor and metabolic syndrome were square roots transformed to be normally distributed. All three health outcomes were standardized to have a mean of 0 and a standard deviation of 1, with ranges of -2.27 to 3.46, -3.25 to 3.68, and -4.80 to 5.60 for mental, cardiometabolic, and respiratory health, respectively (**Figure 4.3**). After transformation and standardization all outcomes were normally distributed with two-sided Kolmogorov-Smirnov test p-values of 0.08, 0.11 and 0.24 for P-factor, MetS and lung function respectively. For P-factor and MetS, higher scores indicate an increase risk and for lung function a decrease risk.

4.3.2 Predictive performances

To address overfitting, improve stability, and compare model performances and generalizability for all children, we implemented a ten-fold iteration scheme for tested algorithms (XGBoost, RF, and LASSO) with cross-validation (CV) (**Figure 4.2**; see Method). This approach generated, for each algorithm, ten fitted sparse models for each outcome. The comparative analysis of all methods' predictive performances is presented in **Figure 4.4**. These performances were obtained after cohort adjustment (see Method, part 2: Modeling). Cohorts accounted for 5 to 14% of the out-of-sample variance (see **Supplementary Figure 2**).

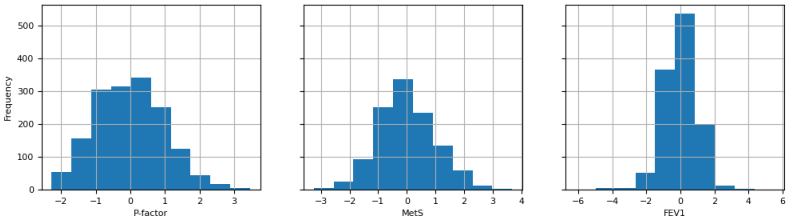


Figure 4.3: Standardized health outcome distributions measured in 6–12-year-old HELIX children.

Each histogram represents the distribution of a health outcome grouped into 10 bins. The horizontal axis shows the range of the outcome, and the vertical axis shows the number of children falling within each bin.

Abbreviations: Metabolic Syndrome (MetS), Forced Expiratory air Volume in 1 second (FEV_1).

LASSO models explained around 12% of the variance in the P-factor, 51% in MetS, and 2% in lung function. In contrast, RF explained 11% of the variance in the P-factor, 41% in MetS, and 3% in lung function. Finally, XGBoost explained 13% of the variance in the P-factor, 50% in MetS, and 4% in lung function (**Figure 4.4**). Across the three outcomes, the differences between LASSO and XGBoost were not statistically significant, as confirmed by 10-fold cross-validated two-sided paired student t-tests with p-values of 0.686, 0.656, and 0.216 for P-factor, MetS, and lung function, respectively. Information about the residuals obtained during the cross-validation procedure is summarized in **Supplementary Table 13**. They were, on average, centered around 0 and normally distributed across all CV folds unless for the respiratory health scores that were considered normal less consistently.

4.3.3 Global feature importance

From the three ECRSs computed with XGBoost, the best-performing non-linear method, we computed Shapley values for each feature and each individual using SHAP. Averaging the absolute values of Shapley values across all individuals gives a measure of the global impact of each variable (or groups of variables) on health outcomes. **Figure 4.5** shows the feature importance for the top 20 variables at the level of each feature and of exposure families for all phenotypes obtained from the 10 fitted cross-validated

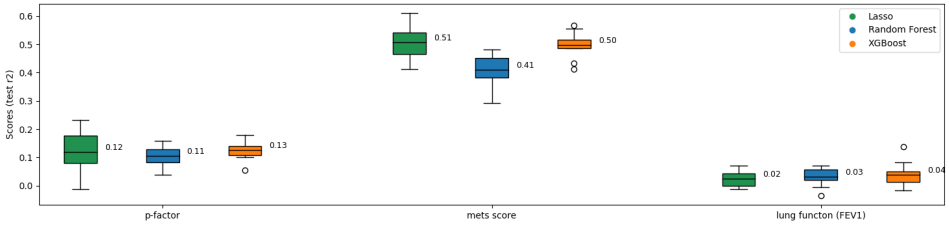


Figure 4.4: Models' performance comparison obtained after cohort adjustment.

The box extends from the first quartile (Q1) to the third quartile (Q3) of the data computed from the ten models (n=10) in the cross-validation procedure, with a line at the mean. The whiskers extend from the box to the farthest data point lying within 1.5x the inter-quartile range (IQR) from the box.

XGBoost models. More exhaustive feature importance lists (top 100) are available in **Supplementary Data 3** and **20** for mental, cardiometabolic and respiratory ECRSs respectively. In addition, the overall importance of all metabolites and proteins compared to exposomic variables, clinical factors and covariates is displayed, for each ECRS, in **Supplementary Figure 3**.

For the P-factor, maternal stress was by large the most important feature, with a mean SHAP of 0.16, followed by noise disturbance from other children with a mean of 0.05 and zBMI with a mean of 0.04. Apart from parental clinical factors with a mean SHAP of 0.16 (mostly driven by the impact of maternal stress), noise disturbance, and lifestyle exposures (such as skipping breakfast, dairy intake and processed meat consumption) were the most important families of exposures, with mean of 0.07 and 0.06, respectively. Other factors not belonging to these families such as tyrosine (urine metabolite) with a mean SHAP of 0.02 and bisphenol A (phenols) with a mean SHAP of 0.02 were also noteworthy.

For MetS, the interleukin-1 beta (IL1B) protein was the most prominent feature, with a mean SHAP of 0.29. Proteins, serum and urine metabolites, as families of variables, exhibited the most impact on the predicted phenotype. They were largely driven by IL1B, Apolipoprotein A1 (APOA1), and the ratio of short-chain acylcarnitines to free carnitine ((C2+C3)/C0). Overall, for this ECRS, metabolites and proteins combined had a mean SHAP of

0.50 while exposures had 0.14 and clinical factors 0.03 (**Supplementary Figure 3**).

Finally, for lung function, although the XGBoost model could explain only a minor part of the outcome (4%) and therefore warrants precaution in the interpretation of the feature importance, no individual features stand out compared to the other health outcomes. The most important features were child zBMI, N-acetylneuraminic acid (Neu5Ac) and the inverse distance to the nearest road during pregnancy (**Figure 4.5**).

On average, childhood measurements were 3.03 to 8.95 times more important than prenatal variables for all risk scores. We found that childhood factors had a mean contribution of 0.23 (standard deviation: 0.01), in comparison with a prenatal mean contribution of 0.06 (standard deviation: 0.01). For cardiometabolic health, the postnatal mean contribution was 0.51 (standard deviation: 0.03), while prenatal factors had a mean contribution of only 0.06 (standard deviation: 0.00). For respiratory health, the postnatal mean contribution was 0.13 (standard deviation: 0.01), while the prenatal mean contribution was only 0.04 (standard deviation: 0.01).

4.3.4 Local feature importance

Unlike linear models where feature coefficients are identical for all individuals, SHAP extracts contributions that are specific to each individual, allowing to assess more complex exposome-health relationships than regression coefficients. The ECRS captured both linear and nonlinear relationships, as shown by SHAP dependence plots (**Figure 4.6** and **Supplementary Figure 4**). For instance, the relationship between maternal stress and noise disturbance from neighbors followed a linear trajectory, while the impact of child zBMI on the P-factor displayed a more complex pattern with a threshold effect. These plots also allowed us to visualize the directions of the distinct associations for each outcome.

A high value in maternal stress was related to an increase in the predicted P-factor, indicating an increased risk for mental health issues. Low levels of noise disturbances and child zBMI slightly reduced the risk of mental health problems, while high values had a particularly harmful impact, especially in the case of child zBMI.

For cardiometabolic health, IL1B was positively associated with MetS, with high values having an important impact on the risk. We observed similar

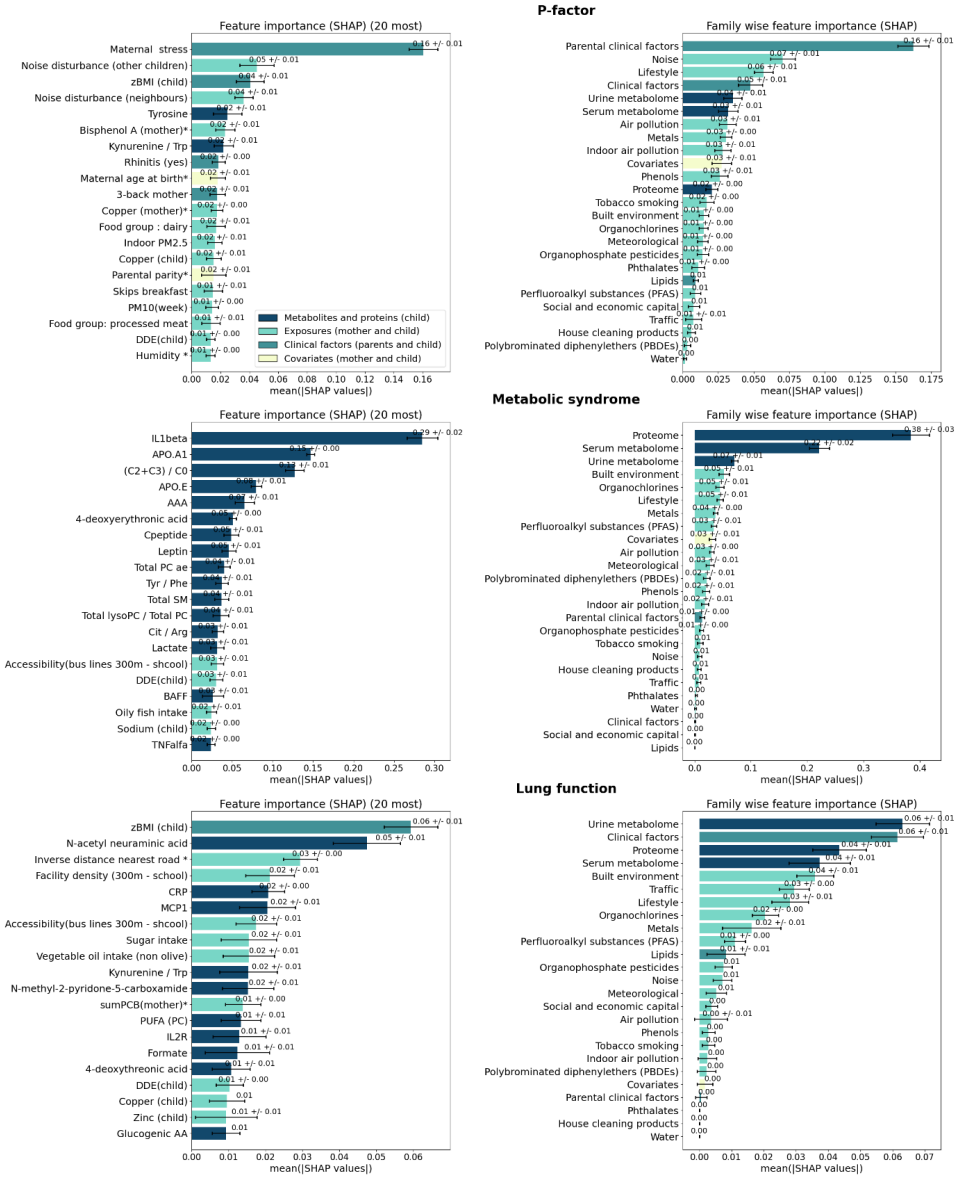


Figure 4.5: Global feature contributions to the three environmental-clinical risk scores in the HELIX mother-child pairs.

Mean contributions are estimated from Shapley values for each individual factor (left column) and each family of factors (right column). The black interval bars represent the standard deviation across the ten models (n=10). Only the top 20 most impactful factors are displayed here. Extended lists of feature contributions for each ERS are available in **Supplementary Data 3**). Variables assessed during pregnancy are indicated by an *.

relationships, to a lesser magnitude for AAA, apolipoprotein-E (APOE) and C-peptide. Conversely, high values of APOA1 and the (C2+C3)/C0 ratio showed a strong protective impact.
The results for the respiratory health scores should be interpreted with caution because of the low variance explained by the model after cohort adjustment (4%). We observed a protective impact on lung function for child BMI and inverse distance from the nearest road during pregnancy. Low BMI values substantially increased the risk, while moderate to high values slightly reduced it. On the contrary, high values of Neu5Ac, facility density near school (300m), CRP, Monocyte Chemoattractant Protein-1 (MCP1), sugar and oil intake were associated with decreased lung function.
Compared with the linear associations obtained with Lasso (**Supplementary Figure 5**), directions of associations were consistent with XGBoost, with some exceptions (e.g., leptin for cardiometabolic health and bus line accessibility for respiratory health). Overall, predictions obtained with XGBoost were more conservative for extreme values of the predictors (e.g., maternal stress for mental health or child zBMI for respiratory health, etc.).

4.3.5 Pairwise interactions

Pairwise interaction effects were derived from Shapley values using SHAP. **Supplementary Figures 6** and **7** show plots for the top ten interactions (according to the mean absolute value of Shapley values) derived from the mental and the cardiometabolic risk scores. For lung function, the predictive power of the risk score was insufficient to extract meaningful information from its captured interactions. Overall, interaction effects on predicted risk were relatively small compared to the marginal effects. We observed a 7.4 to 8.8 ratio between the mean top ten marginal effects and the mean top ten interaction effects, depending on the outcome of interest (respectively, 0.042 and 0.005 for mental health, and 0.093 and 0.013 for cardiometabolic health). This indicates that overall, pairwise interactions had a much smaller impact than marginal relationships on the predicted risk for the two scores. For mental health, the most important captured interactions were between perceived maternal stress during follow-up assessment and factors from diverse exposure families (clinical factors, lifestyle, noise disturbance, etc.). Specifically, the most important interactions were between the following factors: maternal stress and allergic rhinitis, with a mean SHAP value of

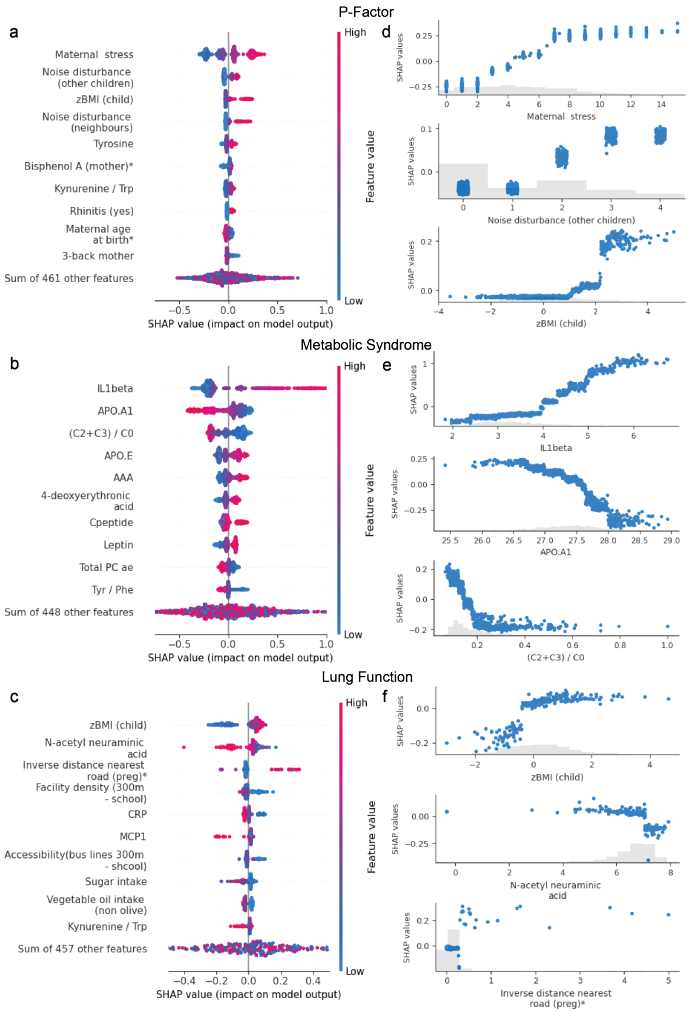


Figure 4.6: Local explanations (SHAP) from the three environmental-clinical risk scores in HELIX mother-child pairs.

a, b, c Beeswarm plots of Shapley values for the ten most important features for mental, cardiovascular and respiratory ECRS respectively. Each dot represents the contribution (Shapley value) of a feature for a given individual in the model's prediction. Dots accumulate along each feature to show density. The feature value for each individual is shown in a colored range from low to high. **d, e, f** Dependence plots of the top three most important features for mental, cardiovascular, and respiratory ECRS, respectively. Each dot represents the contribution (Shapley value), on the y-axis, of a feature, on the x-axis, for a given individual in the model's prediction. Gray bars show the features' distributions. Variables that were assessed during pregnancy are indicated by an *.

0.0070; maternal stress and insulin, with a mean SHAP value of 0.006 and maternal stress with dairy intake, with a mean of 0.006 (**Supplementary figure 6**). For cardiometabolic health, the top ten most relevant captured interactions were between clinical biomarkers (proteins and metabolites) or between IL1B and other factors, such as temperature or child's age. Specifically, the most important captured interactions were between the following factors: IL1B and ratio of short-chain acylcarnitines to free carnitine, with a mean SHAP value of 0.025; APOA1 and APOE, with a mean of 0.016 and APOA1 and the ratio of short-chain acylcarnitines to free carnitine with a mean of 0.013 (**Supplementary Figure 7**). **Figure 4.7** shows two arbitrary interactions, selected from the top ten most important ones, derived from the mental and the cardiometabolic risk scores. The first interaction is between maternal stress and the insulin measured in children and impacts mental health. It indicates a harmful impact of high insulin combined with maternal stress. The second interaction for cardiometabolic health is between IL1B and the ratio of short-chain acylcarnitines to free carnitine. It indicates a significative impact of this ratio on children with high values of IL1beta, with a high ratio value associated with higher risk and vice-versa.

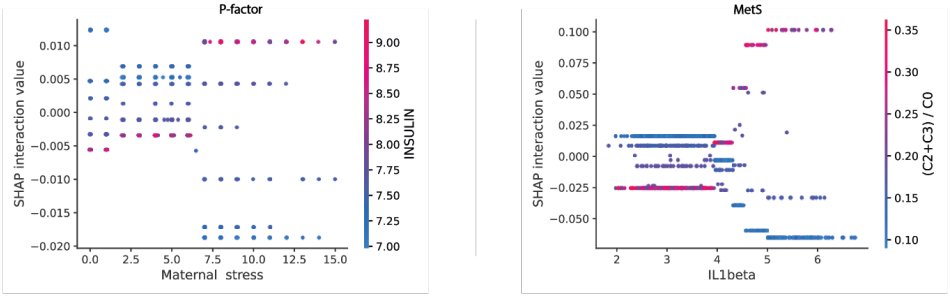


Figure 4.7: SHAP selected interaction effects derived from the mental (P-factor) and the cardiometabolic (MetS) environmental-clinical risk scores.
Each dot corresponds to a child. Ordinate axis is the corresponding Shapley value of the pairwise interaction, representing its contribution to the predicted risk. Feature values are given on the x-axis (first feature) and on the colored scale (second feature). Marginals and interactions effects are additive for each individual and sum to the predicted value.

4.3.6 Generalizability across cohorts

For each of the three ECRSs, performances obtained from the leave-one-cohort-out cross-validation were consistent with those obtained using the ten-fold cross-validation. The explained variance was 13.4% (4% std), 46.8% (12% std), and 2.4% (2% std) for mental, cardiometabolic, and respiratory health, respectively. Differences in predictive performances were observed for all ERCSs depending on the left-out cohort. For instance, the ECRS for the P-Factor was less predictive of the outcome when the KANC cohort was predicted based on all the other cohorts (R2=6.3% versus 13.4 on average). **Table 4.1** shows the variance explained within each cohort.

	Cohorts						**Aggregates**	
	BiB	**EDEN**	**KANC**	**MoBa**	**RHEA**	**INMA**	**Mean**	**SD**
P-Factor	16.9	11.8	6.3	15.5	13.3	16.7	13.4	4
MetS	53.2	47.3	41.6	59.4	56.4	22.7	46.8	12
FEV$_1$	3.4	3.6	2.8	1.0	-0.8	4.6	2.4	2

Table 4.1: Variance explained (R2) by each ECRS in the leave-one-cohort-out cross validation procedure.

Our XGBoost-based machine learning procedure was used to predict each cohort while training on the others.

Abbreviations: Metabolic Syndrome (MetS), Standard Deviation (SD), Forced Expiratory air Volume in one second (FEV$_1$)

4.4 Discussion

This is the first study that computed children's machine learning-based ECRSs (for mental, cardiometabolic, and respiratory health), covering such a wide range of pre- and post-natal factors (including air pollution, noise, urban and social environment, lifestyle, chemical exposures, metabolites, proteins, and clinical factors). The inclusion of data from six different cohorts across different countries allowed us, by adjusting our models for cohorts, to extract non-cohort-specific relationships, thereby increasing their likelihood of being generalizable. Predictive performances were superior for the cardiometabolic risk score (R2: 50%), in particular, driven by the (pre)clinical biomarkers compared to the other two health domains, R2: 13% for general mental health and 4% for lung function. Most important variables were the following: parental clinical factors (mainly maternal stress), noise distur-

bance (mainly from neighbors and other children) and lifestyle exposure for mental health; protein and metabolites (mainly IL1B) for cardiometabolic health; and child BMI and urine metabolites for respiratory health. While our results need to be validated on an external population (e.g., different from the countries available in HELIX), the cohorts-based sensitivity analysis (leave-one-out cross-validation) showed promising results regarding the generalizability of our ECRSs on European populations. Our approach's main benefit lies in its ability to capture complex associations and extract insights at both a global and personal level for each exposure or group of exposures. Results showed that several important relationships were potentially nonlinear.

Our study was performed in a context where current clinical tests often fail to identify children at-risk of developing chronic diseases, notably considering mental and cardiovascular diseases, which is a key challenge to the development of effective prevention and treatment policies. This limitation is largely attributed to the fact that many of these tests were developed and validated primarily in adult populations and to the paucity of longitudinal data on children. In the case of cardiovascular risk, the role of adipokines is a pertinent example. While the association between elevated inflammatory biomarkers and increased cardiovascular risk is well established in adults, there is a lack of comprehensive studies on the onset and progression of these biomarkers in children [Balagopal et al., 2011]. Therefore, more prospective studies focused on children are necessary to provide guidance to the pediatric medical community for more effective interventions and prevention policies.

Polygenic risk scores have substantially improved predictions in comparison with single genetic factors, and their potentials for screening, prevention, treatment, and disease management start to become apparent, suggesting that their environmental analog will be equally valuable in public health prevention [Vassos et al., 2019], for both the identification of individuals and environmental factors of risk. A recent study from the UK biobank [He et al., 2021] showed that, in addition to standard clinical risk factors such as sex, age, blood pressure, or BMI, ERSs provides a greater increase in predictive performances compared with PRSs. Furthermore, ERSs capture holistic individual-level non-hereditary risk associations, informing clinicians about actionable factors in high-risk patients that are indepen-

dent of genetics and provide guidance for prevention and treatment. In this case the environmental factors include, among others: noise disturbance, bisphenol A and humidity for mental health; bus lines accessibility, child's dichlorodiphenyldichloroethylene and oily fish intake for cardiometabolic health; facility density, bus lines accessibility and sugar intake for respiratory health.

A known limitation of most nonlinear machine learning approaches is the model identifiability, which can lead to complications such as overfitting or ambiguity in interpreting the model's parameters [Hastie et al., 2008b, Hastie et al., 2008a]. However, we believe our approach has several strengths that mitigate these concerns. First, our modeling strategy, backed by rigorous cross-validation, limits overfitting and prioritizes generalizability across different data subsets. Second, both nonlinear methods used in this study aggregate results from multiple models, whether from boosting or bagging, which enhances the stability of results. Finally, it is worth remembering that identifiable models such as linear regression or LASSO also have important limitations. In particular, they cannot capture nonlinear relationships, and the coefficients obtained through regression can be interpreted with confidence only if all relationships are linear, which is unlikely to always be the case in real application scenearios.

Our tree-based approach has shown comparable predictive performances to LASSO, which could indicate that the relationships to capture are mostly linear in nature and that there are no interactions. However, the difference in prediction is likely to be due to the ability of simpler models to perform better in situations of small training sample size [Hastie et al., 2008b]. More data would be needed in order to confirm the nature of those relationships and, even if LASSO performs well, it may not capture all the complexities within the data, especially nonlinear relationships and interactions, which other algorithms could reveal. In this study, we explored such relationships using, to our knowledge, a novel approach in this field. We did not aim to assess nor confirm the causality of extracted relationships but rather to identify new associations that previous methods would have missed, and explore the complexity of known ones (e.g., nonlinearities, interactions). Thus, the causality of those findings needs to be validated in a causal inference framework.

We compared our findings to those previously obtained in the literature

and highlighted relationships with poor literature coverage. Besides non-linearities that are rarely assessed in the field, we found our results, when comparing only directions of associations, to be mainly consistent with those observed in other studies, which indirectly validates our approach.

For mental health, previous studies have found that maternal stress can lead to an increased risk of anxiety, depression, and behavioral problems in children [Farewell et al., 2021], while a higher BMI has been associated with depression, anxiety, and low self-esteem [Wang and Veugelers, 2008]. Similarly, exposure to noise has been linked to both internalizing and externalizing behavioral problems in children [Lim et al., 2018]. Overall, while the majority of factors identified in this study have been linked with mental health outcomes in the literature, the exact causal pathways, the potential interactions between these factors, and the nuances of their impacts during specific developmental windows would benefit from further investigation. Specifically, the links involving noise disturbance sources, pollutant exposure to bisphenol A, which is not conventionally studied in mental health research, and certain metabolic markers such as tyrosine levels and the Kynurenine/Tryptophan ratio.

For cardiometabolic health, several of the identified markers are well established in cardiometabolic health (e.g., IL1beta [Esser et al., 2014], APOA1 [Wilkins et al., 2021], Leptin [Tsai, 2017]), and their known relationships are consistent with our findings. Other markers, such as aromatic amino acids or plasmalogens have been linked to several aspects of metabolic and cardiovascular health [Sun et al., 2022, Ding et al., 2023, Novgorodtseva et al., 2011], but their causal pathways are still areas of ongoing research. Finally, 4-deoxyerythronic acid is not well covered in the literature.

At last, concerning respiratory health, we mainly identified a positive (non-linear) association between FEV1 and child BMI, which supports the obesity paradox in chronic obstructive disease [Sun et al., 2019, Köchli et al., 2019]. Unexpectedly, the inverse distance to the nearest road during pregnancy was associated with an increased FEV1. This association was already reported in a previous HELIX study [Agier et al., 2019] and is driven by the RHEA cohort. Overall, while our study identified relationships already well established (e.g., air quality through facility density or accessibility), others might be less direct and require further investigations (e.g., N-acetyl neuraminic acid, sugar and vegetable oil intake).

4.4.1 Strengths and limitations

This study benefited from the richness of the HELIX project data, which used standardized outcomes, clinical biomarkers, and exposure measurement methods across six different European countries. The HELIX project used a wide range of exposure measurement techniques to collect both internal and external exposome data. In contrast to previous studies where ERSs are usually derived from weighted sums of a limited number of exposures, our approach simultaneously assessed the impact of a wide range of exposures, metabolites, and clinical factors on several health outcomes. To address multicollinearities and nonlinearities, we used penalization and recent AI modeling techniques. SHAP allowed us to decompose the complex relationships captured by our ECRSs for each feature at a global and personal level, extracting interactions and marginal effects. Our risk scores were adjusted for the cohorts of inclusion, which is a particularity of the HELIX project. We acknowledge that in the case of highly correlated variables, our approach exposes those with the estimated higher impact on the predicted risk without implying causality. Correlations between exposures are known to present a challenge for exposome research, especially in the ability to differentiate true predictors from correlated covariates [Agier et al., 2016].

The main limitation of our study is the lack of external validation using an independent cohort. While our study benefited from the richness of exposome data not found in any other early-life cohort studies or pre-clinical studies, our sample size was relatively small (n 1500), especially for applying certain machine learning methods (e.g., deep neural networks). This sample size constraint limited our ability to capture and extract complex relationships such as interaction effects, and favored our choice towards tree-based ensemble machine learning methods. Nevertheless, we observed a similar predictive value with LASSO. The other outcomes were both precomputed composite risk scores (P-factor and MetS), possibly suggesting better performances of machine learning methods in predicting raw outcomes. Additionally, as our study is exploratory in nature, we did not assess the causality of extracted exposome-health relationships. Further work is required to validate these relationships using a proper causal inference methodology. For instance, our scores might capture bidirectional cause-and-effect relationships, such as, potentially, the association between maternal stress and

child behavior. The inclusion of internal peripheral markers that reflect the body's response to the measured health outcomes (e.g., IL1B and obesity) could also reinforce this phenomenon. In the case of MetS, the array of proteins and blood metabolites primarily covered biological pathways related to cardiometabolic outcomes, such as blood lipids. By decomposing the importance attributable to metabolites/proteins and the other factors (**Supplementary figure 3**), and further refined into families of factors, we limit this issue. The integration of these preclinical biomarkers aimed to enhance the predictive power of our ECRSs and to investigate their relationships and interactions when combined with a wide variety of factors.
Another limitation is that our study does not account for the risk evolution over time since we used pregnancy and cross-sectional epidemiological data to assess childhood exposures at a single point in time without considering their impact throughout an individual's lifetime. This limitation will be addressed in ATHLETE, the HELIX follow-up project in which the same children were regularly monitored into adolescence with repeated health outcome measures [Vrijheid et al., 2021].
Finally, our study participants are mostly representative of the European population since the data was collected from six different European countries. Therefore, caution must be taken when extrapolating our findings to different populations. As more datasets become available in the future, we will be able to further validate the scores obtained in this study and generalize our findings beyond the populations here analyzed. Beyond the validation of our scores, the associations extracted in this study, which are yet rarely studied in exposome research, could be validated more independently without such a high-dimensional dataset. Some of those factors are easily collectible through questionnaires (e.g., noise disturbance, maternal stress for mental health). Social and perceived environmental factors are yet poorly covered in exposome research [Neufcourt et al., 2022] and this study provides more evidence of their importance.

The combined use of complex machine learning techniques and explainable artificial intelligence methods remains uncommon in the environmental epidemiology field, despite its excellent fit with the exposome paradigm, which aims to capture complex associations within mixtures of environmental factors. Our study revealed results mostly consistent with previous studies while, at the same time, exposing individual-level relationships with nonlin-

earities. Furthermore, we believe that the development of bigger databases and federated analysis tools such as dataShield [Gaye et al., 2014] can unlock the true potential of these approaches to more accurately capture exposome-health associations.

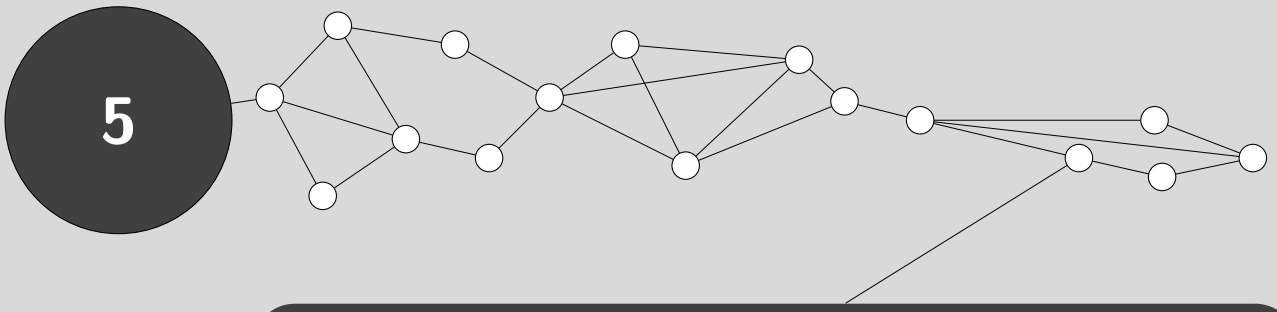
4.5 Conclusion

In this large exposome study, environmental clinical risk scores were computed using linear (LASSO) and nonlinear methods (XGBoost, Random Forest). No significant differences in predictive performances among these methods were found across the computed risks, however, machine-learning-based ERS extracted different effects including non-linearities. From the nonlinear risk scores, we extracted exposome-health relationships from Shapley values, allowing us to derive feature importance at a local and a global level and uncover interactions, which is, to our knowledge, a novelty in the field. The most important predictors included maternal stress, child BMI and noise exposure for mental health; biomarkers such as IL1B and APOA1 for cardiometabolic health; and child BMI and sialic acid (Neu5Ac) for respiratory health.

Besides their usefulness for epidemiological research, our risk scores showed great potential to capture holistic individual-level nonhereditary risk associations that can inform practitioners about actionable factors of high-risk children. As in the post-genetic era, personalized prevention medicine will focus more and more on modifiable factors, we believe that such integrative approaches will be instrumental in shaping future healthcare paradigms.

However, the adoption of such machine learning methods introduces new challenges, particularly concerning 1) the substantial data requirements necessary for training these models to effectively outperform traditional approaches in terms of predictive performance, and 2) the trustworthiness of the derived models. Although our approach did not achieve significantly superior predictive performance, it offered equivalent solutions that captured different relationships. This raises the question of which solution to trust. One possible answer is to choose the solution that best aligns with established domain knowledge when such knowledge is available. In the next chapter, we develop an informed machine learning procedure that integrates this domain knowledge into predictive modeling to supplement the train-

ing data and enhance the trustworthiness of our environmental risk scores approach.



5 A Deep Learning Approach for Informed Environmental Risk Scores

This chapter relates the work performed for the second paper of this doctoral work (not yet published at the time of writing). In this work, we aim to mitigate the limitations encountered in the first contribution, namely: 1) the small available sample size, 2) the presence of noise (e.g., from missingness or measurement errors) that limited our capabilities for modeling complex relationships, and 3) the presence of confounding bias leading to the capture of spurious associations. To that end, we develop a novel method for incorporating domain knowledge into the training of nonparametric methods, this time using deep neural networks, to inform machine learning-based environmental risk scores. We call this method SEANN (Summary Effects Adapted Neural Network). We use synthetic data to illustrate its benefits in a controlled setting.

5.1 Introduction

Various domains of healthcare research (e.g., pharmacology, psychology), and, in particular epidemiology, study the relationships between factors and outcomes of interest using conventional statistical methods such as linear or logistic regression algorithms [Bender, 2009, Stafoggia et al., 2017]. These models typically provide simplistic and easily interpretable representations of potentially complex relationships by encapsulating dose-response estimates in different forms, including odd ratios (ORs), risk ratios (RRs), hazard ratios (HRs), or standardized regression coefficients (SRCs). Such estimates can be easily aggregated across several studies, i.e., in meta-analyses [Borenstein et al., 2021], to derive a more reliable and statistically robust indicator describing the relationship between a variable of interest and a target outcome, namely a pooled effect size (PES) [Pathak et al., 2020], that represents a quantitative formulation of a scientific consensus. In epidemiology, PESs are considered one of the most reliable forms of information [Wallace et al., 2022].

Inspired by risk prediction models, such as the Framingham risk score for coronary heart disease [D'Agostino et al., 2008] or PRSs [Khera et al., 2018] from genetic research, environmental risk scores (ERS) are summary measures of the effects of multiple exposures used to estimate the environmental

liability for a particular health outcome at an individual level [Park et al., 2014]. Those scores are useful tools for screening individuals to select for more expensive testing, and, more importantly, they can be used to study the effect of environmental exposure on health [Park et al., 2014]. ERSs are usually built using interpretable methods as a weighted sum of the individual exposure estimates (cf. **Section 2.2.1.4**), obtained either from previous literature studies or derived through regression models [Pries et al., 2021]. The use of modern ML methods such as DNNs to compute them is still rare (cf. **Section 2.3.2.2**).

PESs have been used to compute literature-only health risk scores (cf. **Section 2.2.1.4**) and informed risk scores (e.g., [Neri et al., 2022]) using traditional biostatistical methods (e.g., logistic and linear regression). In epidemiological studies, data gathering can be challenging as it may require costly measurements and is often subject to ethical and legal regulations. Consequently, risk scores derived from PESs are often more robust compared to those computed from restricted datasets collected for specific populations (e.g., from a specific country, age range, or socioeconomic profile).

In this work, we introduce SEANN (Summary Effects Adapted Neural Network), a method designed to integrate prominent forms of PESs, namely ORs, RRs, and SRCs [Bakbergenuly et al., 2019, Nieminen, 2022] into the training of DNNs for the computation of ERSs. The underlying idea of SEANN is to penalize deviations from integrated PESs measured through differences in prediction when perturbing the inputs. While various methods have been developed to integrate different forms of external knowledge within the learning process of DNNs [Dash et al., 2022], we are, to our knowledge, the first to integrate PESs.

In addition to enhancing the robustness of derived ERSs by leveraging estimates from large and diverse populations, SEANN aims to capture exposure relationships that closely align with known evidence. By incorporating knowledge from well-known relationships, SEANN can better characterize those that are less studied. Additionally, this approach addresses challenges posed by Deep Neural Networks (DNNs) that limit their use in the field. DNNs are not inherently interpretable and require large sample sizes for effective training. SEANN compensates for the limited observational data typically available in epidemiological studies and improves the generalizability and trustworthiness of computed risk indicators.

In **Section 5.2** of this chapter, we introduce SEANN; In **Section 5.3**, we perform a series of experiments on synthetic scenarios illustrating the method's benefits in a controlled environment. More specifically, we refer to improved prediction accuracy in noisy contexts and improved reliability of interpretation using XAI. Finally, **Section 5.4** discusses the significance of those results and concludes.

5.2 Method

This section introduces SEANN. We first define the general setting of integrating PESs as soft constraints via additional terms to the loss function and then detail the implementation of this approach for three types of PESs: standardized regression coefficients, odd ratios, and risk ratios.
Given a set of p observed variables P, and a subset $Q \subseteq P$, with $|Q| = q$ of these variables for which we have an effect size estimate value to use as enforced external knowledge, we define $\mathbf{X}$, an $n \times p$ input matrix of n observations, and V, a vector of q effect size estimates values. Similarly to previous works (e.g., [Muralidhar et al., 2018, Daw et al., 2022]), the general principle of our method, described in **Eq. 5.1**, consists in adding a term to the loss function $\mathcal{L}$ for each meta-heuristic to incorporate.

$$\mathcal{L} = \lambda_0 \mathcal{L}_{pred}(\mathbf{X}, \theta) + \sum_{i=1}^{q} \lambda_i \mathcal{L}_{meta}(\mathbf{X}, \theta, v_i, h_i) \tag{5.1}$$

Where $\mathcal{L}_{pred}$ is the convex function used for the predictive task (e.g., mean squared error, cross-entropy, etc.) and θ the parameters of the model. $\mathcal{L}_{meta}$ is the convex loss function used to enforce the desired soft constraints for the neural network, i.e., to enforce the neural network to respect the PESs vector V. λ_0 and λ_i are weights pondering the importance of each term, namely the predictive task and the i^{th} constraint. They can be treated as hyperparameters and set to values that optimize the obtained predictive performances. However, for cases where learning plausible relationships, i.e., relationships aligned with known associations, is considered equally or more important than raw predictive power (e.g., imperfect input data, trustworthiness), a different approach should be used to settle the tradeoff between learning from the data (and optimizing performances) or learning associations observed in the literature. We propose choosing weight values proportional to

the *confidence* in both the data and the external knowledge. Following a common principle in meta-analyze (e.g., [Borenstein et al., 2021]), we express this confidence using the sample size available in each study[1] as well as the sample size available in the training data.

The proposed weighting is calculated as follows: first, we define a confidence score c_i associated with v_i corresponding to the sample size of the i^{th} meta-analysis. Similarly, we define a confidence score c_0 for the input data $\mathbf{X}$ composed of n rows and p variables to be computed as $n \times p$. Then, for $c > 1$, we estimate the final λ weights using a log scale relative normalization:

$$\lambda_i = \frac{\ln c_i}{\sum_{j=0}^{q} \ln c_j}$$

This scheme ensures that terms associated with small confidence scores have a noticeable impact on the learning process compared to the others and are not entirely ignored.

Depending on the type of PES considered, the loss function $\mathcal{L}_{meta}$ will be implemented differently. As effect estimates in the meta-analysis are typically represented either as OR, RR, or SRC, we express $\mathcal{L}_{meta}$ for those forms below. In all cases, the principle is to generate for each observation a slightly perturbated copy of it with an increment h—called the *perturbation*— for each variable in Q, to measure the difference between the expected change in the target value according to our PESs and the observed change in our model, and to penalize this difference.

5.2.1 Case of a standardized regression coefficient

Let us first consider, for simplicity, a single SRC, called β_i, that we want to integrate into the training of a DNN. This β_i would be either directly extracted in a domain-specific literature study from a uni/multi-variate linear regression model or would summarize several similar effects in a meta-analysis. Considering a multivariate linear regression model defined as:

$$f_{\boldsymbol{\beta}}(\boldsymbol{Z}) = \beta_0 + \sum_{j=1}^{m_2} \beta_j z_j$$

[1]Here we are referring to the total sample size used for estimating each PES.

With $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ an input matrix and $\boldsymbol{\beta}$, a vector of SRCs, $\beta_i \in \boldsymbol{\beta}, 1 \leq i \leq m_2$. Then the expected change in the target values according to β_i when modifying the corresponding input variable z_i with a perturbation step h is:

$$f_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i+h}) = f_{\boldsymbol{\beta}}(\mathbf{Z}) + \beta_i h$$

Where $\mathbf{Z}^{z_i+h}$ denotes the matrix obtained from the input matrix $\mathbf{Z}$ by perturbing its i-th column, denoted as z_i, through the addition of a quantity h, where $h \in \mathbb{R} \setminus \{0\}$.

To integrate β_i within SEANN, we enforce a similar relationship between our model's predicted values $f_\theta(\mathbf{X})$ and predicted values with perturbed inputs $f_\theta(\mathbf{X}^{x_i+h})$ as a soft constraint, i.e., $f_\theta(\mathbf{X}) = f_\theta(\mathbf{X}^{x_i+h}) - \beta_i h$, by penalizing the deviation from this equality.

For a vector V of SRCs derived from the literature, with $v_i \in V$, the i^{th} element of V, the training loss term $\mathcal{L}_{meta}$ is defined as:

$$\mathcal{L}_{meta}(\mathbf{X}, \theta, v_i, h_i) = \frac{1}{n} \sum_{k=1}^{n} \Big(f_\theta(\mathbf{X}_k^{x_i+h}) - v_i h_i - f_\theta(\mathbf{X}_k) \Big)^2 \tag{5.2}$$

Where $\mathbf{X}_k$ denotes the k^{th} row vector of matrix $\mathbf{X}$. f_θ is the output of the neural network with parameters θ, n the number of data points (i.e., batch size), and $h_i \in \mathbb{R} \setminus \{0\}$ a perturbation parameter. In this case, as SRCs (similar to other PESs) are constant regardless of input data $\mathbf{Z}$, we can theoretically use any value other than 0. For simplicity, we use $h_i = 1$ for every SRCs to integrate. In a hypothetical, more general case of a constraint to integrate as a function of $\mathbf{X}$, h would be taken as the smallest possible [2].

5.2.2 Case of an odd-ratio

The approach we proposed in this section, while mathematically correct, can suffer from numerical instability during the training (cf., **Eq. 5.3**). In paper 3, we addressed this issue by generalizing equation 5.2 instead.

Similar to section 5.2.1, let us consider a single OR, referred to as ($OR_i = e^{\beta_i}$), that we want to integrate into the training process of a DNN. This OR would be extracted from logistic regression models in a meta-analysis.

[2] The underlying property is explained more extensively in the next chapter where we extend this equation to any locally differentiable function.

Considering a multivariate logistic regression model defined as:

$$p_\beta(\mathbf{Z}) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{m_2} \beta_j z_j\right)}}$$

With $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ an input matrix and $\boldsymbol{\beta}$, a vector of m_2 log-odds coefficients, $\beta_i \in \boldsymbol{\beta}, 1 \leq i \leq m_2$. We can express the change in $Logit(p_{\boldsymbol{\beta}})$ when modifying the input variable z_i associated with β_i with a perturbation step h. This difference is independent of other variables in $\mathbf{Z}$.

$$log\left(\frac{p_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i - h})}{1 - p_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i - h})}\right) - log\left(\frac{p_{\boldsymbol{\beta}}(\mathbf{Z})}{1 - p_{\boldsymbol{\beta}}(\mathbf{Z})}\right) = -\beta_i h$$

Thus, we can express the corresponding relationship between the predicted values on inputs $\mathbf{Z}$ with and without modifying the corresponding input variable z_i with a perturbation step h:

$$p_{\boldsymbol{\beta}}(\mathbf{Z}) = \frac{e^{\beta_i h} p_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i - h})}{e^{\beta_i h} p_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i - h}) - p_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i - h}) + 1}$$

For a vector V of log-odds coefficients derived from the literature, with $v_i \in V$, the i^{th} element of V, the training loss term $\mathcal{L}_{meta}$ to integrate v_i is defined as in Eq. 5.3

$$\mathcal{L}_{meta}(\mathbf{X}, \theta, v_i, h_i) = \frac{1}{n} \sum_{k=1}^{n} \left(\frac{e^{v_i h_i} p_\theta(\mathbf{X}_k^{x_i - h})}{e^{v_i h_i} p_\theta(\mathbf{X}_k^{x_i - h}) - p_\theta(\mathbf{X}_k^{x_i - h}) + 1} - p_\theta(\mathbf{X}_k) \right)^2 \tag{5.3}$$

Where p_θ is the probability given by the neural network with parameters θ, n the number of data points (i.e., the batch size) and $h \in \mathbb{R} \setminus \{0\}$ a perturbation parameter. Similar to section 5.2.1, as a given OR is constant for every corresponding z, h can theoretically take any values other than 0. However, within SEANN, we fix h to keep the quantities within the exponential terms small and enhance numerical stability during the learning process. We define:

$$h = \begin{cases} 1 & \text{if } v_i = 0, \\ \frac{1}{v_i} & \text{otherwise.} \end{cases}$$

5.2.3 Case of a risk ratio

Following the same principle, let's define the integration of a single PES encoded as a risk ratio. Considering a negative binomial regression model defined as:

$$log\left(\mu_{\boldsymbol{\beta}}(\mathbf{Z})\right) = \beta_0 + \sum_{j=1}^{m_2} \beta_j z_j$$

With $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ an input matrix and $\boldsymbol{\beta}$, a vector of log-estimates, $\beta_i \in \boldsymbol{\beta}, 1 \leq i \leq m_2$. Then the expected change in the target values according to β_i when modifying the corresponding input variable z_i with a perturbation step h is:

$$\mu_{\boldsymbol{\beta}}(\mathbf{Z}^{z_i+h}) = e^{\beta_i h}\mu_{\boldsymbol{\beta}}(\mathbf{Z})$$

Where $\mathbf{Z}^{z_i+h}$ denotes the matrix obtained from the input matrix $\mathbf{Z}$ by perturbing its i-th column, denoted as z_i, through the addition of a quantity h, where $h \in \mathbb{R} \setminus \{0\}$.

To integrate β_i within SEANN, we enforce a similar relationship between our model's predicted values $f_\theta(\mathbf{X})$ and predicted values with perturbed inputs $f_\theta(\mathbf{X}^{x_i+h})$ as a soft constraint, i.e., $f_\theta(\mathbf{X}) = f_\theta(\mathbf{X}^{x_i+h})e^{-\beta_i h}$, by penalizing the deviation from this equality.

For a vector V of log-estimates derived from the literature, with $v_i \in V$, the i^{th} element of V, the training loss term $\mathcal{L}_{meta}$ is defined as:

$$\mathcal{L}_{meta}(\mathbf{X}, \theta, v_i, h_i) = \frac{1}{n}\sum_{k=1}^{n}\left(e^{-v_i h_i} f_\theta(\mathbf{X}_k^{x_i+h}) - f_\theta(\mathbf{X}_k)\right)^2 \tag{5.4}$$

Where $\mathbf{X}_k$ denotes the k^{th} row vector of matrix $\mathbf{X}$. f_θ is the output of the neural network with parameters θ, n the number of data points (i.e., batch size), and $h_i \in \mathbb{R} \setminus \{0\}$ a perturbation parameter. Similar to **Section 5.2.2**, we recommend using:

$$h = \begin{cases} 1 & \text{if } v_i = 0, \\ \frac{1}{v_i} & \text{otherwise.} \end{cases}$$

To demonstrate the approach's potential, we rely on synthetic data that emulates different scenarios. In each experiment, we compare two DNNs,

identical in all aspects but the inclusion of our modified loss and external knowledge. For the sake of simplicity, we use and compare basic multilayer perceptrons. The approach can be directly usable with more complex feed-forward neural architectures (e.g., convolutional networks, residual networks), and the benefits highlighted in this study should apply to other neural configurations. Below, we call SEANN the model that implements our approach, and *agnostic DNN* the reference one.

5.3 Experimental validation

5.3.1 Data scenario

To illustrate the relevance in real applications, we introduce an intuitive fictional example composed of 1) a target variable y representing the risk of developing a disease or the strength of symptoms and 2) several variables contributing to the outcome y according to a dose-response relationship. Those variables are constructed to test specific hypotheses. First, we define two correlated variables, *mercury* (x_1) and *fish intake* (x_2), having an opposite effect on the target variable. Typically, the fish intake reduces the risk, while mercury increases it. Correlation between x_1 and x_2 is designed to emulate a confounding effect [Jager et al., 2008], a common issue in health research. We also define two additional variables, namely *perceived stress* (x_3) and *body mass index* (i.e., BMI, x_4) uncorrelated to the variables x_1 and x_2 but affecting y. x_3 is linear and positively correlated with the outcome, while x_4 has a nonlinear effect.
In this simple scenario, we perform different experiments in which we test the benefits of incorporating PESs to eligible variables (i.e., x_1, x_2 and x_3). We are interested not only in the networks' predictive performances but also in their ability to capture and restitute the input-output relationships that we encoded in the data. The experiments presented focus on SRCs and ORs, but we could use RRs in a similar manner.

5.3.1.1 Standardized regression coefficients

For the case where PESs are encoded as SRCs, we generate an input matrix $\mathbf{X}$ by sampling $m = 1000$ values from a multivariate Gaussian with mean 0 and covariance matrix $\begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. Target vector Y is generated from the

additive function described in Eq. 5.5 with $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = -2$, $\beta_3 = 5$ and $\beta_4 = 10$.

$$Y(\mathbf{X}) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + \beta_4 \times cos(x_4) \tag{5.5}$$

5.3.1.2 Odd-ratios

Similar to the linear case, we generate a data matrix $\mathbf{X}$ with three variables, namely x_1, x_2, and x_3, corresponding to mercury, fish intake, and perceived stress, respectively, to predict an outcome (i.e., a risk to develop a disease). $\mathbf{X}$ was generated by sampling $m = 1000$ values from a multivariate Gaussian with mean 0 and covariance matrix $\left(\begin{smallmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$. Target vector Y was generated from the function described in Eq. 5.6 with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = -2$, $\beta_3 = 5$.

$$Y(\mathbf{X}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 \times x_1 - \beta_2 \times x_2 - \beta_3 \times x_3}} \tag{5.6}$$

5.3.1.3 Experimental design

We use a fully connected neural network (NN) with a single hidden layer for both SEANN and the agnostic model. Both NNs were implemented using Pytorch and trained with a batch size of 64 and a maximum number of epochs of 1000. Parameter optimization was achieved using Adam [Kingma and Ba, 2017]. We standardize and split data into training (n=600), validation (n=200), and test (n=200) datasets. To reduce over-fitting, we use early stopping (with patience 10) on the validation set.

5.3.1.4 Evaluation

To evaluate the correctness of extracted relationships, we propose a score, called $\Delta Shap$, to compute the distance between two dose-response relationships represented with Shapley values [Shapley, 1953]. $\Delta Shap$ is defined by the mean absolute error (MAE) computed across Shapley values for a given marginal relationship that must be computed using the same background dataset. We can use it to compare the distance between a neural network-extracted relationship and a relationship admitted in the literature or, in this work, as we use synthetic data, to compare the distance between a neural network-extracted relationship and the true predictor-outcome relationship.

The smaller the distance, the more we would consider a relationship to be scientifically plausible or in line with the true effect. In this work, we use the generative functions (i.e., **Eq. 5.5** and **Eq. 5.6**) to compute Shapley values representing the reference relationships for $\Delta Shap$. Shapley values are approximated using the SHAP library [Lundberg and Lee, 2017] and systematically computed on the test sets.

To evaluate the performance of models, we use the coefficient of determination (R^2) score for regression tasks and the receiver operating characteristic curve's (ROC) area under the curve (AUC) for binary classification tasks (i.e., odd-ratios).

5.3.2 Experiment 1

In this experiment, we illustrate that SEANN can leverage the information from external expert knowledge to mitigate the poor quality of the data. To simulate the data imperfection, we progressively increase the level of noise on all variables (i.e., x_1, x_2, x_3, x_4) and check that while the performance of the agnostic NN deteriorates quickly, our informed NN can retain most of it. Information was degraded differently for linear coefficient and odd ratios to illustrate two common scenarios. In the linear case, we added Gaussian noise to all input variables, with a mean of 0 and increasing standard deviation. For odd ratios, missing values were generated completely at random with increasing proportions and imputed using a simple mean imputation. External knowledge was integrated on top of training data for every eligible predictor (i.e., $beta_1$,.., $beta_3$ for x_1, ..., x_3 respectively).

Better performances were obtained with SEANN both for the predictive task and the explainability of constrained relationships measured with $\Delta Shap$. No significant gains were observed for the nonlinear unconstrained variable (BMI). Results are displayed in **Figure 5.1** for linear coefficients and summarized in **Table 5.1** for odd-ratios. Results indicate that given PESs encoding correct relationships between input variables and target outcome, performances obtained while training on imperfect data can be more stable with SEANN.

Percent missing	Agnostic DNN		SEANN	
	ROC AUC	Sum Δ Shap	ROC AUC	Sum Δ Shap
0	0.999	0.083	0.998	0.137
25	0.930	0.262	0.975	0.176
50	0.915	0.318	0.95	0.204

Table 5.1: Comparison of Performances depending on the proportion of imputed missing values in training and validation sets (experiment 2).
ROC AUC is a measure of predictive performances, whereas Sum Δ Shap summarizes the quality of captured relationships across all predictors.

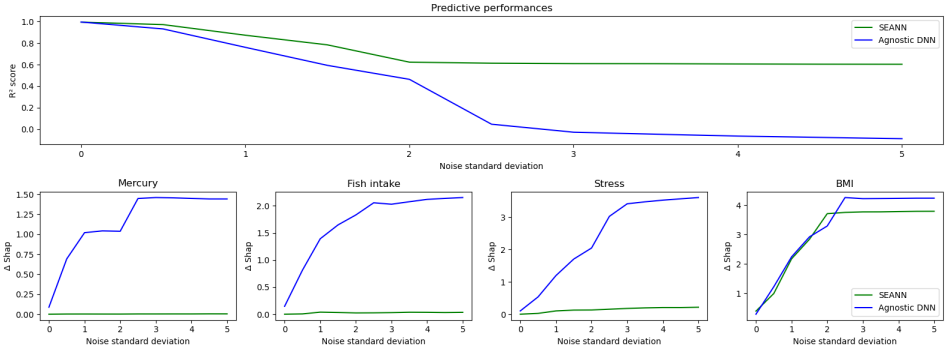


Figure 5.1: Performance comparison of SEANN and agnostic DNN with different noise levels in input features (experiment 1).
The X-axis is the standard deviation of added noise.

5.3.3 Experiment 2

In this experiment, we focus on the quality of the relationships captured when external knowledge is integrated for a single predictor. The objective is to show that variables that do not benefit directly from external knowledge may nevertheless be better captured, thanks to corrections brought to the other variables. Similar to the previous experiment, Gaussian noise was added to all variables with mean 0 and standard deviation 0.75, and 1.5 for the SRCs and ORs respectively, in both training and validation sets.

Figure 5.2 show results with PESs encoded as SRCs. The most significant gain was observed for fish intake, the variable with integrated external knowledge. $\Delta Shap$ (see definition in **Section 5.3.1.4**) measured for this variable was 1.11 with the agnostic DNN and 0.04 with SEANN. A significant gain was also observed for mercury, the variable correlated with the

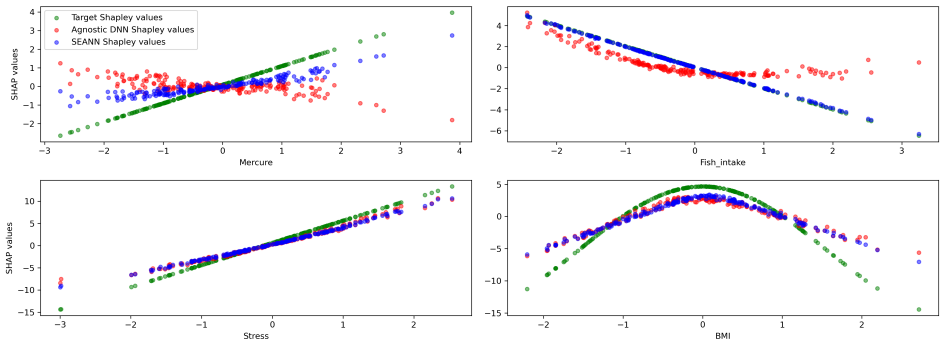


Figure 5.2: Comparison of extracted relationships (Shapley values) between the agnostic DNN and SEANN (experiment 2).
Only the beta coefficient for fish intake was added as external knowledge. However, both mercury and fish intake are better captured by SEANN compared with the agnostic DNN.

informed one (1.02 for agnostic DNN to 0.53 for SEANN). Finally, no significant performance gains were observed for the remaining variables, i.e., those uncorrelated with the informed one (0.99 to 1.04 for perceived stress, 1.74 to 1.64 for BMI, with agnostic DNN and SEANN, respectively). Results show that not only SEANN better captures relationships for features with corresponding external knowledge but also for noninformed features that are correlated with those externally informed.
A similar scenario is observed for odd ratios, with $\Delta Shap$ down from 0.087 with the agnostic DNN to 0.050 with SEANN for mercury.

5.3.4 Experiment 3

In a last experiment, we simulate a setting where a confounding variable is missing from the data. A confounding variable is a predictor impacting both the outcome to predict and other predictor(s) of interest. In numerous contexts, including health science, it is challenging to collect all relevant variables to predict an outcome (and study their effects), and we can expect to have unseen confounders.
We train both NNs with a missing variable (fish intake) and compare both the predictive performance and the quality of extracted relationships on the test set. With SEANN, we integrate external knowledge for mercury alone (i.e., the variable correlated with the missing predictor), and we duplicate the mercury variable in both training and validation datasets, the copy being

an unconstrained variable. The objective of this duplication is that 1) The constrained version captures what is known by external knowledge, and 2) The unconstrained version captures what comes from the unseen confounder.

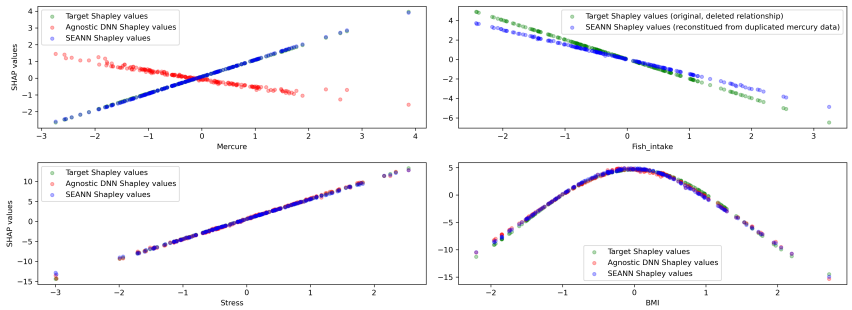


Figure 5.3: Comparison of extracted relationships (Shapley values) between the agnostic DNN and SEANN for the linear case (experiment 3).
Without external knowledge, the interpretation of the Mercury effect is opposed to the ground truth. When adding the constraint, we see that the duplicated variable can capture the confounder, i.e., fish intake.

We observe (**Fig. 5.3**) that without the constraint, mercury is incorrectly captured, with a $\Delta Shap$ of 1.32. In this case, interpreting the Shapley values directly could lead to the misleading conclusion that mercury has a protective effect. On the contrary, with SEANN, the constraint allows the capture of the correct relation ($\Delta Shap$=0.013). Additionally, SEANN was able to capture part of the association with the missing variable (fish intake) using duplicated input data of mercury ($\Delta Shap$: 0.43). Minor improvements were also observed for other variables. Results show that SEANN can be used to better disentangle individual effects while estimating the effect of unknown confounding factors.
For odd ratios, the results are similar. SEANN better captured the mercury predictor, with $\Delta Shap$=0.128 for the agnostic DNN and 0.048 for SEANN. SEANN was also able to capture part of the association with the missing variable (i.e., fish intake) using the duplicated input data ($\Delta Shap$= 0.056).

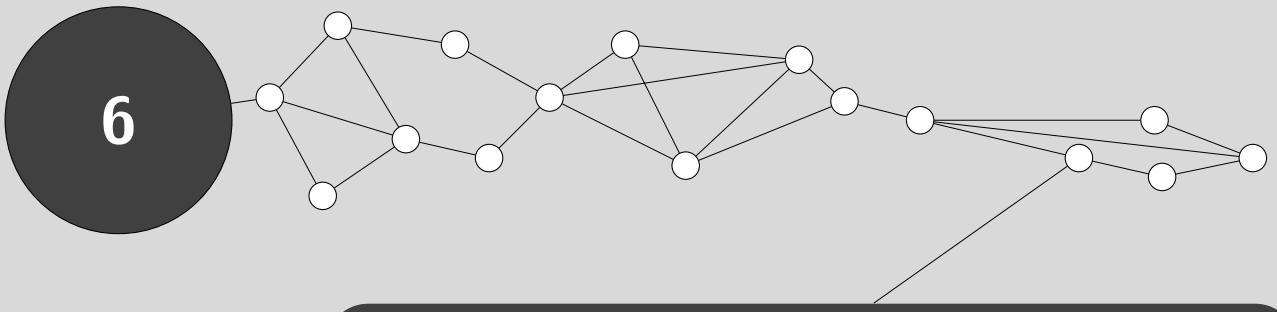
5.4 Conclusion

In this paper, we propose a method to integrate the wealth of knowledge available in the scientific literature encoded as PESs. While these representations are simple estimates unable to express complex relationships, they

are easily understandable, can be aggregated across multiple studies, and are widely used in multiple domains of science, including epidemiology in particular. By integrating traditional statistical measures into the deep learning process, our approach offers a tool to capture complex nonlinear relationships from data while leveraging simpler but well-established knowledge.

Our experimental protocol demonstrates that, theoretically, compared with a standard DNNs, our approach offers two main benefits. First, a better generalization of the predictive performances on unseen data could be obtained under the condition that PESs contain relevant information for the task at hand that is lacking in the available data. This is a common use case, notably in epidemiology, where vast amounts of data are scarce and independent relationships are well studied across multiple studies on different populations. Second, significant improvements were observed in the alignment of extracted relationships with external knowledge when using SEANN, both for informed and uninformed variables. In particular, we demonstrated its potential to better disentangle individual input-output relationships in the presence of collinearity.

A common limitation of approaches that, similar to ours, integrate soft constraints through additional terms in the loss function is the introduction of additional hyperparameters. We propose a way to estimate those, adapted to our application case, based on the relative confidence in each study from which the PESs are taken. In the next chapter, we validate this approach using real data.



6 An Informed Environmental Risk Score for Adult Hypertension

This chapter relates the work performed for the third paper of this doctoral work (not yet published at the time of writing). In this work, we illustrate SEANN's potential in a real case scenario by computing an Informed ERS for hypertension in adults. Those scores were derived from a wide range of exposures measured in the GCAT cohort and "informed" by multiple PESs estimated in meta-analyses. By comparing the relationships extracted with SEANN with those obtained from an agnostic machine learning model, that learned only from the input data, we empirically demonstrate the benefits of our approach. Compared to an agnostic neural network, SEANN learns relationships more aligned with established scientific knowledge and better disentangles each effect. Similarly to Chapter 4, we leverage our risk score to identify actionable factors of diseases, assess their overall impact on the risk, and estimate the nature of their relationships with health.

6.1 Introduction

Defined as elevated blood pressure to an unhealthy level, hypertension is the leading preventable cause of cardiovascular diseases (CVDs) worldwide [Reuter and Jordan, 2019], with direct implications in a wide range of adverse conditions such as heart failure, stroke or atherosclerosis. CVDs are the leading cause of mortality worldwide, accounting for an estimated 17.9 million lives each year, according to the World Health Organization. From those deaths, 8.5 million deaths worldwide (4.5 million men and 4 million women) were attributable to elevated systolic blood pressure (i.e., >115 mmHg) [Zhou et al., 2021].

Well-established CVD factors, easily measurable in clinical settings, such as age, sex, blood pressure, body mass index and current smoking have been used as predictors for diagnostic risk scores for decades, e.g., [D'Agostino et al., 2008, Woodward et al., 2007, Achenbach et al., 2021]. The integration of representative arrays of genetic and environmental factors into these scores represents a more recent advancement [Bhatnagar, 2017, O'Sullivan et al., 2022]. This evolution in approach is critical, as chronic diseases are increasingly recognized as the outcome of a complex interplay between

genetic predispositions and environmental exposures [Hunter, 2005]. However, while the integration of polygenic factors into these risk scores has slightly improved performances of predictions, in particular for early-onsets, the contribution of environmental CVD factors such as sedentary lifestyle, unhealthy diet, social stress, air pollution or traffic noise have not been as systematically assessed but rather studied as single causes. Acknowledging and incorporating the multifactorial nature of non-communicable disease into risk assessments is essential for a more comprehensive and accurate understanding of disease aetiology and the development of more effective preventive and therapeutic strategies.

A growing research paradigm in recent environmental epidemiological studies, called the exposome [Siroux et al., 2016], diverges from the traditional studies focusing on only one environmental exposure-health relationship to endorse a more holistic approach. Classical single exposure analyses may be limited because the studied exposure association could arise from another correlated factor not taken into account and are, moreover, unable to capture interactions or cumulative effects from the exposure mixture. Within this framework, new data analysis approaches are developed, combining machine learning predictive power and explainable AI as we [Guimbaud et al., 2024] and others [Atehortúa et al., 2023] have recently proposed in order to capture those complexities while providing intelligible insights. Those approaches, however, are still limited by the presence of confounding biases [Jager et al., 2008] and multicollinearities, impacting the quality of extracted relationships.

Moreover, the generalizability of data-driven risk scores needs to be validated in external populations, both in terms of predictive performances and robustness of captured relationships. However, compared to polygenic risk scores for which all the genes are measurable at once, the validation of environmental risk scores is more challenging as the diversity of environmental exposures and methods to measure them makes it difficult to find different populations with the same measured exposures. A wealth of knowledge is however available in the literature, regarding certain environmental exposure-health relationships considered independently across several studies and populations. While limited in several aforementioned aspects (i.e., they ignore non-linearities and interactions), those pooled effect size estimates represent one of the strongest currently available evidence in the field

and can be leveraged to guide the training of multifactorial risk scores [Rosner, 2012].

In this work, we applied SEANN, our informed ML approach described in the previous chapter, in a real-life scenario on observational data. With SEANN, we aimed to calculate an environmental risk score for hypertension encompassing a wide range of factors in an adult Spanish population. This approach incorporates literature knowledge, more specifically well-known exposure-health relationship estimates, into the training of deep neural networks. Leveraging the predictive capabilities of deep learning, SEANN is able to capture complex relationships that can be further extracted with explainable AI. More importantly, by learning well-known relationships more aligned with the scientific consensus, SEANN is able to better disentangle each individual effects. By comparing the relationships extracted with SEANN with those obtained from an agnostic machine learning model, that learned only from the input data, we empirically demonstrate the benefits of our approach. Compared with risk scores used as diagnosed tools that focus only on predictive abilities [D'Agostino et al., 2008, Ulusoy, 2013, Ahsan and Siddique, 2022], and are hence built using predictors strongly associated with the outcomes such as body mass index (BMI) or blood pressure for CVDs [1], our score is designed to extract more reliable and informative insights about environmental and socio-economic stressors and would be usable as a better-informed decision-support tool.

6.2 Methods

6.2.1 Study participants

This study uses data from the GCAT project, a prospective cohort study designed to recruit middle-aged adults from the general population of the region of Catalonia in Spain. With the aim of identifying chronic disease events in the mid-term, the project covers 19 209 adults at baseline aged 40-65, from whom written informed consent was obtained. Participants were recruited across all of Catalonia from 11 permanent recruitment centers. A more detailed description of this project is available elsewhere [Obón-

[1] As those factors encode the body response to exposure, not the exposures themselves, including them in a risk score is likely to impact the quality of the captured exposure-health relationships due to confounding.

Santacana et al., 2018].
Within the 18337 total individuals from which a diagnosis of hypertension has been obtained, 59% were females. A vast majority (98%) were born in Spain, are Caucasians (84%) and were married (65%). A vast majority also lived in cities (90%, vs 9% in suburban areas and 1% in rural areas). A more detailed description of the population is given in **Supplementary figure 1**.
All predictors used in this study were collected at baseline (between 2014 and 2017). Diagnoses of hypertension come from follow-up data collected until 2022.

6.2.2 Outcome

Based on EHR data collected until 2022, individuals were considered to have hypertension if they had at least one diagnosis of hypertension at any point or if they were taking hypertensive-related medication (i.e., anatomical therapeutic chemical codes C02, C03, C07, C08, and C09). From 18337 individuals with available information, 4592 were categorized as hypertensive.

6.2.3 Predictors

Figure 6.1 displays an overview of the diversity of exposures included in this study. A complete description is available in **Supplementary Table 1**. Lifestyle information (i.e., sociodemographic and socio-economic status, occupation, diet, and tobacco-alcohol consumption) used in this study was collected from a questionnaire submitted at baseline. A more detailed description has been presented in [Obón-Santacana et al., 2018]. Environmental exposures (i.e., green spaces proximity, air pollution, noise, degree of urbanization) were assessed using a geographical information approach based on the residence localization of participants. This was also described in more detail in [Obón-Santacana et al., 2018].

6.2.4 Ethics approval

Informed consent was obtained from all participants in the study.

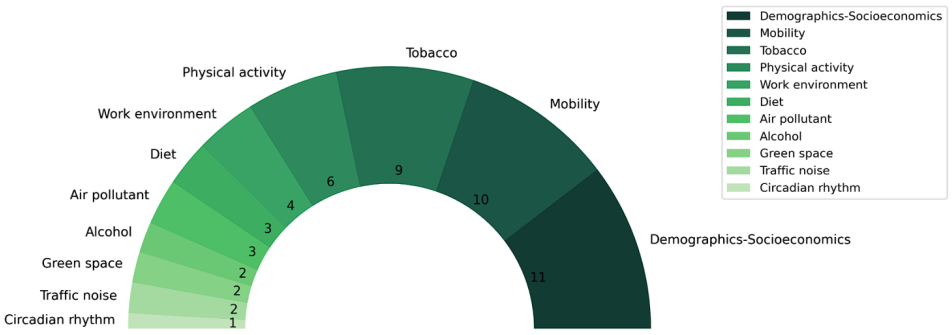


Figure 6.1: Exposome factors and their families assessed in the GCAT cohort, collected at baseline (2014-2017) and used in the Environmental Risk Score for Hypertension.
The counts of exposures within the circle were carried out before one hot encoding for multi-categorical variables (53 exposures vs 99 after this step).

6.2.5 Machine learning pipeline

In the following section, we describe the machine learning procedure performed in this study. This procedure was performed using Python 3.10.13.

6.2.5.1 Data preparation

Several steps were required before training the neural networks and obtaining the risk score models. The initial steps involved preparing the input data. Categorical variables were either one-hot encoded or encoded with floating values, depending on their nature (e.g., frequencies or binned continuous variables). After this encoding step, the final dataset comprised 99 variables. Missing values within the dataset, which accounted for 7.71% of the data overall, were imputed using MissForest [Stekhoven and Bühlmann, 2011], a single imputation algorithm able to handle both categorical and continuous variables while capturing nonlinear relationships. Within the imputed variable, no variable had more than 42% missingness. The input data was finally standardized before being fed into the neural networks.

6.2.5.2 Literature effect sizes

To build an informed risk score, we collected external knowledge, i.e., different from the input data, in the form of PESs reported in the domain

literature. To that end, a literature review was performed in order to extract suitable effect sizes. All relevant studies were identified by searching PubMed and Google Scholar databases on March 2024. The keywords used in the search queries were the name of the exposure of interest (for instance "sedentary behaviour"), "hypertension", "odds ratio" and "meta-analysis". A first selection was made based on the titles' relevance, and a second, based on the selected full texts. **Table 6.1** lists the final 11 PESs included in this work.

Table 6.1: Selected meta-analysis with corresponding pooled effect estimates included as external knowledge in the informed risk score.

Exposure	Estimate (with CI)	Lit. Unit	Input Unit	Sample Size	First Author(s)	Year
Current smoker	1.13 (0.93-1.37)	Binary	Binary	250741	[Guo et al., 2011]	2011
Alcohol consumption	1.06 (1.01-1.11)	10g/d	UBE/wk	414477	[Liu et al., 2020]	2020
Mediterranean Diet	0.87 (0.78-0.98)	Binary	Binary	59001	[Cowell et al., 2020]	2021
Sedentary time	1.04 (1.00-1.07)	1h/d	1h/w	367264	[Guo et al., 2019]	2019
Physical activity	0.94 (0.92-0.96)	10 METh/w	METh/w	330222	[Liu et al., 2017]	2017
NDVI	0.97 (0.96-0.99)	0.1 unit	1 unit	100×10^6	[Liu et al., 2022]	2022
NO_2	1.05 (1.02-1.18)	10µg/m3	1µg/m3	29274	[Yang et al., 2018]	2018
$PM_{2.5}$	1.10 (1.06-1.13)	10µg/m3	1µg/m3	20006	[Yang et al., 2018]	2018
O_3	1.05 (0.98-1.12)	10µg/m3	1µg/m3	27783	[Yang et al., 2018]	2018
Deprivation index	1.14 (1.01-1.30)	N/A	N/A	62×10^6	[Satapathy et al., 2024]	2024
Traffic noise (day)	1.02 (0.98-1.05)	10 dB	1 dB	5.5×10^6	[Dzhambov and Dimitrova, 2018]	2018

Abbreviations: CI - Confidence Interval; NDVI - Normalized Difference Vegetation Index; NO_2 - Nitrogen Dioxide; $PM_{2.5}$ – Particulate Matter 2.5; O_3 – Ozone.

Meta-analyses were accepted only if 1) They assessed the association between hypertension or pre-hypertension and a predictor directly computable within our available input data. 2) If they reported a pooled effect estimate

in the form of odds ratios or risk ratios. Similar to [Liu et al., 2020, Liu et al., 2017, Satapathy et al., 2024], we assumed that risk ratios were approximately ORs for hypertension [Orsini et al., 2011]. 3) Their reported estimates for adults or the general population. 4) The publication date was after 2010. Meta-analyses that used both cross-sectional and longitudinal studies were accepted. When encountering multiple analyses concerning the same predictor, preference was given to the study with the largest sample size.

Prior to standardization, each literature estimate was converted to match the unit of the corresponding input data. For standardized predictors, the standard deviation of each predictor was used to recalibrate the literature estimates into the appropriate units.

6.2.5.3 Machine learning workflow

An overview of the study workflow is presented in **Figure 6.2**. A reference risk score was first obtained using a standard feed forward deep neural network only trained using the input data. The shape of this neural network (i.e., the number of layers and neurons within each layer) has been optimized using the Tree-structured Parsen Estimator [Bergstra et al., 2011] from the Optuna library within a 10-fold cross-validation procedure. **Supplementary table 2** displays the list of hyperparameters used within the study.

Once the structure of the reference model was obtained, we built a second, exactly equivalent that would be trained with additional knowledge extracted from the literature. As the proposed approach is quite generic, in the sense that it can be directly used with most feed-forward neural architectures, this point ensures a fair comparison that focuses only on the external knowledge integration.

Data was randomly split into training, validation and test datasets using 60%, 20% and 20% of the original sample size respectively. The training set was used to train both NNs; the validation set was used to perform early stopping and the test set was used to compute predictive performances (ROC AUC) as well as to approximate Shapley values using the SHAP [Lundberg and Lee, 2017] python package.

Shapley values were used to define exposure-outcome associations encoded in the literature or captured by each of the two neural networks. For literature-

extracted exposure-outcome relationships, we computed Shapley values, using also the SHAP package, by constructing univariate logistic regression models according to the literature's pooled ORs. As those models used fixed ORs as coefficients, they were not trained on the data.
From those Shapley values, we computed a simple measure, delta SHAP (cf. **Section 5.3.1.4**), of the distance between two marginal relationships, between the same factors, captured differently. In this work, we use it to compute the distance between a neural network extracted relationship and the relationship that is admitted in the literature from meta-analysis. The smaller the distance, the more we would consider a relationship to be scientifically plausible.
We present our approach integrating literature effect size estimates within the training of neural networks in the next section. It follows a common principle in constrained learning [Fajemisin et al., 2024], which is the presence of a tradeoff between respecting the constraints (i.e., literature-extracted marginal relationships) and learning from the data. As such, our approach seeks to optimize two objectives: First, minimizing the NN error on the datasets and, hence, indirectly maximizing the predictive performances on the held-out dataset; Second, maximizing the concordance with the literature-extracted relationships. We determined this trade-off using a Pareto frontier combined with an elbow method.

6.2.6 Introducing SEANN

In this section, we present a refined version of SEANN, which was, to the best of our knowledge, the first approach to integrate literature effect size as soft constraints within the training of deep neural networks. Within SEANN, several types of estimates (i.e., linear betas, odds ratios, risk-ratios) are covered (cf., **Section 5.2**). However, in this work, we focused on odd ratios, a commonly reported form in epidemiology.
Similar to what was described in the previous chapter, the general principle of our method described in **Eq. 5.1** consists of adding a term to the loss function L for each meta-heuristic to incorporate.
In this work, we propose a refined approach to determine the values of λ_0 and λ_i in **Eq. 5.1**. Those parameters are weighting the relative importance of the original error term and the penalty terms, and regulate the learning from both input data and external knowledge. Compared with the approach

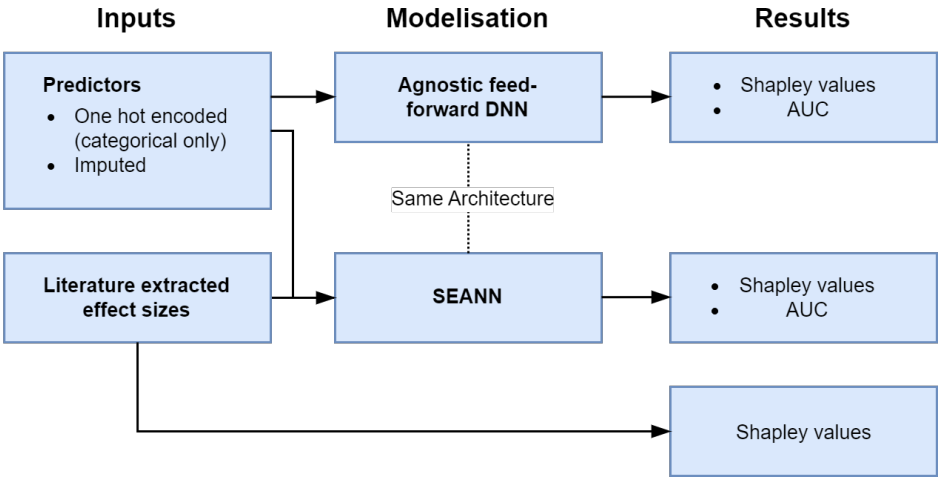


Figure 6.2: Simplified overview of the analysis workflow.
Shapley values are used to extract marginal relationships and to compare them with those extracted in the literature. AUC (Area Under the receiver operator characteristic Curve) is used to compare predictive performances.

proposed in the previous chapter, this one is more flexible and allows the user to choose a preferred trade-off value (between the plausibility of learned relationships and predictive performances) among efficient solutions. We decompose each λ into the following terms:

$$\begin{cases} \lambda_0 = \alpha_0 \times \gamma \\ \lambda_i = \alpha_i \times (1-\gamma) \times \omega_i \end{cases} \tag{6.1}$$

where $\alpha \in \mathbb{R}$ is a standardization coefficient, ensuring each λ is comparable with the others, $\gamma \in [0,1]$ is the trade-off hyperparameter settled by the user that gives the importance of learning relationships from the data vs respecting the constraints, and $\omega \in [0,1]$ is a relative importance coefficient within each constraint. The first step is to determine each α depending on L_{pred} and L_{meta} such as $\alpha_0 = \frac{1}{\mathbb{E}_{\mathbf{X}}[L_{pred}(\mathbf{X},\theta_0)]}$ and $\alpha_i = \frac{1}{\mathbb{E}_{\mathbf{X}}[L_{meta}(\mathbf{X},\theta_0,v_i,h_i)]}$. We simply express $\mathbb{E}[L_{pred}(\mathbf{X},\theta_0)]$ as $L_{pred}(\mathbf{X},\mathbb{E}[f_{\theta_0}(\mathbf{X})])$ and similarly $\mathbb{E}[L_{meta}(\mathbf{X},\theta_0,v_i,h_i)]$ as $L_{meta}(\mathbf{X},\mathbb{E}[f_{\theta_0}(\mathbf{X})],v_i,h_i)$, $f_{\theta_0}(\mathbf{X})$ being the prediction of the neural network with initial parameters θ_0. In this work, we are predicting a binary target. Thus, we set $\mathbb{E}[f_{\theta_0}(\mathbf{X})] = 0.5$. L_{pred} is the binary cross-entropy and L_{meta} is defined below. Then, we determine ω to be proportional to the level

of confidence within each study from which we extract the effect sizes. In this work we used each study sample size to determine its confidence score. Denoting c_i the confidence (i.e., the sample size) of the i-th study, we determine ω as $\omega_i = \frac{c_i}{\sum_q^{j=1} c_j}$. Finally, the hyperparameter ω is determined using an elbow method on a Pareto frontier as described in **Section 6.2.5.3**.

We now explain how to determine the loss function terms L_{meta} in **Eq. 5.1**. Depending on the type of exposure-outcome relationship (e.g., OR, RR, linear coefficient), the terms will be implemented differently. Similar to the previous chapter, the general principle of our approach is to generate, for each observation, a slightly perturbated copy of it with an increment h—called the perturbation—for each variable in Q, in order to measure the difference between the expected change in the target value, according to the literature extracted relationship, and the observed change in our model, and to penalize this difference. To address problems of numerical instability when enforcing certain forms of non-linear estimates such as ORs (cf. **Eq. 5.3**), we now compute this expected change using the local derivative of the literature relationships (as ORs for instance, are logistic estimates, the encoded relationship is derivable).

Let us consider ψ a univariable function we want to enforce that is derivable locally on an input variable $x \in Q$ (the relation encoded by an odd ratio in a meta-analysis for instance). We know that

$$h\frac{\delta\psi}{\delta x_0} = \lim_{x\to 0} \psi(x_0 + h) - \psi(x_0)$$

Then, if $h\frac{\delta\psi}{\delta x}$ exists and is known locally for every x, we can easily construct the following loss function that penalizes the divergence from the relation ψ.

$$\mathcal{L}_{meta}(\mathbf{X}, \theta, v_i, h) = \frac{1}{n}\sum_{k=1}^{n}\left(f_\theta(\mathbf{X}_k^{x_i+h}) - h\frac{\delta\psi(x_i, v_i)}{\delta x_i} - f_\theta(\mathbf{X}_k)\right)^2 \quad (6.2)$$

Where $\mathbf{X}^{x_i+h}$ denotes the matrix obtained from the input matrix $\mathbf{X}$ by perturbating its i-th column, denoted as x_i, through the addition of a small quantity h. $\mathbf{X}_k$ denotes the k-th row vector of input matrix $\mathbf{X}$. f_θ is the output of the neural network with parameters θ, n the number of data points

(i.e., the batch size), and $v_i \in V$ the literature effect estimate associated with variable x_i. For odd ratios, we can directly compute $\frac{\delta\psi(x_i,v_i)}{\delta x_i}$, denoting $\sigma(x_i, v_i) = \frac{1}{1+e^{-v_i \times x_i}}$, as $\frac{\delta\psi(v_i)}{\delta x_i} = v_i \times \sigma(x_i, v_i)(1 - \sigma(x_i, v_i))$.
Despite its simplicity, this novel approach is flexible as it can directly be used to incorporate any complex relationship given the condition of being able to compute its derivative. While being able to incorporate traditional forms of PESs, its potential applications may extend beyond this use case.

6.3 Results

6.3.1 Captured associations

Figure 6.3 displays a comparison of captured associations with SEANN, the agnostic NN, and those reported in the literature. Compared with the agnostic neural network, relationships captured with SEANN are much closer to those reported in the literature. For example, when examining the Shapley values for the variable "smoking habits", the agnostic neural network identified a protective effect of smoking on the predicted risk, while in contrast, the relationship captured with SEANN aligned more closely with the literature-reported association indicating a harmful effect. This pattern of being closer to the integrated literature association is observed for every variable and confirmed by delta SHAP distances. Overall, the mean delta SHAP distance averaged over all variables of the relationships extracted with both models and those observed in the literature, was 0.8×10^{-3} with SEANN compared to 6.1×10^{-3} with the agnostic NN. Delta SHAP distances for all individual variables are displayed in **Supplementary Table 2**. More plausible directions of associations (taking the literature pooled effect sizes as reference) were observed with SEANN for smoking habits, physical activity, no_2, and traffic noise variables.
Figure 6.4 displays a comparison of the response functions extracted with SEANN and the agnostic NN fomer a subset of the remaining variables for the sake of conciseness. Response functions for all remaining variables are displayed in **Supplementary Figure 3-8**. It should be noted that due to the absence of corresponding relationships identified within the existing literature, these relationships were not subjected to constraints during the training of the SEANN model. We however observe, in both **Figure 6.4**

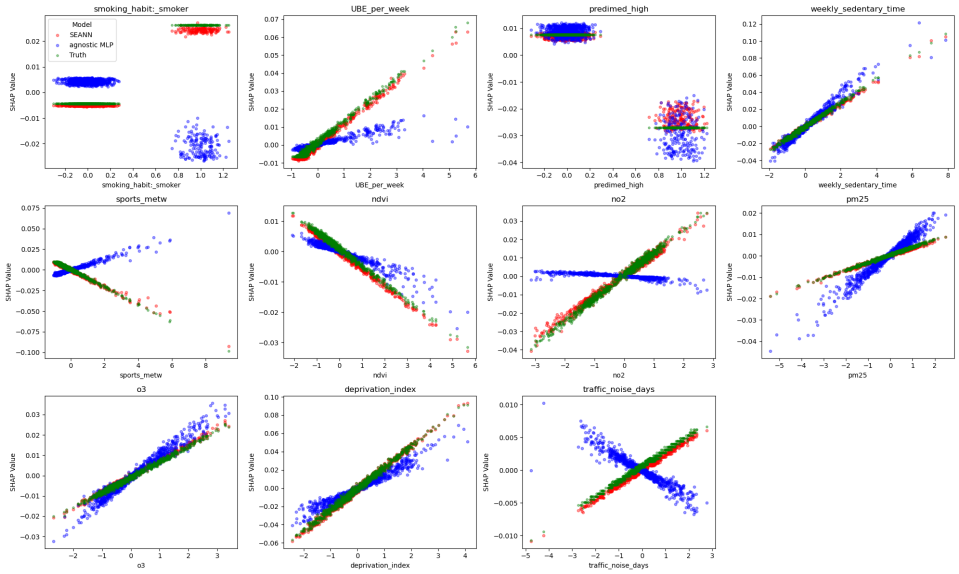


Figure 6.3: Comparison of response functions extracted from the literature, SEANN and the agnostic NN.
Dots are Shapley values, encoding the response of the model (y-axis) for a given exposure value (x-axis).
Abbreviations: UBE - *Unidad de Bebida Estándar* (Standard Beverage Unit); NDVI – Normalized Difference Vegetation Index; NO2 – Nitrogen Dioxide; PM25 – Particulate Matter 2.5, O3 - Ozone; SEANN – Summary Estimates Adapted Neural Network

and **supplementary Figure 3-8**, substantial changes in the extracted relationships of some variables extracted with SEANN compared with those extracted with the agnostic NN, confirming that the model is able to adjust its captured effects. For smoking-related variables notably, we see interesting results, with extracted relationships more inline with the known effects of smoking on hypertension. Unlike the agnostic NN, SEANN correctly captured that being an ex-smoker is generally a factor increasing the risk and more specifically, that being a recent ex-smoker is more detrimental than being 10 years ex-smoker which, in turn, is also more detrimental than being a 20 years ex-smoker. The effect of electronic cigarettes seems also to be less pronounced and the effect of smoking starting age more pronounced. These adjustments are likely to be principally due to the correction of the smoking status variable, which was incorrectly captured by the agnostic NN as a protective factor (c.f., **Figure 6.3**). Similar adjustments are observed for other variables. Nevertheless, although the method has the potential to

mitigate data biases and disentangle correlated associations, it is important to clarify that the approach is not causal. Consequently, it does not offer guarantees in that respect.

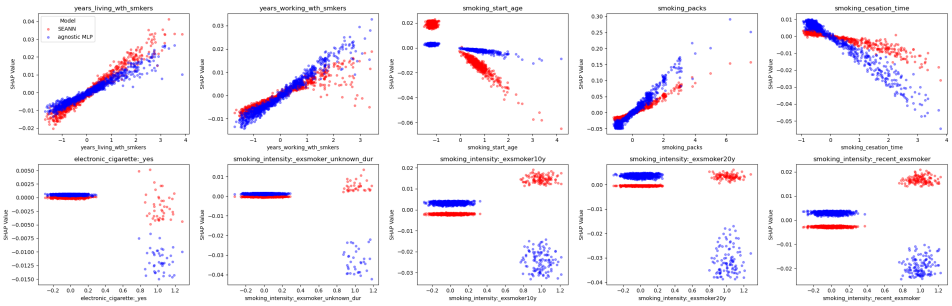


Figure 6.4: Comparison of response functions extracted from a subset of the remaining variables (relative to smoking).
Dots are Shapley values, encoding the model's response (y-axis) for a given exposure value (x-axis).

6.3.2 Predictive performances

Figure 6.5 shows the Pareto frontier used to determine the gamma parameter (cf. **Eq. 6.1**). The objective was to find a solution that both maximizes the AUC and minimizes delta SHAP. Following the elbow method, we chose γ=0.045. Obtained predictive performances (AUC) was 0.702 for the agnostic NN vs 0.695 for SEANN. With our approach, the very principle is to purposely diverge from the optimal solution obtained within the data in order to capture more plausible relationships, closer to those estimated in the literature (from meta-analysis, for instance). Hence predictive performances, while very similar, are, by design, lower than those obtained by the agnostic NN on the held-out dataset. However, performances measured on different populations, would not necessarily be lower as we incorporate reliable estimates that are computed from bigger, more heterogenous data across different studies. For instance, it could hypothetically increase the generalizability of the risk for non-Spanish adults. However, we need access to data encoding similar exposures across different cohorts to accurately test this hypothesis.

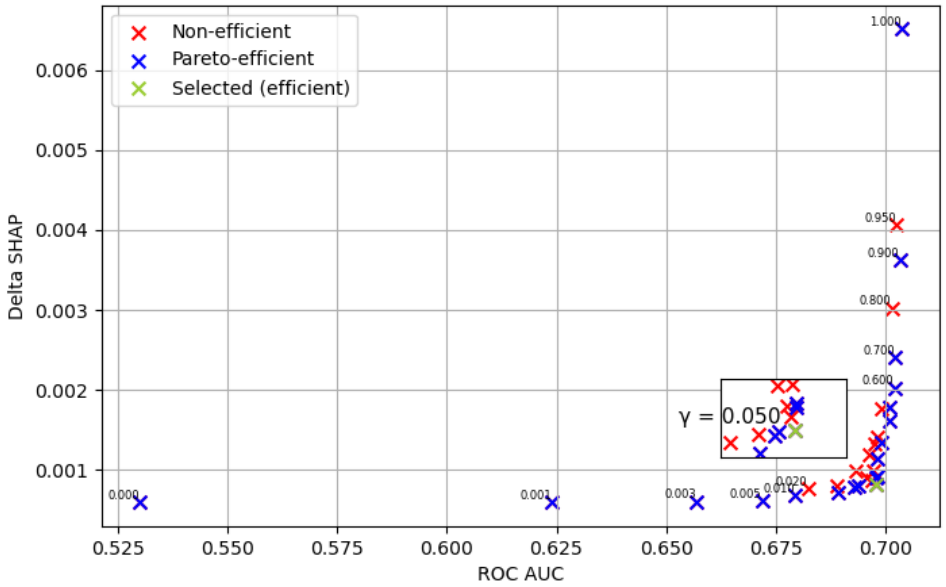


Figure 6.5: Pareto front with several efficient solutions.
The y-axis measures the mean distance between the relationships captured by SEANN and those encoded within the literature. The X-axis is a measure of predictive performance. Displayed numbers are the value of the parameter gamma determining the trade-off between learning from the literature or from the data. A value of 0 is the extreme case where the model is only learning relationships in the literature, while a value of 1 is equivalent to a standard neural network that only learns from the data. The selected point using the elbow method has been highlighted in green and is further displayed within a zoomed-in area frame.

[2]. The risk score is not necessarily less accurate than one obtained with the data only, when trying to compute a score for the general population as the idea is to incorporate reliable estimates that are computed from bigger, more heterogenous data across different studies. Hence, to accurately test enhancements in generalizability, we would need access to data encoding similar exposures within different cohorts.

6.3.3 Important predictors

Figure 6.6 displays the importance of the top 20 predictors in terms of their average impact on the predicted value. Demographic and socio-economic

[2]Here we are referring to a dataset that comes from the same distribution as the training dataset (i.e., with the same biases).

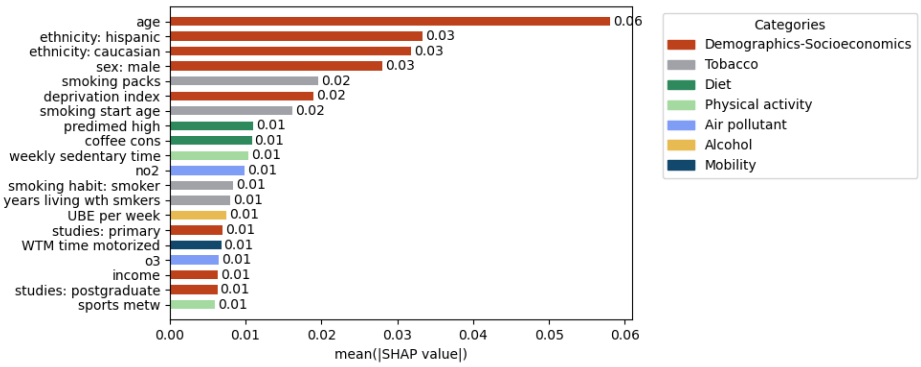


Figure 6.6: Relative importance of features within the informed risk score (top 20).
Measures are obtained by averaging amplitudes of Shapley values over each feature. Abbreviations: UBE - *Unidad de Bebida Estándar* (Standard Beverage Unit); NO2 - Nitrogen Dioxide; O3 - Ozone; WTM - Work Transport Mode.

variables such as age, sex and ethnicity were the most impactful variables. More details are available about the nature of those relationships **Supplementary Figure 3-8** where corresponding response functions are displayed, showing in this case that the risk is substantially increased with age and for males while decreased for individuals of Hispanic or Caucasian ethnicity. Then, smoking-related variables such as the number of smoking packs or smoking start age were the most important, followed by diet (including coffee consumption), physical activity, then air pollution and alcohol consumption.

6.4 Discussion

The impact of certain environmental exposures addressed in this research has been extensively explored for decades across numerous studies and is largely well-understood. However, their effects are mostly studied individually, and may ignore potential interactions among exposures. This study utilizes the exposome holistic approach, considering all exposures as acting together, to draw a simplified map of the nature of each individual effects adjusted from one another, and their averaged, global impact. Furthermore, our approach leverages already well-established knowledge in order to better disentangle

the intricacies of less studied relationships. The computed risk score relies mostly on demographic-socioeconomic factors, including ethnicity, which could be subject to fairness problematics. While it is important to note that it is straightforward, using our approach, to ensure fairness by enforcing a fair relationship between predictors and outcome of interest, for instance using an odd ratio of 1 with a strong associated weight parameter ω, this is beyond the scope of this work that focuses on incorporating literature knowledge.

Results highlight that SEANN is able to successfully learn from the data while capturing relationships substantially closer to the expert knowledge. Preferences between learning relationships from the data or from the literature can be a choice left to the users depending on their particular setting. In this study we employed an elbow method on a Pareto frontier. Another possibility worth considering to determine this trade-off would be to compare the available input data sample size and number of features with each individual literature study incorporated. This second approach could be preferred as it is more systematic but does not allow the users to choose whereas they want to favour scientific plausibility over predictive performances on the input data or vice-versa.

Additionally, exposures from which no corresponding exposure-health relationships have been found in the literature were adjusted from the newly learned relationships within SEANN. In that regard, a particular attention must be given by the user about the choice of the remaining variable. Taking an extreme example for demonstration: let us consider a case of enforcing a relationship between actively smoking tobacco and an increased health risk, as it was done in this work. Let us also consider that the user is including another one hot encoded variable designing people that don't smoke tobacco (hence completely redundant with the first one), then the model will directly adjust what it has been forced to learn on the first variable into the second and extract a relationship were non-smoking is also increasing the related risk. The point is that direct redundancies between the constrained variables and the unconstrained ones should be avoided when possible. Other types of correlation however (variables that are correlated with one another, but different in nature), can be kept and in this case, the adjustment made with SEANN will help the model disentangle the true effect within each correlate. Taking another example to illustrate this last point: let us imagine a

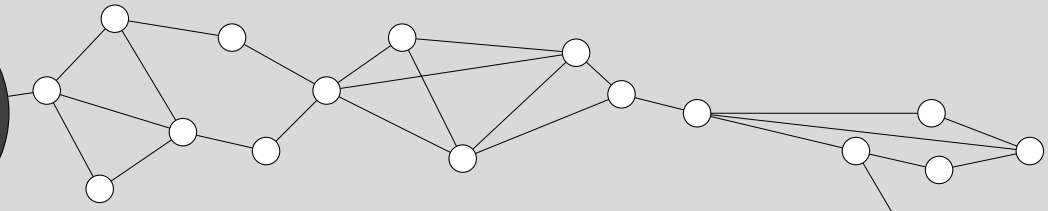
confounding bias where variable A, let's say exposure to mercury, is strongly correlated with variable B, fish consumption, which leads to difficulties for models to disentangle the true effects for each variable. Enforcing one effect using a well-defined literature estimate with SEANN will allow to correctly capture the other.

The score presented in this work does provide predictive capabilities that could be used to identify individuals at risk, understand the risk factors at play and assess the nature of their impact. This provides potential for use in diverse contexts, such as a decision support tool for aiding in the development of preventive policies. However, with an AUC of 0.7, the discriminative accuracy is insufficient to be used as a diagnose tool. This is partially due to the deliberate omission of clinical factors in our score, such as BMI or blood pressure, that are closely related to the outcome, and commonly used in such settings as they greatly enhance the precision. As the inclusion of such factors would have impacted the quality of extracted relationships (potentially hiding effects of exposures), we chose to focus on environmental and socio-economic variables. A promising direction for further refinement of our score involves integrating genomic traits to allow for more precise relationship adjustments.

6.5 Conclusion

In this work we applied SEANN, a new informed neural machine learning method in a real case scenario on cohort data, providing an environmental risk score for hypertension. By integrating literature-extracted exposome-health relationships in the form of pooled effect estimates, SEANN successfully aligned learned exposome-health relationships toward known literature consensus. Furthermore, the remaining variables for which no well-established literature estimate was infused were successfully adjusted. While the approach in itself does not guarantee the extraction of causal relationships, its potential within the exposome framework to provide holistic maps of exposures' effects and better disentangle their effects is certain.

7



Conclusion

7.1 Conclusion

Chronic diseases, the most important cause of death worldwide, are caused by a combination of environmental and hereditary factors. Despite important advancements in the understanding of the effects of environmental exposures on health, there remains a need for further research in this area. Epidemiological studies studying environmental factors, in particular, are still widely relying on simple parametric methods to assess health relationships despite the recent development of more expressive statistical methods. The expressive power of these methods remains limited, particularly when the interactions between environmental factors are not simple correlations but present non-linear effects (threshold effects, multiplicative effects, etc.).

This work aimed to address the complexities of environmental risk assessment by employing advanced nonparametric machine learning methods to develop environmental risk score (ERS) based on a broad spectrum of factors. The adoption of such methods introduces new challenges, particularly concerning 1) the interpretability and trustworthiness of the derived models—a critical aspect in healthcare applications and 2) the extensive data requirement for training these models effectively and the fact that existing observational data from studies on broad exposure panels often prove inadequate. To overcome those challenges, we utilized a model-agnostic explainable AI tool to extract and study exposome health relationships and incorporated literature-based domain knowledge to enhance the robustness and the trustworthiness of our ERS indicators.

More precisely, in the first publication, we highlighted the benefits of complex nonparametric models combined with appropriate tools to study environmental-health relationships using a rich early-life exposome dataset. In the second paper, we developed an informed machine-learning approach to integrate known relationships into the training of deep neural networks and demonstrated the benefits of this approach using synthetic data. In the last paper, we further demonstrated these benefits in a real scenario by computing informed ERS in the GCAT adult population.

7.2 Contributions

In **Chapter 4**, referring to Paper 1, we computed environmental risk scores for the early life exposome on three general health outcomes, namely mental, cardiometabolic, and respiratory health. While the predictive performances obtained were comparable with simpler traditional methods (which may be due to the available sample size), we discovered new associations compared with previous HELIX studies. Moreover, we extracted nonlinear associations, pairwise interactions, local and global feature importance, and family-wise feature importance.
In **Chapter 5**, referring to Paper 2, we developed SEANN, a new approach for the integration of domain knowledge into the learning of deep neural networks. More specifically, we integrated knowledge about known exposome-health relationships in the form of pooled effect size, which are considered to be one of the strongest levels of evidence in epidemiology. In this chapter, we used synthetic data to test our approach in a controlled environment, showing its ability to better disentangle the true effects within correlated variables and to estimate unaccounted confounding effects.
In **Chapter 6**, referring to Paper 3, we further refined our approach, offering better numerical stability in the case of odd ratio. We also proposed new ways to set the hyperparameter values regulating the tradeoff between complying with the literature knowledge or learning from the data. We demonstrated the benefits of our approach on real data by providing informed ERSs for hypertension in a Spanish adult population. We compared informed ERS with agnostic ERS (i.e., obtained with input data only) and showed that we captured relationships that are more in line with the domain knowledge without sacrificing performances. By better adjusting for the known relationships, we better captured those that are less known. Additionally, similar to the first paper, we derived nonlinear relationships and feature importance.

7.3 Benefits and potential applications

In this section, we briefly summarize the main benefits of the approach used in this thesis and point out potential applications. In line with personalized medicine, our risk scores can capture different effect sizes for each

individual which can be used to tailor health interventions based on an individual's unique environmental exposure profile. Traditional ERSs provide insights at the population level. Furthermore, by incorporating literature knowledge, we enhance the reliability and trustworthiness of risk indicators, compared with agnostic approaches, by ensuring compliance with known relationships. In cases where access to data is limited (i.e., in terms of quality, quantity, or population diversity), our informed machine-learning procedure can supplement those deficiencies and improve performances by leveraging literature-reported associations.

The approach developed in this thesis has the potential to impact various sectors related to health. Industries can implement such risk scores to assess and regulate risk factors and their impact on health. For instance, manufacturing or agriculture companies can use ERSs to monitor environmental factors in workers, thereby enhancing workplace safety, reducing healthcare costs, and ensuring regulatory compliance. Researchers can use these ERSs to uncover and refine unknown associations and provide new etiological insights into chronic diseases as we did during this thesis. Public health authorities can leverage these scores to formulate informed and targeted prevention policies to promote health. For instance, local authorities could use such scores to promote regulations on urban planning and justify measures in municipal elections.

7.4 Future works

During the thesis, we provided new evidence regarding the impact of some environmental factors on health. More importantly, we highlighted the potential of using explainable AI tools applied to complex models to study the exposome, but this approach faces multiple challenges, including the need for a large amount of data to be trained efficiently, the discovery of spurious relationships, or the provision of acceptable ethical guarantees. While we help address those by incorporating domain knowledge, this is not sufficient to solve them completely. Informed risk scores are still partially dependent on the input data quantity and quality and our approach provides no guarantee for causal relationships. In this section, we discuss promising research directions for addressing those issues.

7.4.1 Improving data quality and resolution

Several improvements can be made to the quality of the exposure measurements in future observational studies. In the HELIX and GCAT studies, air pollution was measured in a few fixed outdoor areas and modeled using statistical tools to extrapolate measurements at home or school. These measurements may have limited accuracy as they were averaged over various periods (e.g., year, month) and showed little variance within the same city. The HELIX project also measured indoor air exposures using in-situ sensors at home, but this was limited to only 157 individuals out of the 1,600 in the subcohort. Industrial companies like Meersens and its competitors now offer scalable solutions to monitor both outdoor and indoor exposures using deployable sensors and advanced GIS modeling methods. Increasing the number of sensors and placing them in strategic areas can greatly enhance data quality. An even better, though more costly, solution is to use small sensors, smartphones, and other IoT devices to create a more personalized exposure profile. These devices can continually track exposure levels (e.g., physical activity, sleep, air quality) at each time point, providing a comprehensive view of an individual's exposure profile.

7.4.2 Assessing the time dimension

In this work, we did not assess the dynamics of exposures over time, and we acknowledge that understanding their cumulative impact on health is critical for prevention. Echoing the point just mentioned, rich time series data on exposures could be assessed using methods similar to those employed in this thesis. This can be achieved by using adapted tools such as recurrent neural networks or transformers to modelize health trajectories, explainable tools such as SHAP or attention weights to extract exposure-health relationships, and an informed ML method for incorporating appropriate domain knowledge, such as hazard ratios into those models.

7.4.3 Facilitating access to data

Federated data analysis (FDA) is a very promising research direction that can provide the amount of data required by recent machine learning methods such as DNNs in environmental studies. FDA aims to facilitate access to larger and more robust datasets by pooling data from several sources.

As well known in healthcare, the pooling of information from individuals in a central database can raise important ethical and legal questions. Softwares such as DataShield [Wilson et al., 2017] allows the analysis of sensitive individual-level data from one or several studies simultaneously without physically pooling them or disclosing sensitive information. However, those tools principally incorporate traditional biostatistical methods and do not allow the use of more complex approaches yet.

7.4.4 Assessing the causal pathways

More work is needed to incorporate causal inference techniques such as g-computation [Snowden et al., 2011], Mendelian Randomization [Emdin et al., 2017], or Mediation Analyses [MacKinnon et al., 2007] to clarify the potential mechanisms that underpin the statistical associations. Our method, while having the potential to enhance the plausibility of extracted relationships, is an informed machine learning approach not designed for causal inference *per se*. Thus, it does not offer direct guarantees about the causality of extracted relationships. It could however integrate most relationships estimated with causal methods (i.e., treatment effects).

7.4.5 Ensuring equity and fairness

The ethical use of AI-derived health risk scores should consider the diverse socio-economic and demographic contexts in which they are applied to avoid biases that could disproportionately affect vulnerable populations. To that end, explainable AI tools are required to understand the prediction given by complex models and verify their fairness and equity. Those issues were out of the scope of this thesis, however, our informed machine-learning approach can be directly used to enforce the learning of fair exposure-outcome relationships[1].

7.5 Difficulties

We briefly discuss the main difficulties encountered during this doctoral work in this section. Obtaining access to large cohort data was a major issue in

[1] Instead of PESs, we can consider using SEANN to incorporate fair relationships in the form of ORs, SRCs or others (given that they are locally derivable)

this thesis. The original plan was to leverage such data combined with additional data provided by Meersens (from its customer base composed of adults) in a transfer learning procedure designed to improve their industrial solution. However, as we couldn't access sufficient data from the company and had only access to birth cohorts (from the HELIX project) initially, we had to change direction. In the end, after an amount of research, discussions, and unsuccessful trials, we managed to obtain access to a rich database on an adult population (the GCAT cohort) for research purposes only.
Another difficulty worth mentioning concerned the time and effort necessary to perform this industrial PhD supervised by two universities specialized in different domains. While the company has kept the number of solicitations near a minimum, being employed in an industrial company in parallel added extra work, taking part in regular meetings, sharing expertise, and redacting reports. Additionally, the first year of the doctoral program involved a valuable but necessary process of discovery and learning in the field of environmental epidemiology, as the PhD candidate's background was in computer science with no prior expertise in this area.

7.6 Final words

In recent years, there have been important breakthroughs in the AI subfields of generative AI and natural language processing, still transforming numerous industrial and research applications. However, despite these advancements, the need for interpretability remains crucial, especially in domains like healthcare, as the ability to understand AI-based outputs is a requirement to ensure a trustworthy, reliable and ethical use of these technologies.

[Achenbach et al., 2021] Achenbach, S., Aleksandrova, K., Amiano, P., and et al. (2021). Score2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in europe. *European heart journal*, 42(25):2439–2454.

[Achenbach, 1991] Achenbach, T. (1991). Integrative guide for the 1991 cbcl 4-18, ysr, and trf profiles. *Department of Psychiatry University of Vermont, Burlington, VT*.

[Agier et al., 2019] 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., Petraviciene, I., Robinson, O., Roumeliotaki, T., Sunyer, J., Tamayo-Uria, I., Thomsen, C., Urquiza, J., Valentin, A., Slama, R., Vrijheid, M., and Siroux, V. (2019). Early-life exposome and lung function in children in europe: an analysis of data from the longitudinal, population-based helix cohort. *The Lancet Planetary Health*, 3(2):e81–e92.

[Agier et al., 2016] 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., and Vermeulen, R. (2016). A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. *Environmental Health Perspectives*, 124(12):1848–1856.

[Ahsan and Siddique, 2022] Ahsan, M. M. and Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 128:102289.

[Alanazi, 2022] Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30:100924.

[Alsolami et al., 2019] Alsolami, B., Mehmood, R., and Albeshri, A. (2019). *Hybrid Statistical and Machine Learning Methods for Road Traffic Pre-*

diction: A Review and Tutorial, page 115–133. Springer International Publishing.

[Anthimopoulos et al., 2016] Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., and Mougiakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1207–1216.

[Armstrong, 1999] Armstrong, G. L. (1999). Trends in infectious disease mortality in the united states during the 20th century. *JAMA*, 281(1):61.

[Atehortúa et al., 2023] Atehortúa, A., Gkontra, P., Camacho, M., Diaz, O., Bulgheroni, M., Simonetti, V., Chadeau-Hyam, M., Felix, J. F., Sebert, S., and Lekadir, K. (2023). Cardiometabolic risk estimation using exposome data and machine learning. *International Journal of Medical Informatics*, 179:105209.

[Bair et al., 2006] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

[Bakbergenuly et al., 2019] Bakbergenuly, I., Hoaglin, D. C., and Kulinskaya, E. (2019). Pitfalls of using the risk ratio in meta-analysis. *Res. Synth. Methods*, 10(3):398–419.

[Balagopal et al., 2011] Balagopal, P. B., de Ferranti, S. D., Cook, S., Daniels, S. R., Gidding, S. S., Hayman, L. L., McCrindle, B. W., Mietus-Snyder, M. L., and Steinberger, J. (2011). Nontraditional risk factors and biomarkers for cardiovascular disease: Mechanistic, research, and clinical considerations for youth: A scientific statement from the american heart association. *Circulation*, 123(23):2749–2769.

[Barrera-Gómez et al., 2017] 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., and Basagaña, X. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environmental Health*, 16(1).

[Beelen et al., 2009] Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., De Hoogh, K., and Briggs, D. J. (2009). Mapping of background air pollution at a fine spatial scale across the european union. *Science of the total environment*, 407(6):1852–1867.

[Bellinger et al., 2017] Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., and Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, 17(1).

[Bender, 2009] Bender, R. (2009). Introduction to the use of regression models in epidemiology. In *Methods in Molecular Biology*, Methods in molecular biology (Clifton, N.J.), pages 179–195. Humana Press, Totowa, NJ.

[Bergman, 2020] Bergman, D. L. (2020). Symmetry constrained machine learning. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 501–512. Springer.

[Bergstra et al., 2011] Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[Bhatnagar, 2017] Bhatnagar, A. (2017). Environmental determinants of cardiovascular disease. *Circulation research*, 121(2):162–180.

[Bi et al., 2019] Bi, Q., Goodman, K. E., Kaminsky, J., and Lessler, J. (2019). What is machine learning? a primer for the epidemiologist. *American Journal of Epidemiology*.

[Bien et al., 2013] Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3).

[Billionnet et al., 2012] Billionnet, C., Sherrill, D., and Annesi-Maesano, I. (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Annals of Epidemiology*, 22(2):126–141.

[Bondell and Reich, 2012] Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624.

[Borenstein et al., 2021] Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

[Bravo-Moncayo et al., 2019] Bravo-Moncayo, L., Lucio-Naranjo, J., Chávez, M., Pavón-García, I., and Garzón, C. (2019). A machine learning approach for traffic-noise annoyance assessment. *Applied Acoustics*, 156:262–270.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

[Carrico et al., 2014] Carrico, C., Gennings, C., Wheeler, D. C., and Factor-Litvak, P. (2014). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(1):100–120.

[Caspersen et al., 2016] Caspersen, I. H., Kvalem, H. E., Haugen, M., Brantsæter, A. L., Meltzer, H. M., Alexander, J., Thomsen, C., Frøshaug, M., Bremnes, N. M. B., Broadwell, S. L., et al. (2016). Determinants of plasma pcb, brominated flame retardants, and organochlorine pesticides in pregnant women and 3 year old children in the norwegian mother and child cohort study. *Environmental research*, 146:136–144.

[Caspi et al., 2020] Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., Harrington, H., Hogan, S., Poulton, R., Ramrakha, S., Rasmussen, L. J. H., Reuben, A., Richmond-Rakerd, L., Sugden, K., Wertz, J., Williams, B. S., and Moffitt, T. E. (2020). Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the dunedin birth cohort study. *JAMA Network Open*, 3(4):e203221.

[Cequier et al., 2016] Cequier, E., Sakhi, A. K., Haug, L. S., and Thomsen, C. (2016). Development of an ion-pair liquid chromatography–high

resolution mass spectrometry method for determination of organophosphate pesticide metabolites in large-scale biomonitoring studies. *Journal of Chromatography A*, 1454:32–41.

[Cervin et al., 2021] Cervin, M., Norris, L. A., Ginsburg, G., Gosch, E. A., Compton, S. N., Piacentini, J., Albano, A. M., Sakolsky, D., Birmaher, B., Keeton, C., Storch, E. A., and Kendall, P. C. (2021). The p factor consistently predicts long-term psychiatric and functional outcomes in anxiety-disordered youth. *Journal of the American Academy of Child & Adolescent Psychiatry*, 60(7):902–912.e5.

[Cf, 2015] Cf, O. (2015). Transforming our world: the 2030 agenda for sustainable development. *United Nations: New York, NY, USA*.

[Chang et al., 2018] Chang, M., He, L., and Cai, L. (2018). *An Overview of Genome-Wide Association Studies*, page 97–108. Springer New York.

[Chatzi et al., 2017] 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., and Kogevinas, M. (2017). Cohort Profile: The Mother-Child Cohort in Crete, Greece (Rhea Study). *International Journal of Epidemiology*, 46(5):1392–1393k.

[Chawla et al., 2009] Chawla, M., Sharma, S., Sivaswamy, J., and Kishore, L. (2009). A method for automatic detection and classification of stroke from brain ct images. In *2009 Annual international conference of the IEEE engineering in medicine and biology society*, pages 3581–3584. IEEE.

[Chen et al., 2019] Chen, H., Zhang, K., Lyu, P., Li, H., Zhang, L., Wu, J., and Lee, C.-H. (2019). A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific reports*, 9(1):3840.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[Cheng et al., 2020] Cheng, D., Ting, C.-Y., Ho, C. C., and Ho, C.-K. (2020). Performance evaluation of explainable machine learning on non-communicable diseases. *Solid State Technology*, pages 2780–2793.

[Chipman, 1996] Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.

[Chipman et al., 2010] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

[Choi et al., 2010] Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364.

[Cohen, 1988] Cohen, S. (1988). Perceived stress in a probability sample of the united states. *The social psychology of health*, pages 31–67.

[Consortium, 2012] Consortium, I. (2012). Validity of a short questionnaire to assess physical activity in 10 european countries. *European journal of epidemiology*, 27(1):15–25.

[Constantinou et al., 2019] Constantinou, M. P., Goodyer, I. M., Eisler, I., Butler, S., Kraam, A., Scott, S., Pilling, S., Simes, E., Ellison, R., Allison, E., and Fonagy, P. (2019). Changes in general and specific psychopathology factors over a psychosocial intervention. *Journal of the American Academy of Child & Adolescent Psychiatry*, 58(8):776–786.

[Cornier et al., 2008] Cornier, M.-A., Dabelea, D., Hernandez, T. L., Lindstrom, R. C., Steig, A. J., Stob, N. R., Van Pelt, R. E., Wang, H., and Eckel, R. H. (2008). The metabolic syndrome. *Endocrine reviews*, 29(7):777–822.

[Cowell et al., 2020] Cowell, O. R., Mistry, N., Deighton, K., Matu, J., Griffiths, A., Minihane, A. M., Mathers, J. C., Shannon, O. M., and Siervo, M. (2020). Effects of a mediterranean diet on blood pressure: a systematic review and meta-analysis of randomized controlled trials and observational studies. *Journal of Hypertension*, 39(4):729–739.

[Crutzen, 2006] Crutzen, P. J. (2006). The "anthropocene". In *Earth system science in the anthropocene*, pages 13–18. Springer.

[Dash et al., 2022] Dash, T., Chitlangia, S., Ahuja, A., and Srinivasan, A. (2022). A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040. Number: 1 Publisher: Nature Publishing Group.

[Daw et al., 2022] Daw, A., Karpatne, A., Watkins, W. D., Read, J. S., and Kumar, V. (2022). Physics-guided neural networks (pgnn): An application in lake temperature modeling. In *Knowledge Guided Machine Learning*, pages 353–372. Chapman and Hall/CRC.

[Demirel et al., 2021] Demirel, O. B., Yaman, B., Dowdle, L., Moeller, S., Vizioli, L., Yacoub, E., Strupp, J., Olman, C. A., Uğurbil, K., and Akçakaya, M. (2021). 20-fold accelerated 7t fmri using referenceless self-supervised deep learning reconstruction. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3765–3769. IEEE.

[Didan, 2015] Didan, K. (2015). Mod13q1 modis/terra vegetation indices 16-day l3 global 250m sin grid v006.

[Dieterle et al., 2006] Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics. *Analytical Chemistry*, 78(13):4281–4290.

[Ding et al., 2023] Ding, Y., Wang, S., and Lu, J. (2023). Unlocking the potential: Amino acids' role in predicting and exploring therapeutic avenues for type 2 diabetes mellitus. *Metabolites*, 13(9):1017.

[Dzhambov and Dimitrova, 2018] Dzhambov, A. M. and Dimitrova, D. D. (2018). Residential road traffic noise as a risk factor for hypertension in adults: Systematic review and meta-analysis of analytic studies published in the period 2011–2017. *Environmental Pollution*, 240:306–318.

[D'Agostino et al., 2008] D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The framingham heart study. *Circulation*, 117(6):743–753.

[Emdin et al., 2017] Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *Jama*, 318(19):1925–1926.

[Esser et al., 2014] Esser, N., Legrand-Poels, S., Piette, J., Scheen, A. J., and Paquot, N. (2014). Inflammation as a link between obesity, metabolic syndrome and type 2 diabetes. *Diabetes Research and Clinical Practice*, 105(2):141–150.

[Fajemisin et al., 2024] Fajemisin, A. O., Maragno, D., and den Hertog, D. (2024). Optimization with constraint learning: A framework and survey. *European Journal of Operational Research*, 314(1):1–14.

[Farewell et al., 2021] Farewell, C. V., Melnick, E., and Leiferman, J. (2021). Maternal mental health and early childhood development: Exploring critical periods and unique sources of support. *Infant Mental Health Journal*, 42(4):603–615.

[Ferrari and Dunson, 2020] Ferrari, F. and Dunson, D. B. (2020). Bayesian factor analysis for inference on interactions. *Journal of the American Statistical Association*, 116(535):1521–1532.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5).

[Frndak et al., 2023] Frndak, S., Yu, G., Oulhote, Y., Queirolo, E. I., Barg, G., Vahter, M., Mañay, N., Peregalli, F., Olson, J. R., Ahmed, Z., and Kordas, K. (2023). Reducing the complexity of high-dimensional environmental data: An analytical framework using lasso with considerations of confounding for statistical inference. *International Journal of Hygiene and Environmental Health*, 249:114116.

[Fuller et al., 2022] Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., et al. (2022). Pollution and health: a progress update. *The Lancet Planetary Health*, 6(6):e535–e547.

[Gascon et al., 2017] Gascon, M., Guxens, M., Vrijheid, M., Torrent, M., Ibarluzea, J., Fano, E., Llop, S., Ballester, F., Fernández, M. F., Tardón, A., Fernández-Somoano, A., and Sunyer, J. (2017). The inma—infancia y medio ambiente—(environment and childhood) project: More than 10 years contributing to environmental and neuropsychological research. *International Journal of Hygiene and Environmental Health*, 220(4):647–658.

[Gaye et al., 2014] Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd, A. W., Newby, C. J., Nuotio, M.-L., Wilson, R., Butters, O., Murtagh, B., Demir, I., Doiron, D., Giepmans, L., Wallace, S. E., Budin-Ljosne, I., Oliver Schmidt, C., Boffetta, P., Boniol, M., Bota, M., Carter, K. W., deKlerk, N., Dibben, C., Francis, R. W., Hiekkalinna, T., Hveem, K., Kvaløy, K., Millar, S., Perry, I. J., Peters, A., Phillips, C. M., Popham, F., Raab, G., Reischl, E., Sheehan, N., Waldenberger, M., Perola, M., van den Heuvel, E., Macleod, J., Knoppers, B. M., Stolk, R. P., Fortier, I., Harris, J. R., Woffenbuttel, B. H., Murtagh, M. J., Ferretti, V., and Burton, P. R. (2014). Datashield: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*, 43(6):1929–1944.

[Geirhos et al., 2020] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

[Gibson et al., 2019] Gibson, E. A., Goldsmith, J., and Kioumourtzoglou, M.-A. (2019). Complex mixtures, complex analyses: an emphasis on interpretable results. *Current Environmental Health Reports*, 6(2):53–61.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Grazuleviciene et al., 2015] Grazuleviciene, R., Danileviciute, A., Dedele, A., Vencloviene, J., Andrusaityte, S., Uždanaviciute, I., and Nieuwenhuijsen, M. J. (2015). Surrounding greenness, proximity to city parks and pregnancy outcomes in kaunas cohort study. *International Journal of Hygiene and Environmental Health*, 218(3):358–365.

[Grinsztajn et al., 2022] Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data?

[Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

[Guimbaud, 2024] Guimbaud, J. (2024). Ml-based ecrs for european children - python code.

[Guimbaud et al., 2024] Guimbaud, J., Cazabet, R., Maître, L., and et al. (2024). Machine learning-based health environmental-clinical risk scores in european children. *Commun Med*, 4:98.

[Guo et al., 2019] Guo, C., Zhou, Q., Zhang, D., Qin, P., Li, Q., Tian, G., Liu, D., Chen, X., Liu, L., Liu, F., Cheng, C., Qie, R., Han, M., Huang, S., Wu, X., Zhao, Y., Ren, Y., Zhang, M., Liu, Y., and Hu, D. (2019). Association of total sedentary behaviour and television viewing with risk of overweight/obesity, type 2 diabetes and hypertension: A dose–response meta-analysis. *Diabetes, Obesity and Metabolism*, 22(1):79–90.

[Guo et al., 2011] Guo, X., Zou, L., Zhang, X., Li, J., Zheng, L., Sun, Z., Hu, J., Wong, N. D., and Sun, Y. (2011). Prehypertension: a meta-analysis of the epidemiology, risk factors, and predictors of progression. *Texas heart institute journal*, 38(6):643.

[Haddad et al., 2019] Haddad, N., Andrianou, X. D., and Makris, K. C. (2019). A scoping review on the characteristics of human exposome studies. *Current Pollution Reports*, 5:378–393.

[Haltigan et al., 2018] Haltigan, J. D., Aitken, M., Skilling, T., Henderson, J., Hawke, L., Battaglia, M., Strauss, J., Szatmari, P., and Andrade, B. F. (2018). "p" and "dp:" examining symptom-level bifactor models of psychopathology and dysregulation in clinically referred children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 57(6):384–396.

[Hansen et al., 2016] Hansen, V., Oren, E., Dennis, L. K., and Brown, H. E. (2016). Infectious disease mortality trends in the united states, 1980-2014. *JAMA*, 316(20):2149.

[Hao et al., 2018] Hao, N., Feng, Y., and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625.

[Harrington, 2009] Harrington, D. (2009). *Confirmatory factor analysis*. Oxford university press.

[Hart, 1989] Hart, S. (1989). *Shapley Value*, page 210–216. Palgrave Macmillan UK.

[Hastie et al., 2008a] Hastie, T., Tibshirani, R., and Friedman, J. (2008a). *Model Assessment and Selection*, page 219–259. Springer New York.

[Hastie et al., 2008b] Hastie, T., Tibshirani, R., and Friedman, J. (2008b). *Overview of Supervised Learning*, page 9–41. Springer New York.

[Haug et al., 2009] Haug, L. S., Thomsen, C., and Becher, G. (2009). A sensitive method for determination of a broad range of perfluorinated compounds in serum suitable for large-scale human biomonitoring. *Journal of chromatography A*, 1216(3):385–393.

[He et al., 2021] He, Y., Lakhani, C. M., Rasooly, D., Manrai, A. K., Tzoulaki, I., and Patel, C. J. (2021). Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care*, 44(4):935–943.

[He et al., 2023] He, Y., Qian, D. C., Diao, J. A., Cho, M. H., Silverman, E. K., Gusev, A., Manrai, A. K., Martin, A. R., and Patel, C. J. (2023). Prediction and stratification of longitudinal risk for chronic obstructive pulmonary disease across smoking behaviors. *Nature communications*, 14(1):8297.

[Heude et al., 2015] Heude, B., Forhan, A., Slama, R., Douhaud, L., Bedel, S., Saurel-Cubizolles, M.-J., Hankard, R., Thiebaugeorges, O., De Agostini, M., Annesi-Maesano, I., Kaminski, M., Charles, M.-A., 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., and Thiebaugeorges, O, o. b. o. t. E. m.-c. c. s. g. (2015). Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *International Journal of Epidemiology*, 45(2):353–363.

[Hunter, 2005] Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature reviews genetics*, 6(4):287–298.

[Jaffee and Price, 2008] Jaffee, S. R. and Price, T. S. (2008). Genotype–environment correlations: implications for determining the relation-

ship between environmental exposures and psychiatric illness. *Psychiatry*, 7(12):496–499.

[Jager et al., 2008] Jager, K., Zoccali, C., MacLeod, A., and Dekker, F. (2008). Confounding: What it is and how to deal with it. *Kidney International*, 73(3):256–260.

[Jeong et al., 2012] Jeong, C. H., Wagner, E. D., Siebert, V. R., Anduri, S., Richardson, S. D., Daiber, E. J., McKague, A. B., Kogevinas, M., Villanueva, C. M., Goslan, E. H., et al. (2012). Occurrence and toxicity of disinfection byproducts in european drinking waters in relation with the hiwate epidemiology study. *Environmental science & technology*, 46(21):12120–12128.

[Johns et al., 2012] Johns, D. O., Stanek, L. W., Walker, K., Benromdhane, S., Hubbell, B., Ross, M., Devlin, R. B., Costa, D. L., and Greenbaum, D. S. (2012). Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution. *Environmental Health Perspectives*, 120(9):1238–1242.

[Kennedy et al., 2013] Kennedy, E. H., Wiitala, W. L., Hayward, R. A., and Sussman, J. B. (2013). Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical Care*, 51(3):251–258.

[Khera et al., 2018] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224.

[Kingma and Ba, 2017] Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs].

[Köchli et al., 2019] Köchli, S., Endes, K., Bartenstein, T., Usemann, J., Schmidt-Trucksäss, A., Frey, U., Zahner, L., and Hanssen, H. (2019). Lung function, obesity and physical fitness in young children: The examin youth study. *Respiratory Medicine*, 159:105813.

[Kolachalama and Garg, 2018] Kolachalama, V. B. and Garg, P. S. (2018). Machine learning and medical education. *npj Digital Medicine*, 1(1).

[Kondo et al., 2019] Kondo, K., Ishikawa, A., and Kimura, M. (2019). Sequence to sequence with attention for influenza prevalence prediction using google trends. In *Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics*, ICCBB '19. ACM.

[Krall and Strickland, 2017] Krall, J. R. and Strickland, M. J. (2017). Recent approaches to estimate associations between source-specific air pollution and health. *Current Environmental Health Reports*, 4(1):68–78.

[Kumar et al., 2021] Kumar, S., Yu, S., Michelson, A., and Payne, P. (2021). Self-explaining neural network with plausible explanations. *arXiv preprint arXiv:2110.04598*.

[Lau et al., 2018] Lau, C.-H. E., Siskos, A. P., Maitre, L., Robinson, O., Athersuch, T. J., Want, E. J., Urquiza, J., Casas, M., Vafeiadi, M., Roumeliotaki, T., McEachan, R. R. C., Azad, R., Haug, L. S., Meltzer, H. M., Andrusaityte, S., Petraviciene, I., Grazuleviciene, R., Thomsen, C., Wright, J., Slama, R., Chatzi, L., Vrijheid, M., Keun, H. C., and Coen, M. (2018). Determinants of the urinary and serum metabolome in children from six european populations. *BMC Medicine*, 16(1).

[Le Magueresse-Battistoni et al., 2018] Le Magueresse-Battistoni, B., Vidal, H., and Naville, D. (2018). Environmental pollutants and metabolic disorders: The multi-exposure scenario of life. *Frontiers in Endocrinology*, 9.

[Lee et al., 2021] Lee, K., Ray, J., and Safta, C. (2021). The predictive skill of convolutional neural networks models for disease forecasting. *PLOS ONE*, 16(7):e0254319.

[Leiser et al., 2023] Leiser, F., Rank, S., Schmidt-Kraepelin, M., Thiebes, S., and Sunyaev, A. (2023). Medical informed machine learning: A scoping review and future research directions. *Artificial Intelligence in Medicine*, 145:102676.

[Lim et al., 2018] Lim, J., Kweon, K., Kim, H., Cho, S., Park, J., and Sim, C. (2018). Negative impact of noise and noise sensitivity on mental health in childhood. *Noise health*, 20:199–211.

[Lim and Hastie, 2015] Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

[Liu et al., 2020] Liu, F., Liu, Y., Sun, X., Yin, Z., Li, H., Deng, K., Zhao, Y., Wang, B., Ren, Y., Liu, X., Zhang, D., Chen, X., Cheng, C., Liu, L., Liu, D., Chen, G., Hong, S., Wang, C., Zhang, M., and Hu, D. (2020). Race- and sex-specific association between alcohol consumption and hypertension in 22 cohort studies: A systematic review and meta-analysis. *Nutrition, Metabolism and Cardiovascular Diseases*, 30(8):1249–1259.

[Liu et al., 2017] Liu, X., Zhang, D., Liu, Y., Sun, X., Han, C., Wang, B., Ren, Y., Zhou, J., Zhao, Y., Shi, Y., Hu, D., and Zhang, M. (2017). Dose–response association between physical activity and incident hypertension: A systematic review and meta-analysis of cohort studies. *Hypertension*, 69(5):813–820.

[Liu et al., 2022] Liu, X.-X., Ma, X.-L., Huang, W.-Z., Luo, Y.-N., He, C.-J., Zhong, X.-M., Dadvand, P., Browning, M. H., Li, L., Zou, X.-G., Dong, G.-H., and Yang, B.-Y. (2022). Green space and cardiovascular disease: A systematic review with meta-analysis. *Environmental Pollution*, 301:118990.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

[Lundberg et al., 2020] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67.

[Lundberg et al., 2018] Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Ma et al., 2015] Ma, S., Carroll, R. J., Liang, H., and Xu, S. (2015). Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates. *The Annals of Statistics*, 43(5).

[MacKinnon et al., 2007] MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614.

[Magnus et al., 2016] 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., and Stoltenberg, C. (2016). Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International Journal of Epidemiology*, 45(2):382–388.

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

[Maitre et al., 2018] Maitre, L., de Bont, J., Casas, M., Robinson, O., Aasvang, G. M., Agier, L., Andrusaitytė, 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., Grazuleviciene, R., Bjerve Gutzkow, 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., Petraviciene, 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., and Vrijheid, M. (2018). Human early life exposome (helix) study: a european population-based exposome cohort. *BMJ Open*, 8(9):e021311.

[Maitre et al., 2022b] Maitre, L., Guimbaud, J.-B., Warembourg, C., Güil-Oumrait, N., Petrone, P. M., Chadeau-Hyam, M., Vrijheid, M., Basagaña, X., and Gonzalez, J. R. (2022b). State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. *Environment International*, 168:107422.

[Mandreoli et al., 2022] Mandreoli, F., Ferrari, D., Guidetti, V., Motta, F., and Missier, P. (2022). Real-world data mining meets clinical practice: Research challenges and perspective. *Frontiers in Big Data*, 5.

[Mangasarian and Wild, 2008] Mangasarian, O. L. and Wild, E. W. (2008). Nonlinear knowledge-based classification. *IEEE Transactions on Neural Networks*, 19(10):1826–1832.

[Marschner et al., 2021] Marschner, S. N., Lombardo, E., Minibek, L., Holzgreve, A., Kaiser, L., Albert, N. L., Kurz, C., Riboldi, M., Späth, R., Baumeister, P., et al. (2021). Risk stratification using 18f-fdg pet/ct and artificial neural networks in head and neck cancer patients undergoing radiotherapy. *Diagnostics*, 11(9):1581.

[Mas et al., 2020] Mas, S., Boloc, D., Rodríguez, N., Mezquida, G., Amoretti, S., Cuesta, M. J., González-Peñas, J., García-Alcón, A., Lobo, A., González-Pinto, A., Corripio, I., Vieta, E., Castro-Fornieles, J., Mané, A., Saiz-Ruiz, J., Gassó, P., Bioque, M., and Bernardo, M. (2020). Examining gene–environment interactions using aggregate scores in a first-episode psychosis cohort. *Schizophrenia Bulletin*, 46(4):1019–1025.

[Masselot et al., 2022] Masselot, P., Chebana, F., Campagna, C., Lavigne, E., Ouarda, T. B. M. J., and Gosselin, P. (2022). Constrained groupwise additive index models. *Biostatistics*, 24(4):1066–1084.

[Mehta and Majumdar, 2017] Mehta, J. and Majumdar, A. (2017). Rodeo: robust de-aliasing autoencoder for real-time medical image reconstruction. *Pattern Recognition*, 63:499–510.

[Miller and Jones, 2013] Miller, G. W. and Jones, D. P. (2013). The nature of nurture: Refining the definition of the exposome. *Toxicological Sciences*, 137(1):1–2.

[Mosqueira-Rey et al., 2022] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, A. (2022). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

[Muralidhar et al., 2018] Muralidhar, N., Islam, M. R., Marwah, M., Karpatne, A., and Ramakrishnan, N. (2018). Incorporating Prior Domain Knowledge into Deep Neural Networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 36–45, Seattle, WA, USA. IEEE.

[Murray et al., 2021] Murray, G. K., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., and Wray, N. R. (2021). Could polygenic risk scores be useful in psychiatry?: A review. *JAMA Psychiatry*, 78(2):210.

[Narisetty et al., 2018] Narisetty, N. N., Mukherjee, B., Chen, Y., Gonzalez, R., and Meeker, J. D. (2018). Selection of nonlinear interactions by a forward stepwise algorithm: Application to identifying environmental chemical mixtures affecting health outcomes. *Statistics in Medicine*, 38(9):1582–1600.

[Nelder, 1977] Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, 140(1):48.

[Neri et al., 2022] Neri, L., Lonati, C., Titapiccolo, J. I., Nadal, J., Meiselbach, H., Schmid, M., Baerthlein, B., Tschulena, U., Schneider, M. P., Schultheiss, U. T., Barbieri, C., Moore, C., Steppan, S., Eckardt, K.-U., Stuard, S., and Bellocchio, F. (2022). The cardiovascular literature-based risk algorithm (calibra): Predicting cardiovascular events in patients with non-dialysis dependent chronic kidney disease. *Frontiers in Nephrology*, 2.

[Neufcourt et al., 2022] Neufcourt, L., Castagne, R., Mabile, L., Khalatbari-Soltani, S., Delpierre, C., and Kelly-Irving, M. (2022). Assessing how social exposures are integrated in exposome research: A scoping review. *Environmental Health Perspectives*, 130(11).

[Nielsen, 2016] Nielsen, F. (2016). *Hierarchical Clustering*, page 195–211. Springer International Publishing.

[Nieminen, 2022] Nieminen, P. (2022). Application of standardized regression coefficient in meta-analysis. *BioMedInformatics*, 2(3):434–458.

[Nieuwenhuijsen et al., 2014] Nieuwenhuijsen, M. J., Kruize, H., Gidlow, C., Andrusaityte, S., Antó, J. M., Basagaña, X., Cirach, M., Dadvand, P., Danileviciute, A., Donaire-Gonzalez, D., et al. (2014). Positive health effects of the natural outdoor environment in typical populations in different regions in europe (phenotype): a study programme protocol. *BMJ open*, 4(4):e004951.

[Nishimura and Suchard, 2018] Nishimura, A. and Suchard, M. A. (2018). Prior-preconditioned conjugate gradient for accelerated gibbs sampling in "large n & large p" sparse bayesian logistic regression models. *arXiv: Computation*.

[Novgorodtseva et al., 2011] Novgorodtseva, T. P., Karaman, Y. K., Zhukova, N. V., Lobanova, E. G., Antonyuk, M. V., and Kantur, T. A. (2011). Composition of fatty acids in plasma and erythrocytes and eicosanoids level in patients with metabolic syndrome. *Lipids in Health and Disease*, 10(1).

[Obón-Santacana et al., 2018] Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J., Galván-Femenía, I., Alonso, T., Puig, L., Sumoy, L., Duell, E. J., Perucho, M., Moreno, V., and de Cid, R. (2018). Gcat|genomes for life: a prospective cohort study of the genomes of catalonia. *BMJ Open*, 8(3):e018324.

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34.

[Omran, 2005] Omran, A. R. (2005). The epidemiologic transition: A theory of the epidemiology of population change. *The Milbank Quarterly*, 83(4):731–757.

[Onay and Onay, 2020] Onay, A. and Onay, M. (2020). A drug decision support system for developing a successful drug candidate using

machine learning techniques. *Current Computer-Aided Drug Design*, 16(4):407–419.

[Orsini et al., 2011] Orsini, N., Li, R., Wolk, A., Khudyakov, P., and Spiegelman, D. (2011). Meta-analysis for linear and nonlinear dose-response relations: Examples, an evaluation of approximations, and software. *American Journal of Epidemiology*, 175(1):66–73.

[Oskar and Stingone, 2020] Oskar, S. and Stingone, J. A. (2020). Machine learning within studies of early-life environmental exposures and child health: Review of the current literature and discussion of next steps. *Current Environmental Health Reports*, 7(3):170–184.

[O'Sullivan et al., 2022] O'Sullivan, J. W., Raghavan, S., Marquez-Luna, C., Luzum, J. A., Damrauer, S. M., Ashley, E. A., O'Donnell, C. J., Willer, C. J., and Natarajan, P. (2022). Polygenic risk scores for cardiovascular disease: a scientific statement from the american heart association. *Circulation*, 146(8):e93–e118.

[Paatero and Tapper, 1994] Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

[Padmanabhan et al., 2017] Padmanabhan, J. L., Shah, J. L., Tandon, N., and Keshavan, M. S. (2017). The "polyenviromic risk score": Aggregating environmental risk factors predicts conversion to psychosis in familial high-risk subjects. *Schizophrenia Research*, 181:17–22.

[Park et al., 2014] Park, S. K., Tao, Y., Meeker, J. D., Harlow, S. D., and Mukherjee, B. (2014). Environmental risk score as a new tool to examine multi-pollutants in epidemiologic research: An example from the nhanes study using serum lipid levels. *PLoS ONE*, 9.

[Patel et al., 2010] Patel, C. J., Bhattacharya, J., and Butte, A. J. (2010). An environment-wide association study (ewas) on type 2 diabetes mellitus. *PLoS ONE*, 5(5):e10746.

[Pathak et al., 2020] Pathak, M., Dwivedi, S. N., Thakur, B., and Vishnubhatla, S. (2020). Methods of estimating the pooled effect size under

meta-analysis: A comparative appraisal. *Clinical Epidemiology and Global Health*, 8(1):105–112.

[Petersen et al., 2017] Petersen, B.-S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D., and Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, 18(1).

[Poothong et al., 2017] Poothong, S., Thomsen, C., Padilla-Sanchez, J. A., Papadopoulou, E., and Haug, L. S. (2017). Distribution of novel and well-known poly-and perfluoroalkyl substances (pfass) in human serum, plasma, and whole blood. *Environmental science & technology*, 51(22):13388–13396.

[Pries et al., 2021] Pries, L.-K., Erzin, G., Rutten, B. P. F., van Os, J., and Guloksuz, S. (2021). Estimating aggregate environmental risk score in psychiatry: The exposome score for schizophrenia. *Frontiers in Psychiatry*, 12.

[Quadrianto et al., 2011] Quadrianto, N., Kersting, K., and Xu, Z. (2011). *Gaussian Process*, page 428–439. Springer US.

[Quanjer et al., 2012] 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., Zheng, J., and Stocks, J. (2012). Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *European Respiratory Journal*, 40(6):1324–1343.

[Radchenko and James, 2010] Radchenko, P. and James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553.

[Radovanović et al., 2019] Radovanović, S., Delibašić, B., Jovanović, M., Vukićević, M., and Suknović, M. (2019). A framework for integrating domain knowledge in logistic regression with application to hospital readmission prediction. *International Journal on Artificial Intelligence Tools*, 28(06):1960006.

[Rahaman and Hossain, 2013] Rahaman, S. and Hossain, M. S. (2013). A belief rule based clinical decision support system to assess suspicion of

heart failure from signs, symptoms and risk factors. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–6. IEEE.

[Rahman et al., 2023] Rahman, S., Ibtisum, S., Bazgir, E., and Barai, T. (2023). The significance of machine learning in clinical disease diagnosis: A review. *arXiv preprint arXiv:2310.16978*.

[Rappaport, 2012] Rappaport, S. M. (2012). Biomarkers intersect with the exposome. *Biomarkers*, 17(6):483–489.

[Reid and Tibshirani, 2015] Reid, S. and Tibshirani, R. (2015). Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, page kxv049.

[Reuter and Jordan, 2019] Reuter, H. and Jordan, J. (2019). Status of hypertension in europe. *Current Opinion in Cardiology*, 34(4):342–349.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Rijlaarsdam et al., 2021] Rijlaarsdam, J., Barker, E. D., Caserini, C., Koopman-Verhoeff, M. E., Mulder, R. H., Felix, J. F., and Cecil, C. A. (2021). Genome-wide dna methylation patterns associated with general psychopathology in children. *Journal of Psychiatric Research*, 140:214–220.

[Riley, 2001] Riley, J. C. (2001). *Rising life expectancy: a global history*. Cambridge University Press.

[Robinson et al., 2018] Robinson, O., Tamayo, I., de Castro, M., Valentin, A., Giorgis-Allemand, L., Hjertager Krog, N., Marit Aasvang, G., Ambros, A., Ballester, F., Bird, P., Chatzi, L., Cirach, M., Dėdelė, A., Donaire-Gonzalez, D., Gražuleviciene, R., Iakovidis, M., Ibarluzea, J., Kampouri, M., Lepeule, J., Maitre, L., McEachan, R., Oftedal, B., Siroux, V., Slama, R., Stephanou, E. G., Sunyer, J., Urquiza, J., Vegard Weyde, K., Wright, J., Vrijheid, M., Nieuwenhuijsen, M., and Basagaña, X.

(2018). The urban exposome during pregnancy and its socioeconomic determinants. *Environmental Health Perspectives*, 126(7).

[Rodushkin and Axelsson, 2000] Rodushkin, I. and Axelsson, M. D. (2000). Application of double focusing sector field icp-ms for multielemental characterization of human hair and nails. part ii. a study of the inhabitants of northern sweden. *Science of the Total Environment*, 262(1-2):21–36.

[Romano et al., 2024] Romano, D., Novielli, P., Diacono, D., Cilli, R., Pantaleo, E., Amoroso, N., Bellantuono, L., Monaco, A., Bellotti, R., and Tangaro, S. (2024). Insights from explainable artificial intelligence of pollution and socioeconomic influences for respiratory cancer mortality in italy. *Journal of Personalized Medicine*, 14(4):430.

[Rosner, 2012] Rosner, A. L. (2012). Evidence-based medicine: Revisiting the pyramid of priorities. *Journal of Bodywork and Movement Therapies*, 16(1):42–49.

[Rosseel, 2012] Rosseel, Y. (2012). lavaan: Anrpackage for structural equation modeling. *Journal of Statistical Software*, 48(2).

[Rønningen et al., 2006] Rønningen, K. S., Paltiel, L., Meltzer, H. M., Nordhagen, R., Lie, K. K., Hovengen, R., Haugen, M., Nystad, W., Magnus, P., and Hoppin, J. A. (2006). The biobank of the norwegian mother and child cohort study: A resource for the next 100 years. *European Journal of Epidemiology*, 21(8):619–625.

[Sakhi et al., 2018] Sakhi, A. K., Sabaredzovic, A., Papadopoulou, E., Cequier, E., and Thomsen, C. (2018). Levels, variability and determinants of environmental phenols in pairs of norwegian mothers and children. *Environment international*, 114:242–251.

[Sampedro et al., 2014] Sampedro, F., Escalera, S., Domenech, A., and Carrio, I. (2014). A computational framework for cancer response assessment based on oncological pet-ct scans. *Computers in Biology and Medicine*, 55:92–99.

[Samuel, 1959] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.

[Satapathy et al., 2024] Satapathy, P., Khatib, M. N., Gaidhane, S., Zahiruddin, Q. S., Gaidhane, A. M., Rustagi, S., Serhan, H. A., and Padhi, B. K. (2024). Association of neighborhood deprivation and hypertension: A systematic review and meta-analysis. *Current Problems in Cardiology*, 49(4):102438.

[Schlemper et al., 2017] Schlemper, J., Caballero, J., Hajnal, J. V., Price, A., and Rueckert, D. (2017). A deep cascade of convolutional neural networks for mr image reconstruction. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pages 647–658. Springer.

[Schröder et al., 2011] Schröder, H., Fitó, M., Estruch, R., Martínez-González, M. A., Corella, D., Salas-Salvadó, J., Lamuela-Raventós, R., Ros, E., Salaverría, I., Fiol, M., et al. (2011). A short screener is valid for assessing mediterranean diet adherence among older spanish men and women. *The Journal of nutrition*, 141(6):1140–1145.

[Shailaja et al., 2018] Shailaja, K., Seetharamulu, B., and Jabbar, M. A. (2018). Machine learning in healthcare: A review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 910–914.

[Shapley, 1953] Shapley, L. S. (1953). A Value for n-Person Games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press.

[Sinisi and van der Laan, 2004] Sinisi, S. E. and van der Laan, M. J. (2004). Deletion/substitution/addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–38.

[Siroux et al., 2016] Siroux, V., Agier, L., and Slama, R. (2016). The exposome concept: a challenge and a potential driver for environmental health research. *European Respiratory Review*, 25(140):124–129.

[Snowden et al., 2011] Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, 173(7):731–738.

[Stafoggia et al., 2017] Stafoggia, M., Breitner, S., Hampel, R., and Basagaña, X. (2017). Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science. *Current Environmental Health Reports*, 4(4):481–490.

[Stekhoven and Bühlmann, 2011] Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

[Stratakis et al., 2020] Stratakis, N., Conti, D. V., Borras, E., Sabido, E., Roumeliotaki, T., Papadopoulou, E., Agier, L., Basagana, X., Bustamante, M., Casas, M., Farzan, S. F., Fossati, S., Gonzalez, J. R., Grazuleviciene, R., Heude, B., Maitre, L., McEachan, R. R. C., Theologidis, I., Urquiza, J., Vafeiadi, M., West, J., Wright, J., McConnell, R., Brantsaeter, A.-L., Meltzer, H.-M., Vrijheid, M., and Chatzi, L. (2020). Association of fish consumption and mercury exposure during pregnancy with metabolic health and inflammatory biomarkers in children. *JAMA Network Open*, 3(3):e201007.

[Sun et al., 2022] Sun, S., He, D., Luo, C., Lin, X., Wu, J., Yin, X., Jia, C., Pan, Q., Dong, X., Zheng, F., Li, H., and Zhou, J. (2022). Metabolic syndrome and its components are associated with altered amino acid profile in chinese han population. *Frontiers in Endocrinology*, 12.

[Sun et al., 2019] Sun, Y., Milne, S., Jaw, J. E., Yang, C. X., Xu, F., Li, X., Obeidat, M., and Sin, D. D. (2019). Bmi is associated with fev1 decline in chronic obstructive pulmonary disease: a meta-analysis of clinical trials. *Respiratory Research*, 20(1).

[Sun et al., 2013] Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A., and Mukherjee, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health*, 12(1).

[Sweet, 2011] Sweet, L. H. (2011). *N-Back Paradigm*, page 1718–1719. Springer New York.

[Tamayo-Uria et al., 2019] 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., Meltzer, H. M., Petraviciene, I., Robinson, O., Roumeliotaki, T., Sakhi, A. K., Urquiza, J., Vafeiadi, M., Waiblinger, D., Warembourg, C., Wright, J., Slama, R., Vrijheid, M., and Basagaña, X. (2019). The early-life exposome: Description and patterns in six european countries. *Environment International*, 123:189–200.

[Teramoto et al., 2020] Teramoto, A., Yamada, A., Tsukamoto, T., Imaizumi, K., Toyama, H., Saito, K., and Fujita, H. (2020). *Decision Support System for Lung Cancer Using PET/CT and Microscopic Images*, page 73–94. Springer International Publishing.

[Thomas et al., 2023] Thomas, S. A., Browning, C. J., Charchar, F. J., Klein, B., Ory, M. G., Bowden-Jones, H., and Chamberlain, S. R. (2023). Transforming global approaches to chronic disease prevention and management across the lifespan: integrating genomics, behavior change, and digital health solutions. *Frontiers in Public Health*, 11.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[Tibshirani et al., 2016] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

[Tsai, 2017] Tsai, J.-P. (2017). The association of serum leptin levels with metabolic diseases. *Tzu Chi Medical Journal*, 29(4):192.

[Ulusoy, 2013] Ulusoy, Ş. (2013). Assessment of cardiovascular risk in hypertensive patients: a comparison of commonly used risk scoring programs. *Kidney international supplements*, 3(4):340–342.

[Vassos et al., 2019] Vassos, E., Sham, P., Kempton, M., Trotta, A., Stilo, S. A., Gayer-Anderson, C., Di Forti, M., Lewis, C. M., Murray, R. M., and Morgan, C. (2019). The maudsley environmental risk score for psychosis. *Psychological Medicine*, 50(13):2213–2220.

[Volkova et al., 2017] Volkova, S., Ayton, E., Porterfield, K., and Corley, C. D. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLOS ONE*, 12(12):e0188941.

[Von Rueden et al., 2021] Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., et al. (2021). Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633.

[Vrijheid, 2014] Vrijheid, M. (2014). The exposome: a new paradigm to study the impact of environment on health. *Thorax*, 69(9):876–878.

[Vrijheid et al., 2021] Vrijheid, M., Basagaña, X., Gonzalez, J. R., Jaddoe, V. W. V., Jensen, G., Keun, H. C., McEachan, R. R. C., Porcel, J., Siroux, V., Swertz, M. A., Thomsen, C., Aasvang, G. M., Andrušaitytė, S., Angeli, K., Avraam, D., Ballester, F., Burton, P., Bustamante, M., Casas, M., Chatzi, L., Chevrier, C., Cingotti, N., Conti, D., Crépet, A., Dadvand, P., Duijts, L., van Enckevort, E., Esplugues, A., Fossati, S., Garlantezec, R., Gómez Roig, M. D., Grazuleviciene, R., Gützkow, K. B., Guxens, M., Haakma, S., Hessel, E. V. S., Hoyles, L., Hyde, E., Klanova, J., van Klaveren, J. D., Kortenkamp, A., Le Brusquet, L., Leenen, I., Lertxundi, A., Lertxundi, N., Lionis, C., Llop, S., Lopez-Espinosa, M.-J., Lyon-Caen, S., Maitre, L., Mason, D., Mathy, S., Mazarico, E., Nawrot, T., Nieuwenhuijsen, M., Ortiz, R., Pedersen, M., Perelló, J., Pérez-Cruz, M., Philippat, C., Piler, P., Pizzi, C., Quentin, J., Richiardi, L., Rodriguez, A., Roumeliotaki, T., Sabin Capote, J. M., Santiago, L., Santos, S., Siskos, A. P., Strandberg-Larsen, K., Stratakis, N., Sunyer, J., Tenenhaus, A., Vafeiadi, M., Wilson, R. C., Wright, J., Yang, T., and Slama, R. (2021). Advancing tools for human early lifecourse exposome research and translation (athlete): Project overview. *Environmental Epidemiology*, 5(5):e166.

[Wallace et al., 2022] Wallace, S. S., Barak, G., Truong, G., and Parker, M. W. (2022). Hierarchy of evidence within the medical literature. *Hospital Pediatrics*, 12(8):745–750.

[Wang and Veugelers, 2008] Wang, F. and Veugelers, P. J. (2008). Self-esteem and cognitive development in the era of the childhood obesity epidemic. *Obesity Reviews*, 9(6):615–623.

[Wiemken and Kelley, 2020] Wiemken, T. L. and Kelley, R. R. (2020). Machine learning in epidemiology and health outcomes research. *Annual Review of Public Health*, 41(1):21–36.

[Wild, 2005] Wild, C. P. (2005). Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention*, 14(8):1847–1850.

[Wild, 2012] Wild, C. P. (2012). The exposome: from concept to utility. *International Journal of Epidemiology*, 41(1):24–32.

[Wilkins et al., 2021] Wilkins, J. T., Seckler, H. S., Rink, J., Compton, P. D., Fornelli, L., Thaxton, C. S., LeDuc, R., Jacobs, D., Doubleday, P. F., Sniderman, A., Lloyd-Jones, D. M., and Kelleher, N. L. (2021). Spectrum of apolipoprotein ai and apolipoprotein aii proteoforms and their associations with indices of cardiometabolic health: The cardia study. *Journal of the American Heart Association*, 10(17).

[Wilson et al., 2017] Wilson, R. C., Butters, O. W., Avraam, D., Baker, J., Tedds, J. A., Turner, A., Murtagh, M., and Burton, P. R. (2017). Datashield – new directions and dimensions. *Data Science Journal*, 16.

[Woodward et al., 2007] Woodward, M., Brindle, P., and Tunstall-Pedoe, H. (2007). Adding social deprivation and family history to cardiovascular risk assessment: the assign score from the scottish heart health extended cohort (shhec). *Heart*, 93(2):172–176.

[Wray et al., 2021] Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., Murray, G. K., and Visscher, P. M. (2021). From basic science to clinical application of polygenic risk scores: A primer. *JAMA Psychiatry*, 78(1):101.

[Wright et al., 2012] 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., and West, Jane, o. b. o. t. B. i. B. S. C. G.

(2012). Cohort Profile: The Born in Bradford multi-ethnic family cohort study. *International Journal of Epidemiology*, 42(4):978–991.

[Wu et al., 2018] Wu, J.-L., Xiao, H., and Paterson, E. (2018). Physics-informed machine learning approach for augmenting turbulence models: A comprehensive framework. *Physical Review Fluids*, 3(7):074602.

[Wu et al., 2020] Wu, N., Green, B., Ben, X., and O'Banion, S. (2020). Deep transformer models for time series forecasting: The influenza prevalence case. *ArXiv*, abs/2001.08317.

[Xue et al., 2021] Xue, C., Karjadi, C., Paschalidis, I. C., Au, R., and Kolachalama, V. B. (2021). Detection of dementia on voice recordings using deep learning: a framingham heart study. *Alzheimer's Research & Therapy*, 13(1).

[Yan et al., 2016] Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S., Metaxas, D. N., and Zhou, X. S. (2016). Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition. *IEEE Transactions on Medical Imaging*, 35(5):1332–1343.

[Yang et al., 2018] Yang, B.-Y., Qian, Z., Howard, S. W., Vaughn, M. G., Fan, S.-J., Liu, K.-K., and Dong, G.-H. (2018). Global association between ambient air pollution and blood pressure: A systematic review and meta-analysis. *Environmental Pollution*, 235:576–588.

[Yang et al., 2010] Yang, P., Hwa Yang, Y., B. Zhou, B., and Y. Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308.

[Yuan et al., 2009] Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4).

[Yuan and Lin, 2005] Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

[Zewdie et al., 2019] Zewdie, G. K., Lary, D. J., Levetin, E., and Garuma, G. F. (2019). Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *International journal of environmental research and public health*, 16(11):1992.

[Zhang et al., 2024] Zhang, Y., Jiang, B., Gao, Z., Wang, M., Feng, J., Xia, L., and Liu, J. (2024). Health risk assessment of soil heavy metals in a typical mining town in north china based on monte carlo simulation coupled with positive matrix factorization model. *Environmental Research*, 251:118696.

[Zheng et al., 2020] Zheng, Y., Chen, Z., Pearson, T., Zhao, J., Hu, H., and Prosperi, M. (2020). Design and methodology challenges of environment-wide association studies: A systematic review. *Environmental Research*, 183:109275.

[Zhou et al., 2021] Zhou, B., Perel, P., Mensah, G. A., and Ezzati, M. (2021). Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension. *Nature Reviews Cardiology*, 18(11):785–802.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

[Zou and Zhang, 2009] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4).

A.1 Paper 1

Machine Learning based Mental, Cardiovascular and Respiratory Environmental-Clinical risk scores in European Children

Jean-Baptiste Guimbaud, Alexandros P. Siskos, Amrit Kaur Sakhi, Barbara Heude, Eduard Sabidó, Eva Borràs, Hector Keun, John Wright, Jordi Julvez, Jose Urquiza, Kristine Bjerve Gützkow, Leda Chatzi, Maribel Casas, Mariona Bustamante, Mark Nieuwenhuijsen, Martine Vrijheid, Mónica López-Vicente, Montserrat de Castro Pascual, Nikos Stratakis, Oliver Robinson, Regina Grazuleviciene, Remy Slama, Silvia Alemany, Xavier Basagaña, Marc Plantevit, Rémy Cazabet, Léa Maitre

Table of Contents

List of Tables

List of Figures

Supplementary Notes

This note is extracted from [1], we provide the methods used to estimate all exposures included in the exposome for the HELIX subcohort. For the purpose of this document "pregnancy" refers to the period from conception to the day of birth, while "childhood" refers to the period between 6 and 11 years (the exact range varies among cohorts). Part 1 pertains to outdoor and urban exposures, part 2 to contaminant exposure biomarkers, part 3 to water disinfection by-products and indoor air pollutants, and part 4 to lifestyle and other exposures (tobacco smoke, diet, physical activity, alcohol, allergens, sleep, socio-economic capital).

Part 1 - Outdoor and urban exposures

Outdoor and urban exposures were assessed in the following exposure groups: Atmospheric pollutants, ultraviolet (UV) radiation, surrounding natural space, meteorological measures, built environment, traffic, and road traffic noise. Exposure assessment for these exposure groups was conducted within the PostgreSQL (copyright © 1996-2017 The PostgreSQL Global Development Group), PostGIS (Creative Commons Attribution-Share Alike 3.0 License http://postgis.net) and QGIS (QGIS Development Team, 2016. QGIS Geographic Information System) platforms. Source of data for each exposure are summarized in **Supplementary Table 1**. For the pregnancy period, exposure was assessed at the geocoded residential address of each woman. For each woman, assessment of exposure during pregnancy at the geocoded residential address at recruitment was made. For the childhood period, exposure was assessed at the geocoded residential and school addresses of each child as reported at the time of the subcohort visit. In case of multiple addresses, results were averaged by mother or child.

Supplementary Table 1. Exposure data sources

Exposure	BiB	EDEN	INMA	KANC	MoBa	Rhea
Atmospheric pollutants						
NO_2	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR
$PM_{2.5}$	ESCAPE local LUR	ESCAPE European LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR
PM_{10}	ESCAPE local LUR	Local dispersion model[a]	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR
PM_{abs}	ESCAPE local LUR	NA	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR	ESCAPE local LUR
Surrounding natural space						
Major green and blue spaces and landuse	Urbanatlas (2006)	Urbanatlas (2006)	Urbanatlas (2006)	Urbanatlas (2006)	Kartverket (2014)	Urbanatlas (2006)
NDVI	Landsat 4–5 TM, Landsat 7 ETM+, and Landsat 8 OLI/TIRS					
Meteorological measures						
Temperature, Humidity, Pressure[b]	Keighley	Poitiers	Cerdanyola/ Sabadell	Kaunas	Tryvannshogda	Iraklion
UV	TEMIS project	TEMIS project	TEMIS project	TEMIS project	TEMIS project	TEMIS project
Built environment						

Building Density	MasterMap (Ordinance Survey) (2013)	IGN (2014)	ICC (2011)	Open Street Maps (2014)	Open Street Maps (2014)	Greek Statistical Authority (2001)
Street Connectivity and Facilities	Navteq	Navteq	Navteq	Navteq	Navteq	Navteq
Population Density	EEA (2001)	EEA (2001)	INE (2011)	EEA (2001)	Statistics Norway (2005-2013)	EEA (2001)
Public transport (bus stops)	Bradford Metropolitan District Council (2014, 2015)	Grand Poitiers (2013)	Sabadell Municipality (2014)	Open Street Maps (2015)	Company "Ruter" (2015)	Open Street Maps (2015)
Road traffic						
Traffic	City of Bradford metropolitan district, Leeds City Council (2009, 2012)	Atmo Poitou Charentes (2005)	GENCAT (2007)	AudriusDedelé, Vytautas Magnus University (2010)	Municipality of Oslo, Norwegian Public Roads Administration (2006, 2011, 2014)	Fieldwork (2015
Road traffic noise						
Noise	DEFRA GOV. UK (2006)	Mairie de Poitiers (2007-2009)	GENCAT, Barcelona municipality (2006, 2012)	Kaunas Municipality (2007)	Oslo Municipality (2006, 2011)	Fieldwork (2015)

Abbreviations: **DEFRA**, Department of Environment Food and Rural Affairs; **EEA**, European Environment Agency; **ESCAPE**, European Study of Cohorts for Air Pollution Effects; **ETM+**, Enhanced Thematic Mapper Plus; **GENCAT**, Generalitat of Catalonia; **ICC**, Institut Cartogràfic de Catalunya; **IGN**, Institut National de l'Information Géographique et Forestière (http://professionnels.ign.fr); **INE**, Instituto Nacional de Estadística; **LUR**, Land Use Regression; **NA**, not available; **Navteq**: ESRI Street Map for Mobile Navteq 2012; **NDVI**, Normalized Difference Vegetation Index; **NO_2**, nitrogen dioxide; **OLI**, Operational Land Imager; **$PM_{2.5}$**, particulate matter with an aerodynamic diameter of less than 2.5 μm; **PM_{10}**, particulate matter with an aerodynamic diameter of less than 10μm; **PM_{abs}**, absorbance of $PM_{2.5}$ filters; **TEMIS**: Tropospheric Emission Monitoring Internet Service (http://www.temis.nl/uvradiation/archives); **TIRS**, Therma Infrared Sensor; **TM**, Thematic Mapper; **UV**, ultraviolet.
[a] only for pregnancy period; [b]location of weather station.

Atmospheric pollutants

The following atmospheric pollutants were assessed: nitrogen dioxide (NO_2), particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$) and of less than 10 μm (PM_{10}), and absorbance of $PM_{2.5}$ filters (PM_{abs}). These were assessed using land use regression (LUR) or dispersion models (for PM10 in EDEN during pregnancy), temporally adjusted to measurements made in local background monitoring stations and averaged over the periods of interest. In most cases we used site-specific LUR models developed in the context of the European Study of Cohorts for Air Pollution Effects (ESCAPE) project [2–6]. For BiB, assessment for $PM_{2.5}$ and PM_{10} was made based on the ESCAPE LUR model developed in London/Oxford (UK) and adjusted for background PM levels from monitoring stations in Bradford [7]. For EDEN, the ESCAPE European-wide LUR model was applied for $PM_{2.5}$ [8], and ESCAPE local LUR were used to assess NO_2 and PM_{10} exposure (the latter only for the pregnancy period) [9]. Data on daily background concentrations of air pollutants for temporal adjustment were obtained from routine background stations active during the whole study period. Back-extrapolation based on other available pollutants was used when data on a pollutant were not available. In particular, daily PM_{10} was used to adjust NO_2; daily NO_2 or PM_{10} factors to adjust $PM_{2.5}$; daily NO_2 to adjust PM_{10}; and daily NO_x to adjust PM_{abs}. Data availability is summarized in **Supplementary Table 2**. For the pregnancy period the exposure estimates were calculated for the three pregnancy trimesters and as the mean of whole pregnancy period.

Supplementary Table 2. Availability of daily values for each outdoor air pollutant by cohort

Cohort	NO_2	PM_{10}	$PM_{2.5}$	PM_{abs}
MoBa	Daily values available	Daily values available	Daily values available	Back extrapolated (NO_x)
KANC	Daily values available	Daily values available	Back extrapolated (NO_2)	Back extrapolated (NO_x)
BiB	Daily values available	Back extrapolated (NO_2)	Back extrapolated (NO_2)	Back extrapolated (NO_x)
EDEN	Daily values available	Back extrapolated (NO_2)	Back extrapolated (NO_2)	NA
INMA	Daily values available	Back extrapolated (NO_2)	Back extrapolated (NO_2)	Back extrapolated (NO_x)
RHEA	Back extrapolated (PM_{10})	Daily values available	Back extrapolated (PM_{10})	NA

Abbreviations: **NA**, not available; **NO_2**, nitrogen dioxide; **$PM_{2.5}$**, particulate matter with an aerodynamic diameter of less than 2.5 μm; **PM_{10}**, particulate matter with an aerodynamic diameter of less than 10μm**;** **PM_{abs}**, absorbance of $PM_{2.5}$ filters.

Surrounding natural space

We followed the PHENOTYPE protocol [10] to measure the surrounding greenness, i.e. trees, shrubs and parkland, and applied the Normalized Difference Vegetation Index (NDVI)[11] derived from the Landsat 4–5 Thematic Mapper (TM), Landsat 7 Enhanced Thematic Mapper Plus (ETM+), and Landsat 8 Operational Land Imager (OLI)/Thermal Infrared Sensor (TIRS) with 30m × 30m resolution (courtesy of the U.S. Geology Survey). NDVI quantifies greenness by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). NDVI values range from +1.0 to -1.0, with higher numbers indicating more greenness. To achieve maximum exposure contrast, we used available cloud-free Landsat images during

the period between May and August for years relevant to our period of study and calculated greenness within 100, 300 and 500 meter buffers around each address. Negative values in the images have been reclassified to null values previously. Furthermore, an indicator for "residential proximity to major green spaces" was created, as it covers different aspects of natural space exposure, i.e. easy access to recreational space. We calculated access to major green spaces (parks or countryside) and major blue spaces (bodies of water) as the straight line distance from the home or school to nearest blue or green space with an area greater than 5000 m^2 from topographical maps [12,13] or local sources, see table **Supplementary Table 1**. We also created a dichotomous variable to define whether a major green or blue space was present or not within a buffer of 300 m. For the pregnancy period the **presence of a major blue or green space, and NDVI** within a 100 meter buffer, were selected as the main exposure variables.

Meteorological variables

We used meteorological stations in the study area to obtain data on temporal variability in temperature. Daily measurements of temperature and humidity were obtained from a local weather station in each study area and averaged over each period of interest. Atmospheric pressure data were obtained from the ESCAPE project, and were available only for pregnancy trimesters and the entire pregnancy period (**pregnancy mean** was selected as main exposure), not for the childhood period. During the childhood period temperature and humidity were estimated for the home and school address. Daily, weekly and monthly measurements of UV radiation (as erythemal UV, Viamin-D and DNA damaging UV) at home and at school at 0.5 x 0.5 degree resolution were obtained from the Global Ozone Monitoring Experiment onboard the ERS-2 (European Remote Sensing) satellite (Temis), and averaged over the day, week and month before the subcohort follow-up examination.

Built environment

Topological maps for the following built environment indicators were obtained from local authorities or from Europe wide sources (**Table 1**). Building density was calculated within 100 and 300 meters buffer by dividing the area of building cover (m^2) by the area of each buffer (km^2). Population density was calculated as the number of inhabitants per km^2 surrounding the home address. Street connectivity was calculated as the number of street intersections inside 100 and 300 meters buffer, divided by the area (km^2) of each buffer. Facility richness index was calculated as the number of different facility types present divided by the maximum potential number of facility types specified, in a buffer of 300 meters, giving a score of 0 to 1. Facility density index was calculated as the number of facilities present divided by the area of the 300 meters buffer (number of facilities/km^2). A higher value indicates a more availability of different facility types. Landuse Shannon's Evenness Index (SEI) was calculated to provide the proportional abundance of each type of land use in a buffer of 300 meters, giving a score between 0 and 1 [14]. It was calculated by multiplying each proportion of land use type by its logarithm and dividing the sum of all land use type products by the logarithm of the total possible land use types. We developed an indicator of walkability, adapted from the previous walkability indexes [15–17], calculated as the mean and sum of the deciles of population density, street connectivity, facility richness index and land use SEI within 300 meters buffers, giving a walkability score ranging from 0 to 1. Accessibility was measured by BST (bus public transport) lines and stops were obtained from local authorities of each study area and from Open Street Maps ("OpenStreetMap") where local layers were not available. BST lines density was calculated as meters of BST lines inside 100, 300 and 500 meters buffer, divided by the buffer area in square kilometers. BST stop density was

calculated as number of BST inside 100, 300 and 500 meters buffer, divided by the buffer area in square kilometers.

Road traffic

Traffic density indicators (traffic density on nearest road, traffic load on all and major roads within 100 m buffer and inverse distance to nearest road) were calculated from road network maps following the ESCAPE protocol [4,6]. A fieldwork campaign was conducted in Heraklion during 2015, to assess multiple exposures as previously described [19]. Briefly, measurements of manual traffic counts of light and heavy vehicles over 15 minutes, and of noise, averaged over 30 minutes monitoring (Sonometer SC160, CESVA monitors - Spain), were made in 160 sites around the city. Sites were chosen representing multiple types (e.g. traffic, urban background, urban green etc.). During the campaign each monitoring site was measured three times in different seasons (summer, winter and autumn). We applied the LUR methods and GIS predictor variables used within the ESCAPE project and described in Eftens (Eeftens et al. 2012) to develop LUR models of traffic count and road traffic noise (**Supplementary Table 3**).

Road traffic noise

Noise levels, i.e. Lden (annual average sound pressure level of 24h period: day, evening and) and Ln (annual average sound pressure level of night period) were derived from noise maps produced in each local municipality under the European Noise Directive (EC Directive 2002/49/EC [20]). To improve comparability between centers, the values were categorized into six categories (<55; 55-59.9; 60-64.9; 65-69.9; 70-74.9; >75) for analysis. For RHEA, estimates on noise were newly modeled following new fieldwork (see **Supplementary Table 3** and above for details).

Supplementary Table 3. Summary of land use regression models and descriptive statistics of traffic count and road traffic noise exposure within Heraklion.

Exposure	LUR model	R2 model	R2 cross validation	RMSE	Moran's I2 (p value)	Mean Measured levels (range)
Traffic count	1.2 - 0.38 * TypeofRoad + PostCode + Land Use – 0.47 * LOG dist dense road + 0.004 * Buffer 50 m to roads	0.71	0.65	3.2veh/15 mins	-0.04 (0.16)	133 veh/15 mins [0-933]
Road traffic noise	71.7 - 103 * TypeofRoad + SiteType – 39.4 * PavedY	0.45	0.41	55 dB	-0.02 (0.23)	58.6 dB [44.4 - 72.3]

Abbreviations: **dB**, decibel; **LUR**, land use regression; mins, minutes; **RMSE**, Root-mean-square deviation; **dist**, distance; **veh**, vehicles.

Part 2 - Water Disinfection By-Products and Indoor Air Pollutants

Water Disinfection By-Products (DBPs)

We collected data on routine measurements of disinfection by-product (DBP) in water from water companies for all cohorts for the pregnancy period. For KANC, BiB, INMA and RHEA cohorts this was built on the HiWate project (Health Impacts of long-term exposure to disinfection by-products in drinking Water) [21] that previously modelled exposure levels in the water supply of the residence of each participating mother-child pair. For BiB, routine monitoring data on trihalomethanes (THMs) were obtained for the eight water supply zones covering the study area. Each zone was sampled nine times per year on average, giving 374 data points in total [22]. For INMA, levels of THMs were ascertained based on sampling campaigns and regulatory data from local authorities and water companies. Sampling locations were defined to be geographically representative of the study areas, and water samples were collected from taps with no filtration or other treatments that could affect THMs concentration. THMs were determined in 198 places [23]. For RHEA, the city was divided into six zones according to the source of underground water used in each area, corresponding to six different water treatment plants. In total, 18 sampling points were selected (12 areas in Heraklion and 6 in rural areas), which covered geographically the residences of participating mother-child pairs [24]. For KANC, tap water THM concentration, derived as the average of quarterly sample values over the time that the pregnancy occurred from all sampling sites located in the each distribution system, and geocoded maternal address at birth to assign the individual women's residential exposure index [25]. Routine DBP measurements were acquired for MoBa and EDEN cohorts as these cohorts were not part of the HiWate project. THMs exposure levels were modelled for each residence, following the protocol developed within HiWate [21].

Indoor Air Pollutants

Indoor air concentrations of nitrogen dioxide (NO_2), particulate matter <2.5μm ($PM_{2.5}$), particulate matter absorbance (PM_{abs}), benzene, and toluene, ethylbenzene, xylene (TEX) were estimated through a prediction model that combined measurements in the homes of a subgroup of children with questionnaire data from the subcohort.

Measurements of indoor NO_2, benzene and TEX were conducted in the homes of 157 participants as part of the child panel study , which was nested within the HELIX subcohort in all cohorts except MoBa. $PM_{2.5}$ and PM_{Abs} were measured in INMA, BiB, and EDEN. Participants in the child panel study were followed for one week in two seasons, and the last day of the first week coincided with the subcohort examination, including the completion of the main HELIX questionnaire. NO2, benzene and TEX sampling lasted 7 days, and $PM_{2.5}$ and PM_{Abs} sampling lasted 24 hours.

NO_2 short-term diffusive Passam samplers were used to measure indoor NO_2 concentrations. The samplers were composed of polypropylene housing with a 20 mm diameter opening, covered with a removable plastic cap and protected from wind disturbance by a teflon membrane. Triethanolamine was used as absorbent material inside the tube. NO_2 was collected by molecular diffusion to the absorbent and its concentration was determined spectrophotometrically by the Saltzmann method. The detection limit (DL) for a week's sampling for the NO_2 sampler was 0.3 $\mu g/m^3$. Passam ORSA5 diffusion tubes were used to measure indoor levels of benzene, toluene, ethylbenzene and ortho-, para- and metaxylenes. The DL for a week's sampling for each compound was 0.4 $\mu g/m^3$. The samplers were placed in the living rooms of the participating homes, away from the sources of ventilation. After collection, the NO_2, Benzene and TEX samplers were

hermetically sealed and kept in zip-lock bags in boxes, in a cool and dark place and shipped to the analyzing laboratory within 3 months of the end of the sampling campaign.

For indoor PM modeling, active $PM_{2.5}$ cyclone pumps were placed in the living room. After 24 hours the samplers were collected and sent to laboratory. $PM_{2.5}$ mass was collected gravimetrically using 37-mm Teflon filters held in a cyclone (model GK2.05 SH, BGI Inc., Waltham MA, USA) with an aerodynamic cut point of 2.5 μm and connected to a BGI/Mesa Labs A4004 pump working at 3.5L/min. Filter weighing and reflectance measurements were conducted with a microbalance of 1 μg accuracy (Model MX5, Mettler-Toledo International Inc., Switzerland) and a Smoke Stain Reflectometer (SSR) (Model 43D, Diffusion Systems Ltd., UK), respectively. Measurement procedures, quality control, as well as $PM_{2.5}$ mass concentration and absorbance estimations followed the ESCAPE project protocols (both available at www.escapeproject.eu/manuals).

Statistical analyses were performed separately for each of the exposure variables. A TEX variable was created by summing the concentrations of each TEX compound. The HELIX main questionnaire (Maitre et al., under revision) was used to identify housing and participant characteristics as input for the prediction model; these characteristics included: exposure to environmental tobacco smoke, cooking and heating methods at the home, cleaning products between others.

After extracting potential predictor variables from the questionnaires, bivariate analyses were run by either Kruskal-Wallis or Wilcoxon rank sum tests, as all of the potential predictors were categorical and the exposure variables were not normally distributed. The variables that yielded a p value lower than 0.2 in bivariate analyses were selected to enter into the multiple linear regression models. Prior to that, univariate linear regressions were performed for each of the predictors selected in the bivariate analysis in order to assess the adjusted determination coefficient (adjusted R^2) for each of them individually. To ensure normality of the distributions of the outcome variables, the univariate linear regression models and subsequent multiple linear regression models were built using log-transformed.

Supervised forward stepwise procedure was employed to build multiple linear regression models. In all cases the starting point for the regression was the variable which yielded the highest adjusted R^2 in the univariate linear regressions. Then the other predictors were added one-by-one and additional increase in the adjusted R^2 was recorded. The variable which increased the adjusted R^2 by a highest value was retained in the model and the procedure was repeated until none of the variables increased the adjusted R^2 by at least 1%. In case any of the variables included into the model had an individual p value equal or higher than 0.05, it was removed from the model. All statistical analyses were performed using R Statistic Software (version 3.4.1).

The best explained pollutant was NO_2 with an R^2 of 57%, followed by PM_{Abs} with 50%. **Supplementary Table 4** shows the efficiency of the models and the statistically significant variables. For example, cohort, natural gas oven, type of hob and boiler, butane in the living room, and the number of people living in the house, were the statistically significant variables in the NO_2 model; all of these were positively correlated.

Supplementary Table 4. Summary of the models of indoor air pollutions.
*** < 0.001 ** < 0.005 * < 0.05. For negative coefficients (-) sign is included.

Exposure	NO_2	Benzene	TEX	PM2.5	PMAbs
Explained variability (R^2)	**57%**	**31%**	**31%**	**47%**	**50%**
Cohort	***			*	**
Oven with natural gas	***				
Type of hob	***				
Type of boiler	**				
Butane in living room	**			*	
How many people live at home?	***				
Garage connected to the house?		***	*		
$PM_{2.5}$ outdoor		***			
Does air pollution bother you?		*			
Does your family manage financially?		**			
Number of floors of the house		(-)*			
How many cigarettes per week do you smoke (mother)?		*			
How often do you use degreasing sprays?			**		
Presence of central heating?			(-)*		
How often do you use perfumed cleaning products?			(-)**		
How many cigarettes smoke (mother's partner)?			**		
Calendar month			**		
NO_2 outdoor			*		***
How many cigarettes last week (mother)?				***	***
Family has a car?				(-)**	
Stay at home parent?				**	
How often do you use glass cleaning sprays?				*	

Part 3 - Lifestyle and other exposures

Tobacco smoke

Tobacco smoke exposure was assessed in pregnancy via questionnaire for active and passive smoking, as well as based on cotinine measurements (as described in part 3). Pregnancy questions on tobacco smoke from the cohorts were harmonized as part of the ESCAPE project. Tobacco smoke exposure of the mother at any point during pregnancy was categorised into: no exposure, only passive smoke exposure, active smoking. Active smoking was also measured by the number of cigarettes per day on average during pregnancy

For children, in addition to the cotinine-based classification (see part 3), the following two variables were created based on the questionnaires completed by the parents:

- The global exposure of the child to ETS with two categories: "no exposure", no exposure at home neither in other places; “exposure”: exposure in at least one place, at home or outside.
- Active smoking of the parents: “1” none of the parents, “2” one parent or “3” both parents.

Diet

Diet during pregnancy was assessed through food frequency questionnaires by each cohort and harmonized *a posteriori* for the HELIX project. Harmonisation was possible for eight main food groups (average consumption in times/week) and folic acid supplementation intake (yes/no) in the first trimester for five of the six cohorts (KANC not available).

In early childhood years information about breastfeeding duration (in weeks) was collected by the cohorts and then harmonized as part of HELIX.

Information on the child’s diet was collected through the standardized HELIX subcohort questionnaire. The child’s diet was then summarized in 15 food groups (times/week) and dietary habits such as eating organic food (see **Supplementary Table 5**). We also included the KIDMED index, a dietary score representative of healthy eating and based on the principles of Mediterranean dietary patterns. The KIDMED index consists of 16 questions with questions denoting a negative connotation with respect to the Mediterranean diet assigned a value of -1, and those with a positive aspect scored +1 (Serra-Majem et al., 2004). Further, we analysed as separate variables few factors that contribute to the KIDMED index including fast food visits, organic food and ready-made supermarket meal consumption.

Supplementary Table 5. Diet variables included in the exposome for pregnancy and childhood periods.

	Pregnancy	**Childhood**
Cereals	Yes	Yes
Dairyproducts	Yes	Yes
Fish and seafood	Yes	Yes
Fruits	Yes	Yes
Meat	Yes	Yes
Vegetables	Yes	Yes
Visits a fast food restaurant/take away	Yes	Yes

Folic acid supplementation (yes/no)	Yes	-
Legumes	Yes	-
Breastfeeding duration (in weeks)	-	Yes
Bakery products	-	Yes
Breakfast cereal	-	Yes
Bread (white and whole wheat)	-	Yes
Potatoes	-	Yes
Sweets	-	Yes
Yogurt and probiotics	-	Yes
Processed meat	-	Yes
Total added lipids (butter, margarine and vegetable oils)	-	Yes
Beverages (sodas)	-	Yes
Caffeinated drinks	-	Yes
Organic food	-	Yes
Ready-made supermarket meal	-	Yes
KIDMED score	-	Yes

Physical activity

Physical activity during pregnancy (3rd trimester only) was estimated based on the harmonization of the respective cohort questionnaire data. Two variables were created: (1) moderate activity corresponding to walking and/or cycling activity (expressed in frequency categories: never or sometimes; often; very often); and (2) vigorous activity (in two frequency categories: low and medium/high) corresponding to exercise or sport activity.

For children, the moderate-to-vigorous physical activity variable was created based on questionnaire data. It was defined as the amount of time children spent doing physical activities with intensity above 3 metabolic equivalent tasks (METs), and is expressed in units of min/day. Physical activity over-reporting was corrected based on the accelerometer (Actigraph) correlation with questionnaire answers, using the data from three cities involved in the HELIX panels (nested study of the HELIX project where participants wore accelerometers for two non-consecutive weeks).

A variable representing sedentary behavior in the children was created based on the questionnaire and corresponds to the duration of time spent watching TV, playing computer games or other sedentary games. This variable is a new concept which is commonly defined as ‘‘any waking behavior characterized by an energy expenditure <1.5 metabolic equivalent tasks (METs) while in a sitting or reclining posture’’ by the Sedentary Behaviour Research Network [46]. Sedentary behavior has been shown to be a health risk factor independently from physical activity.

Alcohol

Alcohol consumption during pregnancy was harmonized based on questionnaire data from the cohorts and classified as whether or not any alcohol was consumed during pregnancy (except in the KANC cohort where the lowest exposure category included women with less than 1 glass a month).

Allergens

For allergen exposure only pet ownership of the child was added to the exposome. There was no prenatal information on this. Three variables were created based on the HELIX

questionnaire as follow: (1) if the child had any cats that live mainly in his home (2) or dogs, or (3) any other pets than dogs and cats.

Sleep

Sleep duration was available for the subcohort children, not for the mothers, and corresponds to the average sleep duration at night during an entire week (weighted average of weekdays and weekend sleep duration). This variable was calculated based on the questionnaire taking the average bedtime and wake-up time (earliest and latest bedtime/wake-up times available) during weekdays and weekends.

Socio-economic capital

Questions related to socio-economic position (maternal education and others) were collected during the pregnancy in all cohorts and harmonized for use as covariates in analyses; they were not included in the exposome as separate exposure variables. In the childhood exposome, the Family Affluence Score (FAS) was included based on questions from the subcohort questionnaire [47]. A composite FAS score was calculated based on the responses to the next four items: (1) Does your family own a car, van or truck? (2) Do you have your own bedroom for yourself? (3) During the past 12 months, how many times did you travel away on holiday with your family? (4) How many computers does your family own? (Liu et al, 2012).A three point ordinal scale was used, where FAS low (score 0,1,2) indicates low affluence, FAS medium (score 3,4,5) indicates middle affluence, and FAS high (score 6,7,8,9) indicates high affluence FAS [48]. The FAS score in this study had only a maximum value of 7 instead of 9 because of the smaller number of possible answers for certain items.

Further social capital-related questions were included in the HELIX questionnaire to capture different aspects of social capital, relating both to the cognitive (feelings about relationships) and structural (number of friends, number of organizations) dimensions and to bonding capital (close friends and family), bridging capital (neighborhood connections, looser ties) and linking capital (ties across power levels; for example political membership). Two summary variables were selected for the exposome analysis: social participation (membership of organizations: 0, 1, or 2) and contact with friends and family (daily, once a week, less than once a week). In addition, house crowding was included, representing the number of persons living in the house with the child.

Part 4 - Contaminant exposure biomarkers

For all the 1,301 children in the subcohort, biomarker the determinations of a set of chemical contaminants (organochlorine compounds, brominated compounds, perfluorinated alkylated substances (PFAS), metals and elements, phthalate metabolites, phenols, and organophosphate (OP) pesticide metabolites) were performed at the Department of Environmental Exposure and Epidemiology at the Norwegian Institute of Public Health (NIPH), in Norway or in collaboration with their contract laboratories. This was also the case for the majority of the maternal samples collected during pregnancy or at birth and stored in cohort biobanks; however, for some maternal samples in some cohorts, measurements were already completed at thus we used these results (**Supplementary Table 6**). Here we provide a summary of the methods used to determine biomarker levels for the chemical contaminants; more detailed information can be found in Haug et al.[26]

Supplementary Table 6. Concentrations of chemical contaminants previously analyzed in other labs.

	Total maternal samples analyzed	Analyzed in NIPH as part of HELIX	Previously analyzed in other labs
Organochlorine compounds	1078	657	INMA: 223 RHEA: 198
Brominated compounds	855	657	RHEA: 198 (only PBDE-47 available)
Perfluorinated alkylated substances	1240	1032	INMA: 208
Metals and essential elements	1020	833	INMA: 223 (only Hg available)
Phthalate metabolites	1089	914	INMA: 175
Phenols	1085	1023	EDEN: 62
Organophosphate pesticide metabolites	1086	1086	-
Cotinine	1093	883	INMA: 210
Creatinine	1093	870	INMA: 223
Lipids	1075	654	INMA: 223 RHEA: 198

Quality assurance

The sample collections for the children were performed in a completely harmonized way, using the same protocols and equipment for sample collection and processing in all the six cohorts (Maitre et al, under revision). The children's samples were randomized into batches before chemical analyses, aiming at a minimum of three cohorts to be included in each batch. However, this was not feasible for the maternal samples as the cohorts shipped the maternal samples at different time points to the laboratories for analysis.

Chemical analysis

SupplementaryTable 7 shows the fifty-eight environmental chemicals measured in the HELIX subcohort. **Supplementary Tables 8** and **9** show the collection time points and the biological matrices, respectively.

Supplementary Table 7. Chemical contaminants and number of samples analyzed from mothers and children in the HELIX subcohort.

	Abbreviation	Children's samples N=1,301		Maternal samples N=1,294	
Compound		**n analysed**	**% quantifiable samples**	**n analysed**	**% quantifiable samples**
Organochlorine compounds (OCs)					
2,3',4,4',5-Pentachlorobiphenyl	PCB 118	1296	99.8	1078	79.1
2,2',3,4,4',5'-Hexachlorobiphenyl	PCB 138	1296	99.8	1078	96.5
2,2',4,4',5,5'-Hexachlorobiphenyl	PCB 153	1296	100	1078	99.6
2,2',3,3',4,4',5-Heptachlorobiphenyl	PCB 170	1296	90.7	855	99.5
2,2',3,4,4',5,5'-Heptachlorobiphenyl	PCB 180	1296	99.2	1078	97.6
4,4'dichlorodiphenyltrichloroethane	DDT	1296	79.8	1078	65.6
4,4'dichlorodiphenyldichloroethylene	DDE	1296	100.0	1078	99.9
Hexachlorobenzene	HCB	1296	99.9	1078	99.1
Brominated compounds (PBDEs)					
2,2',4,4'-Tetrabromodiphenyl ether	PBDE 47	1296	90.8	855	80.9
2,2',4,4',5,5'-Hexabromodiphenyl ether	PBDE 153	1296	54.4	657	72.9
Perfluoroalkyl substances (PFASs)					
Perfluorohexanesulfonate	PFHxS	1301	99.7	1240	97.5
Perfluorohexanesulfonate	PFOS	1301	99.8	1240	100
Perfluorooctanoate	PFOA	1301	100	1240	99.7
Perfluorononanoate	PFNA	1301	99.5	1240	97.9
Perfluoroundecanoate	PFUnDA	1301	68.6	1032	95.4
Metals and essential elements					
Mercury	Hg	1298	97.7	1020	98.9
Cadmium	Cd	1298	86.5	833	99.6
Lead	Pb	1298	100	833	100
Arsenic	As	1298	67.1	833	58.5
Cesium	Cs	1298	100	833	100
Copper	Cu	1298	100	833	100
Thallium	Tl	1298	7.2	833	1.1
Manganese	Mn	1298	100	833	100
Zinc	Zn	1298	100	833	100
Cobalt	Co	1298	99.9	833	100
Molybdenum	Mo	1298	99.5	833	100
Sodium	Na	1298	100	833	100
Potassium	K	1298	100	833	100
Magnesium	Mg	1298	100	833	100
Phthalate metabolites					
Monoethyl phthalate	MEP	1301	100	1089	99.0
Mono-iso-butyl phthalate	MiBP	1301	100	1089	99.9
Mono-n-butyl phthalate	MnBP	1301	100	1089	100
Mono benzyl phthalate	MBzP	1301	99.9	1089	99.7
Mono-2-ethylhexyl phthalate	MEHP	1301	96.8	1089	99.5
Mono-2-ethyl-5-hydroxyhexyl phthalate	MEHHP	1301	99.8	1089	100
Mono-2-ethyl-5-oxohexyl phthalate	MEOHP	1301	99.9	1089	100
Mono-2-ethyl 5-carboxypentyl phthalate	MECPP	1301	99.9	914	99.9
Mono-4-methyl-7-hydroxyoctyl phthalate	oh-MiNP	1301	100	914	92.6
Mono-4-methyl-7-oxooctyl phthalate	oxo-MiNP	1301	100	914	95.7
Phenols					
Methyl paraben	MEPA	1301	99.7	817	99.8
Ethyl-paraben	ETPA	1301	99.3	817	97.4
Propyl-paraben	PRPA	1301	67.3	1085	97.3
N-Butyl paraben	BUPA	1301	96.6	1085	97.0
Bisphenol-A	BPA	1301	98.3	1085	99.4
Oxybenzone	OXBE	1301	100	1085	98.5
Triclosan	TCS	1301	99.9	1085	99.3
Organophosphate (OP) pesticide metabolites					
Dimethyl phosphate	DMP	1301	49.3	1086	90.8
Dimethyl thiophosphate	DMTP	1301	90.4	1086	88.9
Dimethyl dithiophosphate	DMDTP	1301	18.2	1086	41.6
Diethyl phosphate	DEP	1301	80.9	1086	97.8
Diethyl thiophosphate	DETP	1301	43.5	1086	50.0
Diethyl dithiophosphate	DEDTP	1301	1.5	1086	1.7
Other compounds					
Cotinine		1301	17.4	1093	43.7
Creatinine		1301	100	1093	100
Phospholipids		1284	100	1052	62.4
Total cholesterol		1284	100	1052	100
Triglycerides		1284	100	1052	100
High-density lipoprotein cholesterol	HDL	1284	100	830	100
Low-density lipoprotein cholesterol	LDL	1284	99.8	830	100

n analysed: samples with biomarker measurements

% quantifiable samples: % of the biomarker measurements with concentrations reported

Supplementary Table 8. Collection time points of maternal and child blood and urine samples (mean, SD).

	Cohort					
	BiB	**EDEN**	**KANC**	**INMA**	**MoBa**	**RHEA**
Mother, gestational weeks	26.6 (1.4)	26.1 (1.2)	39.4 (1.3)	13.7 (2.0) / 34.2 (1.3) a	18.7 (0.9)	14.1 (3.7)
Child, years	6.6 (0.2)	10.8 (0.6)	6.5 (0.5)	8.8 (0.6)	8.5 (0.5)	6.5 (0.3)

Abbreviations: **SD**: standard deviation

[a]In INMA, blood was collected in the first trimester whereas urine was collected in the third trimester of pregnancy.

Supplementary Table 9. Biological matrices of maternal and child samples.

	Cohort					
Chemicals	**BiB**	**EDEN**	**KANC**	**INMA**	**MoBa**	**RHEA**
OCs and PBDEs						
Mother	serum/plasma	serum	-	serum	plasma	serum
Child	serum	serum	serum	serum	serum	serum
PFASs						
Mother	serum/plasma	serum	whole blood	plasma	plasma	serum
Child	plasma	plasma	plasma	plasma	plasma	plasma
Metals						
Mother	whole blood	whole blood	whole blood	cord whole blood	whole blood	whole blood
Child	whole blood	whole blood	whole blood	whole blood	whole blood	whole blood
Phthalate metabolites, phenols, OP pesticide metabolites, cotinine, and creatinine						
Mother	urine	urine	-	urine	urine	urine
Child	urine	urine	urine	urine	urine	urine
Lipids						
Mother	serum/plasma	serum	-	serum	plasma	serum
Child	plasma	plasma	Plasma	plasma	plasma	plasma

Abbreviations: **OC**: organochlorine; **OP**: organophosphate pesticides; **PBDEs**: polybrominateddiphenyl ethers; **PFASs**: per- and polyfluoroalkyl substances.

Organochlorine compounds (OCs)

Concentrations of OCs were determined in serum or plasma according to Caspersen et al (2016) except that gas chromatography–mass spectrometry (GC-MS/MS) was used instead of gas chromatography/high-resolution mass spectrometry (GC-HRMS). The limit of detection (LOD) was in the range of 0.3 to 1.5 pg/g. OCs concentrations in maternal samples (serum) of INMA and RHEA were determined according to Goñi et al (2007) with a LOD of 67.0 pg/g and Koponen et al (2013) with LODs between 1.7 and 14.3 pg/g, respectively. We also calculated the sum of PCBs by summing the concentrations of the 5 PCBs in pg/g.

Brominated compounds (PBDEs)

Concentrations of PBDEs were determined in serum or plasma following the method described in Caspersen et al (2016) also using GC-MS/MS for detection. The LOD ranged

from 0.15 to 0.3 pg/g. In RHEA only PBDE-47 was determined in maternal samples (serum) following the method described in Koponen et al (2013) with a LOD of 2.85pg/g.

Perfluorinated alkylated substances (PFAS)

Concentrations of PFASs were determined in serum or plasma using the method by Haug et al (2009), while the method by Poothong et al (2017a) was applied for the whole blood samples. The LOD was 0.02 µg/L for all PFASs. In the majority of INMA maternal samples (plasma), PFASs were determined according to Manzano-Salgado et al (2015) and with LODs between 0.05 and 0.1 µg/L. Only five maternal samples from INMA were analyzed at NIPH. In order to know whether concentrations measured in both labs were comparable we performed an inter-laboratory comparison of 10 samples with low to high PFOS concentrations as reference selected from all analyzed in the Institute for Occupational Medicine, RWTH Aachen University (Germany) [32]. NIPH was blinded to the concentrations of samples. PFOS and PFHxS plasma concentrations determined in both laboratories were highly correlated (Spearman r=0.83 and 0.93, respectively) whereas PFOA and PFNA were less correlated (Spearman r=0.70 and 0.55, respectively). The three samples with low PFOS concentrations had levels between the LOD and the LOQ or close to the LOQ for PFHxS, PFOA, and PFNA. Considering that concentrations between the LOD and the LOQ have higher uncertainty, we excluded these samples and the spearman correlations became higher: PFOA r=0.96, PFHxS r=0.93, and PFNA r=0.86. Due to the high correlations the NIPH concentrations for subjects included in the comparison have been used. For the PFASs, 1:1 ratios were assumed for serum and plasma, while 1:2 ratios were used for whole blood vs serum/plasma [33]. Thus, for PFASs all whole blood concentrations were multiplied by two.

Metals and essential elements

Concentrations of 15 metals and elements in whole blood were performed at ALS Scandinavia, Sweden according to Rodushkin et al (2000). The LOD ranged from 0.003–3.03 µg/L except for sodium (Na), potassium (K) and magnesium (Mg) for which the LOD ranges from 0.06-0.15 mg/L. Mercury in INMA was determined in cord whole blood following the procedure described in Ramon et al (2011) with a LOD of 2.0 µg/L. Cord blood Hg concentrations were be divided by 1.7 to be comparable with maternal whole blood concentrations [36]. Ten of these metals and elements (Hg, Cd, Pb, As, Cs, Cu, Tl, Mn, Co, and Mo) were included in the exposome analyses because of their potential toxicity. Zn, Na, K, Mg, and Se were not considered toxic and were included as covariates. This classification was based on expert judgment (Joan Grimalt, personal communication) and literature review [37]

Phthalate metabolites

Concentrations of ten phthalate metabolites were determined in urine according to Sabaredzovic et al (2015). The LOD ranged from 0.06 to 0.61µg/L. In the majority of INMA maternal samples, phthalates were determined according to Valvi et al (2015) with LOD ranged from 0.5-1.0 µg/L except 37 INMA samples that were analyzed at NIPH. For comparability, we analyzed 10 samples with low to high monoethyl phthalate (MEP) concentrations as reference selected from all analyzed in the Bioanalysis Research Group at the Hospital del Mar Medical Research Institute (Barcelona, Spain) [39]. NIPH was blinded to the concentrations of samples. Urinary concentrations of the phthalate metabolites determined in both laboratories were highly correlated (Spearman ranging from r=0.69 to 0.97). Due to the high correlations the NIPH concentrations for subjects included in the comparison have been used. We also calculated the total concentration of di-2-ethylhexyl phthalate (DEHP) by summing the molar concentrations of mono-2-

ethylhexyl phthalate (MEHP), mono-2-ethyl-5-hydroxyhexyl phthalate (MEHHP), mono-2-ethyl-5-oxohexyl phthalate (MEOHP), and mono-2-ethyl 5-carboxypentyl phthalate (MECPP). The molar concentrations (in µmol/L) were calculated by dividing the concentration of every metabolite by its molecular weight.

Phenols

Concentrations of phenols were determined in urine according to Sakhi et al (2018) with the LOD ranged from 0.03-0.06 µg/L. In EDEN, phthalate metabolites were determined in urine samples according to Philippat et al (2011) with the LOD ranged from 0.2-2.3 µg/L. We performed an inter-lab comparison of 12 samples selected from all analyzed in the I National Center for Environmental Health laboratory at the CDC in Atlanta, Georgia, USA [40]. NIPH was blinded to the concentrations of samples. Phenols urinary concentrations determined in both laboratories were strongly correlated (Spearman ranging from r=0.90 to 1.0). Due to the high correlations the NIPH concentrations for subjects included in the comparison have been used.

Organophosphate (OP) pesticide metabolites

Analysis of OP pesticide metabolites in urine was made according to Cequier et al (2016) and with the LOD ranged from 0.06-0.36µg/L. DMDTP in children was detected in less than 20% of samples and DEDTP in children and mothers was detected in less than 2% of samples (Table 7); therefore, categorical variables were created categorizing urinary DMTDP and DEDTP levels as detected or not detected considering the limits of detection of 0,19 and 0,05 µg/L, respectively. However, the DEDTP variable in mothers and children and the DMDTP variable in mothers had too few subjects in the "detected" category (less than 30) and were finally removed from the exposome analyses.

Cotinine

Concentrations of cotinine in urine were determined using The Immulite® 2000 Nicotine Metabolite (Cotinine) 600 Test on an Immulite 2000 XPi from Siemens Healthineers at Fürst Medisinsk Laboratorium, Norway. The LOD was 3.03µg/L. Cotinine in maternal urine samples from INMA were determined according to Aurrekoetxea et al (2013) and with a LOD of 1.21µg/L. We performed an interlab-comparison of 10 urine samples with low to high cotinine concentrations selected from all analyzed in the Public Health Laboratory of Bilbao - LSPPV (Spain) [42]. NIPH was blinded to the concentrations of samples. Cotinine urinary concentrations determined in both laboratories were highly correlated (Spearman r=0.95).

For maternal smoking, a categorical variable was created based on the urinary cotinine levels to distinguish non-smokers, second-hand-tobacco smokers, and smokers [43]:

- Non-smokers: values <LOD or cotinine levels <18.5 µg/L
- Second-hand-tobacco smokers: cotinine levels ≥18.5-50 µg/L
- Smokers: cotinine levels >50 µg/L

In the children, a categorical variable was created categorizing urinary cotinine levels as detected or not detected considering the limit of detection of 3.03µg/L.

Adjustments for total fat percentage and creatinine

Concentrations of lipids were determined in the Fürst Medical Analysis Laboratory in serum or plasma using the FS kit from DiaSys for phospholipids and the ADVIA® Chemistry XPT System for the other lipids. LODs ranged from 0.003 to 0.08 mmol/L. In maternal samples (serum) of INMA and RHEA total cholesterol and triglycerides were determined using the Cobas Mira self-analyzer (Roche Diagnostic, Basel, Switzerland) using an enzymatic-colorimetric method with spin react reagents and a standard

enzymatic method, respectively. Phospholipid concentrations in maternal samples from INMA and RHEA were calculated based on the formula of Covaci et al (2006). Total fat percentage was calculated considering the molecular weight of phospholipids, total cholesterol, and triglycerids, and calculated according to the method described in [45]. Concentrations of OCs and PBDEs were then adjusted in respect to total fat percentage and expressed in ng/g of lipids. Concentrations of creatinine in urine were performed on an AU680 Chemistry System form Beckman Coulter using DRI® Creatinine-Detect® Test at Fürst Medisinsk Laboratorium, Norway with a LOD of 0,03mmol/L. Creatinine in maternal samples of INMA and EDEN were determined by using the Jaffé method - Beckman Coulter© AU5400 and an enzymatic reaction using a Roche Hitachi 912 chemistry analyzer (Roche Hitachi, Basel, Switzerland), respectively. Urinary concentrations of phthalate metabolites, phenols, OP pesticide metabolites, and cotinine were adjusted in respect to creatinine and expressed in μg/g of creatinine.

Supplementary Methods

This section details the data selection procedure performed in this study.

Part 1 - Initial data selection

From all variables available in the HELIX sub-cohorts, we made minimal selection decisions among groups of related variables, selecting representatives in order to reduce the dimensionality of the dataset with minimal loss of information. More specifically, we filtered single representatives from groups of corelated variables identified in previous HELIX studies [49]. In total, we dropped 122 variables from 598 variables, for a remaining total of 476 variables. See full description below:

1. Chemical exposures

For those pollutants, we used the sum of pollutants instead of the single entities, as we are interested in the overall effect of those toxicant exposure.

For Polychlorinated biphenyl exposures (PCB), both in mother and child, we only used a summary variable that aggregates the measure of all types of PCBs (namely 118, 138, 153, 170, 180). In total, we dropped 10 variables and added 2.

Similarly, for phthalates (DEHP), we used a summed variable to resume the measure of all types of DEHP (namely mono benzyl phthalate, mono-2-ethyl 5-carboxypentyl phthalate, mono-2-ethylhexyl phthalate, Mono-2-ethyl-5-hydroxyhexyl phthalate, mono-2-ethyl-5-oxohexyl phthalate, mono-iso-butyl phthalate, Mono-n-butyl phthalate, mono-4-methyl-7-hydroxyoctyl phthalate and mono-4-methyl-7-oxooctyl phthalate). In total, we dropped 18 variables and added 2.

2. Built environment

For build environment variables that were measured at different radius (100-meter, 300-meter and 500-meter radius), following selection made on previous studies using those data, we selected 300m when available or else 100m. As a result, we selected: 100m for NDVI values, 300m area for amount of public transport lines, 300m area for the number of bus public transport mode stops, 300m area for building density and 300m area for connectivity density (number of intersections / km2)). In total, we dropped 14 variables.

3. Outdoor air pollutants

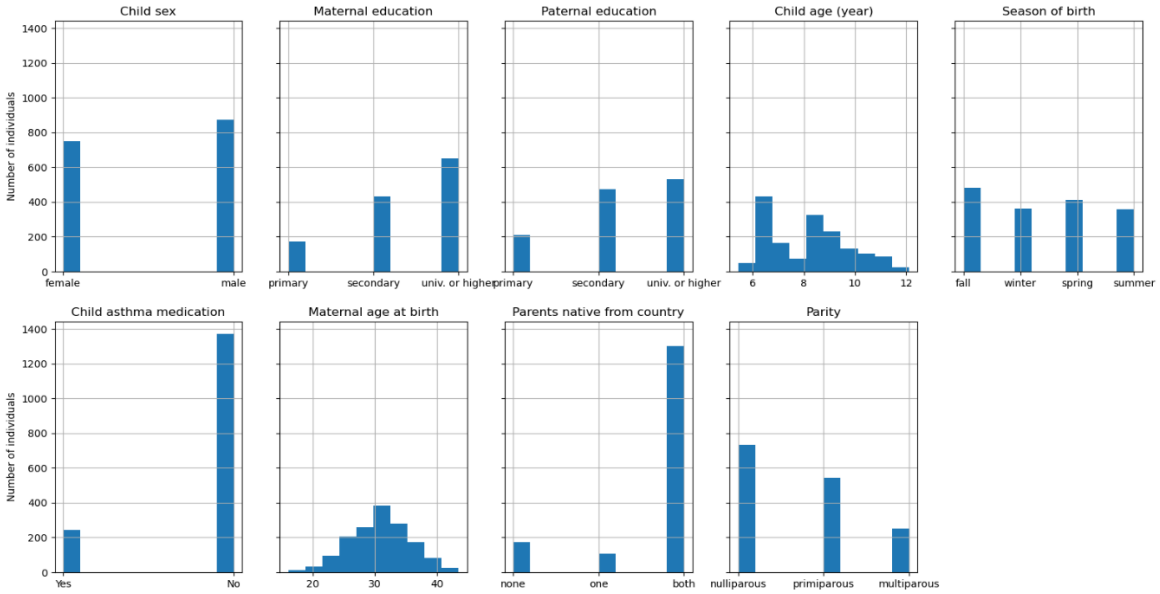
For outdoor air pollutants variables measured at different areas (home, school, commuting, other places), we selected pollutants measured at home (namely for NO2, NOX, pm10, pm25, pm absorbance and pm coarse). We dropped a total of 72 variables.

4. Meteorological variables

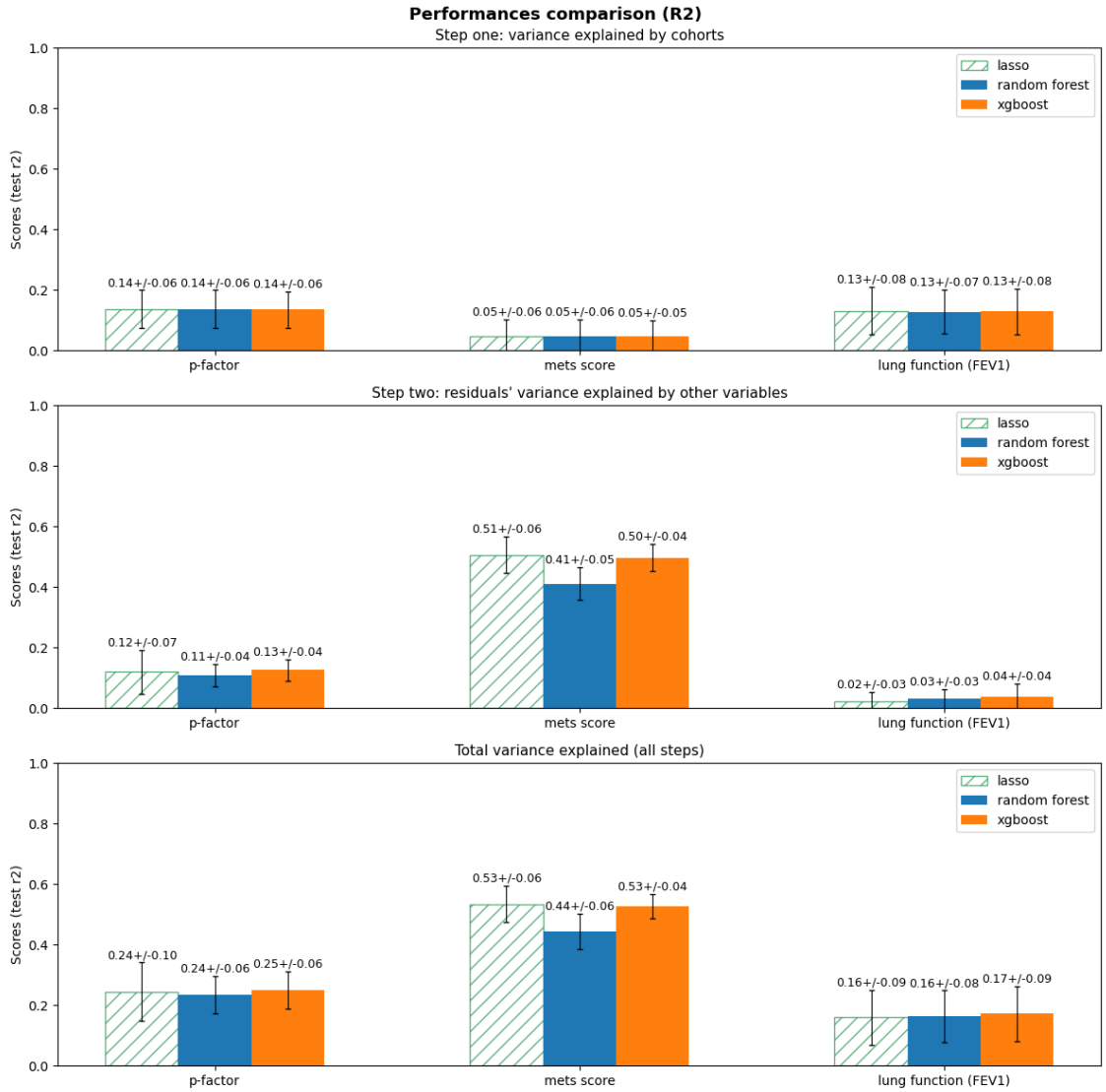
For temperature and humidity, we only selected averaged values across several periods (day, week and month) and discarded min/max values. Additionally, yearly averaged values were discarded as they were encoding only the cohort information. In total, we dropped 12 variables.

Part 2 - Data driven selection

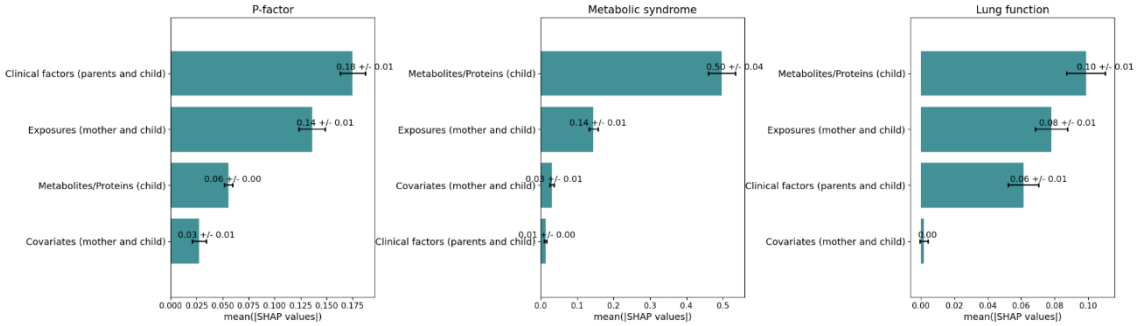
We additionally filtered remaining groups of very strongly correlated variables (r > 0.9) to reduce dimensionality without losing information. In total, from 476, we discarded 28 variables for 448 remaining variables. The rules for selecting variables among correlated groups were designed to retain features that are likely to be more informative and more universally applicable. Namely, 1. if correlation were between the same variable averaged on different time frames (e.g., day, week, year), keep the longest; 2. if correlation were between the same variables computed at home and at school, or other places, keep the variable computed at home. In any other cases, the default rule was simply to keep the first variable in the order they appear. Those rules are really simple but, as Pearson correlation r is > 0.9, we are dropping variables that mostly encode redundant information, and thus, impact on the performance is likely to be low.



Supplementary Figure 1. Description of all covariates.
Shows the distributions of variables used as covariates in the study.

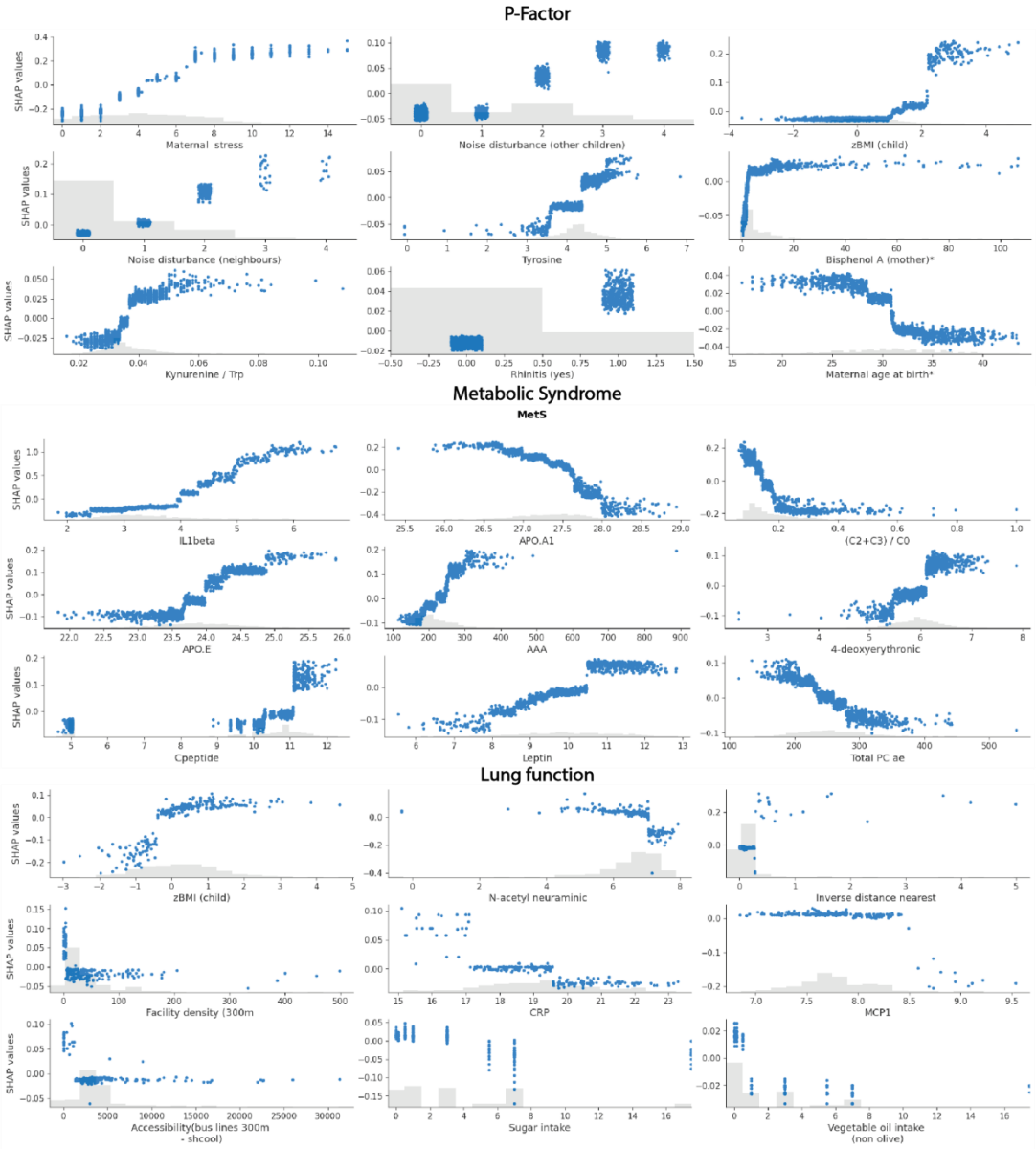


Supplementary Figure 2. Explained variance comparison.
First plot shows variance explained (R2 score) by original cohorts (first step of the modelling) for mental (P-Factor), cardiometabolic (MetS) and respiratory (lung function) risk scores. Second is variance explained by other variables after cohort adjustment (second step). Final plot shows total variance explained by modelling. The black interval bars represent the standard deviation across the ten models (n=10)

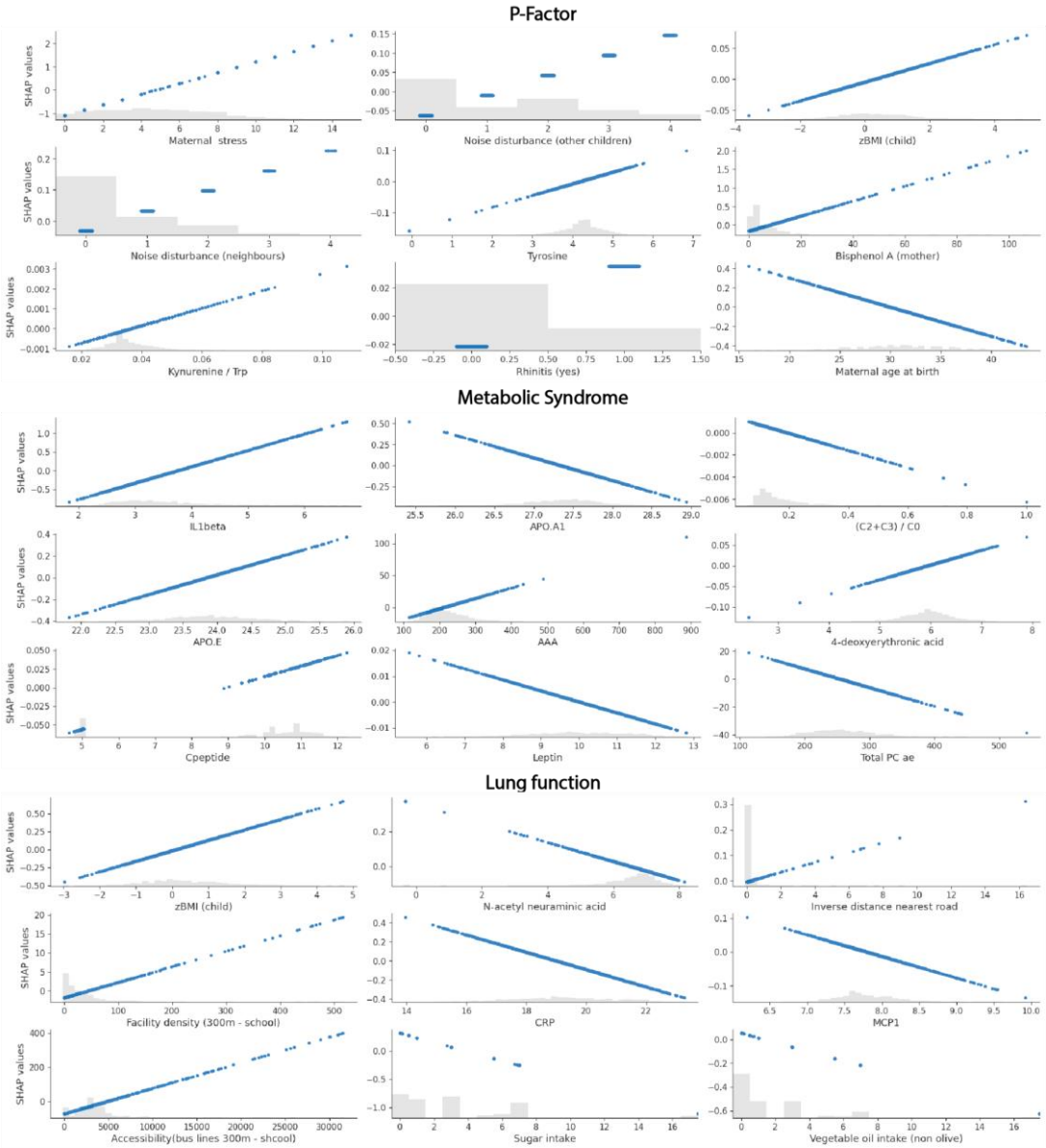


Supplementary Figure 3. Global feature importance across all Exposures, Metabolites/Proteins, Clinical Factors and Covariates.

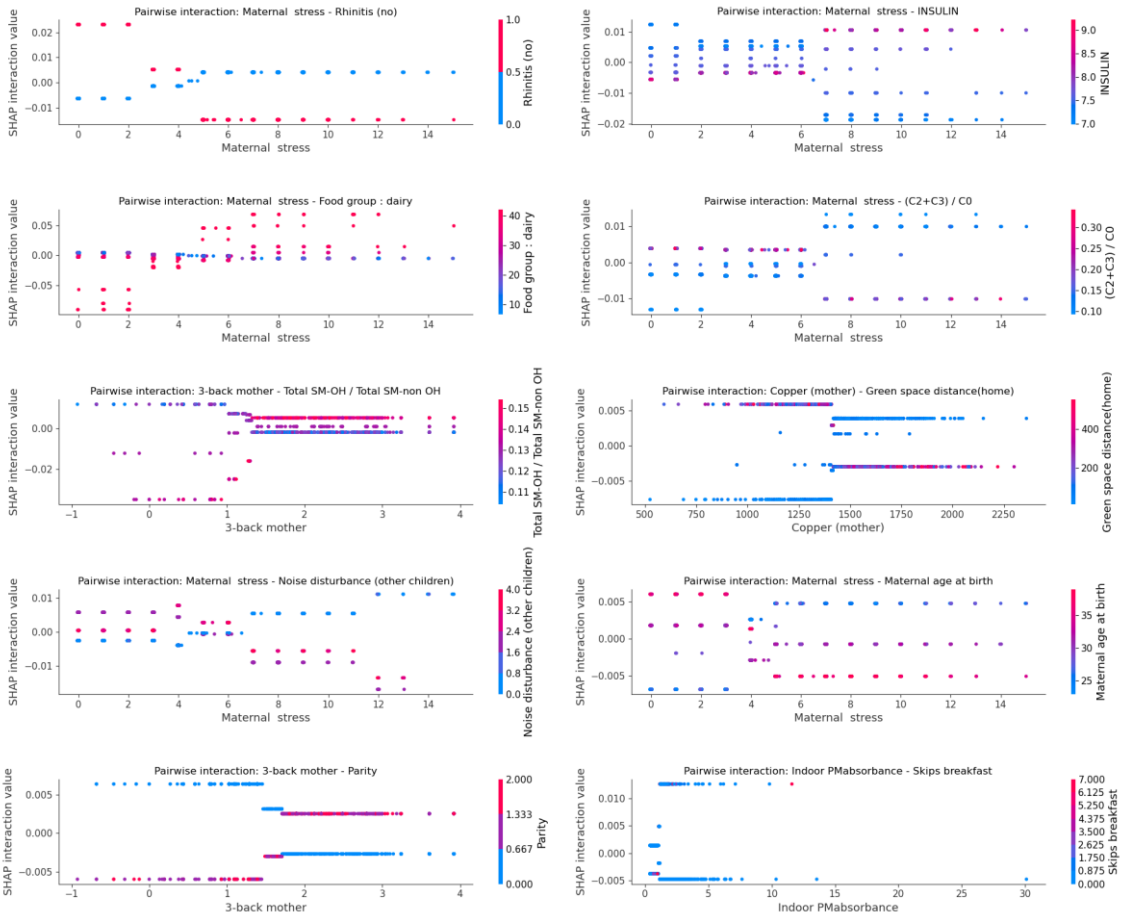
Shapley values were aggregated within each category, with the mean absolute value then computed for each group across all participants. The black interval bars represent the standard deviation across the ten models (n=10)



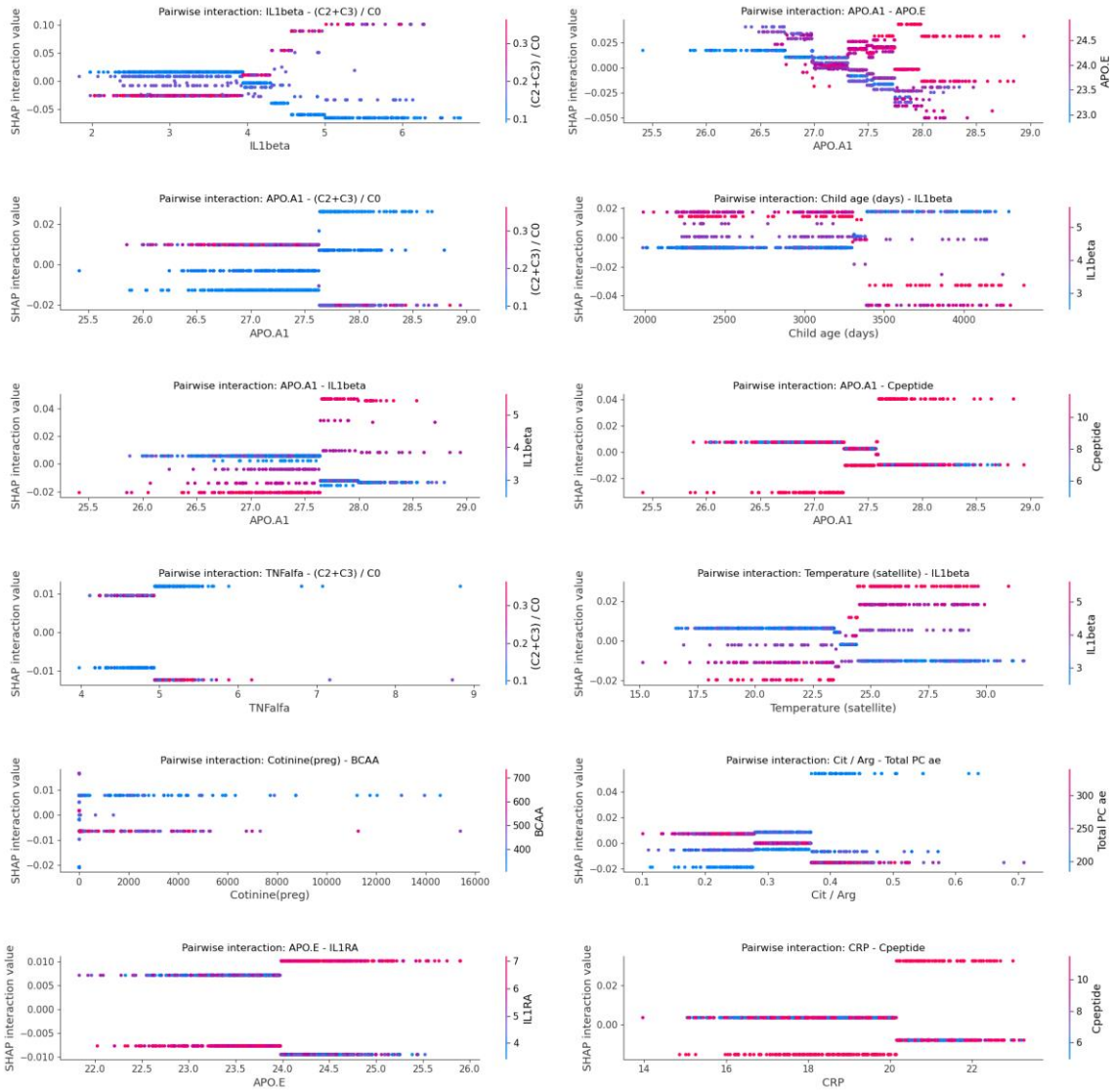
Supplementary Figure 4. SHAP dependence scatter plots (XGBoost).
Shows how the models respond to variations in the nine most impactful features for mental (P-Factor), cardiometabolic (MetS) and respiratory (lung function) risk scores. Grey bars show the features' distribution.



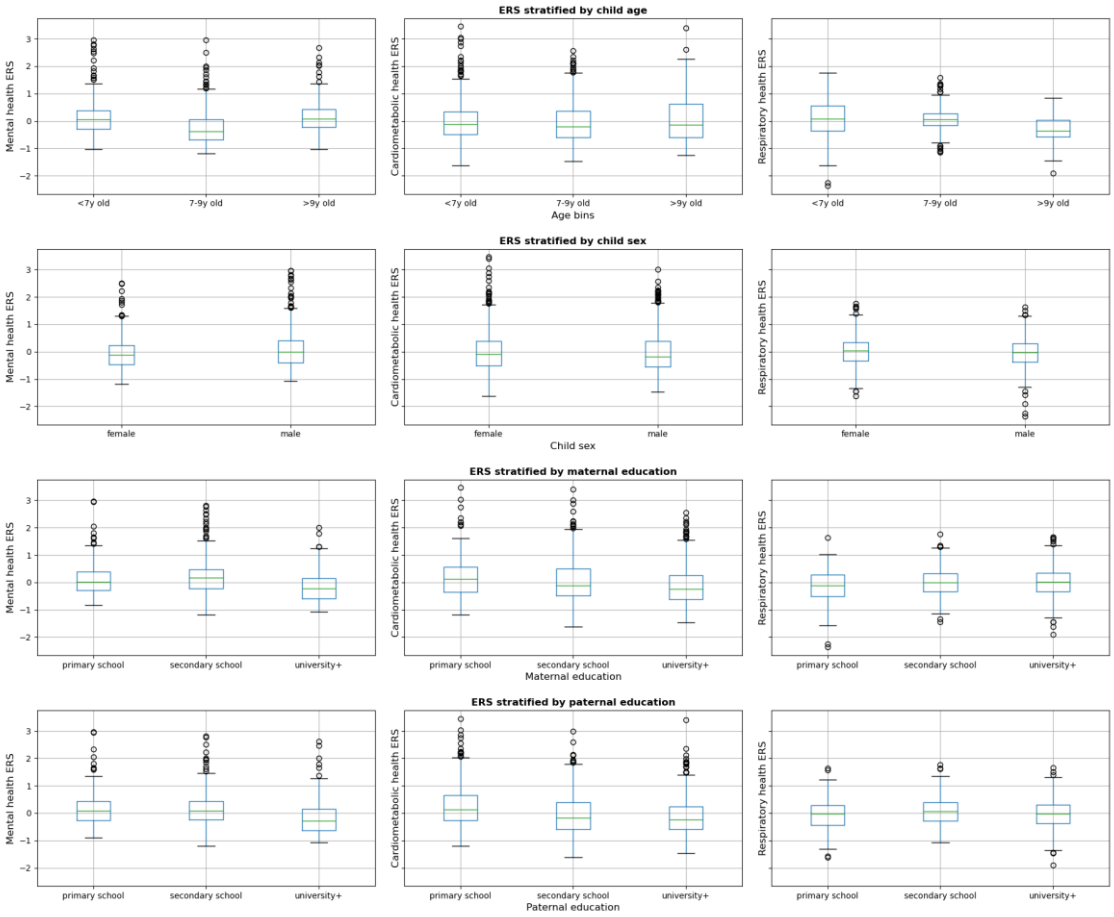
Supplementary Figure 5. SHAP dependence scatter plots (Lasso).
Shows how the lasso models respond to variations in the nine most impactful features for mental (P-Factor), cardiometabolic (MetS) and respiratory (lung function) risk scores. Grey bars show the features' distribution.



Supplementary Figure 6. SHAP interactions effects (P-Factor).
Shows the ten most impactful pairwise interaction effects derived from the mental (P-Factor) risk score (according to the mean absolute value of the Shapley values of all individuals for a given interaction) sorted by decreasing order.



Supplementary Figure 7. SHAP interactions effects (MetS).
Shows the ten most impactful pairwise interaction effects derived from the cardiometabolic (MetS) risk score (according to the mean absolute value of the Shapley values of all individuals for a given interaction) sorted by decreasing order.



Supplementary Figure 8. ECRS stratification with age, sex and parental education.

Supplementary Table 10. Data selection process

		Population selection	Variable selection
Step 0	available data	n=1622	p=478
Step 1	Selection of (one among) strongly correlated variables (r>0.9)	…	p=451
Step 2	Selection of individuals with sufficient non missing values (>50%)	n=1520	…
Step 3	Selection of features with sufficient non missing values (>40%)	…	p=448
Step 4	Selection of individuals with non missing outcomes.	p-factor = 1513 mets score = 1151 lung function = 1176	…

Abbreviations: n: number of individuals, p: number of variables, r: Pearson correlation coefficient.

Supplementary Table 11. Hyperparameters (step 1).

	P-Factor	MetS	Lung function
XGBoost			
learning_rate	0.1 [1e-2, 1e-1]	0.1 [1e-2, 1e-1]	0.1 [1e-2, 1e-1]
n_estimators	140 [50, 200]	80 [50, 200]	100 [50, 200]
max_depth	2 [0, 4]	3 [0, 4]	3 [0, 4]
objective	reg:squarederror	reg:squarederror	reg:squarederror
booster	gbtree	gbtree	gbtree
seed	0	42	42
Random Forest			
n_estimators	100 [50, 200]	100 [50, 200]	100 [50, 200]
min_sample_leaf	90 [0, 100]	70 [0, 100]	90 [0, 100]
max_leaf_nodes	8 [0, 10]	6 [0, 10]	8 [0, 10]
max_depth	5 [2, 8]	5 [2, 8]	5 [2, 8]
random_state	42	42	42
LASSO			
alpha	0	0	0

Supplementary Table 12. Hyperparameters (step 2).

	P-Factor	MetS	Lung function
XGBoost			
learning rate	0.024 [1e-3, 5e-1]	0.0842 [1e-3, 5e-1]	0.0395 [1e-3, 5e-1]
n_estimators	382 [50, 450]	317[50, 400]	318 [50, 400]
max_depth	2 [1, 10]	2 [1, 8]	1 [1, 8]
min_child_weight	1 [1, 100]	9 [1, 75]	6 [1, 100]
subsample	0.725 [0.5, 1]	0.948 [0.5, 1]	0.944 [0.5, 1]
colsample_bytree	0.839 [0.5, 1]	0.552 [0.5, 1]	0.915 [0.5, 1]
reg_alpha	0.3 [0, 10]	1.42 [0, 10]	0.2 [0, 10]
reg_lambda	7.577 [0, 10]	4.688 [0, 10]	0.595 [0, 10]
gamma	0 [0, 5]	0.3 [0, 5]	2.2 [0, 5]
objective	reg:squarederror	reg:squarederror	reg:squarederror
booster	gbtree	gbtree	gbtree
seed	42	42	42
Random Forest			
n_estimators	158 [50, 300]	232 [50, 300]	50 [50, 300]
min_samples_leaf	4 [0, 150]	6 [0, 150]	15 [0, 150]
max_leaf_nodes	61 [0, 10]	74 [0, 10]	17 [0, 10]
max_depth	13 [1, 12]	11 [1, 12]	8 [1, 12]
min_sample_split	20 [2, 50]	14 [2, 50]	48 [2, 50]
max_features	0.84 [0.5, 1]	0.72 [0.5, 1]	0.7 [0.5, 1]
random_state	42	42	42
LASSO			
alpha	0.03 [1e-2, 1]	0.02 [1e-2, 1]	0.05 [1e-2, 1]
random_state	42	42	42

Supplementary Table 13. Summary of residuals statistics obtained in the held out sets within the 10 fold cross-validation procedure

	Mean of residuals across all folds	**Number of normality distributed residuals over 10 folds** (Shapiro-Wilk test p-value > 0.05)
Mental ECRS		
Lasso	0.002	10
Random Forest	0.002	9
XGBoost	0.002	9
Cardiometabolic ECRS		
Lasso	-0.001	8
Random Forest	0.001	9
XGBoost	0.002	8
Respiratory ECRS		
Lasso	-0.001	5
Random Forest	0.002	5
XGBoost	0.002	5

Supplementary References

1. Léa Maitre, Jordi Julvez, Monica López-Vicente, Charline Warembourg, Ibon Tamayo-Uria, Claire Philippat, Kristine B. Gützkow, Monica Guxens, Sandra Andrusaityte, Xavier Basagaña, Maribel Casas, Montserrat de Castro, Leda Chatzi, Jorunn Evandt, Juan R. Gonzalez, Regina Gražulevičienė, Line Smastuen Haug, Barbara Heude, Carles Hernandez-Ferrer, Mariza Kampouri, Dan Manson, Sandra Marquez, Rosie McEachan, Mark Nieuwenhuijsen, Oliver Robinson, Remy Slama, Cathrine Thomsen, Jose Urquiza, Marina Vafeidi, John Wright, Martine Vrijheid, Early-life environmental exposure determinants of child behavior in Europe: A longitudinal, population-based study, Environment International, Volume 153, 2021, 106523, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2021.106523.

2. Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K, Briggs DJ. Mapping of background air pollution at a fine spatial scale across the European Union. *Sci Total Environ*. 2009;407(6):1852-1867. doi:10.1016/j.scitotenv.2008.11.048.

3. Cyrys J, Eeftens M, Heinrich J, et al. Variation of NO2 and NOx concentrations between and within 36 European study areas: Results from the ESCAPE study. *Atmos Environ*. 2012;62:374-390. doi:10.1016/j.atmosenv.2012.07.080.

4. Eeftens M, Beelen R, de Hoogh K, et al. Development of Land Use Regression Models for $PM_{2.5}$, $PM_{2.5}$ Absorbance, PM_{10} and PM_{coarse} in 20 European Study Areas; Results of the ESCAPE Project. *Environ Sci Technol*. 2012;46(20):11195-11205. doi:10.1021/es301948k.

5. Eeftens M, Tsai M-Y, Ampe C, et al. Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 – Results of the ESCAPE project. *Atmos Environ*. 2012;62(N/A):303-317. doi:10.1016/j.atmosenv.2012.08.038.

6. Beelen R, Hoek G, Vienneau D, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe - The ESCAPE project. *Atmos Environ*. 2013;72:10-23. doi:10.1016/j.atmosenv.2013.02.037.

7. Schembari A, de Hoogh K, Pedersen M, et al. Ambient Air Pollution and Newborn Size and Adiposity at Birth: Differences by Maternal Ethnicity (the Born in Bradford Study Cohort). *Environ Health Perspect*. 2015;123(11). doi:10.1289/ehp.1408675.

8. Wang M, Beelen R, Bellander T, et al. Performance of multi-city land use regression models for nitrogen dioxide and fine particles. *Environ Health Perspect*. 2014;122(8):843-849. doi:10.1289/ehp.1307271.

9. Rahmalia A, Giorgis-Allemand L, Lepeule J, et al. Pregnancy exposure to atmospheric pollutants and placental weight: An approach relying on a dispersion model. *Environ Int*. 2012;48:47-55. doi:10.1016/J.ENVINT.2012.06.013.

10. Nieuwenhuijsen MJ, Kruize H, Gidlow C, et al. Positive health effects of the natural outdoor environment in typical populations in different regions in Europe (PHENOTYPE): a study programme protocol. *BMJ Open*. 2014;4(4):e004951. doi:10.1136/bmjopen-2014-004951.

11. Herring JW and D. Measuring Vegetation (NDVI & EVI) : Feature Articles. August 2000.

12. Urban Atlas — European Environment Agency.

13. Smargiassi A, Goldberg MS, Plante C, Fournier M, Baudouin Y, Kosatsky T. Variation of daily warm season mortality as a function of micro-urban heat islands. *J Epidemiol Community Health*. 2009;63(8):659-664. doi:10.1136/jech.2008.078147.

14. Shannon CE, E. C. A mathematical theory of communication. *ACM SIGMOBILE Mob Comput Commun Rev*. 2001;5(1):3. doi:10.1145/584091.584093.

15. Duncan DT, Aldstadt J, Whalen J, Melly SJ, Gortmaker SL. Validation of Walk Score® for Estimating Neighborhood Walkability: An Analysis of Four US Metropolitan Areas. *Int J Environ Res Public Health*. 2011;8(12):4160-4179. doi:10.3390/ijerph8114160.

16. Frank LD, Sallis JF, Conway TL, Chapman JE, Saelens BE, Bachman W. Many Pathways from Land Use to Health: Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality. *J Am Plan Assoc*. 2006;72(1):75-87. doi:10.1080/01944360608976725.

17. Walk Score Terms of Use.

18. OpenStreetMap.

19. van Nunen E, Vermeulen R, Tsai M-Y, et al. Land Use Regression Models for Ultrafine Particles in Six European Areas. *Environ Sci Technol*. 2017;51(6):3336-3345. doi:10.1021/acs.est.6b05920.

20. EUR-Lex. EUR-Lex - 31992L0055 - EN - EUR-Lex.

21. Jeong CH, Wagner ED, Siebert VR, et al. Occurrence and Toxicity of Disinfection Byproducts in European Drinking Waters in Relation with the HIWATE Epidemiology Study. *Environ Sci Technol*. 2012;46(21):12120-12128. doi:10.1021/es3024226.

22. Smith RB, Edwards SC, Best N, Wright J, Nieuwenhuijsen MJ, Toledano MB. Birth Weight, Ethnicity, and Exposure to Trihalomethanes and Haloacetic Acids in Drinking Water during Pregnancy in the Born in Bradford Cohort. *Environ Health Perspect*. 2015;124(5). doi:10.1289/ehp.1409480.

23. Villanueva CM, Gracia-Lavedán E, Ibarluzea J, et al. Exposure to Trihalomethanes through Different Water Uses and Birth Weight, Small for Gestational Age, and Preterm Delivery in Spain. *Environ Health Perspect*. 2011;119(12):1824-1830. doi:10.1289/ehp.1002425.

24. Stayner LT, Pedersen M, Patelarou E, et al. Exposure to Brominated Trihalomethanes in Water During Pregnancy and Micronuclei Frequency in Maternal and Cord Blood Lymphocytes. *Environ Health Perspect*. 2013;122(1):100-106. doi:10.1289/ehp.1206434.

25. Danileviciute A, Grazuleviciene R, Vencloviene J, Paulauskas A, Nieuwenhuijsen M. Exposure to Drinking Water Trihalomethanes and Their Association with Low Birth Weight and Small for Gestational Age in Genetically Susceptible Women. *Int J Environ Res Public Health*. 2012;9(12):4470-4485. doi:10.3390/ijerph9124470.

26. Haug L, Sakhi A, Cequier E, et al. In-utero and early life chemical exposome in six European mother-child cohorts. In preparation.

27. Caspersen IH, Kvalem HE, Haugen M, et al. Determinants of plasma PCB, brominated flame retardants, and organochlorine pesticides in pregnant women and 3 year old

children in The Norwegian Mother and Child Cohort Study. *EnvironRes*. 2016;146(1096-0953 (Electronic)):136-144.

28. Goni F, Lopez R, Etxeandia A, Millan E, Amiano P. High throughput method for the determination of organochlorine pesticides and polychlorinated biphenyls in human serum. *JChromatogrB Anal Sci*. 2007;852(1570-0232 (Print)):15-21.

29. Koponen J, Rantakokko P, Airaksinen R, Kiviranta H. Determination of selected perfluorinated alkyl acids and persistent organic pollutants from a small volume human serum sample relevant for epidemiological studies. *JChromatogrA*. 2013;1309(1873-3778 (Electronic)):48-55.

30. Haug LS, Thomsen C, Becher G. A sensitive method for determination of a broad range of perfluorinated compounds in serum suitable for large-scale human biomonitoring. *JChromatogrA*. 2009;1216(0021-9673 (Print)):385-393.

31. Poothong S, Lundanes E, Thomsen C, Haug LS. High throughput online solid phase extraction-ultra high performance liquid chromatography-tandem mass spectrometry method for polyfluoroalkyl phosphate esters, perfluoroalkyl phosphonates, and other perfluoroalkyl substances in human serum, plasma, and w. *Anal Chim Acta*. 2017;957:10-19. doi:10.1016/j.aca.2016.12.043.

32. Manzano-Salgado CB, Casas M, Lopez-Espinosa MJ, et al. Transfer of perfluoroalkyl substances from mother to fetus in a Spanish birth cohort. *EnvironRes*. 2015;142(1096-0953 (Electronic)):471-478.

33. Poothong S, Thomsen C, Padilla-Sanchez JA, Papadopoulou E, Haug LS. Distribution of Novel and Well-Known Poly- and Perfluoroalkyl Substances (PFASs) in Human Serum, Plasma, and Whole Blood. *Environ Sci Technol*. 2017;51(22):13388-13396. doi:10.1021/acs.est.7b03299.

34. Rodushkin I, Axelsson MD. Application of double focusing sector field ICP-MS for multielemental characterization of human hair and nails. Part II. A study of the inhabitants of northern Sweden. *Sci Total Environ*. 2000;262(1-2):21-36.

35. Ramon R, Murcia M, Aguinagalde X, et al. Prenatal mercury exposure in a multicenter cohort study in Spain. 2011;37:597-604. doi:10.1016/J.ENVINT.2010.12.004.

36. Stern AH, Smith AE. An assessment of the cord blood:maternal blood methylmercury ratio: implications for risk assessment. *EnvironHealth Perspect*. 2003;111(0091-6765 (Print)):1465-1470.

37. Padilla MA, Elobeid M, Ruden DM, Allison DB. An examination of the association of selected toxic metals with total and central obesity indices: NHANES 99-02. *Int J Environ Res Public Health*. 2010;7(9):3332-3347. doi:10.3390/ijerph7093332.

38. Sabaredzovic A, Sakhi AK, Brantsæter AL, Thomsen C. Determination of 12 urinary phthalate metabolites in Norwegian pregnant women by core-shell high performance liquid chromatography with on-line solid-phase extraction, column switching and tandem mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2015;1002:343-352. doi:10.1016/j.jchromb.2015.08.040.

39. Valvi D, Monfort N, Ventura R, et al. Variability and predictors of urinary phthalate metabolites in Spanish pregnant women. *Int J Hyg Environ Health*. 2015;218(2):220-231. doi:10.1016/j.ijheh.2014.11.003.

40. Philippat C, Mortamais M, Chevrier C, et al. Exposure to Phthalates and Phenols during Pregnancy and Offspring Size at Birth. *EnvironHealth Perspect*. 2011;(1552-9924 (Electronic)).

41. Cequier E, Sakhi AK, Haug LS, Thomsen C. Development of an ion-pair liquid chromatography-high resolution mass spectrometry method for determination of organophosphate pesticide metabolites in large-scale biomonitoring studies. *J Chromatogr A*. 2016;1454:32-41. doi:10.1016/j.chroma.2016.05.067.

42. Aurrekoetxea JJ, Murcia M, Rebagliato M, et al. Determinants of self-reported smoking and misclassification during pregnancy, and analysis of optimal cut-off points for urinary cotinine: a cross-sectional study. *BMJ Open*. 2013;3(2044-6055 (Electronic)).

43. Sunyer J, Garcia-Esteban R, Castilla AM, et al. Exposure to second-hand smoke and reproductive outcomes depending on maternal asthma. *Eur Respir J*. 2012;40(2):371-376. doi:10.1183/09031936.00091411.

44. Covaci A, Voorspoels S, Thomsen C, van Bavel B, Neels H. Evaluation of total lipids using enzymatic methods for the normalization of persistent organic pollutant levels in serum. *SciTotal Environ*. 2006;366(0048-9697 (Print)):361-366.

45. Grimvall E, Rylander L, Nilsson-Ehle P, et al. Monitoring of polychlorinated biphenyls in human blood plasma: methodological developments and influence of age, lactation, and fish consumption. *Arch Environ Contam Toxicol*. 1997;32(3):329-336.

46. Sedentary Behaviour Research Networ SBR. Letter to the Editor: Standardized use of the terms “sedentary” and “sedentary behaviours.” *Appl Physiol Nutr Metab*. 2012;37(3):540-542. doi:10.1139/h2012-024.

47. Liu Y, Wang M, Villberg J, et al. Reliability and Validity of Family Affluence Scale (FAS II) among Adolescents in Beijing, China. *Child Indic Res*. 2012;5(2):235-251. doi:10.1007/s12187-011-9131-5.

48. Boyce W, Torsheim T, Currie C, Zambon A. The Family Affluence Scale as a Measure of National Wealth: Validation of an Adolescent Self-Report Measure. *Soc Indic Res*. 2006;78(3):473-487. doi:10.1007/s11205-005-1607-6.

49. Maitre, L. et al. Multi-omics signatures of the human early life exposome. Nat. Commun. 13, 7024 (2022).

A.2 Paper 3

An Informed Machine Learning based Environmental Risk Score for Hypertension in European Adults: results from the GCAT cohort

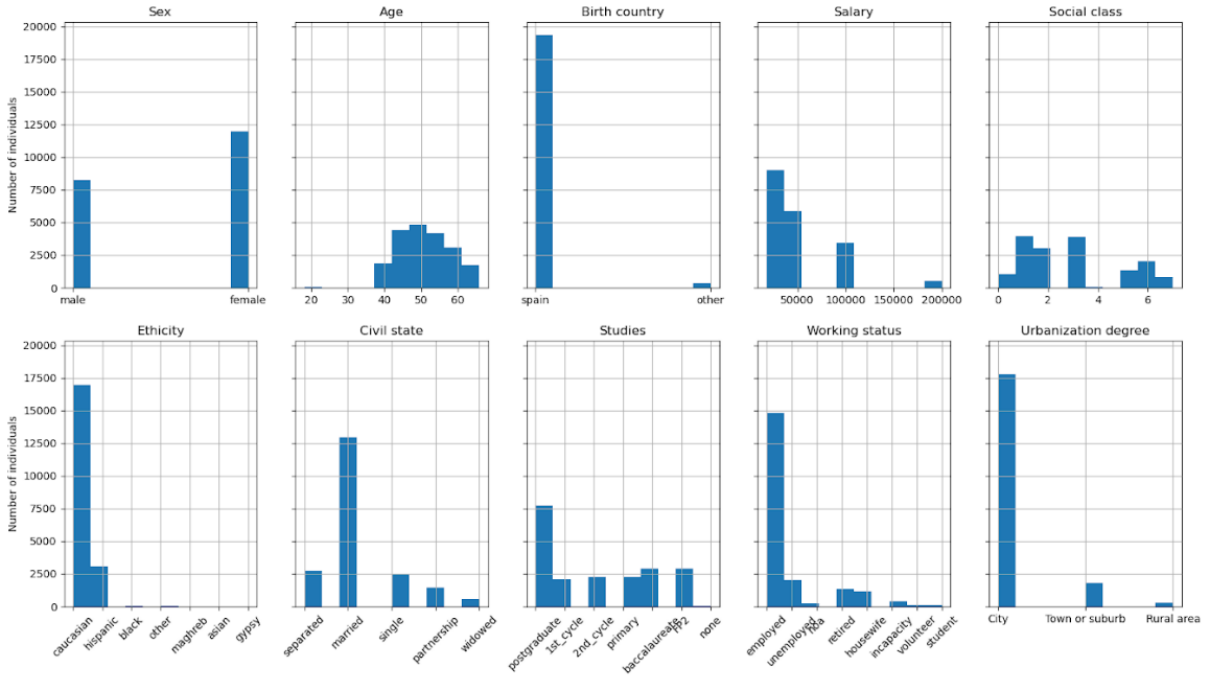
Jean-Baptiste Guimbaud, Rafael de Cid Ibeas, Emilie Calabre, Rémy Cazabet, Léa Maître

Name	Agnostic NN	SEANN
Hidden layers	3	3
Size layer 0	89	89
Size layer 1	8	8
Size layer 2	2	2
Learning rate	1.9e-3	1.5e-3

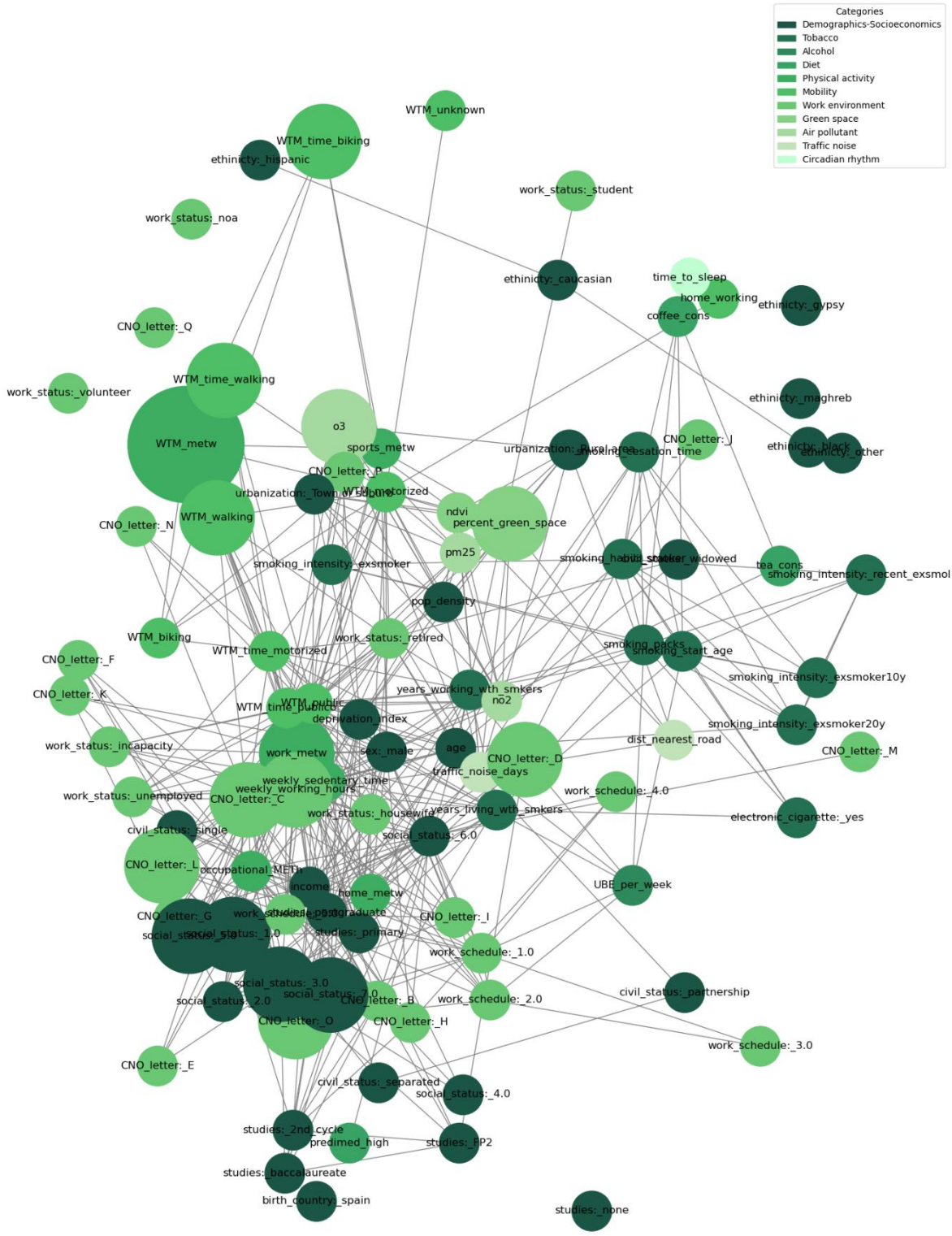
Supplementary table 1: Hyperparameter sets for SEANN and its agnostic counterpart. Both use the same architecture, only the training procedure is different.

Variable name	Delta SHAP Agnostic NN	Delta SHAP SEANN
Smoking habit: smoker	15.2×10^{-3}	1×10^{-3}
UBE per week	5×10^{-3}	1.4×10^{-3}
Predimed high	2.6×10^{-3}	1.4×10^{-3}
Weekly sedentary time	3.3×10^{-3}	0.4×10^{-3}
Sports (Metw)	10.1×10^{-3}	0.8×10^{-3}
NDVI	2.5×10^{-3}	1×10^{-3}
NO2	11.7×10^{-3}	1×10^{-3}
PM25	3.6×10^{-3}	0.3×10^{-3}
O3	2.3×10^{-3}	0.4×10^{-3}
Deprivation index	7.1×10^{-3}	0.6×10^{-3}
Traffic noise days	3.4×10^{-3}	0.6×10^{-3}
Average	6.1×10^{-3}	0.8×10^{-3}

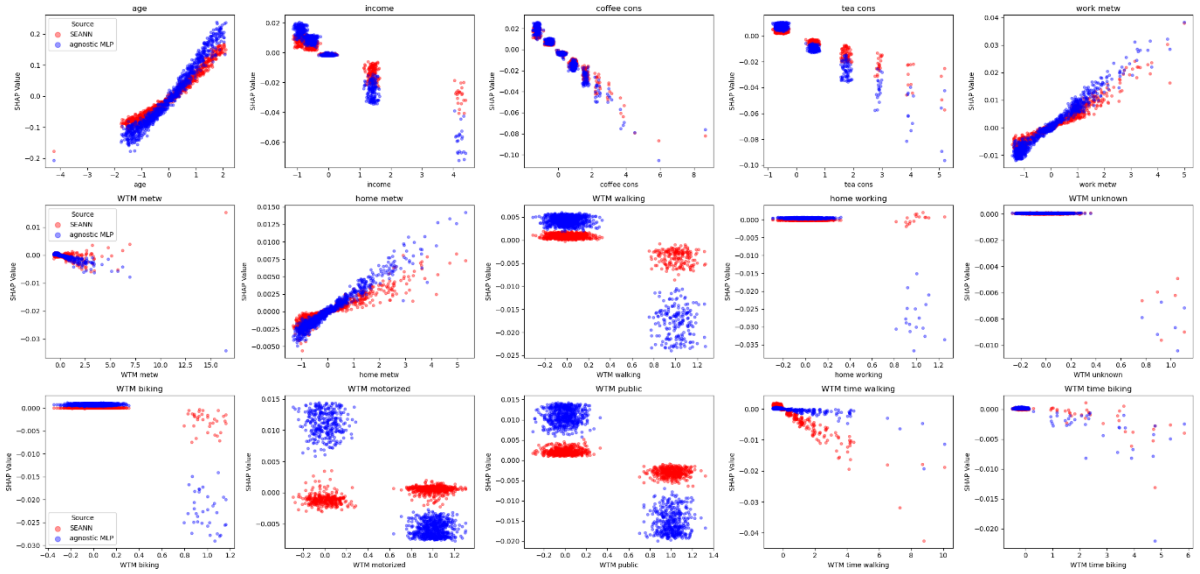
Supplementary table 2: Mean absolute error (i.e., delta SHAP) between ground truth Shapley values and those obtained with the agnostic NN and SEANN respectively.



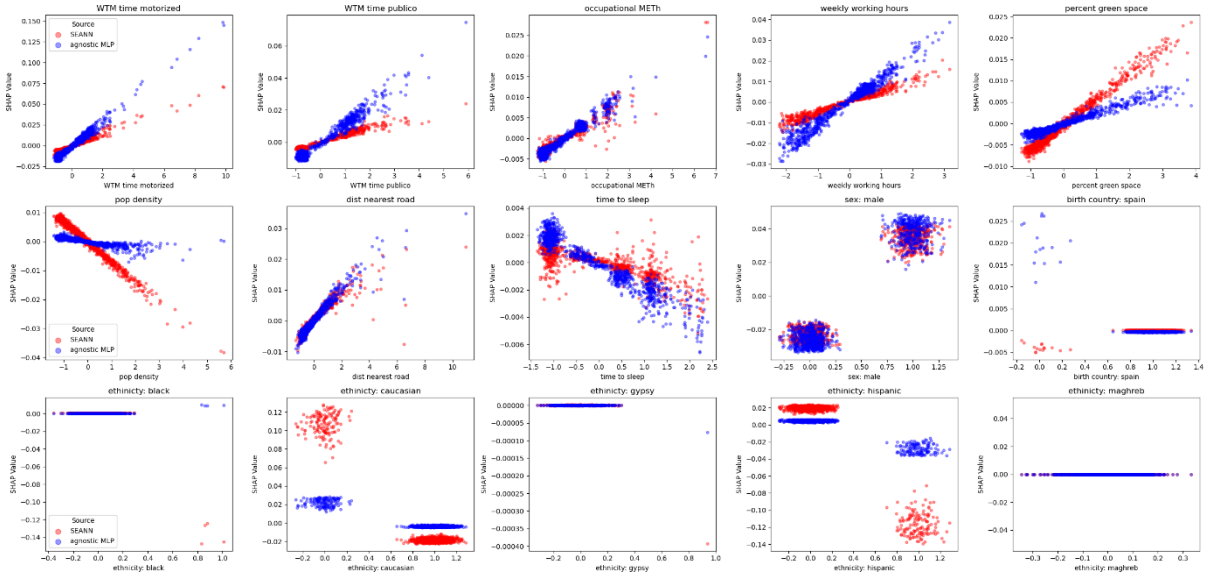
Supplementary Figure 1. *Distributions of the main covariates.*



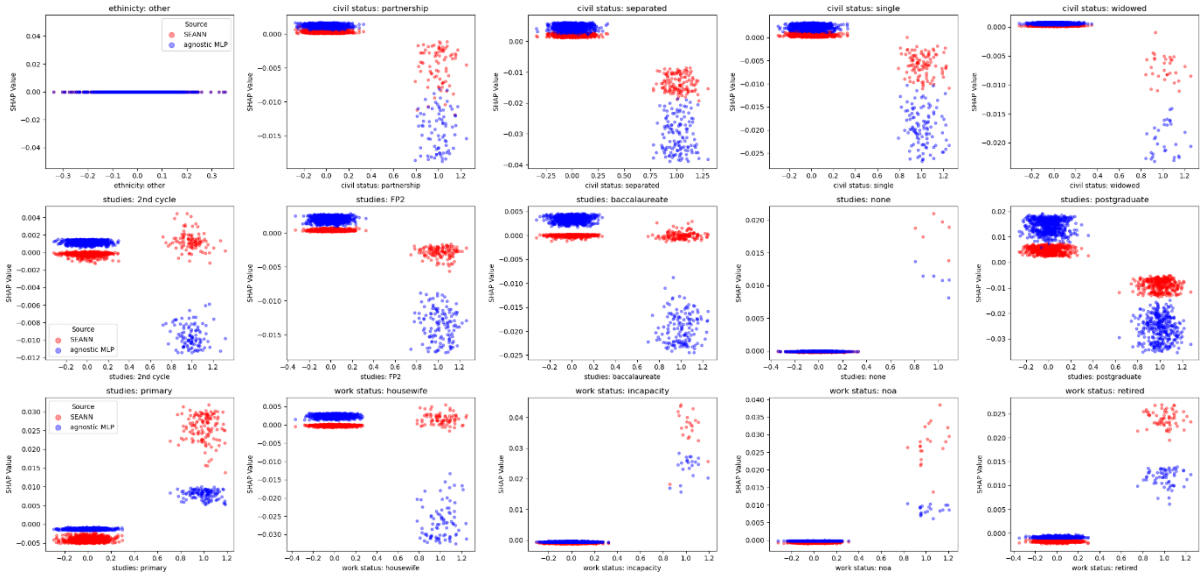
Supplementary Figure 2: Correlation graph of the exposome. *The size of the nodes is proportional to the number of correlations were >0.5 outside the exposure group and the length of the edges is proportional to the inverse of the correlation (the higher the correlation, the shorter the edge length) between exposures. The colour of the nodes represents the pre-defined exposure groups. The minimum absolute correlation to create an edge was 0.10.*



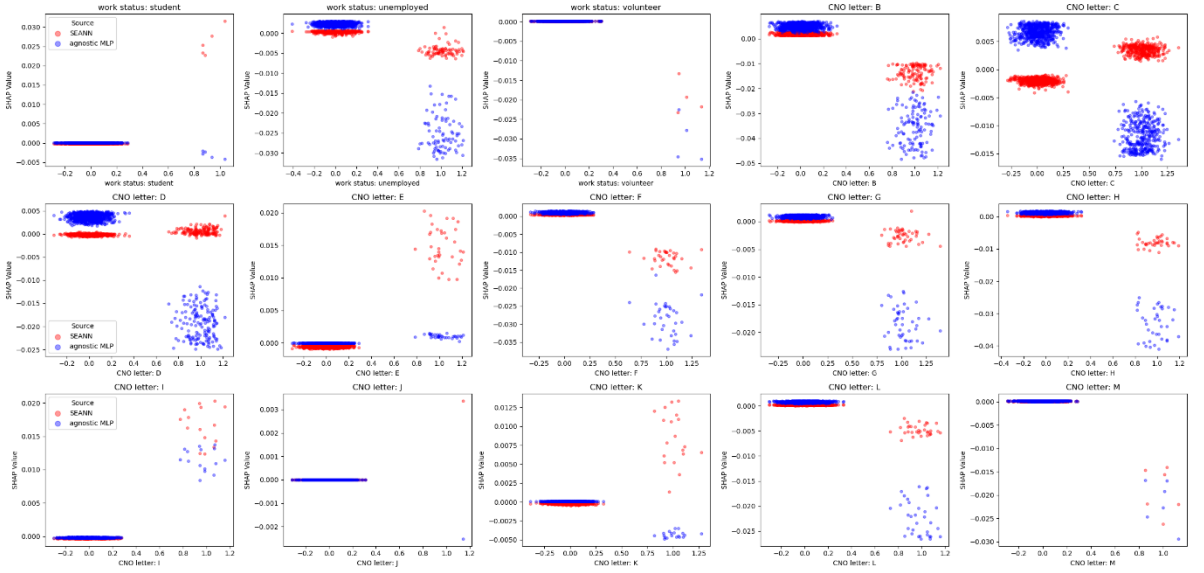
Supplementary Figure 3: Comparison of response functions extracted from a subset of the remaining variables (1-15). *Dots are Shapley values, encoding the model's response (y-axis) for a given exposure value (x-axis).*



Supplementary figure 4: Comparison of response functions extracted from a subset of the remaining variables (15-30). *Dots are Shapley values, encoding the model's response (y-axis) for a given exposure value (x-axis).*



Supplementary figure 5: Comparison of response functions extracted from a subset of the remaining variables (30-45). *Dots are Shapley values, encoding the model's response (y-axis) for a given exposure value (x-axis).*



Supplementary figure 6: Comparison of response functions extracted from a subset of the remaining variables (45-60). *Dots are Shapley values, encoding the model's response (y-axis) for a given exposure value (x-axis).*

B.1 About the author

Jean-Baptise Guimbaud graduated in Mathematics and Computer Science from the Claude Bernard University of Lyon (UCBL) (2018), where he also completed his Master of Computer Science and Artificial Intelligence (2020). He joined Meersens in 2020 as a data scientist and shortly after (2021) joined the LIRIS and ISGlobal laboratories as a PhD researcher under the supervision of Rémy Cazabet and Léa Maitre.

B.2 Research activities

A summary of the research activities performed during the three years of the author's PhD is provided below.

B.2.1 Other co-authored paper(s)

Léa Maitre, Jean-Baptiste Guimbaud, Charline Warembourg, Nuria Güil-Oumrait, Paula Marcela Petrone, Marc Chadeau-Hyam, Martine Vrijheid, Xavier Basagaña, Juan R. Gonzalez; State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. *Environment International*, **168**, pp.107422. 2022.

B.2.2 Oral presentations

The PhD Candidate has participated in international scientific conferences and meetings, presenting the work conducted during his PhD.

B.2.2.1 International conferences and events

- The Exposome Data Challenge Event, April 2021, Online. *Leveraging machine learning and explainable AI to better understand exposomic data*. Recordings of the presentations are available on YouTube.
- EHEN conference Leuven 2023, Belgium, *Exposome-omics risk scores for cardiometabolic, respiratory, and mental health in European children*. A presentation about Paper 1.
- ATHLETE 6th Consortium Meeting, 23-24th January 2024, Grenoble, France. *Machine Learning Based Health Environmental-Clinical Risk Scores in European Children*. Another presentation about Paper 1 as part of the ATHLETE project.

B.2.2.2 Internal presentations

- Meersens scientific board. Bi-annual meeting with meersens' business partners. Presentations about papers 1 and 2.
- Internal presentations within ISGlobal and LIRIS about papers 1 and 2.

B.2.3 Formation and training

- **Foundamentals of Epidemiology**. Organized by ISGlobal and the UPF, this class was an initiation to the field of epidemiology. The duration was 30 hours.
- **International Summer School on Advanced Methods in Global Health**. Organized by IsGlobal, this summer school was about statistical methods for assessing the exposome. The duration was 30 hours.
- **Science in action**. Organized by the UPF, this was a class about general research practices. The duration was 10 hours.
- **Seminars in biomedical research**. Compulsory activity organized by ISglobal and the UPF for a duration of 20 hours.
- **Lectures on scientific publications**. The duration was 3 hours.

B.2.4 Attended conferences and scientific meetings

- **Explain'ai** workshop, EGC 2023 (jan 17), Lyon, France.
- **H2020 ATHLETE** symposium, Jan 2022.
- **EHEN** scientific meetings 2022, 24-25 may, Barcelona, Spain.
- **Exposome hub** working group on advanced statistical methods for the exposome.

B.2.5 Reviews for peer-reviewed journals

- KDD2024, Barcelona, Spain (conference). 2 papers.
- Environment International (journal). 1 paper.