ID N. 17

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

UNIVERSITÀ DEGLI STUDI DI CATANIA

DEPARTMENT OF MEDICAL AND SURGICAL SCIENCES AND ADVANCED TECHNOLOGIES "GF INGRASSIA"

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Application of Artificial Intelligence models in Public Health for the study of the interaction between exposome, biological aging, and maternal-child health

Supervisors
Prof. Martina Barchitta
Prof. Antonella Agodi
Prof. Sebastiano Battiato
Prof. Concetto Spampinato

Candidate
Claudia La Mastra

2025

To my family and friends.

Acknowledgements

The research was conducted at the Department of Medical and Surgical Sciences and Advanced Technologies GF Ingrassia, University of Catania (Catania, Italy), in collaboration with the Obstetrics and Gynecology Unit, ARNAS Garibaldi-Nesima of Catania, and the Obstetrics and Gynecology Unit, Azienda Ospedaliero Universitaria Policlinico 'G. Rodolico - San Marco' (Catania, Italy). The research was carried out under the scientific direction of Professor Antonella Agodi, the principal investigator of both birth cohorts (the 'MAMI-MED' and 'Mamma & Bambino' cohorts) involved in the project, and the Director of the Department of Medical and Surgical Sciences and Advanced Technologies at the University of Catania.

Abstract

My PhD thesis focuses on exploring how maternal and environmental exposures influence maternal and child health, with particular attention to early biological aging and adverse outcomes. From the very beginning, an individual's exposure starts in utero, shaped by environmental factors and social determinants that interact with the individual's epigenetic mechanisms. Within the framework of precision medicine, this project aims to characterize the exposome during the peri-conceptional period and develop models to assess its interaction with early biological aging and health outcomes in mother-child pairs. The primary objective of this research is to integrate epidemiological data with advanced Artificial Intelligence (AI) techniques and insights from biomolecular and translational research. This integration seeks to evaluate environmental risk profiles, behavioral factors, and genetic susceptibility markers while also monitoring molecular markers such as those related to aging and epigenetics. Specifically, the research aims to understand how the exposome influences the early biological aging of the fetus, explore causal relationships between maternal exposures and adverse pregnancy or neonatal outcomes, and leverage AI to estimate the risk of these outcomes. Furthermore, it aims to assess the potential effectiveness of public health strategies and identify the most effective AI techniques for analyzing causal relationships in the context of population health. Epidemiological data and biological samples were collected from two Sicilian birth cohorts. Traditional linear and logistic regression models were used for the analysis, complemented by innovative causal machine learning models to explore the causal links between maternal exposures – such as Body Mass Index (BMI) and Gestational Weight Gain (GWG) – and adverse pregnancy outcomes.The results of this research will contribute to a deeper understanding of how the exposome impacts maternal and child health, particularly regarding early biological aging and adverse pregnancy outcomes. By identifying the key risk factors and causal mechanisms, this work will provide valuable insights for developing more effective public health interventions and strategies. Moreover, the integration of epidemiological, genetic, and epigenetic data with AI techniques represents a significant step forward in advancing precision medicine and optimizing maternal and child health outcomes.

Contents

List of Pubblications

This PhD thesis is based on the following original publications:

- Favara, G.; Maugeri, A.; Barchitta, M.; Magnano San Lio, R.; La Rosa, M.C.; La Mastra, C.; Galvani, F.; Pappalardo, E.; Ettore, C.; Ettore, G.; et al. Social and Nutritional Profiles of Pregnant Women: A Cluster Analysis on the "MAMI-MED" Cohort. Nutrients 2024, 16, 3975. https://doi.org/10.3390/nu16233975.

- Barchitta M, Magnano San Lio R, La Rosa MC, La Mastra C, Favara G, Ferrante G, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A, Maugeri A. The Effect of Maternal Dietary Patterns on Birth Weight for Gestational Age: Findings from the MAMI-MED Cohort. Nutrients. 2023 Apr 16;15(8):1922. doi: 10.3390/nu15081922. PMID: 37111140; PMCID: PMC10147093.

- Barchitta M, Maugeri A, La Mastra C, Favara G, La Rosa MC, Magnano San Lio R, Gholizade Atani Y, Gallo G, Agodi A. Pre-pregnancy BMI, gestational weight gain, and telomere length in amniotic fluid: a causal graph analysis. Sci Rep. 2024 Oct 8;14(1):23396. doi: 10.1038/s41598-024-74765-y. PMID: 39379607.

- Magnano San Lio R, Barchitta M, Maugeri A, Campisi E, Favara G, Ojeda Granados C, La Mastra C, La Rosa MC, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Sex differences in delivery and neonatal characteristics of new-borns from the "MAMI MED" cohort. Submitted to the Journal of Personalized Medicine in December 2023.

- Application of artificial intelligence to study the causality in public health: a systematic review. Manuscript in preparation.

The publications have been adapted with the permission of the copyright owners.

Abbreviations

- Adjusted Rand Index (ARI)
- Alcohol Use Disorder (AUD)
- Amyotrophic Lateral Sclerosis (ALS)
- Appropriate for Gestational Age (AGA)
- Area Under the Curve (AUC)
- Artificial Intelligence (AI)
- Artificial Neural Networks (ANN)
- Atrial Fibrillation (AF)
- Attention-Deficit/Hyperactivity Disorder (ADHD)
- Autism Spectrum Disorder (ASD)
- Average Rate of Change (ARCs)
- Balancing Covariates Automatically Using Supervision (BCAUS)
- Bayesian Additive Regression Trees (BART)
- Bayesian Belief Network (BBN)
- Bayesian Constraint-based Causal Discovery (BCCD)
- Bayesian Networks (BNs)
- Bias Loss (LBIAS)
- Binary Cross-Entropy Loss (LBCE)
- Body Mass Index (BMI)
- cardiovascular diseases (CVDs)
- Cardiovascular Lost Years (CVLY)
- Cardiovascular Risk Scores (CVRS)

- Causal Analysis Using Structural and Conditional Associations for Detecting Effects (CASCADE)
- Causal Forest (CF)
- Causal Tree (CT)
- Chest Radiography (CXR)
- Chi-square Test (χ^2 test)
- Chronic Obstructive Pulmonary Disease (COPD)
- classification and regression trees (CART)
- Cluster Features Tree (CF Tree)
- Clustering on Principal Components (CPC)
- Confidence Intervals (CIs)
- Coronary Artery Calcium (CAC)
- Cross-Validated Targeted Maximum Likelihood Estimation (CV-TMLE)
- Decision Tree (DT)
- Directed Acyclic Graphs (DAG)
- Double Machine Learning (DML)
- Drinking To Cope (DTC)
- Dynamic Bayesian Network (DBN)
- Dynamic Uncertainty Causal Graphs (DUCG)
- Electronic Health Records (EHRs)
- Fast Greedy Equivalence Search (FGES)
- Food Frequency Questionnaire (FFQ)
- Gestational diabetes mellitus (GDM)
- Gestational Weight Gain (GWG)

- Good Clinical Practice (GCP)
- Granger causality (GC)
- Greedy Fast Causal Inference (GFCI)
- Hepatitis B Virus (HBV)
- Hepatitis C Virus (HCV)
- Hepatocellular Carcinoma (HCC)
- Heterogeneous Treatment Effects (HTEs)
- Individual Case Safety Reports (ICSRs)
- Individual Treatment Effect (ITE)
- Institute of Medicine (IOM)
- Instrumental Variable Causal Forest Algorithm (IV-CFA)
- Interquartile Ranges (IQR)
- K Nearest Neighbors (KNN)
- Kaiser-Meyer-Olkin (KMO)
- Large for Gestational Age (LGA)
- Linear Non-Gaussian Acyclic Model (LiNGAM)
- Local Causal Discovery (LCD2)
- Low Atherosclerotic CVD (ASCVD)
- Low Birth Weight (LBW)
- Low-Dose Computed Tomography (LDCT)
- Machine Learning (ML)
- Major Adverse Cardiovascular Events (MACE)
- Mediterranean Diet (MD)
- Mediterranean Diet Score (MDS)

- Metabolic Syndrome (MetS)
- Multi-Ethnic Study of Atherosclerosis (MESA)
- Multiple Sclerosis (MS)
- Non-Alcoholic Fatty Liver Disease (NAFLD)
- Non-vitamin K Antagonist Oral Anticoagulants (NOACs)
- Obsessive-Compulsive Disorder (OCD)
- Odds Ratios (OR)
- Postoperative Nausea and Vomiting (PONV)
- Preferred Reporting Items for a Systematic Review and Meta-analysis (PRISMA)
- Preterm birth (PTB)
- Principal Component Analysis (PCA)
- Principal Components (PCs)
- Quantitative Polymerase Chain Reaction (qPCR)
- Random Forests (RF)
- Randomized Clinical Trials (RCTs)
- Schwarz's Bayesian Information Criterion (SBIC)
- Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)
- Shapley Values (SV)
- Small for Gestational Age (SGA)
- Socioeconomic status (SES)
- Statistical Package for Social Sciences (SPSS)
- Support Vector Machines (SVM)
- Targeted Bidirectional EHR Transformer (T-BEHRT)
- Targeted Maximum Likelihood Estimation (TMLE)

- Telomere (T)
- Telomere Length (TL)
- Total Events Avoided (TEA)
- Urban and Rural Resident Basic Medical Insurance (URRBMI)
- Various Cardiovascular Risk Factors (CVRF)
- World Health Organization (WHO)

Chapter 1

Introduction

1.1 Exposome and Maternal and Child Health

In recent decades, precision medicine has gained increasing attention, thanks to the rise of "omic" sciences, which have enabled the development of personalized approaches to disease prevention and treatment [1]. These approaches consider individual genetic and epigenetic variability, providing a more comprehensive and targeted view of health conditions. The core idea of precision medicine is to tailor healthcare strategies to the specific characteristics of each individual, optimizing the prevention and treatment of diseases by considering not only genetic factors but also environmental and behavioural influences [2]. In this context, the concept of the exposome has emerged as a powerful tool to better understand the interaction between the environment, genetics, and epigenetics in health and disease. The exposome is defined as the totality of all environmental exposures an individual experiences throughout life, including factors such as air pollution, diet, lifestyle, occupational exposures, and social determinants [3]. These exposures can influence health through complex mechanisms that interact with an individual's genetic and epigenetic makeup, ultimately determining the risk of developing chronic or acute diseases [4]. The exposome begins to develop early in life, starting in utero, where maternal environment and exposures to chemicals, nutrients, stress, and other factors influence fetal health. Understanding how the exposome impacts maternal and child health is crucial for developing strategies for prevention and intervention to improve health outcomes [5]. Scientific evidence suggests that exposures during pregnancy can have long-lasting effects, extending through childhood and even into subsequent generations, a phenomenon known as "fetal programming" [6]. Thus, genetics is not the only determinant of health; it interacts with the environment in ways that are not yet fully understood. The peri-conceptional period, which includes pregnancy and the immediate

pre- and post-conception periods, is particularly critical for maternal and child health. During this period, environmental exposures, maternal lifestyle choices such as smoking, diet, and alcohol consumption, can influence pregnancy outcomes, birth weight, and predispose neonates to chronic diseases like diabetes and cardiovascular diseases, which can manifest even in adulthood [7]. These effects may be mediated through epigenetic changes that alter gene expression without modifying the DNA sequence but influencing long-term health. The concept of the exposome innovatively integrates these factors, proposing a complex model that considers not only environmental and behavioural exposures but also the body's ability to respond to these stimuli through epigenetic and genetic modifications. Recent studies have highlighted how epigenetic changes, such as DNA methylation, play a crucial role in determining the response to environmental exposures, including those occurring during pregnancy [8]. For instance, research has shown how exposure to air pollutants and industrial chemicals during pregnancy can alter the epigenetic profile of newborns, increasing the risk of respiratory and cardiovascular diseases in both infancy and adulthood [9]. The interactions between the exposome and early biological aging are also of great relevance in the context of maternal and child health. Biological aging refers to the deterioration of physiological functions at the cellular and tissue level, which does not always correspond to chronological age. While aging is a natural process, environmental factors such as pollution, stress, diet, and physical activity can accelerate it. In particular, exposure to toxic substances during pregnancy may affect the early aging process in both the mother and the child through DNA damage and telomere shortening, which is an indicator of cellular aging [10]. In this context, the exposome offers a unique opportunity to identify and understand the factors contributing to early aging and to design more targeted prevention interventions.

1.2 Molecular Biomarkers of Cellular Aging

Cellular aging, an inevitable process affecting all living organisms, plays a fundamental role in the progression of various age-related diseases [11]. In recent decades, advances in molecular biology have led to the identification of molecular biomarkers of aging that help understand the underlying mechanisms of cellular senescence, tissue degeneration, and the onset of chronic diseases. These biomarkers provide crucial information on how environmental factors, genetic predisposition, and lifestyle choices influence the aging process [12]. The application of Artificial Intelligence (AI) models in public health can improve the understanding of how these biomarkers interact with the exposome (the totality of environmental exposures) and influence maternal and child health outcomes [13]. Epigenetics refers to the study of heritable changes in gene expression or cellular phenotype that do not involve alterations

to the underlying DNA sequence. DNA methylation, histone modification, and non-coding RNA molecules are key components of the epigenetic landscape that influence aging [14]. With age, the epigenome undergoes modifications that alter gene expression patterns, often leading to a decline in the body's ability to repair tissues and maintain cellular function [15]. For example, DNA methylation patterns change with age, with specific regions of the genome becoming predictably hypermethylated or hypomethylated. These changes can serve as epigenetic clocks to predict biological age, independent of chronological age. The "epigenetic clock," developed by Horvath [16], uses DNA methylation profiles to estimate an individual's biological age and has been widely applied in aging research. This approach holds great potential in public health applications, particularly in understanding how environmental factors and lifestyle choices can accelerate or slow the aging process. One of the most studied biomarkers of cellular aging is telomere length (TL). Telomeres are repetitive DNA sequences located at the ends of chromosomes that protect genetic material from degradation during cell division [17]. However, with each cell division, telomeres progressively shorten, a process that is accelerated by oxidative stress, inflammation, and lifestyle factors such as smoking and an unhealthy diet. This shortening limits the cell's ability to divide and ultimately leads to cellular senescence or apoptosis, which are central processes in aging [18]. Recent studies have shown that telomere shortening is associated not only with aging but also with various age-related diseases, including cardiovascular diseases, diabetes, and neurodegenerative disorders [19, 20]. Furthermore, in the context of maternal and child health, the effect of maternal exposures (e.g., stress, smoking, and nutrition) on telomere length has been widely studied, with implications for fetal development and long-term health outcomes for children [21]. AI models can assist in analyzing large datasets on telomere length and its association with health outcomes, improving risk prediction and early intervention strategies in Public Health [22].

1.2.1 Early Biological Aging: The Role of Telomeres as Biomarkers

Telomeres are repetitive sequences of TTAGGG located at the ends of chromosomes, forming DNA-protein structures that play a critical role in maintaining genomic stability by protecting chromosomes from degradation and fusion [23]. Over the past few decades, TL has emerged as a potential biomarker for biological aging, as it gradually shortens in somatic cells throughout an individual's lifespan. This shortening occurs due to the lack of telomerase activity, an enzyme that normally replenishes telomeres [24]. Although telomeres primarily serve as protective caps, preventing genomic instability, their progressive shorten-

ing is increasingly seen as both a consequence and a cause of biological aging. Specifically, telomere attrition is considered indicative of cumulative cellular damage over time, which ultimately influences the lifespan and functionality of cells. Telomere length has been shown to vary significantly between individuals, even those of the same chronological age, highlighting the influence of both genetic and environmental factors. While genetic heritability plays a role in determining TL, accounting for approximately 30-80% of the variation, only a small fraction of this variance is attributed to known telomere maintenance genes [25]. Recent research has revealed that DNA methylation at over 800 CpG sites is associated with TL, linking these genetic variations to biological processes like circadian rhythm, coagulation, and wound healing [26]. These findings suggest that environmental exposures, behaviors, and diseases may also influence telomere dynamics, contributing to the observed variation in TL across populations. Telomere shortening is considered a marker of age-related diseases, including cancer and cardiovascular disorders, due to its presumed role in accelerating biological aging. As such, TL has garnered attention as a potential early biomarker for these conditions. However, there is growing consensus that lifestyle factors—such as physical inactivity, smoking, and poor diet—may impact TL even before the onset of these diseases. In this context, alcohol consumption has been identified as a significant risk factor that could influence TL, with studies linking alcohol abuse to a range of health issues. According to the World Health Organization (WHO), excessive alcohol consumption is responsible for millions of deaths and disability-adjusted life years globally [27]. Consequently, several studies have investigated the association between TL and alcohol consumption, often considering alcohol as a covariate or mediator in the relationship between TL and age-related diseases [28, 29, 30]. More recently, however, some research has begun to focus on the direct impact of alcohol consumption on TL, revealing mixed results. These inconsistencies are often attributed to differences in study design, population characteristics, and the methods used to measure TL. Despite these challenges, the emerging body of evidence points to the potential role of alcohol-induced telomere shortening in various health conditions, including those related to mental health, such as depression and anxiety disorders [31]. Interestingly, the impact of alcohol consumption on TL during pregnancy remains an underexplored area of research. Previous studies have demonstrated that adverse exposures during pregnancy, such as maternal stress and smoking, can lead to telomere shortening in both cord blood and placenta samples [32, 33, 34, 35, 36]. Moreover, conditions like gestational diabetes and obesity are associated with telomere attrition in pregnant women, which increases the risk of cardiovascular diseases for both mothers and their offspring [37, 38]. While the molecular mechanisms underlying the effects of alcohol consumption during pregnancy are not yet fully understood, telomere shortening is a plausible factor contributing to the observed

adverse outcomes. Furthermore, the potential impact of maternal alcohol use on the TL of newborns remains largely unexplored, despite evidence that alcohol consumption during pregnancy, particularly in the first trimester, can have severe consequences for both maternal and child health, including an increased risk of miscarriage, premature birth, and low birth weight [39, 40]. Given the potential implications for maternal and child health, further research is urgently needed to understand how alcohol consumption during pregnancy affects telomere length and contributes to long-term health risks.

1.2.2 Artificial Intelligence in the Study of Aging Biomarkers

The application of AI and machine learning techniques has dramatically transformed the way aging biomarkers are studied and applied in public health [41]. By integrating large datasets, AI models can identify complex patterns and interactions between genetic, environmental, and lifestyle factors that contribute to aging [42]. Traditional methods, often limited by their inability to process and analyze vast amounts of data, may overlook crucial biomarkers and metabolic pathways. However, AI techniques, such as random forests, support vector machines, and deep neural networks, offer new opportunities to identify these markers with a higher degree of accuracy [43]. A key example of AI's transformative role in aging research is the analysis of telomere length data, epigenetic clocks, and inflammatory biomarkers. Telomeres, repetitive DNA sequences at the ends of chromosomes, shorten with each cell division, and their length serves as a significant indicator of cellular aging [44]. AI can enhance the analysis of telomere dynamics by processing large-scale data to uncover previously hidden associations between telomere length and environmental exposures, lifestyle factors, and age-related diseases [45]. For instance, studies have demonstrated that oxidative stress and environmental pollutants accelerate telomere shortening, leading to cellular senescence and age-related pathologies [46]. AI algorithms, when applied to these data, enable a more nuanced understanding of how such exposures influence aging at the molecular level, paving the way for improved diagnostic tools and preventive strategies. Moreover, AI's potential is evident in the integration of epigenetic clocks into aging research. The concept of the "epigenetic clock" has gained significant attention as a tool for predicting biological age, independent of chronological age. This approach relies on DNA methylation patterns, which are known to change with age [47]. Recent studies have shown that AI-based models can use these methylation patterns to provide even more precise estimates of biological age and link these estimates to health outcomes [48]. For example, AI algorithms can analyze epigenetic data to identify specific methylation sites that are predictive of age-related diseases, such as cardiovascular disease, diabetes, and neurodegenerative disorders [49]. By integrating these

biomarkers into a unified model, AI can improve risk prediction and provide insights into the molecular mechanisms of aging. In addition to these molecular biomarkers, inflammatory biomarkers have also become crucial in understanding aging and its associated diseases. Chronic inflammation is widely recognized as a key driver of age-related diseases, including cardiovascular disease, Alzheimer's disease, and type 2 diabetes [50]. AI models, particularly those based on machine learning, can analyze complex data sets of inflammatory markers (e.g., C-reactive protein, interleukins) to identify novel patterns that may predict the onset of these diseases. For example, a recent study applied machine learning algorithms to identify novel inflammation-related biomarkers linked to aging and found that certain inflammatory markers were predictive of both frailty and early mortality [51]. AI's integration with molecular biomarkers can lead to the development of predictive models that assess individual risks for age-related diseases. Such models can facilitate personalized public health interventions, providing early detection of at-risk individuals and allowing for tailored preventative strategies [52]. For instance, by integrating AI with telomere length data, epigenetic clocks, and inflammatory biomarkers, researchers can generate multi-dimensional profiles of biological aging that are specific to individual patients, helping clinicians to design personalized interventions [53]. Moreover, AI can assist in identifying early biomarkers of aging in maternal and child health, offering the potential for early-stage interventions that could reduce the impact of age-related diseases across the lifespan. The influence of prenatal factors on biological aging is a critical area of research. Studies have shown that maternal stress, smoking, and nutritional deficiencies can influence fetal development and may lead to long-term health effects in children, including premature aging. AI models that incorporate maternal and fetal data could offer powerful tools for predicting the long-term health outcomes of children and help identify strategies for minimizing the impact of these exposures [54, 55].

1.3 Social and Environmental Determinants of Inequalities in Neonatal Health

Neonatal health outcomes are influenced by a complex interplay of social, environmental, and biological factors. Socioeconomic disparities, maternal lifestyle, and environmental exposures contribute significantly to perinatal health inequalities [56]. In addition to these determinants, sex-related differences in neonatal outcomes are a well-documented phenomenon in perinatal epidemiology. Understanding these disparities can provide valuable insights into the biological and environmental mechanisms influencing perinatal health. Evidence suggests that male and female neonates may exhibit different vulnerabilities to perinatal complica-

tions due to hormonal, genetic, and physiological factors. Investigating these disparities, alongside social and environmental determinants, is essential to developing targeted interventions aimed at reducing neonatal health inequalities. Socioeconomic status (SES) is one of the most influential social determinants of neonatal health, with robust literature demonstrating its impact on a wide range of neonatal health outcomes, including birth weight, preterm birth, and neonatal mortality. Low SES is often associated with an increased risk of exposure to factors such as inadequate maternal nutrition, insufficient prenatal care, and higher stress levels, all of which can have negative effects on neonatal health [57]. Research has shown that infants born to mothers with low SES are more likely to experience adverse health outcomes, including higher rates of preterm birth and low birth weight, which are significant predictors of neonatal morbidity and mortality [58]. Environmental factors also play a crucial role in neonatal health inequalities. Exposure to environmental pollutants, such as air pollution, hazardous chemicals, and secondhand smoke, has been linked to adverse pregnancy outcomes and neonatal health issues [59]. For example, maternal exposure to air pollution has been associated with Preterm birth (PTB), Low Birth Weight (LBW), and respiratory problems in newborns [60]. The interaction between environmental exposures and other social determinants, such as housing quality and neighborhood conditions, can exacerbate these risks. For instance, living in areas with high levels of pollution and poor housing conditions can increase maternal exposure to environmental toxins, which in turn can compromise pregnancy and neonatal outcomes. Maternal nutrition and lifestyle factors are also significant contributors to inequalities in neonatal health. Inadequate maternal nutrition, including deficiencies in essential nutrients like folic acid and iron, can increase the risk of adverse pregnancy outcomes, such as preterm birth and low birth weight [61]. Lifestyle factors, such as smoking, alcohol consumption, and physical inactivity, are equally crucial determinants of neonatal health. For example, smoking during pregnancy is associated with an increased risk of PTB, LBW and neonatal respiratory issues [62].

1.3.1 The Use of Artificial Intelligence in Studying the Social and Environmental Determinants of Neonatal Health

In recent years, advances in AI and machine learning have provided new tools to examine the social and environmental determinants that influence neonatal health. By processing large amounts of data and analyzing complex patterns, AI enables the identification of risk factors that may be overlooked by traditional methods, providing a deeper understanding of the causes of health inequalities [63]. A recent study used advanced machine learning techniques to analyze the relationship between SES and neonatal health outcomes, highlighting how AI

models can identify patterns and risk factors previously unknown. By integrating data on maternal education, income, and access to healthcare, these models were able to more accurately predict neonatal health risks, offering new insights into how socioeconomic inequalities contribute to negative health outcomes. This study underscores the importance of considering SES in public health policies and the potential of AI in supporting the development of more targeted and effective interventions [64]. Furthermore, the use of AI models to analyze the cumulative effects of environmental exposures on neonatal health has gained increasing attention. Machine learning algorithms, for example, are able to integrate data on air pollution levels, geographical location, and health outcomes, allowing for the identification of populations most vulnerable to the harms caused by environmental exposures [65]. A recent study applied this methodology to map geographical inequalities in neonatal health, highlighting how environmental factors, such as air pollution, significantly contribute to health inequalities between different areas. These approaches enable the development of targeted interventions that address specific environmental risk factors, improving the effectiveness of public health strategies [66]. Finally, AI models have also been used to analyze the impact of nutrition and lifestyle factors on neonatal health [67]. By integrating data from various sources, such as maternal health surveys, dietary assessments, and medical records, AI algorithms can identify high-risk behaviors and conditions, offering opportunities for targeted intervention. For example, a study used predictive models to assess how maternal nutrition impacts neonatal outcomes, suggesting that adequate nutrition during pregnancy could significantly reduce the risk of adverse neonatal outcomes. These models, therefore, offer the possibility of adopting more personalized preventive approaches, improving maternal-infant health, and reducing health inequalities [68].

1.4 Application of Artificial Intelligence in Epidemiology: Traditional and Innovative Approaches

Epidemiology, the study of the distribution and determinants of health events in populations, has made significant progress in recent decades. With the rise of digital data and computational technologies, the application of AI in epidemiology is emerging as a transformative force, capable of reshaping the way public health data is analyzed and interpreted [69]. AI methods, including machine learning, deep learning, and causal inference techniques, offer powerful tools to uncover complex patterns and relationships in large, diverse datasets. These approaches enhance traditional epidemiological methods by providing insights that were previously difficult, if not impossible, to obtain with conventional statistical techniques

[70]. Before the advent of AI methods, epidemiology relied on traditional statistical methods, such as regression analysis, survival analysis, and cohort studies, to identify associations between risk factors and health outcomes. While these methods were fundamental, they often struggled to address the complexity of modern data, which includes high-dimensional variables, nonlinear relationships, and the presence of numerous confounding factors [71]. Traditional AI methods, particularly machine learning techniques like Decision Trees (DT), Random Forests (RF), and Support Vector Machines (SVM), have gradually been integrated into epidemiological research. These techniques allow researchers to build models that can adapt to complex interactions between multiple variables and predict disease outcomes with greater accuracy compared to conventional methods [72]. In the context of disease prediction, machine learning algorithms are frequently used to analyze clinical and demographic data to identify individuals at high risk for conditions such as cardiovascular diseases (CD), diabetes, and cancer [73]. By leveraging historical patient data, machine learning models can provide personalized risk assessments, helping guide early intervention and preventive strategies. Furthermore, AI methods have also been employed to improve traditional epidemiological study designs, such as cohort and case-control studies [74]. By applying algorithms to identify hidden patterns in large datasets, researchers can uncover previously overlooked factors that contribute to the onset and progression of disease. These advancements allow for a more detailed understanding of complex diseases and provide practical insights for Public Health policies and clinical practice [63].

1.4.1 Machine Learning Techniques and Causal Models in Epidemiology

In recent years, the integration of machine learning (ML) techniques with causal modeling has proven to be a powerful tool in epidemiology, especially for understanding complex interactions between risk factors and diseases [42]. Bayesian Networks (BNs), for instance, are increasingly utilized to manage uncertainty in public health and clinical applications. These probabilistic graphical models represent relationships between variables, enabling the analysis of intricate interactions between risk factors and health outcomes, such as cardiovascular diseases (CVDs) and metabolic syndrome. With the rising global prevalence of CVDs, these models are becoming essential in public health research, as they can help identify the impact of controllable risk factors, such as smoking and obesity, and estimate the potential for disease prevention [75]. Among the ML techniques, the Bayesian Additive Regression Trees (BART) method stands out as an innovative approach for estimating causal effects in complex epidemiological scenarios, where exposures may be multivariate and continuous. This method

overcomes the limitations of traditional epidemiological models by providing a more accurate understanding of causal relationships, particularly in the presence of high-dimensional data. Furthermore, BART can be applied to estimate the "Total Events Avoided" (TEA), which helps assess the effectiveness of public health policies, such as those aimed at reducing environmental pollutants [76]. Other causal inference models, such as the Linear Non-Gaussian Acyclic Model (LiNGAM), are valuable for distinguishing between correlation and causality. These models have been successfully applied in various epidemiological studies, such as those involving non-alcoholic fatty liver disease (NAFLD) and metabolic syndrome [77]. Additionally, Directed Acyclic Graphs (DAGs) are frequently employed to visualize and infer causal relationships in uncertain settings, like medical diagnoses or analysis of psychopathological conditions. The use of DAGs enables researchers to better understand the complex interdependencies between variables and guide decision-making in public health strategies. Techniques like causal forest analysis provide further depth by identifying causal effects at the individual level, addressing treatment effect heterogeneity [78]. Such methods have been applied to evaluate the outcomes of preventive programs, including hospital initiatives like the "Transitions Program" designed to prevent readmissions. Moreover, advanced causal inference methods, including G-computation and Targeted Maximum Likelihood Estimation (TMLE), are used to estimate causal effects in the presence of complex models, offering a more nuanced understanding of the impact of various interventions [79]. Lastly, the use of AI in Public Health is becoming increasingly important for understanding the socio-economic determinants of health. Methods like causal Shapley values (SV), which are used to analyze the impact of socioeconomic disparities on disease distribution, are gaining prominence in the analysis of diseases like Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) - COVID-19. These AI-driven models, including probabilistic graphical models such as DUCG (Dynamic Uncertainty Causal Graphs), have shown great potential for improving diagnostic accuracy and the formulation of more effective health policies [80]. By combining machine learning techniques with traditional epidemiological approaches, these methods are revolutionizing the way causal dynamics are analyzed in Public Health, paving the way for more personalized treatments and improved health outcomes at the population level [81].

1.4.2 Causal vs. Association Modeling in Epidemiology

In this study, we apply a causal modeling framework to explore the relationships between maternal exposures and adverse pregnancy outcomes. While causal models allow us to structure potential cause-effect relationships, we acknowledge that our findings should primarily be interpreted in terms of associations. Observational data, even when analyzed using causal inference techniques, cannot fully establish causality without experimental validation.

The use of directed acyclic graphs (DAGs) and statistical adjustment methods helps mitigate confounding and provides a structured way to infer relationships consistent with a causal hypothesis. However, it is crucial to emphasize that our approach does not establish definitive causal effects but rather identifies associations aligned with a causal framework. This distinction is particularly relevant when using machine learning (ML) techniques in epidemiology, as ML models can detect complex patterns but do not inherently differentiate between correlation and causation.

Given these considerations, our study aims to identify associations that may suggest potential causal relationships, providing a basis for future experimental research to validate these findings.

Chapter 2

Objectives

2.1 Rationale and specific objectives

In line with the current state of the art in maternal and child health, integrating epidemiological data management with advanced Artificial Intelligence techniques and insights from biomolecular and translational research is essential. This approach aims to assess environmental risk profiles, behavioural factors, and genetic susceptibility markers, as well as to monitor markers of effect, aging, and epigenetics. Such integration remains a key priority for Public Health. Specifically, studying the molecular mechanisms underlying the transgenerational effect on early biological aging through the evaluation of environmental and behavioral risk profiles and monitoring genetic susceptibility, effect, aging, and epigenetic markers using biological samples from Sicilian birth cohorts constitutes the general aim of this thesis. Although various conventional statistical approaches are widely used, recent advances in health informatics have introduced more accurate models employed in various areas of epidemiology. The availability of large-scale health data has enabled the application of ML methods to predict specific adverse outcomes. For these reasons, the primary goal of my PhD thesis is to evaluate the interplay between epidemiological data, including lifestyle factors, and biomarker characterization by applying artificial intelligence models to estimate the risk of adverse outcomes in mother-child pairs from two Sicilian birth cohorts. The specific objectives of this PhD thesis are:

i) the integration of clinical, epidemiological, and genetic/epigenetic/genomic data to investigate the effect of the exposome on premature biological aging of the fetus, through molecular signatures such as telomere length.

ii) the identification and application of causal machine learning models to investigate causal relationships between maternal exposures and adverse pregnancy/neonatal outcomes.

iii) the application of traditional data analysis models alongside innovative artificial intelligence techniques to estimate the risk of adverse pregnancy outcomes and predict the effectiveness of potential Public Health strategies. iv) the exploration of the main AI techniques used to identify causal relationships in public health applications, with a focus on analyzing the most effective methods, challenges, and the potential of AI in improving the understanding of population health determinants and optimizing prevention and intervention policies.

As a case study, these models were tested to explore causality between Body Mass Index (BMI), Gestational Weight Gain (GWG), and Telomere Length (TL) in amniotic fluid, considering other confounding factors. The data obtained will therefore allow for the evaluation of the effects of the exposome on maternal and child health, with the aim of identifying more appropriate interventions to develop effective primary prevention and maternal-child health promotion strategies.

Chapter 3

Methods

3.1 Study Design and Inclusion Criteria

The project "Diet and Epigenetic Markers: The Role of Maternal Diet on Epigenetic Patterns in Mothers and Newborns," which is still ongoing, was approved by the Ethics Committee of the Azienda Ospedaliero Universitaria Policlinico "G. Rodolico-San Marco" with letter Prot. N. 47/2014/VE on April 29, 2014, and the research has been coordinated at the University of Catania with Professor Antonella Agodi as the scientific coordinator. The study population consists of pregnant women referred to the U.O.C. of Prenatal Diagnosis and Medical Genetics at the Azienda Ospedaliero Universitaria Policlinico "G. Rodolico-San Marco," who requested karyotype analysis through amniocentesis as per the protocols of the Ministry guidelines. The study has been extended until November 2025. The fetal karyotype examination was conducted at the Medical Genetics laboratory of the U.O.C. of Prenatal Diagnosis and Medical Genetics at the Azienda Ospedaliero Universitaria Policlinico "G. Rodolico-San Marco." As part of the study, a reserve culture used for cytogenetic analysis was sent to the Molecular Epidemiology Laboratory of the GF Ingrassia Department for subsequent DNA extraction and the determination of genetic polymorphisms and methylation profiles of the newborn. The project "The Mamma & Bambino Cohort: A Multisectoral Approach to Maternal and Infant Health Through Assessment of the Exposome in Women - MAMI-MED," which is still ongoing, was approved by the Catania 2 Ethics Committee during the session held on 28/07/2020 (minutes no. 71/2020/CECT2), note protocol no. 487/C.E. dated 04.08.2020, with Professor Antonella Agodi as the scientific coordinator. The research has been coordinated at the University of Catania. The study population consists of pregnant women referred to the U.O.C. of Obstetrics and Gynecology, ARNAS Garibaldi-Nesima in Catania. Recruitment occurred during the first prenatal visit and within the first trimester

of pregnancy and involved all women who expressed their willingness to give birth at the aforementioned unit. The study has been extended until September 2, 2025. All recruited participants were invited to sign an informed consent form for participation in the study and an additional informed consent form for genetic studies. Data were processed in accordance with GDPR No. 2016/679 and the Authorization for the Processing of Genetic Data - Official Gazette No. 302 of December 27, 2013. Furthermore, each woman, as well as the child's father, signed the information notice and consent form for the processing of their personal data and that of the newborn.

3.1.1 Description of the Birth Cohorts ('MAMI-MED' and 'Mamma & Bambino')

The ongoing projects "The Mamma & Bambino Cohort: A Multisectoral Approach to Maternal and Infant Health through the Evaluation of the Exposome in Women - MAMI-MED" and "Diet and Epigenetic Markers: The Role of Maternal Diet on Epigenetic Patterns in Mothers and Newborns" involved the collection of biological samples, clinical, and epidemiological data within the framework of two birth cohorts—already established and still ongoing ("MAMI-MED" and "Mamma & Bambino", respectively)—included in the European biobank network birthcohort.net. These birth cohorts actively engaged mother-child pairs attending ARNAS Garibaldi Nesima and the University Hospital Policlinico "G. Rodolico-San Marco" in Catania. Women were invited to participate during their first prenatal visit and within the first trimester of pregnancy. The scientific coordinator of both birth cohorts (the "MAMI-MED" cohort and the "Mamma & Bambino" cohort) is Professor Antonella Agodi, Director of the Department of Medical and Surgical Sciences and Advanced Technologies at the University of Catania. "The Mamma & Bambino Cohort: A Multisectoral Approach to Maternal-Child Health Through the Evaluation of the Women's Exposome - MAMI-MED" is a prospective study, funded by the University of Catania in the framework of the PIACERI project, currently ongoing, that involves the recruitment of pregnant women at ARNAS Garibaldi-Nesima in Catania during the first prenatal visit and within the first trimester of pregnancy, with scheduled follow-ups until the children reach four years of age. This cohort is included and registered in the international network of birth cohorts (birthcohorts.net) and the Lifecycle project (https://lifecycle-project.eu/). Information is collected at recruitment, at birth, and at 12, 24, and 48 months after birth and includes: i) anthropometric measurements, ii) socio-demographic and behavioral characteristics, iii) dietary data obtained through a FFQ, iv) pregnancy and neonatal outcomes, and v) molecular data on genetic, epigenetic, and aging biomarkers obtained via Real-Time PCR and Pyrosequencing.

The study also collects the following biological samples: maternal blood, placental samples, and umbilical cord blood. The "Mamma & Bambino Cohort: Diet and Epigenetic Markers: The Role of Maternal Diet on Epigenetic Patterns in Mothers and Newborns" is an ongoing prospective study that involves the recruitment of pregnant women at the University Hospital Policlinico "G. Rodolico-San Marco" in Catania during prenatal genetic counselling. It includes scheduled follow-ups through the first two years of the child's life. This cohort is included and registered in the international network of birth cohorts (birthcohorts.net) and the Lifecycle project (https://lifecycle-project.eu/). Information is collected at recruitment, at birth, and at 12 and 24 months after birth, and includes: i) anthropometric measurements, ii) socio-demographic and behavioural characteristics, iii) dietary data obtained through a Food Frequency Questionnaire (FFQ), iv) pregnancy and neonatal outcomes, and v) molecular data on genetic, epigenetic, and aging biomarkers obtained via Real-Time PCR and Pyrosequencing. The study also collects the following biological samples: maternal blood; amniotic fluid and amniocytes; and umbilical cord blood.

3.1.2 Collection of Biological Samples

In the framework of the two cohorts various biological samples from the mother and the newborn are collected. Specifically, at the time of recruitment, each woman included in the study underwent a blood draw, which was sent to the Molecular Epidemiology Laboratory of the "GF Ingrassia" Department for the determination of maternal genetic and epigenetic biomarkers. At the time of delivery, a blood sample was collected from the umbilical cord (5 ml in EDTA) along with a placenta sample. These were also sent to the Molecular Epidemiology Laboratory of the GF Ingrassia Department for the determination of neonatal genetic and epigenetic biomarkers. Women who underwent amniocentesis were invited to donate an amniotic fluid sample and the corresponding reserve culture of amniocytes. The biological samples obtained and classified in this way were used for nucleic acid extraction using commercially available kits. All analyses were performed according to the study protocol, in compliance with the principles of Good Clinical Practice (GCP), the Declaration of Helsinki, and current regulations. Specifically: i) Single Nucleotide Polymorphisms were analyzed using allelic discrimination techniques in Real-Time PCR; ii) Epigenetic markers were analyzed following protocols described in the literature, based on the latest sequencing techniques; iii) Telomere length was analyzed through relative quantification using Real-Time PCR; The collected biological samples were stored at -20°C/-80°C for subsequent analyses. Genetic polymorphisms were analyzed using allelic discrimination techniques in Real-Time PCR [82]. For the analysis of methylation levels, as well as other epigenetic biomarkers,

protocols described in the literature were employed [83, 84, 85, 86]. To date, for the two birth cohorts "Mamma & Bambino" and "MAMI-MED," which are still ongoing, a total of 2605 women have been recruited, and a total of 2512 maternal blood samples were collected at the time of recruitment, along with 527 cord blood samples and 500 placenta samples at the time of delivery. Specifically, for the "Mamma & Bambino" cohort, which involves recruiting women during amniocentesis, a total of 497 women have been enrolled, of whom 443 underwent amniocentesis. At the time of enrollment, 404 maternal blood samples were collected. Additionally, at the time of delivery, 27 cord blood samples were collected. For women who underwent amniocentesis, 204 amniotic fluid samples were collected. For the "MAMI-MED" cohort, which involves recruiting pregnant women during prenatal screening (bi-test), a total of 2108 women have been enrolled, and 2108 maternal blood samples were collected at the time of recruitment. Furthermore, at the time of delivery, 500 cord blood and placenta samples were collected.

3.1.3 Telomere Length Assessmennt

Biological samples were collected from mothers and newborns participating in the "Mamma & Bambino" cohort during recruitment and at delivery. Specifically, women who underwent amniocentesis were invited to donate a sample of amniotic fluid. For the current analysis, the amniotic fluid samples were used for cfDNA extraction. A 1 ml aliquot of uncultured amniotic fluid was centrifuged at 12,500 g to remove any residual cells. The supernatant was then used for cfDNA extraction using the QIAamp Blood Kit (Qiagen, Milan, Italy) according to the manufacturer's protocol. DNA purification was automated using the QIAcube instrument (Qiagen, Milan, Italy). The concentration and purity of the extracted DNA were assessed using the NanoDrop 1000 spectrometer and the Qubit 3.0 Fluorometer with the dsDNA HS Assay Kit (Thermo Fisher Scientific, Carlsbad, CA, USA). The relative TL of cfDNA was evaluated using the Relative Human TL Quantification Assay Kit (ScienCell Research Laboratories, Carlsbad, CA, USA) according to the manufacturer's protocol. Real-time quantitative polymerase chain reaction (qPCR) was performed on a QuantStudio 7 Flex Real-Time PCR System (Thermo Fisher Scientific, Carlsbad, CA, USA) using two sets of primers: the telomere (T) primer set for amplifying telomere sequences and the single-copy reference (S) primer set for amplifying a 100 bp region on human chromosome 17, used as a reference for data normalization. Each reaction included 1 µl of DNA (5 ng/µl), 2 µl of primer solution (telomere or SCR), 10 µl of 2X GoldNStart TaqGreen qPCR master mix (ScienCell Research Laboratories, Carlsbad, CA, USA), and 7 µl of nuclease-free water. The PCR conditions included an initial denaturation at 95°C for 10 minutes, followed by 32 cycles

of 95°C for 20 seconds, 52°C for 20 seconds, and 72°C for 45 seconds. All reactions were performed in duplicate, and the relative TL was expressed as the average telomere/single-copy reference (T/S) ratio.

3.1.4 Collection of Epidemiological and Biological Data

For each mother-child pair, epidemiological data were collected through the administration of ad hoc questionnaires and follow-up telephone interviews. Specifically, at the time of recruitment, demographic information, lifestyle habits, and dietary patterns were gathered using a tailored questionnaire and a validated Food Frequency Questionnaire. At the end of pregnancy, women were contacted again to collect information on maternal vaccination status, antibiotic use, and pregnancy outcomes. Furthermore, to monitor the child's health up to two years post-delivery, mothers are contacted again by phone to gather data regarding parental health, socio-economic status, and lifestyle, as well as information on the child's health, vaccinations, potential antibiotic use, and diet. The project also included a 48-month follow-up from the time of delivery. At twelve and twenty-four months after the child's birth, the recruited women were contacted by phone to collect data regarding parental health, socio-economic status, lifestyle, as well as the child's health, habits, and diet. To support this, ad hoc tools were developed to gather epidemiological, behavioral, clinical, socio-psychological, and contextual data for the assessment of concurrent events. In particular, anthropometric measurements, clinical data, nutritional status, and genetic and epigenetic biomarkers were assessed. For all the women recruited in the study, anthropometric data were collected; specifically, data regarding height and weight, both current and pre-pregnancy (according to internationally recommended procedures), were gathered to assess nutritional status. The evaluation of the women's nutritional status was carried out by calculating the Body Mass Index based on the criteria of the World Health Organization [87]. To verify if the study populations are in Hardy-Weinberg equilibrium, for each of the genetic traits studied, the chi-square test and the G-test (also called the "likelihood ratio test") were performed, which compare the observed genotypes with the expected ones. The participants in the study were also given a FFQ specifically designed to assess the consumption of foods and beverages that are the major contributors to micronutrient intake in the study population. The administration of the FFQ allowed for the evaluation of the dietary profile for the month preceding recruitment. The specific and validated FFQ consists of a defined list of foods and beverages that are the main contributors to the intake of macro- and micronutrients. In particular, approximately 130 foods are listed, divided into 9 food categories: cereals, bread, and snacks; red meat, chicken, fish, and eggs; dairy products; pasta, mixed dishes, and soups; vegetables

and grains; condiments and sauces; sweets; fruits; and beverages. For each food, women were asked to indicate the frequency of consumption and portion sizes. To improve the quality of the surveys, a photographic atlas of portion sizes (small, medium, and large) for each food was used. The dietary information obtained through the FFQ was converted into monthly and daily intake values for various nutrients by multiplying the frequency of consumption for each food by the respective portion size (g) [88, 89, 90]. The translation into energy and nutrients was performed using food composition tables and specific databases [91, 92, 93]. Adherence to the Mediterranean Diet (MD) was evaluated using the Mediterranean Diet Score (MDS), which assigns each woman recruited a score from 0 to 9 [94, 95]. The foods consumed are divided into nine categories. Cereals, fruits, vegetables, legumes, fish, and the ratio of unsaturated to saturated fats are considered protective foods, while meat and dairy are considered non-protective. Each food is assigned a score of 0 or 1. Specifically, for protective foods, a score of 1 is assigned if the consumption in grams is above the median consumption value of the study population, while a score of 0 is assigned if the consumption is below the median. For non-protective foods (meat and dairy), a score of 1 is assigned if their consumption in grams is below the median consumption value, and a score of 0 is assigned if the consumption is above the median. The adherence profile to the MD was calculated with a score ranging from 0 to 9, according to the Mediterranean Diet Score model [96]. Based on the obtained score, adherence to the MD can be classified as: low adherence (score 0-3); moderate adherence (score 4-6); and high adherence (score 7-9) [97].

3.1.5 Methods of Data Processing

The data, processed using both electronic and other tools, were disseminated only in a strictly anonymous form, such as through scientific publications, statistics, and scientific conferences. Participation in the study implies that the Ethical Committee and health authorities, both Italian and foreign, may access data regarding individual patients, which may also be contained in the original clinical documentation, in a manner that ensures the confidentiality of their identity. The enrolled women can exercise the rights specified in Article 7 of the Code (e.g., accessing, supplementing, updating, correcting their personal data, opposing their processing for legitimate reasons, etc.) by directly contacting the study center. They can also withdraw from the study at any time, without providing any justification: in this case, the biological samples and related data will be destroyed. Furthermore, no additional data will be collected, although previously collected data may still be used to determine the results of the research without altering them.

3.2 Statistical Analysis

The data related to the enrolled subjects – including analysis of anthropometric data; analysis of genetic polymorphisms and methylation profiles; analysis of consumption data obtained from FFQs; and analysis of data related to micronutrient intake (obtained through the use of specific databases) – were entered into a specially designed electronic database and processed using Statistical Package for Social Sciences (SPSS) (version 26.0, SPSS, Chicago, IL, USA). Additionally, for causal inference modeling, including the application of the do-operator and graphical causal models, analyses were performed using Python. The application of the do-operator was performed within this framework to estimate the causal effects of gestational weight gain (GWG) and pre-pregnancy BMI on telomere length (TL). To test the associations and interactions between different exposures and markers of early biological aging, both univariate and multivariate analyses were conducted. In particular, after evaluating the individual associations for each exposure, linear and logistic regression models were applied to test for confounding factors and interactions between the various exposures, as well as the potential involvement of mediators in the relationship. The association between different variables was tested using the chi-square test (χ^2 test) or Fisher's exact test, and the Student's t-test, with a significance level set at $p < 0.05$. For variables with a non-normal distribution, the equivalent non-parametric tests were applied. Furthermore, associations between variables were determined through estimates of odds ratios (OR) and their corresponding confidence intervals. For factors/variables that were significant in the univariate analysis ($p - value < 0.05$), multivariate analyses were conducted using linear and logistic regression models. In particular, in an analysis conducted with the aim of delineating sex disparities in delivery and neonatal characteristics within the "MAMI MED" cohort, descriptive statistics, including means with interquartile ranges (IQR) or frequencies with percentages (%), were used in the univariate analysis to describe the characteristics of the study population. The distribution of quantitative variables was assessed using the Kolmogorov–Smirnov test. For bivariate analysis, the Mann–Whitney U test was used for quantitative variables, while the chi-square test was used for categorical variables. Multivariable linear regression analyses were applied to assess the association between sex and birth length and weight, adjusting for maternal age, gestational age at delivery, pre-pregnancy BMI, GWG, and total energy intake. Additionally, multivariable logistic regression analyses were applied to assess the association between sex and PTB and birth weight according to gestational age (Small for Gestational Age (SGA)vs Appropriate for Gestational Age (AGA), and Large for Gestational Age (LGA) vs AGA). Results from the linear regression models were reported as beta regression coefficients, while logistic regression was reported as OR and their respec-

tive 95% Confidence Intervals (CIs). All statistical tests were two-sided and performed at a significance level of $\alpha = 0.05$. Finally, machine learning and deep learning algorithms were applied to examine the predictors that have the most influence on the aforementioned models. In particular, unsupervised models were used for managing and reducing the dataset under examination (e.g., Principal Component Analysis (PCA) and Clustering Algorithms), as well as other machine learning and deep learning algorithms (e.g., Decision Tree, Support Vector Machine, etc.), which were trained and tested based on the nature of the predictors and outcomes of interest, to identify those with the best predictive performance. Initially, traditional unsupervised models, such as cluster analysis and PCA, were used to characterize the study population, focusing on dietary habits and identifying subgroups of participants with specific characteristics. In particular, PCA was conducted for the analysis of nutritional profiles. Additionally, to improve the accuracy of estimates and identify the most significant predictors, machine learning and deep learning models were applied. Additionally, to further investigate the predictors influencing gestational weight gain (GWG), binary classification models were applied, including **Decision Tree, Random Forest, and XGBoost**. These models were trained to classify GWG as adequate or non-adequate, using maternal and neonatal characteristics as predictors. To interpret the contribution of each predictor in the classification, the **SHapley Additive Explanations (SHAP) method** was employed. SHAP analysis revealed that **the length of telomeres in amniotic fluid was the most significant predictor of GWG classification**, highlighting its potential role as a biomarker for gestational weight adequacy.

3.2.1 Application of Machine Learning and Deep Learning Models

To enhance the identification of key predictors and improve classification accuracy, machine learning and deep learning algorithms were applied to the dataset. The analytical workflow included the following steps:

1. Dimensionality Reduction and Clustering: - Principal Component Analysis (PCA) was used to identify major components explaining dietary variation in the study population. - Two-Step Cluster Analysis was applied to classify participants based on education level, employment status, pre-gestational nutritional status, and adherence to the Mediterranean Diet Score (MDS).

2. Predictive Modeling with Machine Learning: - Supervised learning models, including Decision Tree, Support Vector Machine (SVM), were trained and tested to assess their predictive performance on key study outcomes. - Model selection and tuning were performed based on standard performance metrics.

3. Binary Classification and SHAP Analysis: - Decision Tree, Random Forest, and XGBoost classifiers were applied to classify GWG as adequate or non-adequate based on maternal and neonatal predictors. - SHapley Additive Explanations (SHAP) analysis was used to assess the relative importance of each predictor in the classification models. - Results showed that telomere length in amniotic fluid was the strongest predictor of GWG classification, suggesting a potential link between telomere biology and gestational weight adequacy.

3.2.2 Cluster Analysis

The clustering method is an exploratory approach used to identify natural groupings within a dataset that would otherwise remain hidden. In general, this method allows similar observations to be organized into groups, ensuring that members within the same cluster share common characteristics. To achieve this, the Two-Step Cluster Analysis was applied to divide the original dataset into distinct clusters with high internal homogeneity and significant variability between clusters. The process involves creating a Cluster Features Tree (CF Tree) by positioning the first case at the root and adding each new case to an existing node or creating a new node based on their similarity. Subsequently, an agglomerative clustering algorithm is used to group the CF Tree nodes of clustered variables, generating various grouping options. This algorithm is designed to handle both categorical and continuous variables, and it can automatically determine the optimal number of clusters by comparing the values of specific model selection criteria across different clustering solutions [98]. In particular, in an analysis conducted to evaluate and explore maternal and neonatal characteristics based on socioeconomic status, cluster analysis was used to uncover natural groupings that are not easily evident within the MAMI-MED cohort dataset. This method aims to achieve high similarity within the clusters and significant variability between clusters. The two-step cluster analysis method was used to identify distinct groups of pregnant women based on their level of education, employment status, pre-gestational nutritional status, and MDS. The determination of the optimal number of clusters was automated using Schwarz's Bayesian Information Criterion (SBIC), which allow model selection based on the likelihood function. The two-step cluster analysis applied the log-likelihood distance measure, enabling the integration of both categorical and continuous variables, assuming independence among the variables within the cluster model. Variables were assessed for their predictive importance, with scores ranging from 0.1 (indicating minimal predictive relevance) to 1.0 (indicating maximum predictive relevance). To enhance the reliability of the clustering solution, variables with predictive importance scores below 0.2 were excluded, and the analysis was re-evaluated to confirm

consistency and validity [98, 99].

3.2.3 Principal Component Analysis

PCA has been used in nutritional epidemiology to simplify the complexity of high-dimensional datasets obtained through FFQ, dietary records, or dietary history questionnaires [100]. Specifically, PCA is commonly employed to reduce a dietary dataset consisting of correlated variables into a smaller number of dimensions that reflect distinct dietary patterns. These dimensions, generally called PCs, are ranked by the total variance explained, are uncorrelated with each other, and are fewer in number than the original variables [101]. Before applying PCA to a dataset, certain assumptions should be checked. First, the sample size must be large enough to provide reliable results. Second, there should be sufficient correlation between variables to reduce them to a few PCs. The method used to assess sampling adequacy is the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy, while Bartlett's test of sphericity is applied to test the hypothesis that the correlation matrix is an identity matrix. Generally, high values of KMO (i.e., close to 1) and small values for Bartlett's test (less than 0.05) indicate that the dataset is suitable for PCA. Without delving into the mathematical details of PCA, for each PC, the eigenvalue and eigenvector are derived from the covariance matrix, which respectively represent the total variance explained and its orientation. In the case of untransformed data with different scales, the alternative is to use the correlation matrix as input for PCA [101]. The number of PCs to retain for further analysis is typically determined by the eigenvalues and the amount of variance explained [102, 103]. Several rules of thumb exist to determine an acceptable number of PCs (e.g., those that reach a cumulative variance of approximately 80% or those with eigenvalues greater than 1). However, most of these criteria do not apply well to nutritional epidemiology. For example, the explained variance percentage typically ranges between 10% and 30%, while the cutoff for eigenvalues is around 1.6 [104, 105]. This helps retain a small number of PCs for interpretation and analysis. A third criterion involves visually inspecting the Scree plot (i.e., a line plot of eigenvalues of the PCs) to identify an elbow, beyond which subsequent PCs contribute little to the variance explained. Each PC can be interpreted in terms of correlations with the original variables, represented by the PC loadings. To simplify the interpretation of PCs, the varimax rotation is typically applied, and individual PC scores are generated [106, 107, 108]. In the current analysis, PCA with varimax rotation was applied to the covariance matrix of standardized and energy-adjusted dietary data. In the main analysis, the first two PCs were retained based on the examination of the Scree plot. However, we also assessed the changes associated with retaining the first four PCs or all PCs with eigenvalues greater than 1.

3.2.4 Clustering and Consolidation

The next step in our analysis involves applying hierarchical clustering to the selected PCs to identify distinct clusters based on the hierarchical tree. In this agglomerative approach, data points are progressively merged based on their pairwise distances [109]. While multiple methods exist for measuring these distances, we recommend using Ward's Linkage, as it is based on multidimensional variance, similar to PCA. Several strategies are available for determining the optimal number of clusters. The first method examines inertia (i.e., the mean squared distance between each point and its nearest centroid) as the number of clusters increases. The second method identifies the number of clusters that maximizes the silhouette score [109]. The clustering solution from hierarchical clustering is then refined using K-means clustering. Since K-means requires a predefined number of clusters, the algorithm is applied with the number of clusters determined by the hierarchical clustering process. It is important to note that the two methods might result in slightly different clustering solutions [109]. The consistency between hierarchical and K-means clustering can be evaluated using the Adjusted Rand Index (ARI).

3.2.5 Clustering on Principal Components

Clustering on Principal Components (CPC) is a methodology that combines two main techniques: PCA and clustering. Its purpose is to identify natural groupings within a multidimensional dataset, while simultaneously reducing the data's dimensionality and preserving its main variance. Initially, PCA is applied to transform the original variables into a reduced number of principal components (PCs), which represent the directions of maximum variance in the data. Subsequently, clustering is used to group cases based on the similarity of their principal component scores, identifying clusters of similar observations. This approach allows for the analysis of large, complex datasets, such as those with many correlated variables, making it easier to identify meaningful patterns and improving the interpretability of results compared to using clustering or PCA alone [110]. In particular, in an analysis conducted on 667 women from the "MAMI-MED" cohort to derive dietary patterns and assess associations with birth weight for gestational age, we applied this approach. In brief, the approach involves two multivariate techniques to reduce the dimensionality of the dietary dataset and provide the clustering solution. The combination of these methods allows us to leverage their strengths, achieving a better solution than using PCA or clustering alone [111]. First, PCA was applied to reduce the dietary dataset obtained through the FFQ, which was characterized by a set of correlated dietary variables. This common data reduction method was applied to the daily intake of 39 food categories, energy-adjusted. In this

phase, varimax rotation was used on the covariance matrix to improve the interpretability of the PC scores. The number of PCs to retain was chosen by visually inspecting the eigenvalue plot and using the elbow method. The absolute values of the factor loadings were used to determine the contribution of each food category to the PCs. For each PC, factor scores were calculated as the sum of the products between the energy-adjusted intakes and the factor loadings. To obtain independent clusters of participants, hierarchical clustering was then applied, an agglomerative method where data points are progressively joined based on their distance. In our study, hierarchical clustering was applied to the retained PCs, and distance was measured using the Ward linkage criterion. The reason for choosing Ward's linkage was to work with multidimensional variance, as already done with PCA. Finally, the number of clusters to consider was chosen according to the silhouette method [110, 111]. Specifically, we selected the clustering solution that maximized the silhouette score, a metric based on intra-cluster distance and the separation between clusters.

3.2.6 A Causal Graph Analysis

To analyze the causal relationships between pre-pregnancy BMI, GWG, and TL, we used causal graph analysis, a method that allows for exploring causal pathways and underlying mechanisms between complex variables. This approach is particularly useful for identifying risk and protective factors, improving the effectiveness of public health interventions, especially when addressing issues involving interactions between social, environmental, behavioral, and biological factors. In our study, we hypothesized that pre-pregnancy BMI influences both GWG and TL, with GWG acting as a mediator. For the present analysis, we specifically used data from mothers with a pre-pregnancy BMI ≤ 30, who completed a singleton pregnancy and had available information regarding GWG at delivery and TL in amniotic fluid. During the recruitment process, women were asked to provide their height and pre-pregnancy weight, which were then used to calculate their pre-pregnancy BMI in kg/m^2. According to the WHO criteria, women were classified into three categories: underweight, normal weight, or overweight, based on their pre-pregnancy BMI. To assess GWG, the maternal weight at delivery was collected from clinical records. The total GWG was then calculated by subtracting the self-reported pre-pregnancy weight from the weight at delivery. Following the guidelines from the Institute of Medicine (IOM), GWG was classified as insufficient, adequate, or excessive, considering the pre-pregnancy BMI. This classification took into account the recommended weight gain ranges for each BMI category. In addition to anthropometric measures, our analysis considered various covariates that could influence the relationship between pre-pregnancy BMI, GWG, and TL [112].

3.2.7 Causal Modeling Approach

To analyze the relationships between maternal exposures and pregnancy outcomes, we employed causal modeling techniques based on directed acyclic graphs (DAGs) and the do-operator framework. These methods allow us to hypothesize potential causal pathways and control for confounding factors. However, it is essential to clarify that, despite using causal inference tools, this study remains observational in nature.

The application of machine learning (ML) techniques in epidemiology provides an opportunity to identify complex associations that may suggest underlying causal mechanisms. However, ML models, by themselves, do not establish causality. Instead, we used ML as a complementary tool to improve variable selection, refine model specification, and explore nonlinear relationships that traditional statistical models might overlook.

Therefore, while our results are structured within a causal framework, they should be interpreted as associations consistent with a causal hypothesis rather than as definitive causal effects.

3.2.8 Causal Analysis using "do"-operator

The "do"-operator plays a crucial role in validating a causal graph [112, 113]. This operator broadens the conventional idea of "a posteriori" probability within a causal framework. In the classical concept of "posterior" probability, represented as $P(Y = y \mid X = x)$, we assess the likelihood of observing value y for the variable Y, given that the value x has been observed for the variable X. However, this formula does not fully account for the direct effect of X on Y when other confounding factors influence both variables. In contrast, the "do"-operator allows us to calculate the direct effect of X on Y within a robust cDAG model. By utilizing this operator, we can use the "adjustment formula" and various criteria such as "back-door," "front-door," and "mediation" to examine direct causal effects between variables. A key strength of this approach lies in the established assertion that, in the absence of unmeasured confounders, all causal effects are identifiable [112, 114, 115, 116, 117].

3.2.9 Estimating Causal Effects using Directed Acyclic Graphs (DAGs)

Graphical causal models provide a structured approach to represent causal relationships among variables within a dataset. These models are often depicted as Directed Acyclic Graphs (DAGs), where nodes represent variables, and directed edges illustrate causal influences. A key feature of DAGs is that information flows along the arrows, shaping potential

causal pathways.

DAGs can contain three fundamental structures:

- **Chain ($\mathbf{X \rightarrow Y \rightarrow Z}$):** Both X and Z are dependent through Y, and conditioning on Y can block this dependence.
- **Fork ($\mathbf{X \leftarrow Y \rightarrow Z}$):** Variables X and Z share a common cause (Y) and are dependent unless conditioned on Y.
- **Collider ($\mathbf{X \rightarrow Y \leftarrow Z}$):** Unlike the previous cases, X and Z are independent unless we condition on Y, which introduces a spurious association.

The do-operator framework is applied within DAGs to estimate causal effects by simulating interventions. By leveraging DAG structures and their joint probability distributions, we can determine how a specific variable influences an outcome under hypothetical interventions.

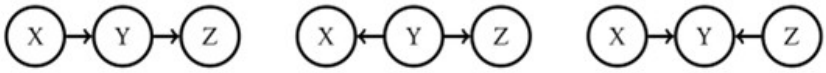


Figure 3.1: Examples of key DAG structures: (Left) Chain, (Middle) Fork, (Right) Collider.

3.2.10 Estimating the Causal Effect of GWG on Telomere Length

The first step of the analysis involves estimating the causal impact of GWG on TL. In this context, the causal effect formula (1) used to quantify the relationship between GWG and TL is as follows:

$$CE = \mathrm{P}(TL = t \mid \mathrm{do}(GW \mid G = g_2)) - \mathrm{P}(TL = t \mid \mathrm{do}(GW \mid G = g_1))$$

This formula enables us to estimate the effect when GWG shifts from g1 to g2. A positive causal effect suggests that changing GWG from g1 to g2 raises the likelihood of TL taking the value t. To assess the impact of GWG on TL, let us examine the initial graph, G1. Based on the principles of causal graph theory, we identify two "open" paths: a. GWG $\rightarrow$ TL, representing a direct path; b. GWG $\leftarrow$ BMI $\rightarrow$ TL, creating a fork. A fork is a specific pattern in a causal graph where two variables are influenced by a common cause. In this context, the fork illustrates the relationship between GWG, BMI, and TL in the graphical models. For the graphical model G1, when GWG is fixed, path (b) becomes "closed," simplifying the calculation of the "do"-operator to computing the conditional probability$P(TL = t \mid GWG = g)$. A similar approach can be used when evaluating the graphical models G2 and G3. In the

case of graphical model G2, the paths from GWG to TL include: a. GWG → TL, represent a direct path; b. GWG← BMI → TL, creating a fork; c. GWG ← E → BMI → TL; d. GWG ← Age → TL; e. GWG ← BMI ← Age → TL; f. GWG ← E → BMI ← Age → TL. When the value of GWG is held constant, all paths, except for path (a), are considered "closed." In this situation, the causal effect of GWG on TL can be estimated in a way similar to the previous case. For the graph G3, there are three distinct paths from GWG to TL: a. GWG → TL, representing a direct path; b. GWG← BMI → TL, forming a fork; c. GWG ← E → BMI → TL; As in the previous cases, when the value of GWG is fixed, paths (b) and (c) are considered "closed" [112].

Causal Effect of GWG on TL

The causal impact of gestational weight gain (GWG) on telomere length (TL) can be estimated using interventional probability. Given a graphical model (e.g., G_1 in Figure 3.2), the joint probability distribution is:

$$P(T, G, B) = P(T|G, B)P(G|B)P(B) \tag{3.1}$$

To determine the probability of TL (T) given a specific value of GWG (G):

$$P(T = t|G = g) = \sum_b P(T = t, B = b|G = g) = \sum_b \frac{P(T = t, B = b, G = g)}{P(G = g)} \tag{3.2}$$

Following the intervention $do(G = g)$, the manipulated probability becomes:

$$P(T = t|do(G = g)) = \sum_b P(T = t|B = b, G = g)P(B = b) \tag{3.3}$$

This transformation removes any confounding influence of B, providing a direct estimate of the causal effect [112].

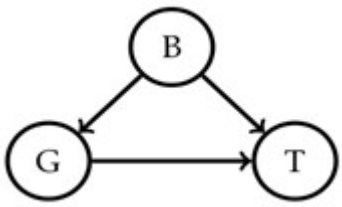


Figure 3.2: Graphical model G_1 before (left) and after (right) applying the do-operator to GWG.

3.2.11 Estimating the Causal Effect of pre-pregnancy BMI on Telomere Length

The second phase of the analysis is centered on estimating the direct causal effect of pre-pregnancy BMI on TL. This can be done by conditioning on GWG and calculating the difference between P($TL = t \mid \text{do}(BMI = b1), GWG = g$) and P($TL = t \mid \text{do}(BMI = b2), GWG = g$). In particular, for graphical model G1, there are two paths from BMI to TL: a. BMI → TL, representing a direct path; b. BMI → GWG → TL, creating a chain in which GWG serves as a mediator A mediator is a variable that is influenced by one variable and, in turn, has an impact on another variable. In this context, GWG acts as a mediator in the relationship between BMI and TL. This means that BMI affects GWG, which then impacts TL. In the case of graphical model G1, when the BMI value is fixed, both paths remain "open." As a result, estimating the direct causal effect of BMI on TL necessitates stratification by GWG. For graphical model G2, there are multiple paths from BMI to TL: a. BMI→ TL, representing a direct path; b. BMI → GWG → TL, creating a chain in which GWG serves as a mediator c. BMI ← Age → TL, creating a fork; d. BMI → GWG ← Age → TL; e. BMI ← EN → GWG → TL; f. BMI ← E → GWG ← Age → TL. When BMI is fixed, paths (c), (e), and (f) are considered "closed," meaning they no longer contribute to estimating the direct causal effect. In contrast, the remaining paths (a), (b), and (d) stay open and are included in the calculation of the "do"-operator. However, path (b) becomes "closed" when GWG is conditioned, while path (d) becomes "open" as a result. To address this, an intervention on GWG is required. By intervening on GWG, the direct causal effect of BMI on TL can be estimated. Thus, the direct causal effect of BMI on TL can be calculated using the controlled direct effect formula (2):

$$\text{CDE} = \text{P}(TL = t \mid \text{do}(\text{BMI} = b_2), \text{do}(\text{GWG} = g)) - \text{P}(TL = t \mid \text{do}(\text{BMI} = b_1), \text{do}(\text{GWG} = g))$$

As with the causal effect described in Eq. (1), a positive controlled direct effect suggests that altering BMI from b1 to b2 increases the likelihood of TL reaching the value t within the GWG = g category. In the case of graphical model G3, the paths from BMI to TL are as follows: a. BMI→ TL, representing a direct path; b. BMI → GWG → TL, creating a chain where GWG acts as a mediator; c. BMI ← EN → GWG → TL; When the BMI value is held constant, both paths (a) and (b) remain "open," while path (c) is considered "closed." As a result, estimating the direct causal effect of BMI on TL necessitates stratifying by GWG [112].

Causal Effect of BMI on TL

Examining the causal pathways between pre-pregnancy BMI (B) and TL, we identify both direct and indirect effects. In the initial model (G_1), BMI influences TL through two routes:

- **Direct path:** $B \rightarrow T$
- **Indirect path:** $B \rightarrow G \rightarrow T$

Applying the **do-operator** to fix BMI (B), the probability estimation follows:

$$P(T = t|do(B = b)) = \sum_{g} P(T = t|B = b, G = g)P(G = g|B = b) \tag{3.4}$$

However, since GWG acts as a **collider** between BMI and Age (A), conditioning on GWG could introduce spurious associations. To mitigate this, an additional intervention on GWG ($do(G = g)$) is required:

$$P(T = t|do(B = b), do(G = g)) = \sum_{a,e} P(T = t|B = b, A = a, G = g)P(A = a)P(E = e) \tag{3.5}$$

This methodology enables an unbiased estimation of the direct causal effect of BMI on TL [112].

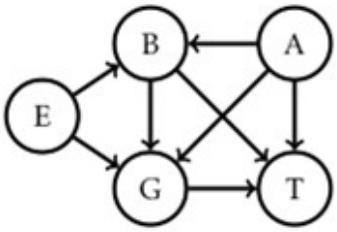


Figure 3.3: Graphical model G_2 with intervention on BMI and GWG.

3.2.12 Implications and Future Directions

The causal inference approach using DAGs (Directed Acyclic Graphs) and the do-operator provides essential insights into the mechanisms linking maternal factors to neonatal outcomes. This approach allows for the isolation of direct and indirect effects and testing the effectiveness of interventions on specific variables, such as pre-pregnancy BMI. However, it is important to note that the conclusions rely on causal assumptions, which depend on the correctness of the underlying DAG structure. Figure 3.4 illustrates how an intervention on

BMI can influence the causal pathways between BMI and telomere length (TL), showing the graphical model G3 before and after the intervention. As highlighted by the model, intervening on BMI alters the causal paths, allowing for a more accurate estimation of BMI's direct effects on TL, separating them from the effects mediated by other variables such as GWG.

The validity of these findings, however, depends on the correctness of the model structure and causal assumptions. Therefore, future research should integrate longitudinal data and randomized interventions to validate and refine the existing models. The use of such techniques would allow for testing the reliability of these approaches and exploring any mechanisms not yet identified.

Moreover, the causal approach described here provides a solid foundation for future investigations into the effect of other maternal factors, suggesting that, in addition to BMI, other interventions may be equally critical for improving neonatal outcomes. The integration of randomized interventions and the analysis of longitudinal data would allow for more robust results and provide more precise guidelines for evidence-based health policies.

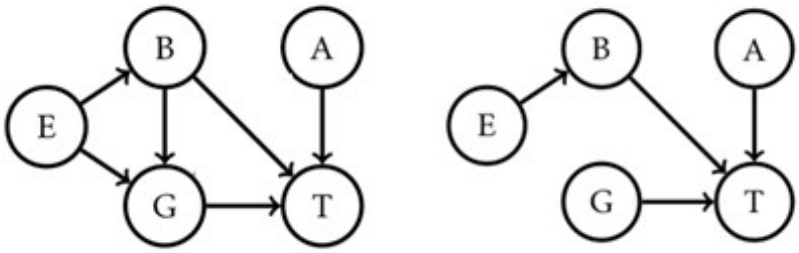


Figure 3.4: Graphical model G_3 before and after intervention on BMI.

3.3 Systematic Review: Methodologies and Inclusion Criteria

At the same time, a comprehensive systematic review of the scientific literature was conducted to identify significant causal machine learning models applied in the field of epidemiology. This systematic review included three scientific databases (PubMed, Web of Science, and Scopus) and was carried out in accordance with specific inclusion and exclusion criteria, aligned with Cochrane guidelines. The models identified through this review, tailored to the nature of the available data, will be useful for further analyzing the causal relationships between pre-pregnancy BMI, GWG, and telomere length in birth cohort samples.

3.3.1 Literature Search

From its inception until August 2023, we conducted a comprehensive literature search across the Pubmed, Web of Science and Scopus databases. The primary focus of this search was to identify studies encompassing causal inference techniques utilizing ML and AI methodologies in public health. Two authors (C.L. and A.M) independently carried out the literature search using the following terms: (Causalit*[Title/abstract] OR Causation*[Title/abstract] OR Causal[Title/abstract]) AND ("Machine Learning"[Title/abstract] OR "Deep Learning" [Title/abstract] OR "Artificial Intelligence" [Title/abstract] OR Algorithm* [Title/abstract]). The Preferred Reporting Items for a Systematic Review and Meta-analysis (PRISMA) guidelines [118] were followed as a framework for the methodology of this review.

3.3.2 Study selection on data extraction

This systematic review covers publications that satisfied the criteria listed below: (1) leveraged an approach associated with AI and ML Algorithms (such as SVM, KNN, decision trees , artificial neural networks [ANN], BBN, and regression models) to uncover causal relationships in public health applications, and (2) was published in English. The exclusion criteria were used to exclude studies that did not meet the review's focus and the following types of documents were excluded: (i) abstracts lacking full-text content and articles not in English; (ii) case reports or case series; (iii) comments, letters, editorials, and reviews; (iv) unpublished studies. The eligibility criteria were used to analyze the titles and abstracts of the publications, and then the full texts of all eligible publications were obtained and independently reviewed. Disagreements over study inclusion were resolved through full-text discussion with a third author (A.A.). The following information was gathered from the articles systematically: title, authors, date of publishing, country, objectives, study population, AI methods utilized, applications of ML and any relevant outcome reported.

Chapter 4

Results

4.1 Cluster Analysis of Social and Nutritional Profiles in the "MAMI-MED" Cohort

In this analysis, a cluster analysis was used to identify groups of pregnant women with similar social and behavioral characteristics and to explore their variability in relation to neonatal outcomes. The analysis included 1,512 women enrolled in the "MAMI-MED" cohort, which comprises all pregnant women attending the Azienda di Rilievo Nazionale e di Alta Specializzazione (ARNAS) Garibaldi Nesima in Catania during their first-trimester neonatal visit and who completed their pregnancy. The median gestational age at delivery was 39 weeks (IQR = 2). The women had a median age of 31 years (IQR = 7), and approximately half (54.4%) were primiparous. In terms of socioeconomic status, 75.2% of the women reported a high-medium level of education, and 51.6% were employed. Regarding lifestyle factors during pregnancy, 90.3% of women were non-smokers, 62.1% reported medium-high adherence to the Mediterranean diet, and 84.8% did not follow any dietary restrictions, with a median total energy intake of 1,714 kcal (IQR = 564). Prior to pregnancy, 58.3% of women were of normal weight, with a median BMI of 23.4 kg/m^2 (IQR = 6.1). In terms of GWG (median = 11 kg; IQR = 7), 39.5% of women experienced reduced weight gain, 28.2% excessive weight gain, and 32.3% adequate weight gain. Regarding neonatal characteristics, 6.3% of newborns were preterm, 6.5% had low birth weight, and 4.4% had macrosomia. The median birth weight was 3.3 kg (IQR = 0.6), and the median birth length was 50 cm (IQR = 2). Accordingly, 79.2% of newborns were AGA, 10.1% were SGA, and 10.7% were LGA.

4.1.1 Study population

Comparing women in the MAMI-MED cohort based on employment status, it emerged that employed women were slightly older than their unemployed counterparts, with a median age of 32 years compared to 30 years ($p < 0.001$). Additionally, a higher proportion of employed women were primiparous ($p = 7.6668 \times 10^{-15}$) and had a medium-to-high educational level ($p = 9.4161 \times 10^{-50}$). In terms of lifestyle, employed women were more likely to be non-smokers ($p = 1.7 \times 10^{-5}$) and to report medium-to-high adherence to the MD ($p = 1.54 \times 10^{-4}$). They also had a lower pre-pregnancy BMI compared to unemployed women ($p = 5.0 \times 10^{-6}$) . Employed women had a slightly higher GWG (12.0 kg vs. 11.0 kg, $p = 5.0 \times 10^{-6}$). Regarding pregnancy outcomes, employed women had a lower proportion of preterm births (p = 0.021) and low birth weight newborns (p = 0.003). The birth weight of infants born to employed women was significantly higher (3.32 kg vs. 3.24 kg; $p = 4.4 \times 10^{-5}$). Statistically significant differences were also observed in gestational age at delivery and birth length between the two groups ($p = 2.96 \times 10^{-4}$ and p = 0.003, respectively). Women with a medium-to-high educational level were more likely to be older ($p < 0.001$), primiparous ($p = 3.1097 \times 10^{-10}$), and non-smokers ($p = 9.9142 \times 10^{-13}$) compared to those with a low educational level. Furthermore, adherence to the Mediterranean Diet (p = 0.014) and the MDS (p = 0.003) were higher in the medium-to-high education group. Additionally, women with medium-to-high education were more likely to follow dietary restrictions (p = 0.008). However, no significant difference in total energy intake was observed between the two groups (p = 0.380). A higher proportion of employed women were in the medium-to-high education group ($p = 9.4161 \times 10^{-50}$). Women with medium-to-high education were more likely to report a lower pre-pregnancy BMI and have normal weight compared to those with lower educational levels ($p = 1.1 \times 10^{-5}$) . These women also showed a higher GWG (12.0 kg vs. 11.0 kg; p = 0.014) and were more likely to have adequate GWG (p = 0.010). Regarding pregnancy outcomes, women with medium-to-high education had a lower proportion of preterm births (p = 0.022). Moreover, significant differences were found in gestational age at delivery (p = 0.013) and birth length (p = 0.017). No significant differences were observed in the proportion of macrosomia, low birth weight newborns, or birth weight based on educational level.

4.1.2 Characteristics of Clusters

Cluster 1 (n = 739) was characterized by a higher proportion of women with lower educational levels, who were unemployed, overweight or obese, and had a lower mean MDS. In contrast, Cluster 2 (n = 773) consisted mainly of women with medium-to-high education,

who were employed, had normal weight, and had a higher average MDS. We compared the characteristics of women based on their cluster membership (Table 3). Women in Cluster 1 were younger ($p < 0.001$) and less likely to be primiparous ($p = 7.4006 \times 10^{-16}$) than those in Cluster 2. In terms of lifestyle, women in Cluster 1 had a lower proportion of non-smokers ($p = 2.5879 \times 10^{-8}$), lower adherence to the Mediterranean Diet ($p = 5.0 \times 10^{-6}$) , and a slightly lower average MDS ($p = 6.0572 \times 10^{-7}$) compared to those in Cluster 2. However, there were no statistically significant differences between the clusters in terms of dietary restriction adherence or total energy intake. Additionally, Cluster 1 had a significantly higher proportion of unemployed women compared to Cluster 2 (90.7% vs. 8.0%, $p = 1.1029 \times 10^{-226}$). Women in Cluster 1 also had a higher pre-pregnancy BMI ($p = 3.0 \times 10^{-6}$) and were more likely to be overweight or obese than those in Cluster 2 ($p = 1.42 \times 10^{-4}$). GWG was lower in Cluster 1 (11.0 kg) compared to Cluster 2 (12.0 kg, p = 0.017). Women in Cluster 1 reported a significantly higher proportion of preterm births (p = 0.004) and low birth weight newborns (p = 0.002) than those in Cluster 2. Additionally, Cluster 1 had a higher proportion of large-for-gestational-age (LGA) newborns, although this difference was not statistically significant (p = 0.074). Statistically significant differences between the clusters were also found in terms of week of delivery ($p = 2.00 \times 10^{-4}$), birth weight ($p = 4.67 \times 10^{-4}$) , and newborn length (p = 0.004). However, no significant differences were observed regarding macrosomia (p = 0.352). Table 1 shows the characteristics of the study sample by clusters [119].

4.1.3 Table and figures

Table 4.1.3.1. Characteristics of the study sample by employment status [119]

Characteristics	Employed (n = 780)	Not Employed (n = 732)	*p*-Value [a]
Age (years) [b]	32.0 (5.0)	30.0 (7.0)	<0.001
Primiparous	64.3%	44.4%	<0.001
Non-smoker	93.4%	86.9%	<0.001
Gestational week at delivery (weeks) [b]	39.0 (1.0)	39.0 (2.0)	<0.001
Adherence to MD			
Low	33.7%	42.5%	<0.001
Medium–high	66.3%	57.6%	
MDS [b]	4.0 (2.0)	4.0 (2.0)	<0.001
Dietary restrictions (% yes)	15.8%	14.6%	0.533
Total energy intake (kcal/day) [b]	1716 (545)	1716 (583)	0.607
Educational level			
Low	8.8%	41.8%	<0.001
Medium–high	91.2%	58.2%	
Pre-pregnancy BMI (kg/m^2) [b]	22.7 (5.3)	24.1 (7.0)	<0.001
Pre-pregnancy BMI classification			
Underweight	5.8%	5.6%	<0.001
Normal weight	64.1%	52.0%	
Overweight	19.9%	25.1%	
Obese	10.3%	17.2%	
GWG (kg) [b]	12.0 (7.0)	11.0 (8.2)	<0.001
GWG classification			
Reduced	37.7%	41.4%	0.345
Adequate	33.1%	31.5%	
Excessive	29.2%	27.1%	
Preterm birth	4.9%	7.8%	0.021
Low birth weight (% yes)	4.6%	8.6%	0.003
Macrosomia (% yes)	4.5%	4.2%	0.802
Birth weight (kg) [b]	3.32 (0.55)	3.24 (0.55)	<0.001
Birth length (cm) [b]	50.0 (2.0)	50.0 (2.0)	0.003

a p-values are obtained through the chi-squared test for qualitative variables and the Mann–Whitney U-test for quantitative variables. b Data are reported as the median (IQR). Abbreviations: MD, Mediterranean diet; BMI, body mass index; GWG, gestational weight gain

Table 4.1.3.2. Characteristics of the study sample by clusters [119]

Characteristics	Cluster 1 (n = 739)	Cluster 2 (n = 773)	*p*-Value [a]
Age (years) [b]	29.0 (7.0)	32.0 (5.0)	<0.001
Primiparous	44.1%	64.8%	<0.001
Non-smoker	85.9%	94.4%	<0.001
Gestational week at delivery (weeks) [b]	39.0 (2.0)	39.0 (2.0)	<0.001
Adherence to MD			
Low	43.4%	32.7%	<0.001
Medium–high	56.6%	67.3%	
MDS [b]	4.0 (2.0)	4.0 (2.0)	<0.001
Dietary restrictions (% yes)	13.9%	16.4%	0.177
Total energy intake (kcal/day) [b]	1715 (579)	1715 (540)	0.909
Employment status			
Employed	9.3%	92.0%	<0.001
Not employed	90.7%	8.0%	
Pre-pregnancy BMI (kg/m^2) [b]	23.9 (7.1)	22.8 (5.4)	<0.001
Pre-pregnancy BMI classification			
Underweight	5.5%	5.8%	<0.001
Normal weight	53.2%	63.1%	
Overweight	24.4%	20.6%	
Obese	16.9%	10.5%	
GWG (kg)[b]	11.0 (8.0)	12.0 (6.0)	0.017
GWG classification			
Reduced	40.6%	38.4%	0.619
Adequate	31.2%	33.3%	
Excessive	28.1%	28.2%	
Preterm birth	8.1%	4.6%	0.004
Low birth weight (% yes)	8.7%	4.5%	0.002
Macrosomia (% yes)	4.9%	3.9%	0.352
Birth weight (kg) [b]	3.2 (0.56)	3.3 (0.58)	<0.001
Birth length (cm) [b]	50.0 (3.0)	50.0 (2.0)	0.004
Birth weight for gestational age			
SGA	11.6%	8.6%	0.074
AGA	76.9%	81.5%	
LGA	11.5%	9.9%	

a p-values are obtained through the chi-squared test for qualitative variables and the Mann–Whitney U-test for quantitative variables. b Data are reported as the median (IQR)

4.2 Analysis of the Effect of Maternal Dietary Patterns on Birth Weight for Gestational Age

4.2.1 Study population

In this analysis, we applied principal component clustering to identify dietary patterns in pregnant women and assess their association with birth weight for gestational age. This approach, described in the methods, combines two multivariate techniques to reduce the dimensionality of dietary data and obtain a cluster solution. The analysis was conducted using data from the "MAMI-MED" cohort, which includes women recruited during the first-trimester visit at the Azienda di Rilievo Nazionale e di Alta Specializzazione (ARNAS) Garibaldi Nesima in Catania. This analysis included 667 women who completed their pregnancies and met the inclusion criteria. The median age was 31 years (IQR = 7), and 51.1% of the women were primiparous. In terms of socioeconomic factors, 24.9% had a high level of education, and 50.7% were employed. Nearly all women reported being non-smokers during pregnancy (91%) and had a median total energy intake of 1703 kcal (IQR = 508). Prior to pregnancy, the median BMI was 23.2 kg/m^2 (IQR = 5.8), with more than half of the women classified as normal weight (60.4%). Based on GWG (median = 12 kg; IQR = 8), 38.5% of women experienced reduced weight gain, and 28.7% had excessive weight gain. The median gestational week at delivery was 39 (IQR = 2), with 94% of births occurring at term. Regarding neonatal characteristics, the median birth weight was 3.3 kg (IQR = 0.6) and the median birth length was 50 cm (IQR = 2). Consequently, 81.9% of newborns were AGA, 7.1% were SGA, and 11.0% were LGA.

4.2.2 Derivation of Clusters Reflecting Distinct Dietary Patterns

After verifying the sampling adequacy and sphericity assumptions (KMO = 0.741 and p-value for Bartlett's test < 0.001), we applied PCA to the standardized and energy-adjusted dietary data. The PCA identified 16 PCs with eigenvalues greater than 1. Based on the scree plot inspection and the elbow method (Figure S1), we selected the first three PCs, which together explained 17.2% of the total variance. Figure 1 presents the factor loadings for each PC: PC1 was primarily associated with the intake of potatoes, cooked vegetables, legumes, and soup; PC2 was characterized by the consumption of potatoes, cooked and raw vegetables, fruit, and offal; PC3 was mainly linked to the intake of processed meats, dipping sauces, salty snacks, and fries. Subsequently, we conducted hierarchical clustering on PC1, PC2, and PC3, generating the dendrogram shown in Figure S2. Silhouette scores

were calculated for different clustering solutions, and the two-cluster solution was chosen (Figure S3). Figure 2 illustrates the distribution of participants according to the three main PCs and the clustering classification, while Figure 3 compares the average z-scores for each dietary category between clusters. Cluster 1 (n = 158) was characterized by higher intakes of potatoes, cooked and raw vegetables, legumes, fruits, nuts, yogurt, rice, wholemeal bread, white meat, offal, fish, eggs, butter and margarine, coffee, tea, and soup. Cluster 2 (n = 509) had higher intakes of milk, pasta, white bread, shellfish, vegetable and olive oils, sweets, fruit juices, dipping sauces, salty snacks, and fries. In terms of nutrients, participants in cluster 1 had higher intakes of folate, magnesium, and vitamins A, B6, and C, whereas those in cluster 2 had higher intakes of saturated and unsaturated fatty acids, calcium, and vitamin B1 (Figure S4).

4.2.3 Differences in Maternal Characteristics and Birth Outcomes according to Dietary Patterns

We then compared the characteristics of women based on their membership in one of the two clusters. Women in cluster 2 were younger ($p < 0.001$) and had a lower educational level (p = 0.018) compared to those in cluster 1. Additionally, they reported a higher total energy intake ($p < 0.001$). No other significant differences were observed between the two clusters. Similarly, there were no differences in birth outcomes, including gestational week at delivery, preterm birth rates, birth weight, and birth length (Table 1). However, a significant difference was observed in birth weight for gestational age. Specifically, a higher proportion of LGA infants were born to women in cluster 2, while women in cluster 1 had higher proportions of SGA and AGA infants (p = 0.030; Figure 4).

4.2.4 Factors Associated with Birth Weight for Gestational Age

Finally, we assessed the main factors associated with birth weight for gestational age (Table 2). For SGA, the key factors identified were employment status and primiparity. Newborns of employed women had lower odds of being SGA compared to those of unemployed women (OR = 0.359; 95% CI = 0.168–0.769; p = 0.008). In contrast, primiparous women had higher odds of having SGA infants compared to those with prior pregnancies (OR = 2.681; 95% CI = 1.293–5.558; p = 0.008). For LGA, the main factors associated were cluster membership and pre-pregnancy BMI. Specifically, women in cluster 2 had higher odds of delivering LGA infants compared to women in cluster 1 (OR = 2.213; 95% CI = 1.047–4.679; p = 0.038). Additionally, the odds of LGA increased by approximately 11% for each unit increase in pre-pregnancy BMI (OR = 1.107; 95% CI = 1.053–1.163; $p < 0.001$) [111].

4.2.5 Table and figures

Figure 4.2.5.1. Factor loadings of the three main principal components. Green and red bars represent food items that positively or negatively characterized the principal components [111]

Figure 4.2.5.2. Distribution of participants by principal components and clusters [111]

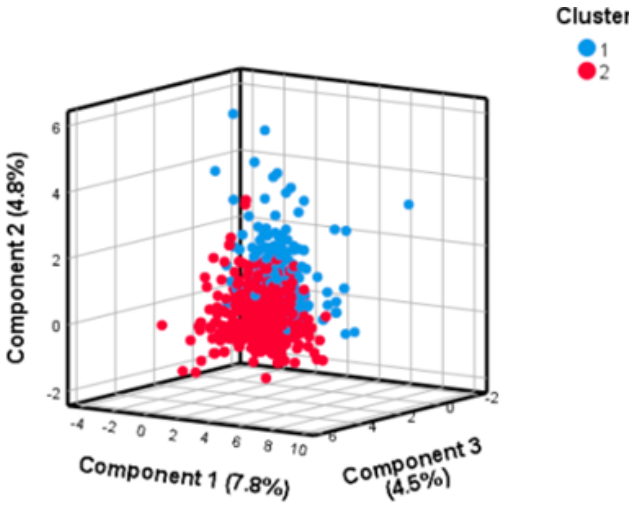


Figure 4.2.5.3. Comparison of dietary intakes between clusters (158 vs. 509 participants). Green bars represent food items that positively characterized the cluster, while red bars represent food items that negatively characterized the cluster. According to Student's t-test, results are reported as *** for p-values < 0.001; ** for p-values < 0.01; * for p-values < 0.05 [111]

Figure 4.2.5.4. Distribution of birth weight for gestational age by clusters [111]

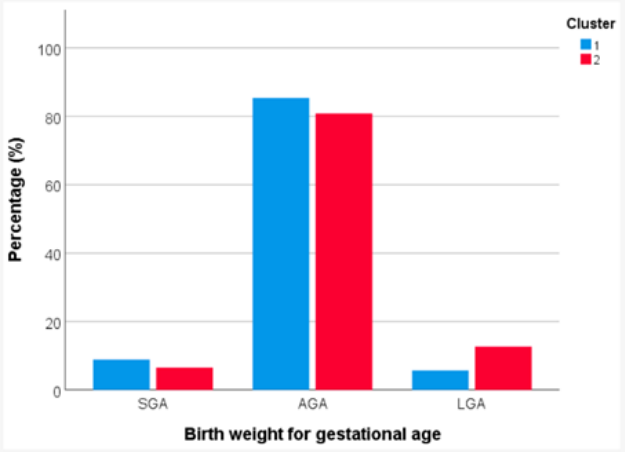


Table 4.2.5.1. Characteristics of the study population by clusters reflecting dietary patterns [111]

Characteristics	Cluster 1 (*n* = 158)	Cluster 2 (*n* = 509)	*p*-Value [a]
Age (years) [b]	32.0 (5.0)	30.0 (7.0)	<0.001
High educational level	29.7%	23.4%	0.018
Employed	55.1%	49.3%	0.207
Non-smoker	94.9%	89.8%	0.055
Primiparous	46.5%	52.5%	0.191
Total energy intake (kcal/day) [b]	1567 (486)	1749 (503)	<0.001
Pre-pregnancy BMI (kg/m^2) [b]	23.5 (5.4)	23.2 (5.9)	0.373
Pre-pregnancy BMI classification			
Underweight	5.7%	5.3%	
Normal weight	58.6%	60.9%	0.965
Overweight	22.3%	21.2%	
Obese	13.4%	12.6%	
GWG (kg) [b]	11.0 (8.3)	12.0 (8.0)	0.272
GWG classification			
Reduced	42.9%	37.2%	
Adequate	27.9%	34.3%	0.289
Excessive	29.2%	28.5%	
Gestational week at delivery (weeks) [b]	39.0 (2.0)	39.0 (2.0)	0.489
Preterm birth	8.3%	5.3%	0.174
Birth weight (kg) [b]	3.2 (0.6)	3.3 (0.6)	0.171
Birth length (cm) [b]	50.0 (2.0)	50.0 (2.0)	0.233

a p-values are obtained through the Chi-squared test for qualitative variables and the Mann–Whitney U-test for quantitative variables. b Data are reported as median (IQR).

Abbreviations: BMI, Body Mass Index; GWG, Gestational Weight Gain.

Table 4.2.5.2. Factors associated with birth weight for gestational age [111]

Characteristics	SGA		LGA	
	OR (95%CI)	*p*-Value	OR (95%CI)	*p*-Value
Cluster 2 vs. Cluster 1	0.537 (0.262–1.104)	0.091	2.213 (1.047–4.679)	0.038
Age (continuous)	0.965 (0.894–1.041)	0.356	0.955 (0.899–1.014)	0.132
Pre-pregnancy BMI (continuous)	1.003 (0.939–1.071)	0.934	1.107 (1.053–1.163)	<0.001
GWG (continuous)	0.966 (0.928–1.005)	0.089	1.030 (0.997–1.064)	0.075
High educational level	1.060 (0.617–1.821)	0.834	1.154 (0.754–1.767)	0.509
Employed	0.359 (0.168–0.769)	0.008	0.745 (0.414–1.341)	0.327
Primiparous	2.681 (1.293–5.558)	0.008	0.980 (0.563–1.704)	0.942
Smoker	1.841 (0.697–4.865)	0.218	0.352 (0.102–1.214)	0.098
Total energy intake (continuous)	1.000 (1.000–1.001)	0.207	1.000 (1.000–1.001)	0.474

Abbreviations: SGA, Small for Gestational Age; LGA, Large for Gestational Age.

Results are obtained by applying logistic regression models including dietary patterns (cluster 2 vs. Cluster 1), age (continuous), pre-pregnancy BMI (continuous), GWG (continuous), educational level (high vs. low-medium), employment status (employed vs. unemployed), primiparity (primiparous vs. non-primiparous), smoking status (smoker vs. non-smoker), and total energy intake (continuous).

4.3 Analysis of Causal Graph to investigate the causal connection between pre-pregnancy BMI, GWG, and telomere length in amniotic fluid

4.3.1 Characteristics of the study population

Before conducting the main analysis using causal graph modeling, a preliminary analysis was performed using binary classification models to explore the most influential predictors of gestational weight gain (GWG). Specifically, Decision Tree, Random Forest, and XGBoost classifiers were trained to classify GWG as adequate or non-adequate, using maternal and neonatal characteristics as predictors. The models exhibited varying accuracy rates, with XGBoost achieving the highest performance (65%), followed by Random Forest (63%) and Decision Tree (60%). To interpret the contribution of each predictor in the classification, the SHapley Additive Explanations (SHAP) method was employed. SHAP analysis revealed that telomere length in amniotic fluid was the most influential factor in GWG classification, even after adjusting for highly correlated variables such as pre-pregnancy BMI, gestational weight, and adherence to the Mediterranean Diet. This finding suggests a potential link between telomere biology and gestational weight adequacy, providing a basis for the subsequent causal analysis. Based on these preliminary results, a causal analysis using causal graph modeling was then conducted to further explore the potential causal relationships between BMI, GWG, and telomere length in amniotic fluid, accounting for confounding factors such as age and total energy intake. The study included 136 mothers from the "Mamma & Bambino" cohort who met the inclusion criteria. This cohort invites all pregnant women receiving prenatal genetic counseling at the Azienda Ospedaliero Universitaria Policlinico "G. Rodolico - San Marco" in Catania. Mothers are recruited between the 4th and 20th weeks of gestation. Exclusion criteria include multiple pregnancies, pre-existing autoimmune and/or chronic diseases, pregnancy complications (e.g., preeclampsia, gestational hypertension, and gestational diabetes), intrauterine fetal death, and congenital malformations. The analysis included a total of 136 mothers who met the inclusion criteria. Table 1 outlines the characteristics of the study population, while Figure 1 illustrates the distribution of women based on pre-pregnancy BMI and GWG categories. Specifically, the median pre-pregnancy BMI was 22.3 kg/m2 (IQR = 4.0), with women classified as underweight (6.6%), normal weight (77.9%), or overweight (15.4%). At delivery, the median gestational weight gain (GWG) was 12.0 kg (IQR = 6.0). Considering both pre-pregnancy BMI and GWG, the women were further categorized into reduced (29.4%), adequate (42.6%), or excessive GWG (27.9%).

Additionally, in this study, reduced and excessive GWG were occasionally combined into a single category referred to as non-adequate GWG. A significant association between pre-pregnancy BMI and GWG was observed ($p = 0.023$), with normal weight women being more likely to have adequate GWG. In contrast, the proportion of reduced and excessive GWG was higher among underweight and overweight women, respectively (Fig. S1).

4.3.2 Relationships between pre-pregnancy BMI, GWG, and TL

Figure 2 illustrates the bivariate associations between pre-pregnancy BMI and GWG with TL. In both cases, the relationship follows a characteristic inverted U-shaped pattern, which was more pronounced for GWG (Fig. 2 A,B). This finding was previously highlighted in our earlier study, as cited in Ref.32. When considering the BMI categories, overweight women showed a lower TL compared to underweight and normal weight individuals, although this difference was not statistically significant (Fig. 2C). For the GWG categories, women with reduced and excessive GWG had a lower TL compared to those with adequate GWG. However, similar to the previous finding, this difference was not statistically significant (Fig. 2D). Figure S2 shows the relationships between TL and both maternal age and total energy intake.

4.3.3 Causal Graph Model definition

Building on previous research (7,30,32,48,49) and the current findings, we developed three distinct cDAGs to represent potential causal relationships among the variables of interest (Fig. 3). These models were rigorously tested using dependency and conditional dependency tests, confirming their compatibility with the observed data. Importantly, all three models align with the data and do not present any inconsistencies. The analysis initially began with the simplest graphical model (G1), which included the variables for GWG, BMI, and TL. However, from a biological perspective, it was acknowledged that factors such as total energy intake and age cannot be overlooked in relation to GWG, BMI, and TL (7,30,32,48,49). As a result, the analysis was extended to include graphical models G2 and G3. These models differ primarily in how they incorporate the effects of age on pre-pregnancy BMI and GWG. This distinction is due to existing limitations and conflicting evidence regarding these relationships. For all three graphical models, two different approaches were used to analyze the GWG variable: one treated GWG as a binary factor, distinguishing between adequate and non-adequate GWG, and the other categorized GWG into three groups: reduced, adequate, and excessive. Additionally, TL was consistently categorized as "long" if its value exceeded the median and as "short" otherwise.

4.3.4 Potential Causal Effect of GWG on TL

By assessing the differences in probabilities, denoted as Prob (TL = t | do(GWG = g)), across all GWG categories, we successfully estimated the potential causal effect of GWG on TL. The causal effect estimation, with GWG divided into two categories (adequate and non-adequate), is shown in Fig. 4. The estimated causal effect values are provided for all three graphical models (G1, G2, and G3). As seen in Fig. 4, transitioning from non-adequate to adequate GWG positively influences the probability of having a "long" TL. In model G1 (Fig. 4A), this increase in probability is quantified at 15%. In models G2 and G3 (Fig. 4B, C), the same increase is estimated at 14% and 12%, respectively. The estimated causal effect of GWG on TL, considering three categories (reduced, adequate, and excessive GWG), is shown in Fig. 5 for all three models. Consistent with the results in Fig. 4, the plots in Fig. 5 illustrate that transitioning from reduced or excessive GWG to adequate GWG has a positive impact on the likelihood of a "long" TL.

4.3.5 Potential Causal Effect of pre-pregnancy BMI on TL

We then estimated the potential causal effect of pre-pregnancy BMI on TL in graphical models G1 and G3 by calculating the differences in probabilities, denoted as Prob(TL = t| do(BM = b), GWG = g), where b represents all BMI categories (underweight, normal, overweight) and g represents the adequate/non-adequate GWG classes. For graphical model G2, we estimated the causal effect of pre-pregnancy BMI on TL by evaluating the probability value Prob(TL = t| do(BMI = b), do(GWG = g)). The results for all three models (G1, G2, and G3) are presented in Fig. 6. These findings show that the impact of BMI on TL differs based on the inclusion of Age and Total energy intake. In Fig. 6A, we observe that transitioning from underweight to normal and then to overweight BMI decreases the likelihood of having a "long" TL, regardless of the GWG class. However, Fig. 6B and C reveal different causal effects of BMI on TL within the two GWG categories. In the non-adequate GWG class, moving from underweight to overweight reduces the probability of having a "long" TL. In contrast, in the adequate GWG class, transitioning from underweight or overweight to normal BMI increases the probability of a "long" TL [112].

4.3.6 Table and figures

Table 4.3.6.1. Characteristics of the study population [112]

Characteristics	Median (interquartile range) or frequency (percentage)
Maternal age, years	37 (3)
Gestational age at sampling, weeks	16 (5)
Total energy intake, kcal	1862 (631)
Pre-pregnancy BMI, kg/m^2	22.2 (4.0)
Underweight	9 (6.6%)
Normal weight	106 (77.9%)
Overweight	21 (15.4%)
Gestational age at delivery, weeks	39 (2)
Gestational Weight Gain, Kg	12.3 (6.0)
Reduced GWG	40 (29.4%)
Adequate GWG	58 (42.6%)
Excessive GWG	38 (27.9%)

Figure 4.3.6.1. Distribution of Pregnant Women Based on Pre-Pregnancy BMI and Gestational Weight Gain (GWG): (A) BMI Categories and (B) GWG Categories [112]

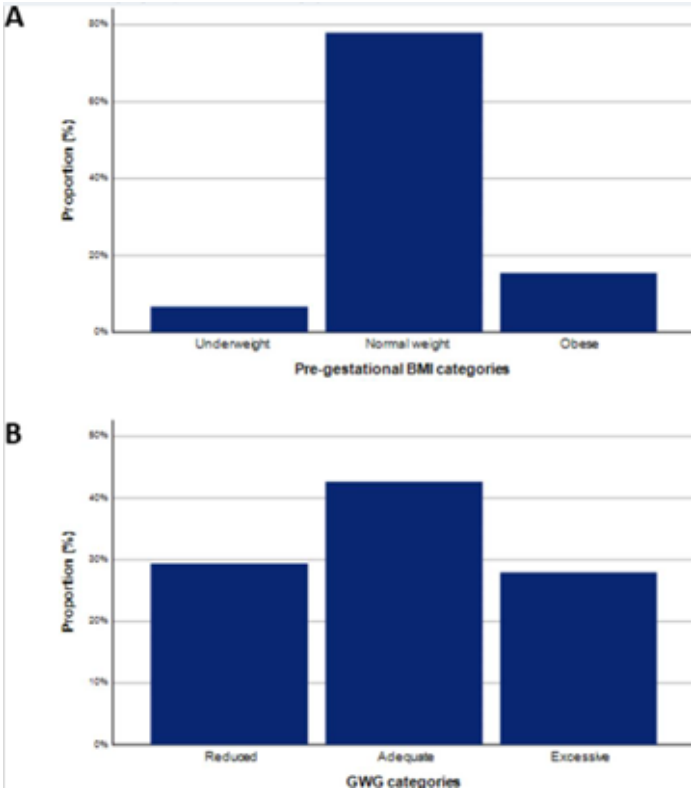


Figure 4.3.6.2. Relationships Between Pre-Pregnancy BMI, Gestational Weight Gain (GWG), and Telomere Length: Continuous and Categorical Analyses [112]

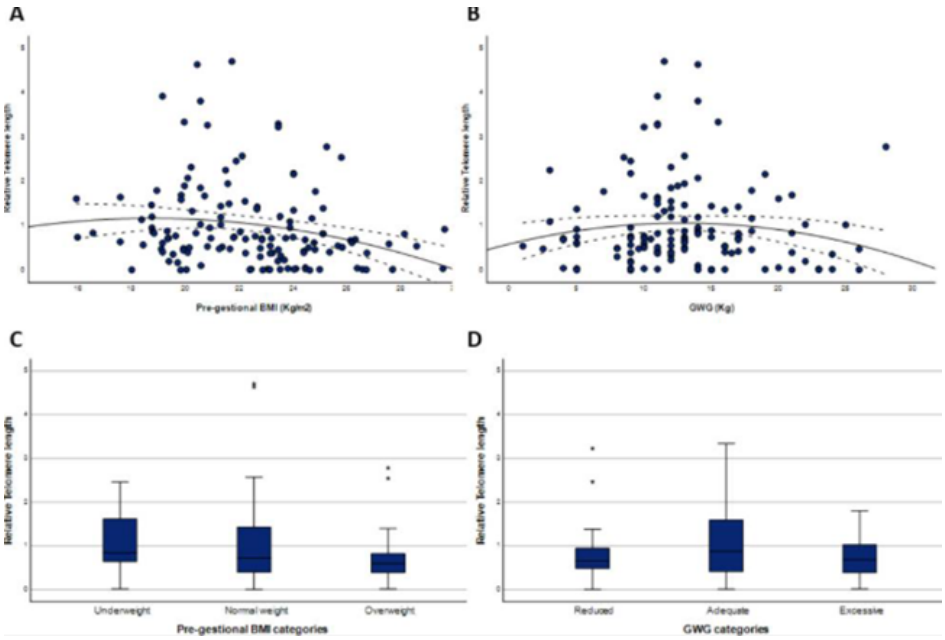
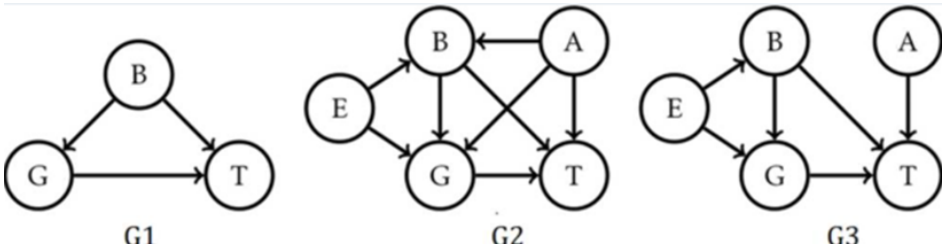
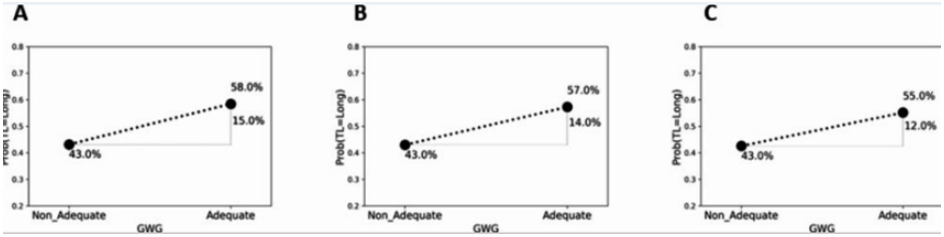


Figure 4.3.6.3. Directed Acyclic Graphs (cDAG) Models to Represent Causal Relationships Between Maternal Variables, Gestational Weight Gain, and Telomere Length [112]



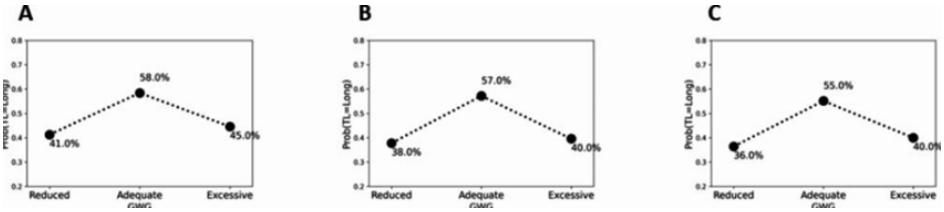
Three graphical models (G1, G2, and G3) were developed, where the variables GWG, BMI, TL, age, and total energy intake were denoted as G, B, T, A, and E, respectively. These models provided a structured representation of the causal connections among the variables in a direct and acyclic manner.

Figure 4.3.6.4. Causal effect of adequate gestational weight gain (GWG) on telomere length (TL) using graphical models, comparing adequate vs. non-adequate GWG [112]



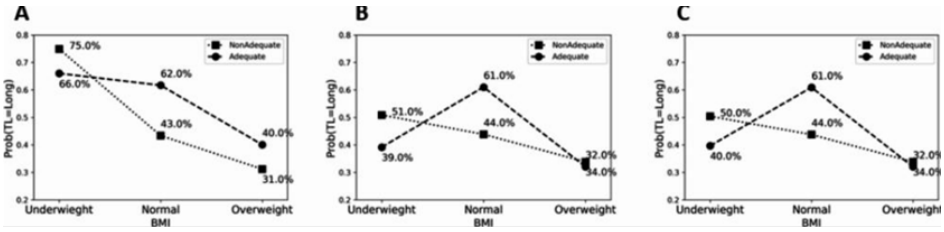
The three graphical models, labelled as G1 (A), G2 (B), and G3 (C), depict the impact of GWG categorized as adequate and non-adequate on TL.

Figure 4.3.6.5. Causal effect of gestational weight gain (GWG) on telomere length (TL) using graphical models, comparing reduced, adequate, and excessive GWG [112]



The three graphical models, labelled as G1 (A), G2 (B), and G3 (C), depict the impact of GWG categorized as reduced, adequate, and excessive on TL.

Figure 4.3.6.6. Causal effect of pre-pregnancy BMI on telomere length (TL) using graphical models [112]



The three graphical models, labelled as G1 (A), G2 (B), and G3 (C), showcase the causal effects of BMI on TL. The effects are examined for two categories of gestational weight gain (GWG): adequate (represented by a dashed line) and non-adequate (represented by a dotted line).

4.4 Analysis of sex differences in delivery and neonatal characteristics of new-borns from the "MAMI-MED" cohort

4.4.1 Characteristics of the study population

This analysis is based on data from the "MAMI-MED" cohort and aims to evaluate the impact of social, environmental, behavioral, and molecular factors on maternal and child health. Specifically, multivariable linear regression analyses were conducted to examine the relationship between sex and birth length and weight, accounting for potential confounding factors, including maternal age, gestational age at delivery, pre-pregnancy BMI, GWG, total energy intake, smoking status, and mode of delivery. Understanding these disparities can provide valuable insights into the biological and environmental mechanisms influencing perinatal health, as sex-related differences in neonatal outcomes are a well-documented phenomenon in perinatal epidemiology. The analysis involved 1,103 mothers who completed their pregnancies and were recruited into the cohort during the first-trimester visit at the Azienda di Rilievo Nazionale e di Alta Specializzazione (ARNAS) Garibaldi Nesima in Catania. Table 1 presents the main characteristics of these women, with a median age of 31 years (IQR = 7) and a median gestational age at recruitment of 12 weeks (IQR = 0). In terms of socioeconomic status, 50.3% of the women had a medium level of education, while 25.7% and 24.0% reported low and high education levels, respectively. Additionally, 51.4% of the women were employed. Based on pre-gestational BMI, 6.0% were underweight, 58.3% had normal weight, 22.3% were overweight, and 13.4% were obese. Regarding pre-gestational BMI and gestational weight gain (GWG), 39.3% and 27.9% of women had reduced or excessive GWG, respectively. In total, 13 women successfully completed a twin pregnancy. Therefore, the analysis included delivery and neonatal data for 1,116 newborns. Of these, 50.7% were male (N=567). Regarding delivery data, the median gestational age at birth was 39.0 weeks (IQR=2), with 7.8% preterm births. Neonatal characteristics showed a median birth weight of 3.3 kg (IQR=0.6) and a median birth length of 50.0 cm (IQR=2). Furthermore, macrosomia was observed in 3.9% of the newborns. Notably, 80.6% of the newborns were AGA (appropriate for gestational age), while 8.6% and 10.8% were SGA (small for gestational age) and LGA (large for gestational age), respectively. Delivery and neonatal characteristics by sex are shown in Figure 1. The birth weight distribution comparison revealed some sex differences, with males having higher values (Median birth weight = 3.3 kg; IQR=0.6) compared to females (Median birth weight = 3.2 kg; IQR=0.6; $p < 0.001$)

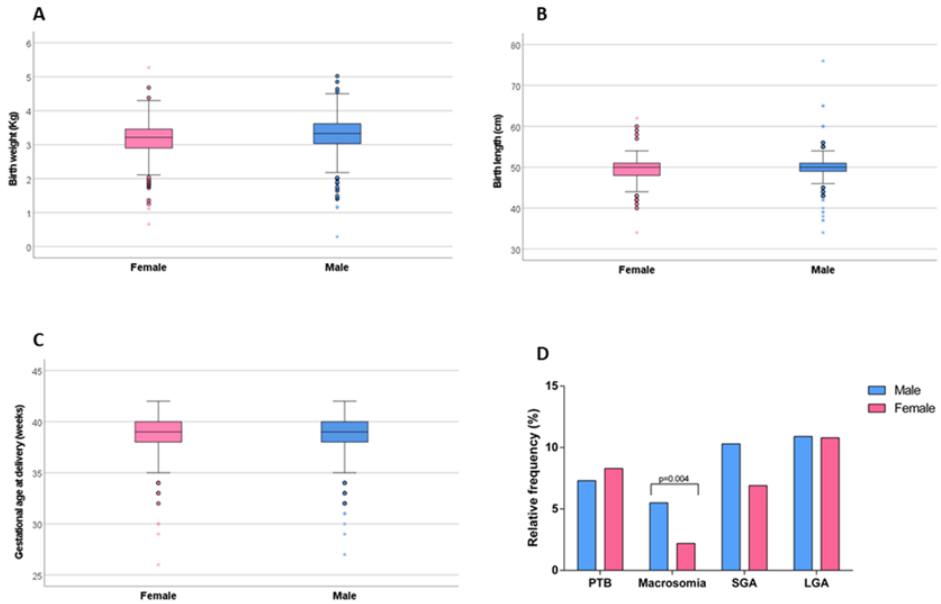
(Figure 1A). The birth length distribution also showed significant differences between sexes ($p < 0.001$), although the median value was the same for both males and females (50.0 cm) (Figure 1B). For birth weight, a higher proportion of males had macrosomia compared to females (5.5% vs. 2.2%; p=0.004). No sex differences were found for gestational age at delivery, preterm births (PTB), or weight for gestational age (Figures 1C-1D). To further assess sex differences in delivery and neonatal characteristics, we performed regression analyses. Regarding neonatal anthropometric measures, we first observed that males had greater birth length than females ($\beta = 0.631$; 95% CI = 0.296 – 0.965; $p < 0.001$). This result was confirmed after adjusting for maternal age, gestational age at delivery, pre-gestational BMI, GWG, and total energy intake ($\beta = 0.644$; 95% CI = 0.353 – 0.936; $p < 0.001$). Similarly, the linear regression model showed that male newborns had higher birth weight compared to females ($\beta = 0.112$; 95% CI = 0.052 – 0.173; $p < 0.001$). The positive association between male sex and birth weight remained significant after adjusting for maternal age, gestational age at delivery, pre-gestational BMI, GWG, and total energy intake ($\beta = 0.121$; 95% CI = 0.071 – 0.170; $p < 0.001$). Logistic regression models were then performed to assess sex differences in delivery outcomes. Comparing SGA to AGA, the likelihood of being SGA was nearly 1.6 times higher in males than females (OR=1.569; 95% CI = 1.016 – 2.422; p = 0.042). This positive association was confirmed after adjusting for potential confounders (OR=1.628; 95% CI = 1.045 – 2.537; p = 0.031). In contrast, no association with LGA was found. Similarly, no sex differences were observed for gestational age at delivery or preterm births.

4.4.2 Table and figures

Table 4.4.2.1. Characteristics of mothers enrolled in the "MAMI-MED" cohort

Characteristics (N=1103)	Median (IQR) or frequency (%)
Age (years)	31.0 (7)
Gestational age at recruitment (weeks)	12.0 (0)
Educational level	
Low	25.7%
Medium	50.3%
High	24.0%
Employed (% yes)	51.4%
Smokers (% yes)	9.7%
Parity (% yes)	47.1%
Pre-gestational Body Mass Index	23.3 (6.0)
BMI-categories	
Underweight	6.0%
Normal weight	58.3%
Overweight	22.3%
Obese	13.4%
Gestational weight gain (GWG)	11.0 (7.0)
GWG categories	
Reduced	39.3%
Adequate	32.8%
Excessive	27.9%
Total energy intake (Kcal)	1691.7 (568.9)

Figure 4.4.2.1. Sex differences in delivery and neonatal characteristics. These plots show the comparison between sex for birth weight (A), birth length (B), gestational age at delivery (C) and other characteristics (D)



4.5 Systematic review on the application of artificial intelligence in studying causality in public health

4.5.1 Study Selection

From a total of 28,230 articles published up to August 2023, after removing duplicates (11,158) and articles that did not meet the inclusion criteria (16,168 post title/abstract screening, a total of 904 articles were identified from the databases. Following full-text screening, 856 studies were further excluded based on our selection criteria. Thus, a total of 48 articles are included in the present systematic review. This systematic study presented a general review of AI-driven causal methodologies used in various domains within Public Health, encompassing disease prediction, evaluation of treatment efficacy, identification of risk factors, analysis of health behaviors, precision Public Health initiatives, and environmental health assessment.

4.5.2 General Characteristics of Included Studies

The included studies were ranged from the year 2004 to 2020. The included studies presented different AI-driven causal methodologies: Bayesian Additive Regression Trees (BART) [120, 121, 122, 123, 124, 125, 126], Bayesian network (BN) models [127, 128, 129, 130, 131, 132, 133, 134], causal forests analysis [135, 136, 137, 138, 139, 140, 141, 142], Graphical causal models [143, 144, 145, 146, 147, 148, 149], Causal discovery algorithm [150, 151], Balancing Covariates Automatically Using Supervision (BCAUS) [152], G-computation causal inference method [153, 154], Causal Shapley values [155], Decision tree [156], The Granger causality (GC) test algorithm [157, 158], Maximum likelihood estimation (TMLE) technique and Super Learner [159, 160, 161, 162, 163], LiNGAM [164, 165, 166] and Greedy Fast Causal Inference [167]. The following sections explore various approaches proposed in the literature, highlighting their potential applications and key findings from Public Health studies.

4.5.3 AI algorithms

Probabilistic Models

Bayesian Network (BN) models

BN models are probabilistic graphical models that integrate probability theory and graph theory to represent relationships between variables. Structured as directed acyclic graphs (DAG), the nodes represent random variables, while the directed edges represent dependencies or conditional probability distributions. BNs can model complex structures with

causal dependencies, especially in contexts of uncertainty and missing data. Among their advantages are transparency and interpretability, which facilitate both diagnostic and causal analysis. BN models have been utilized to predict the likelihood of a causal association between drug exposure and adverse events in individual case safety reports (ICSRs). This application demonstrated high performance, with an Area Under the Curve (AUC) of 0.924, indicating that probabilistic modeling is effective in aiding causality assessments and improving safety decision-making in drug-event pair evaluations [127]. In a detailed study of cardiovascular diseases, BN models were employed to explore the relationships among various cardiovascular risk factors (CVRF). The analysis focused on epidemiological features such as cardiovascular lost years (CVLY), cardiovascular risk scores (CVRS), and metabolic syndrome (MetS). The model helped identify both direct and indirect relationships, revealing that smoking significantly influenced CVLY, while physical activity had a strong impact on MetS [128]. The Dynamic Bayesian Network (DBN) model has found application in biomedical informatics to uncover probabilistic causal chains from temporal datasets. This approach assists in the diagnosis, therapy, and prognosis of diseases. For instance, causal graphs were constructed from entropy and causal tendency measures to predict future trends, such as the progression of Amyotrophic Lateral Sclerosis (ALS). The model, tested on the PRO-ACT dataset, demonstrated comparable or superior performance to other machine learning methods [130]. Another significant use of BN models is in the analysis of prognostic factors for acute leukemia. A decision model was developed based on the causal relationships among these factors, aiming to assist medical specialists. This model helps in predicting and supporting diagnoses prior to costly and time-consuming procedures like bone marrow sampling and pathological tests [131]. In research focused on localized prostate cancer, the BN model was employed to analyze the treatment effects of active treatment versus observation. By utilizing structure learning and excluding non-causal relationships, the model was able to identify confounders such as the year of diagnosis and age through graph analysis, proving valuable for estimating treatment effects [132]. Lastly, the BN model was used to analyze the potentially causal relationships between the symptoms of obsessive-compulsive disorder (OCD) and depression in adult patients. Two networks were created to identify key factors of comorbidity, revealing that the distress associated with obsessions and compulsions is a primary driver, with sadness acting as a bridge between the two symptom clusters [133]. In addition, one study employed the Bayesian network method through the BLCD algorithm to identify causal relationships related to infant mortality in the United States. Analyzing the U.S. Linked Birth/Infant Death dataset from 1991, the algorithm identified six potential causal relationships, three of which appeared plausible. The authors plan to further explore novel causal pathways using the full sample [129]. Another study used a modified Local

Causal Discovery (LCD2) algorithm, also within the BN framework, to investigate factors contributing to infant mortality. This study analyzed the same dataset and identified nine potential causal relationships, eight of which seemed plausible, with plans to refine the algorithm for better efficiency and explore additional causal pathways [134]. Through these diverse applications, Bayesian Network models showcase their versatility and effectiveness across various fields, from healthcare decision-making to understanding complex disease interactions.

Bayesian Additive Regression Trees

Bayesian Additive Regression Trees is a Bayesian machine learning algorithm that uses an ensemble of regression trees to flexibly and non-parametrically model conditional expectations. This model is particularly effective for estimating causal effects by reducing parametric assumptions and capturing complex, nonlinear relationships and interactions. BART is especially useful in causal inference, as it can impute missing counterfactual outcomes and provide more accurate estimates with precise uncertainty intervals. Several studies have applied BART in various contexts. BART has been applied in various research contexts, demonstrating its versatility and effectiveness in handling complex, non-linear relationships. In one study, BART was employed to assess the public health impacts of the Clean Air Act Amendments of 1990. By analyzing population-level data, BART provided a precise estimation of the number of adverse health events prevented due to reduced pollution exposure, offering a more comprehensive and data-driven analysis of multiple pollutants compared to traditional methods [120]. In another study, BART was integrated into the stan4bart algorithm to address the challenges of estimating treatment effects in grouped data. This combination of BART's flexibility with Stan's computational efficiency allowed for more accurate estimates of both average and heterogeneous treatment effects, outperforming other methods that do not account for the multilevel structure of the data [121]. BART was also used to evaluate the effects of different dexamethasone doses in severe COVID-19 patients. The analysis revealed that a 12 mg/day dose was more beneficial than 6 mg/day for certain subgroups, such as older patients and those without diabetes, highlighting the potential for personalized treatment strategies in this context [122]. In pediatric research, BART was applied to estimate the short-term effectiveness of 13 common orthopedic and neurological treatments for children with cerebral palsy. The study found that these treatments had moderate effects on body structures but minimal impact on gait and functional mobility, pointing to the need for further investigation into the sources of heterogeneous treatment effects in this population [123]. BART also helped assess the relationship between chronic cannabis use and the risk of postoperative nausea and vomiting (PONV). The analysis showed that daily cannabis use was associated with a 19% increase in the relative risk of PONV, with

results validated across both internal and external datasets, illustrating BART's capacity for accurate risk quantification [124]. Finally, BART was used to study the causal effects of altered motor control on gait in children with cerebral palsy undergoing multi-level orthopedic surgery. While altered motor control had a strong effect on preoperative gait patterns, its impact on post-surgical changes was minimal, suggesting the involvement of other factors in determining post-surgical outcomes [125]. In the context of lung cancer screening, BART was specifically applied to assess the treatment effects of low-dose computed tomography (LDCT) compared to chest radiography (CXR) for lung cancer mortality. Using the accelerated failure time BART model, the study found no significant mortality benefit for LDCT over CXR overall. However, specific subpopulations, particularly Asian and Black individuals with a significant smoking history, demonstrated enhanced mortality benefits from LDCT. These findings underscore the value of flexible machine learning approaches like BART in informing personalized treatment decisions and clinical trial planning [126].

Causal Shapley values

Causal Shapley values, derived from cooperative game theory and enhanced with do calculus, provide a framework for explaining the contributions of different variables in a causal context. This method is particularly useful in understanding the causal connections and variable importance within complex systems, such as those modeled by machine learning. In this study, causal Shapley values were applied to analyze socioeconomic disparities linked to the spread of COVID-19 in the USA. By examining various phases of the disease's progression, the researchers highlighted how causal relationships evolved over time. They utilized non-linear machine learning models to capture complex, non-linear correlations that traditional linear models might miss. The incorporation of causal Shapley values allowed for a nuanced understanding of how different socioeconomic factors contributed to the pandemic's spread, while also providing a way to validate findings against random effects models. Overall, the study demonstrated the effectiveness of causal Shapley values in quantifying and explaining the importance of various socioeconomic variables in relation to COVID-19, showcasing the advantages of non-linear modeling in revealing intricate causal relationships [155].

Linear Non-Gaussian Acyclic Model (LiNGAM)

The Linear Non-Gaussian Acyclic Model is a statistical method for causal inference in machine learning that distinguishes between correlation and causation. Unlike traditional models, LiNGAM identifies causal relationships by using non-Gaussian probability distributions and structural equation models, often represented as DAGs. It assumes that each variable is a linear combination of its causal predecessors plus an exogenous variable, making it particularly effective for analyzing complex data sets. The LiNGAM-beta extension allows for quantitative assessment of causal strengths and directions, which is especially useful

in health data analysis. In a study on NAFLD, LiNGAM-beta was used to estimate how various checkup items influence disease progression. The approach integrated collaborative filtering and sampling techniques to handle missing values, revealing significant causal links to metabolic syndrome and age-related factors [164]. Another application involved analyzing extensive health checkup data from Osaka prefecture between 2012 and 2017 using the DirectLiNGAM algorithm. This study uncovered meaningful causal relationships among various health indices across different age groups and genders, demonstrating LiNGAM's effectiveness in health data analysis [165]. A study explored the causal link between television viewing time and weight gain, particularly waist circumference and BMI, utilizing LiNGAM among other methods. The findings indicated that high TV time causally contributes to increased waist and BMI, highlighting the health implications of prolonged viewing habits [166]. Overall, LiNGAM has proven to be a versatile tool in identifying and quantifying causal relationships in various health-related contexts.

Graphical Models

Graphical causal models

Graphical Causal Models, particularly those based on DAGs, serve as powerful tools for representing and analyzing causal relationships among variables. In these models, each node signifies a variable, while directed edges illustrate causal links, ensuring a clear direction of causality without feedback loops. Such models adhere to specific conditions, like the causal Markov condition, allowing researchers to incorporate prior knowledge and make both qualitative and quantitative causal inferences, even in the presence of incomplete information. A study used a dynamic uncertain causality graph to develop a diagnostic system for jaundice, achieving an accuracy of 99.01% through a chaining inference algorithm, demonstrating greater explainability compared to traditional methods [143]. Similarly, another study created a computer-aided diagnostic tool for dyspnea, also based on a dynamic uncertain causality graph with 132 variables, reaching a diagnostic accuracy of 96.5% and improving the efficiency of differential diagnoses [70]. In one article, Graphical Causal Models (specifically DAGs) were employed to visually and conceptually represent the relationships between treatment, mediator, and outcome in a causal mediation analysis. The focus was on estimating both the indirect effect of a binary treatment (via a mediator) and the direct effect of the treatment itself, while controlling for confounders in a high-dimensional setting. Using double machine learning, the study aimed to address potential model misspecifications and prevent overfitting by leveraging data splitting. The method was applied to the US National Longitudinal Survey of Youth, where the indirect effect of health insurance on general health through routine checkups was assessed, alongside the direct effect of health insurance [148]. A causal inference framework enhanced with machine learning was used to investi-

gate the relationship between COVID-19 severity and environmental factors in 166 Chinese cities, revealing only one significant causal relationship regarding air temperature [145]. An article used graphical causal models to explore causal relationships between gender, birth weight, waist circumference, and blood glucose levels in 4,081 participants, demonstrating the methodology's potential in revealing causal links in observational data [146]. Additionally, a study examined the impact of early BMI rebound on subsequent cardiometabolic outcomes in 649 children, using targeted maximum likelihood estimation to reveal protective effects of certain interventions, highlighting the importance of rigorous causal inference [147]. Finally, employing a credal network, research analyzed factors influencing the place of death for terminal cancer patients, emphasizing the significant impact of family preferences and advocating for better practices in palliative care [149].

Greedy Fast Causal Inference (GFCI)

The Greedy Fast Causal Inference (GFCI) algorithm is a two-step method for making causal inferences and representing them in DAGs. It begins by exploring potential causal relationships among measured variables using Fast Greedy Equivalence Search (FGES), which evaluates how well each model fits the data. After identifying preliminary causal links, GFCI refines the model through conditional independence tests to eliminate unsupported relationships, resulting in a more accurate causal graph. This method is efficient for large datasets and has been applied in various fields. In a study focusing on internalizing psychopathology and alcohol use disorder (AUD) in patients with co-occurring anxiety disorders, GFCI was used to analyze data from 362 adult AUD treatment patients. The researchers identified two distinct causal paths leading to drinking behavior, revealing that drinking to cope (DTC) with negative affect served as a central hub. One path was direct, stemming from social anxiety, while the other was indirect, linked to perceived stress. This application of GFCI enhanced understanding of how internalizing disorders influence drinking behaviors in AUD patients [62].

Causal discovery algorithms

Causal Discovery algorithms are computational methods designed to uncover the underlying causal relationships within datasets by distinguishing cause-and-effect links through statistical tests, going beyond traditional regression analyses that only measure associations. These algorithms are particularly valuable in fields like epidemiology and biomedical research, where understanding causal mechanisms is crucial. They build a network of causes and effects, identifying which variables influence others. In two studies, the Bayesian Constraint-based Causal Discovery (BCCD) algorithm was applied: one explored relationships between Autism Spectrum Disorder (ASD) and Attention-Deficit/Hyperactivity Disorder (ADHD), identifying key pathways linking impulsivity, hyperactivity, and social difficulties [150]. The

second study used BCCD to model the connections between ADHD and comorbidities like conduct problems, substance use, and gaming habits, finding that conduct problems mediate nicotine use and that ADHD inattentive symptoms are linked to gaming dependence [151].

Causal Analysis

Causal forests analysis

Causal Forests Analysis is an advanced causal inference method that utilizes a specialized form of random forests to identify heterogeneity in treatment effects. Unlike traditional random forests, causal forests apply a unique loss function to estimate treatment effects at the individual tree level, providing flexibility in the analysis of both observational and randomized data. This method allows for more accurate estimation of individual effects without the need to define reference classes a priori and uses techniques such as grid search and cross-validation to optimize model parameters. The Causal Forests Analysis method has been applied in three distinct studies, each with specific objectives and contexts, all aiming to enhance the understanding of treatment effects on an individual basis [146, 147, 148]. In the first study, the method was used to assess the effectiveness of an intervention aimed at preventing hospital readmissions, known as the Transitions Program. Researchers analyzed data from the electronic health records of Kaiser Permanente Northern California, conducting a retrospective analysis of patient outcomes post-discharge. By utilizing Causal Forests, they were able to estimate the program's effects on 30-day readmissions. This analysis revealed significant differences in treatment effects among various patients, thus allowing for the characterization of heterogeneity in effects and the identification of subgroups that might benefit the most from the intervention [139]. In the second study, the Causal Forest (CF) and causal tree (CT) methods were used to identify heterogeneous treatment effects (HTEs). First, the CF algorithm was applied to determine the variable with the greatest influence on predicting treatment effects, addressing limitations in traditional CT regarding variable selection. After identifying the key variable with CF, the CT algorithm was used to partition participants into subgroups with different Average Rate of Change (ARCs) through recursive partitioning. This two-step approach aimed to more effectively identify treatment effect moderators across trials [135]. The third study employed the Instrumental Variable Causal Forest Algorithm (IV-CFA) to investigate personalized treatment effects in an observational context, focusing on Medicare beneficiaries with proximal humerus fractures. By applying IV-CFA to a large dataset, researchers discovered substantial heterogeneity in patient responses to early surgery. The estimated effects varied based on algorithm parameters, but the use of classification and regression trees (CART) allowed for the creation of consistent reference classes. This approach demonstrated the utility of IV-CFA in revealing more complex treatment effects, helping to identify which patients might derive the greatest benefit

from surgery [140]. In another study, the Causal Forests method was used to identify heterogeneous treatment effects (HTE) of intensive glycemic control compared to standard control on major adverse cardiovascular events (MACE) in patients with type 2 diabetes. Individual data from two randomized trials (ACCORD and VADT, with a total of 12,042 participants) were analyzed to find subgroups of patients who may benefit more from intensive glycemic control. Using variable prioritization obtained through Causal Forests, a summary decision tree was built based on five variables, highlighting differences in MACE risks between treatment groups [136]. In yet another application, the Causal Forest method was used to analyze the heterogeneity in the association between coronary artery calcium (CAC) and incident cardiovascular disease among adults in the Multi-Ethnic Study of Atherosclerosis (MESA). The method was applied after propensity score matching to identify how individual characteristics influence the increase in CVD risk when CAC is greater than zero. The Causal Forest model revealed that even individuals with low atherosclerotic CVD (ASCVD) risk showed significant increases in CVD risk when CAC was present, particularly among men, Hispanics, and those with unfavorable CVD risk factors. This highlighted the potential benefit of CAC screening for individuals traditionally considered low-risk [137]. Additionally, the Causal Forest method was used to evaluate the impact of Urban and Rural Resident Basic Medical Insurance (URRBMI) on the health of preschool and school-age children in rural China, using data from the 2018 China Family Panel Studies. The study found significant health improvements for preschool children, especially those with disadvantaged mothers (lower wealth, education, or rural areas). For school-age children, health improvements were limited to obese children, but this effect was not statistically significant. Causal Forest helped identify heterogeneity in the policy's impact, offering insights for better policy design [138]. Finally, the last study applied the Causal Forests method to estimate the causal effect of an intervention on exacerbation rates of chronic obstructive pulmonary disease (COPD). Using data from 8,151 patients involved in the SUMMIT trial, researchers developed machine learning models to predict the Individual Treatment Effect (ITE) of treatment with fluticasone furoate/vilanterol compared to a placebo. In this context, a new metric, the Q-score, was created to measure the effectiveness of causal inference models. The findings highlighted that machine learning models for causal inference can identify individual responses to COPD treatments, suggesting that such tools could become useful in personalized clinical decision-making and disease management [141]. Moreover, in a study on postoperative nausea and vomiting, machine learning-based algorithms were used to estimate the CATE to assess the treatment response heterogeneity of dexamethasone. Algorithms like double machine learning (DML), doubly robust learner, and generalized random forest analyzed data from 2026 adult patients. The findings revealed that only a

small subset of patients responded to dexamethasone, with many being non-responders. Importantly, estimated CATE did not correlate with predicted risk, suggesting that predicting treatment response through CATE models may be more effective than traditional risk-based strategies for clinical decision-making [142]. In summary, the application of Causal Forests Analysis in these studies has underscored the ability to identify heterogeneity in treatment effects and provide personalized evidence in clinical contexts, thereby contributing to more informed therapeutic decisions.

G-computation causal inference method

The G-computation causal inference method is a flexible approach used for causal mediation analysis, especially useful when dealing with multiple ordered mediators. Unlike traditional regression-based methods, which can struggle with the complexity of multiple mediators and estimation issues, G-computation leverages simulation-based techniques to enhance model flexibility and expand the outcome scale. In the study utilizing this method, researchers applied G-computation to analyze data from the Taiwanese Cohort Study. Their objective was to investigate the mediating role of early and late Hepatitis B Virus (HBV) viral load in the relationship between Hepatitis C Virus (HCV) infection and hepatocellular carcinoma (HCC) in HBV seropositive patients. The analysis involved a sample of 2,878 patients, including 123 HCV carriers. The G-computation causal inference method provided valuable insights into the complex relationships among HCV infection, HBV viral load, and HCC incidence, demonstrating its efficacy in handling mediation analysis involving multiple mediators [**?**]. In another article, the G-computation causal inference method was used to estimate the effects of indoor environmental exposures on children with asthma, taking confounders into account. Specifically, it allowed for the analysis of how geospatial factors, such as average household sizes and neighborhood characteristics, influence the presence of triggering allergens like cockroaches and rodents within homes. This approach provided a way to assess the impact of environmental exposures on health in a clinical context, using data from electronic health records and geospatial analyses [154].

Maximum likelihood estimation (TMLE) technique

The targeted maximum likelihood estimation (TMLE) method is a multistep procedure designed to produce estimators with robust inference and optimal asymptotic properties. It is particularly useful for estimating marginal causal parameters at the population level. TMLE addresses model misspecification by combining statistical models with machine learning, allowing for more accurate estimation of causal effects. In the first study, TMLE was used to investigate the causal association between recombinant HBV vaccination and the risk of multiple sclerosis (MS). The sample included 110 confirmed MS cases, matched for age, sex, and nativity with 110 controls. Researchers collected data on sociodemographic fac-

tors, environmental factors, and vaccination history through face-to-face interviews. The application of TMLE provided evidence of a nonspecific protective effect of HBV vaccination on the risk of developing MS, suggesting the need for further research to confirm these findings [161]. In a second study, the TMLE method was employed to explore the "obesity paradox" in critically ill patients, assessing whether non-obese patients would have better survival outcomes if they had been obese. The study included 6,557 adult intensive care patients, utilizing both a traditional regression approach and a robust TMLE approach that managed missing data through machine learning and multiple imputation. While the traditional method suggested a potential survival benefit associated with obesity, the TMLE approach did not find evidence to support an improvement in survival for non-obese patients had they been obese, effectively addressing confounding biases and providing a clearer causal analysis [162]. Furthermore, the Targeted Bidirectional EHR Transformer (T-BEHRT) and Cross-Validated Targeted Maximum Likelihood Estimation (CV-TMLE) methods were used in another study to improve causal inference from electronic health records (EHRs) when randomized clinical trials (RCTs) are not feasible or applicable. The study aimed to model the causal association between antihypertensive classes and incident cancer risk, leveraging the strengths of deep learning and doubly robust estimation. The T-BEHRT model was developed to provide accurate estimates of average risk ratios while addressing confounding factors. The findings demonstrated that T-BEHRT outperformed benchmark statistical and deep learning models in estimating relative risk, even under conditions of limited data, and produced estimates consistent with RCT results regarding antihypertensive effects on cancer risk [159]. Additionally, the Hybrid Causal Tree and TMLE methods were applied to analyze the effectiveness of non-vitamin K antagonist oral anticoagulants (NOACs) compared to warfarin in patients with atrial fibrillation (AF). The study aimed to identify variability in treatment benefits and risks across different patient subgroups. By analyzing claims and medical data from 34,569 patients, the Hybrid CT method helped discover distinct patient subgroups that characterized head-to-head treatment effects on a primary composite outcome, including ischemic stroke, intracranial hemorrhage, and all-cause mortality. TMLE was utilized to estimate the causal effects accurately, revealing that certain subgroups showed different outcomes with specific anticoagulants. The findings highlighted the heterogeneity of OAC effects across AF patient subgroups, suggesting potential for personalized treatment strategies [160].

Balancing and Variable Selection Methods

Balancing Covariates Automatically Using Supervision (BCAUS)

The Balancing Covariates Automatically Using Supervision (BCAUS) method is an innovative neural network-based approach designed to enhance causal inference by balancing co-

variates. It employs a dual loss function strategy: the Binary Cross-Entropy Loss (LBCE), which ensures that the model accurately predicts treatment assignments, and the Bias Loss (LBIAS), which works to minimize discrepancies in covariate distributions between treatment and control groups. This combination allows BCAUS not only to predict treatment assignments effectively but also to ensure that covariates are balanced, which is crucial for making valid causal inferences. In the context of a study on Type-2 diabetes treatment, BCAUS was applied to analyze data from over one million high-risk patients across the United States. This study aimed to evaluate the effectiveness of more than 80 different anti-hyperglycemic treatment strategies. By utilizing the BCAUS method, the researchers were able to make personalized treatment recommendations based on comparative effectiveness analyses. The findings indicated an average reduction in HbA1c levels of 0.69% between patients receiving higher-ranked treatments compared to those receiving lower-ranked ones. This significant result underscores the utility of BCAUS not only in diabetes management but also suggests its potential applicability in optimizing treatment strategies for other chronic conditions [152].

Super Learner

Super Learner is an advanced ensemble learning method that aims to optimize prediction accuracy by creating an optimal weighted combination of various machine learning algorithms. The process involves selecting multiple candidate algorithms and then applying a meta-learning algorithm to identify the best combination that minimizes prediction error through cross-validation. This methodology ensures that the resulting "super learner" performs at least as well as, if not better than, any individual algorithm, particularly in large datasets where diversity in predictions can significantly enhance overall accuracy. The Super Learner technique has been effectively employed in several recent studies, showcasing its versatility and power in clinical research. In a study focusing on Gestational diabetes mellitus (GDM), Super Learner was utilized to estimate propensity scores when comparing the perinatal outcomes associated with glyburide and insulin treatment. This study integrated Super Learner with the inverse probability weighting approach to robustly address both baseline and time-dependent confounding factors. By employing this rigorous analytical framework, the researchers aimed to determine whether glyburide offered any advantages over insulin in managing GDM. The results indicated no statistically significant differences in perinatal complications between the two treatments, suggesting that glyburide does not provide additional benefits compared to insulin [163]. In another article examining the relationship between HBV vaccination and the risk of MS, Super Learner was combined with TMLE. This dual approach allowed researchers to analyze the causal association between HBV vaccination and MS risk by comparing 110 confirmed MS cases with 110 matched controls. Data collected through interviews included detailed sociodemographic and clinical

factors. The integration of Super Learner significantly enhanced the precision of causal parameter estimates, revealing a significant protective effect of HBV vaccination against the risk of developing MS. These findings emphasized the importance of further investigation to confirm the results [162]. The third study also employed a combination of TMLE and Super Learner to explore the so-called "obesity paradox" among critically ill patients. In this context, TMLE served as a robust causal inference method that effectively handled confounding factors and missing data through multiple imputation. Simultaneously, Super Learner was integrated to enhance predictive accuracy by synthesizing various machine learning models. This powerful combination led to a more reliable analysis of the data, indicating that the TMLE results did not support the hypothesis that non-obese critically ill patients would have improved survival outcomes if they had been obese. By addressing potential biases often seen in traditional methodologies, this study provided clearer insights into the complex relationship between obesity and survival in critically ill populations [149]. In summary, the Super Learner technique, particularly when used in conjunction with TMLE, has demonstrated significant effectiveness across various clinical studies. This pairing not only enhances causal inference but also improves predictive accuracy in examining critical health-related outcomes, thereby contributing valuable insights to the field of medical research.

Decision Models

Decision tree

Decision Trees are a powerful machine learning model that simplifies complex decision-making processes by breaking them down into a series of sequential, manageable questions. Visually, they resemble a tree structure, where each node represents a specific question or condition, and each branch illustrates the possible outcomes that lead either to another question or to a final decision represented by leaf nodes. This structure allows for transparent and interpretable decision-making, as each path taken can be traced back to the specific questions answered along the way. The Causal Analysis Using Structural and Conditional Associations for Detecting Effects (CASCADE) method enhances the traditional decision tree approach by specifically focusing on causality assessment in safety evaluations. By decomposing complex problems into smaller, well-defined steps, CASCADE facilitates clearer decision-making regarding causal relationships. Each node in the CASCADE tree addresses a particular question about causality, guiding the analysis through a sequence of logical steps until a final conclusion is reached. The clarity and interpretability of this model are significant advantages, as they allow users to easily understand how each outcome is linked to the decisions made throughout the process. In a recent article investigating the relationship between sleep problems and depressive mood, researchers employed the CASCADE technique to explore causality between these two sets of symptoms. Recognizing that sleep issues often

precede depressive episodes, the authors noted that mere temporal precedence is insufficient for establishing causation. Instead, they used advanced statistical causal-discovery algorithms designed to estimate causality from cross-sectional data, applying the CASCADE framework to dissect the relationships between sleep problems and depressive symptoms. Overall, the application of the CASCADE technique in this study provided valuable insights into the causal relationships between sleep problems and depressive mood, aligning with emerging epidemiological and biological evidence in the field [127].

Time Series Analysis

The Granger causality (GC) test algorithm

The Granger causality (GC) test algorithm is a method used to determine if one time series (X) can predict another time series (Y). The test assesses whether the past values of X provide significant information for forecasting future values of Y. If removing X from the model reduces its predictive power for Y, then X is said to "Granger-cause" Y. The method uses statistical tests (T-tests and F-tests) to evaluate the relationship and reject the null hypothesis that X does not Granger-cause Y if the p-value is less than 0.05. However, this approach works best when the system has low coupling, meaning variables should be separable, which can be challenging in complex real-world systems where variables are interrelated. The GC test has been applied, for instance, in assessing the predictive power of Twitter sentiment on future trends. The GC test method was used in one article to address the limitations of traditional methodologies, such as Bayesian networks, in discovering causal effects between variables. The authors proposed using prior knowledge iteration and time series trend fitting between causal variables to identify bidirectional causal relationships. Specifically, the Granger test was employed to analyze the temporal relationships between variables and help construct more accurate causal graphs, improving the accuracy in modeling causal effects. In this context, the PC+ and DCM algorithms were developed to overcome high computational costs and enhance precision compared to traditional approaches. These innovations enabled the creation of more reliable causal models, using data from the COVID-19 pandemic as a case study [157]. In another article, the Granger causality (GC) test algorithm was used to investigate the predictive relationship between social media sentiment scores about vaccination and vaccination rates during the COVID-19 pandemic. Specifically, the test helped determine whether changes in Twitter sentiment could forecast subsequent changes in vaccination rates [158].

Chapter 5

Discussion

The growing scientific evidence has highlighted the importance of numerous maternal factors, including diet, stress levels, exposure to chemicals and toxins, and parity, as potential influences on telomere length in newborns [45,46]. TL, a key indicator of cellular aging, represents a fundamental biomarker for understanding the dynamics of early aging and its effects on long-term health. The interactions between variables such as pre-gestational BMI, GWG, and TL in amniotic fluid may provide crucial insights for the development of intervention strategies aimed at optimizing maternal health and fetal well-being. However, research in this area remains controversial and imprecise, due to the heterogeneity of studies, the variety of methodological approaches, the exposure factors considered, and the diverse characteristics of the studied samples. Before conducting the causal graph analysis, a preliminary investigation using binary classification models (Decision Tree, Random Forest, and XGBoost) was carried out to identify the most influential predictors of gestational weight gain (GWG). The SHAP analysis confirmed that telomere length in amniotic fluid was the most significant predictor, even after adjusting for confounding factors such as pre-pregnancy BMI and adherence to the Mediterranean Diet. These preliminary findings provided the rationale for conducting the causal analysis, allowing us to explore the direct and indirect relationships between BMI, GWG, and telomere length using causal graphs. The consistency between machine learning models and causal inference techniques strengthens the hypothesis that telomere length may play a role in fetal development and pregnancy outcomes. The study conducted during this doctoral thesis is the first attempt to apply causal graph analysis to explore the complex interactions between pre-gestational BMI, GWG, and TL in amniotic fluid, a source considered ideal for prenatal biomarker diagnosis. Despite the potential of amniotic fluid as a relatively pure fetal sample, its use is limited by the invasive nature of amniocentesis. However, our approach leverages amniotic fluid samples collected through amniocentesis for specific clinical indications, aiming to provide insights

for future research on maternal factors influencing neonatal health from the earliest stages of life [112]. A key aspect of our study is the use of causal graph analysis, a methodology that has gained recognition for its ability to model complex dynamics in public health [144-150]. The application of causal graphs allowed us to explore the presumed causal pathways between pre-gestational BMI, GWG, and TL, treating these factors as nodes and their causal relationships as directed edges. The results confirmed that both high pre-BMI and excessive GWG are associated with shorter TL, suggesting an acceleration of the cellular aging process and potential negative effects on fetal well-being. Conversely, maintaining an adequate weight gain during pregnancy was found to be a protective factor, with positive effects on TL, particularly compared to insufficient or excessive weight gain [112, 144]. One of the most innovative aspects of this study is the inclusion of confounding factors such as maternal age and total energy intake, which could influence the relationship between exposure variables and TL. The results were consistent with the simplified model, but significant differences emerged when examining the direct effect of pre-BMI on TL. In particular, when weight gain was adequate, normal-weight women showed a greater likelihood of having a TL above the median compared to underweight or overweight women. However, in the case of insufficient GWG, underweight women showed the highest probability of TL above the median, followed by normal-weight women, with overweight women showing the lowest probability. This suggests that the adequacy of weight gain during pregnancy may modify the effect of high BMI on TL. One of the strengths of this study is the integration of machine learning models to explore key predictors before conducting causal inference. However, while the preliminary classification models helped identify telomere length as a relevant factor in GWG, they do not establish causality. Although the causal graph approach attempts to infer causal relationships, the observational nature of the data limits definitive conclusions. Future research should validate these findings through longitudinal studies or experimental designs to confirm the role of telomere length in pregnancy outcomes. The results of this study have significant implications for public health. Controlling pre-gestational weight and ensuring adequate weight gain during pregnancy could counteract fetal biological aging, with positive effects also in the long term. TL in newborns could serve as an early indicator of cellular health and aging, influenced by maternal factors such as BMI and GWG. These results not only support the IOM guidelines on weight gain during pregnancy but also suggest that such recommendations could have broader benefits, including the reduction of fetal biological aging. It is important to emphasize that pre-gestational BMI and GWG not only influence immediate neonatal health but could also determine early aging, paving the way for preventive interventions beyond traditional approaches for weight management during pregnancy. Sex-related differences in neonatal outcomes are a well-documented phenomenon

in perinatal epidemiology. Understanding these disparities can provide valuable insights into the biological and environmental mechanisms influencing perinatal health. Our findings highlight the importance of considering sex as a potential factor influencing neonatal outcomes, which may have implications for both clinical practice and public health policies. Another significant contribution of our study is the application of the "do"-operator, which simulates an intervention on a specific variable (e.g., GWG or BMI) to estimate the causal effects of these variables on TL, controlling for confounding factors. This methodological approach has the advantage of going beyond simple correlations, providing a clearer view of how changes in one variable can directly influence another [112, 113]. The use of the "do"-operator helps isolate causal effects and provides a more solid foundation for targeted interventions, overcoming typical limitations of observational studies such as selection bias and confounding [115, 116, 117]. Additionally, by conducting a systematic review of the literature, we were able to study the major applications of Artificial Intelligence methods currently used in public health. The integration of machine learning models, particularly binary classification techniques, helped uncover relationships that were later explored through causal analysis. The agreement between predictive models and causal graphs strengthens the reliability of our findings, emphasizing the potential role of telomere length as a key biomarker in pregnancy-related health outcomes. The application of AI models in public health has transformed our way of analyzing complex interactions between environmental factors, biological aging, and maternal-infant health. Advanced probabilistic models such as BN, BART, Shapley Causal Values, and the LiNGAM provide sophisticated tools for causal inference and predictive evaluation. These models improve not only diagnostic accuracy but also the ability to make personalized clinical decisions, promoting more targeted and effective treatments [120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167]. BNs, in particular, have proven useful in analyzing pharmaceutical safety and complex diseases, allowing for clear and intuitive representation of causal relationships [127, 128, 129, 130, 131, 132, 133, 134]. The integration of BNs with more advanced models, such as LiNGAM models, allows for distinguishing between correlation and causality, a crucial aspect for understanding complex phenomena like biological aging and the exposome. The LiNGAM model has shown great potential in analyzing disease progression such as non-alcoholic fatty liver disease and childhood obesity, successfully identifying causal connections between biological and environmental variables. These models have proven essential for long-term studies, as they allow for the isolation of the direct effects of risk factors and identify their influence on child development and maternal health [165-167]. A further promising development in the use of AI models in public health is

the personalized approach to maternal-infant health care. Tools like Causal Forest Analysis and TMLE allow for identifying subgroups of patients with different responses to specific treatments, paving the way for personalized therapeutic interventions. Personalization of interventions is particularly useful in the prenatal context, where individual characteristics such as genetic background, maternal age, and socio-economic conditions significantly influence pregnancy outcomes. Moreover, techniques such as Super Learner, which integrates various machine learning algorithms, have been successfully applied to improve the accuracy of predictions in clinical contexts, such as in the treatment of gestational diabetes, with clear implications for preventing complications during pregnancy [161, 162, 163]. A key limitation of this study is the observational nature of the data, which prevents us from making definitive causal claims. While we employed causal modeling techniques such as DAGs and statistical adjustments to infer potential causal relationships, residual confounding and unmeasured variables cannot be entirely ruled out.

Additionally, machine learning (ML) algorithms were used to enhance model accuracy and detect complex associations. However, ML models do not inherently establish causality; they identify patterns that require further validation through experimental or quasi-experimental studies.

Future research should aim to validate our findings using longitudinal designs, instrumental variables, or randomized controlled trials (RCTs) to strengthen causal inference. Despite these limitations, our study provides valuable insights into the potential relationships between maternal exposures and pregnancy outcomes, contributing to the growing body of knowledge in maternal-child health epidemiology.

Chapter 6

Conclusion and future perspectives

The application of AI models in public health has significantly transformed our understanding of the complex interactions between environmental factors, biological aging, and maternal-child health. The results presented in this thesis highlight the crucial role of AI, particularly advanced probabilistic models such as BN, BART, Causal Shapley Values, and the LiNGAM, in elucidating the causal relationships between the exposome, biological aging, and maternal-neonatal health outcomes [120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167]. These models have provided new insights into how maternal factors, such as pre-gestational BMI and GWG, influence TL in newborns, offering a promising pathway for identifying early indicators of fetal and neonatal health. The study conducted during this doctoral research focused on the causal pathways between maternal weight factors and TL in amniotic fluid as a biomarker for prenatal diagnosis [112]. The innovative use of causal graph analysis advanced our ability to explore these complex interactions, providing valuable insights into the long-term consequences of early exposures. The application of the "do" operator in causal inference allowed us to model the impact of interventions, simulating how changes in BMI or GWG might influence TL, offering a clearer understanding of causal relationships that traditional statistical methods might overlook [112, 115, 116, 117]. The findings confirm not only the importance of managing maternal weight during pregnancy but also suggest that appropriate weight gain may act as a protective factor against fetal biological aging, with potential benefits extending beyond immediate neonatal health [112]. In addition to the application of causal graph models, the integration of AI methodologies such as Super Learner and TMLE holds great potential in personalizing prenatal care. These tools allow for a more detailed understanding of how maternal characteristics such as genetic predisposition, socioeconomic status, and lifestyle factors can influence pregnancy outcomes, paving the way for

tailored interventions. The ability to identify patient subgroups with differing responses to specific treatments offers a more individualized approach to maternal and child health care, which is crucial for preventing long-term health issues, such as obesity and metabolic disorders, which may arise from early exposures [161, 162, 163]. However, despite promising results, there are some limitations in this study and areas for future research. A significant challenge lies in the heterogeneity of the available data, as variations in study designs, sample characteristics, and exposure factors may limit the generalizability of the results. The current use of amniotic fluid as a biomarker for TL is also constrained by the invasive nature of amniocentesis, and future research could focus on non-invasive methods for obtaining fetal biomarkers. Furthermore, while the application of AI models has provided a clearer understanding of causal relationships, these models need to be further validated in diverse populations and contexts. Future studies should aim to incorporate larger and more diverse datasets to better capture the complexity of interactions between maternal-neonatal health and the environment. An additional important area for future research involves expanding the scope of AI models to include broader environmental and genetic factors within the exposome. The exposome, which encompasses all the environmental exposures experienced by an individual throughout their lifetime, is increasingly recognized as a key determinant of health outcomes [3, 4, 5, 6]. Integrating AI models with data from multiple sources, such as environmental monitoring, genetic screening, and lifestyle assessments, could provide a more comprehensive understanding of how environmental stressors contribute to biological aging and maternal-child health outcomes. Additionally, exploring the role of the exposome in mediating the effects of socio-economic inequalities on health could provide crucial insights into how disparities in exposures may contribute to health disparities between populations. The application of AI in public health research also holds significant potential to improve predictive modeling in maternal-child health [13]. By integrating machine learning algorithms into decision-making processes, healthcare providers could gain access to more accurate and timely predictions, improving their ability to identify high-risk pregnancies and intervene before negative outcomes occur [53]. For example, AI-based tools could be used to predict the likelihood of preterm birth, fetal growth restriction, or gestational diabetes, allowing for timely interventions and more personalized care [168, 169]. Future developments in AI may also lead to the creation of real-time monitoring systems that track maternal health indicators during pregnancy, alerting healthcare providers to potential risks and enabling a more proactive approach to maternal-fetal health management. In conclusion, the application of AI models in the study of the exposome, biological aging, and maternal-child health represents an exciting frontier in public health research [44, 49, 52, 53]. The insights gained from this thesis highlight the potential of these models to uncover hidden relationships and

offer new opportunities for data-driven, personalized interventions. While challenges remain in terms of data heterogeneity and model validation, the future of AI in public health appears promising, with the potential to revolutionize our understanding and management of maternal-child health, leading to healthier and longer lives for mothers and children. Looking to the future, it is essential that research continues to refine AI methodologies, expand the range of datasets used, and validate findings in diverse populations. The ultimate goal is to ensure that the insights provided by AI translate into concrete improvements in health outcomes, addressing the complex interaction of genetic, environmental, and socio-economic factors that shape maternal and child health. With continued advancements in AI, we are poised to make significant strides in preventing and mitigating the effects of early exposures on long-term health, promoting a healthier future for all.

Chapter 7

Other Research Activities

- Teaching activities carried out during the PhD program: "Preventive Interventions in Primary Care: Optional and Mandatory Vaccinations" for the 1st-year curriculum (2021–2024 cycle) of the School of Specialized Training in General Medicine (Course Code: MMG2022CT1; CdC: 03011101) at Azienda Ospedaliera Cannizzaro, Catania.

- Training activities carried out during the PhD program: 2022 PhD AI School (area Healthcare and Life sciences); Winter School – Deep-Learning and HPC to Boost Biomedical Application for Health; Introduction to PhD Research; Probabilistic Graphical Models in Intelligent Systems; AI in Computer-aided drug design; Behavioural biometrics for healthcare, security and related fields; Approximate Bayesian computation; Modeling and managing medical processes; AI complexity: open challenges on digital society; Law and Science.

- Qualified tutoring activities for the Bachelor's Degree in Biological Sciences – Course "Hygiene and Statistics".

- From 13 to 17 April 2024 participation as a learner in the LXI COURSE "Adopting a One Health strategy to combat HAIs and AMR: the priorities and challenges of Public Health" held in Erice at the "Ettore Majorana" Foundation Center for Scientific Culture. The course aimed to present the current epidemiological scenario of HAIs and AMR at European and national level, the main factors determining these phenomena and the counteractions identified in the PNCAR 2022-2025. It also described the state of the art on the One Health strategy to combat AMR, reflecting on the challenges and priorities for Public Health.

- Junior tutoring activities for the Bachelor's Degree in Biological Sciences – Course titled "Hygiene and Statistics".

- Participation in the 57th National Congress of SItI, "Public Health for the Future of the Country: Innovation, Alliances, and Institutional Synergies for Prevention," at the University of Palermo, October 23-26, 2024.

- Oral presentation titled “Dieta e profili nutrizionali nelle donne in gravidanza: risultati della coorte MAMI-MED” at the National Public Health Conference Extraordinary SItI, Cernobbio, Como, Italy, 12-14 October 2023.

List of scientific publications

Scientific articles from November 2021

- Favara G, Maugeri A, Barchitta M, Magnano San Lio R, La Rosa MC, La Mastra C, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Social and Nutritional Profiles of Pregnant Women: A Cluster Analysis on the "MAMI-MED" Cohort. Nutrients. 2024 Nov 21;16(23):3975. doi: 10.3390/nu16233975. PMID: 39683369; PMCID: PMC11643402.

- Barchitta M, Maugeri A, La Mastra C, Favara G, La Rosa MC, Magnano San Lio R, Gholizade Atani Y, Gallo G, Agodi A. Pre-pregnancy BMI, gestational weight gain, and telomere length in amniotic fluid: a causal graph analysis. Sci Rep. 2024 Oct 8;14(1):23396. doi: 10.1038/s41598-024-74765-y. PMID: 39379607.

- Magnano San Lio R, Barchitta M, Maugeri A, Campisi E, Favara G, Ojeda Granados C, La Mastra C, La Rosa MC, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Sex differences in delivery and neonatal characteristics of new-borns from the "MAMI MED" cohort. Submitted to the Journal of Personalized Medicine in December 2023.

- Agodi A, Montineri A, Manuele R, Noto P, Carpinteri G, Castiglione G, Grassi P, Lazzara A, Mattaliano AR, Granvillano G, La Mastra C, La Rosa MC, Maugeri A, Barchitta M. Molecular Typing and Resistance Profile of Acinetobacter baumannii Isolates during the COVID-19 Pandemic: Findings from the "EPIRADIOCLINF" Project. Antibiotics (Basel). 2023 Oct 19;12(10):1551. doi: 10.3390/antibiotics12101551. PMID: 37887252; PMCID: PMC10603994.

- Barchitta M, Magnano San Lio R, La Rosa MC, La Mastra C, Favara G, Ferrante G, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A, Maugeri A. The Effect of Maternal Dietary Patterns on Birth Weight for Gestational Age: Findings from the MAMI-MED Cohort. Nutrients. 2023 Apr 16;15(8):1922. doi: 10.3390/nu15081922. PMID: 37111140; PMCID: PMC10147093.

- Maugeri A, Barchitta M, Favara G, La Mastra C, La Rosa MC, Magnano San Lio R, Agodi A. The Application of Clustering on Principal Components for Nutritional Epidemiology: A Workflow to Derive Dietary Patterns. Nutrients. 2022 Dec 30;15(1):195. doi: 10.3390/nu15010195. PMID: 36615850; PMCID: PMC9824338. - Maugeri A, Magnano San Lio R, Favara G, La Rosa MC, La Mastra C, Riela PM, Guarnera L, Battiato S, Barchitta M, Agodi A. Impact of Eating Context on Dietary Choices of College Students: Evidence from the HEALTHY-UNICT Project. Nutrients. 2022 Oct 21;14(20):4418. doi: 10.3390/nu14204418. PMID: 36297101; PMCID: PMC9609717.

- Ojeda-Granados C, Barchitta M, La Rosa MC, La Mastra C, Roman S, Panduro A, Agodi A, Maugeri A. Evaluating Dietary Patterns in Women from Southern Italy and Western Mexico. Nutrients. 2022 Apr 12;14(8):1603. doi: 10.3390/nu14081603. PMID: 35458165.

- Maugeri A, Barchitta M, Magnano San Lio R, Favara G, La Mastra C, La Rosa MC, Agodi A. The Relationship between Body Mass Index, Obesity, and LINE-1 Methylation: A Cross-Sectional Study on Women from Southern Italy. Dis Markers. 2021 Dec 3;2021:9910878. doi: 10.1155/2021/9910878. PMID: 34900031; PMCID: PMC8664509.

- Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Murolo G, Corrao G, Agodi A. Burden of Healthcare-Associated Infections in Sicily, Italy: Estimates from the Regional Point Prevalence Surveys 2016-2018. Antibiotics (Basel). 2021 Nov 8;10(11):1360. doi: 10.3390/antibiotics10111360. PMID: 34827298; PMCID: PMC8614974.

Contributions to national and international conferences and congresses since November 2021

- La Mastra C, Barchitta M, Maugeri A, Agodi A. Application of artificial intelligence to study the causality in Public Health: a systematic review. Accepted at The 17th European Public Health Conference, 12-15 November 2024, Lisbon, Portugal.

- La Mastra C, Gholizade Atani Y, Barchitta M, Maugeri A, Magnano San Lio R, Favara G, La Rosa MC, Gallo G, Agodi A. L'importanza predittiva della lunghezza dei telomeri per l'incremento di peso gestazionale: applicazione di algoritmi di machine learning e analisi di SHAP sulla coorte "Mamma & Bambino". The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024.

- Barchitta M, Signoriello E, Schiavetti I, La Mastra C, Guarnera L, Canu M, Battiato S, Agodi A. Sviluppo di una web-app per valutare stili di vita e dieta in pazienti con sclerosi multipla. The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024.

- Agodi A, Campisi E, Barchitta M, Maugeri A, Di Liberto E, Branciforte G, Cavallaro ME, Favara G, Ojeda Granados C, La Mastra C, La Rosa MC, Magnano San Lio R, Grasso Leanza F, Requirez S, D'Ancona F. Il contrasto all'antimicrobico-resistenza in Sicilia: i risultati regionali del progetto SPiNCAR. Oral presentation at The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024.

- Favara G, Barchitta M, Maugeri A, Magnano San Lio R, Campisi E, Di Liberto E, Ojeda Granados C, La Mastra C, La Rosa MC, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Applicazione della Cluster analysis per l'identificazione di donne in gravidanza a rischio di outcome neonatali avversi: risultati della coorte MAMI-MED. The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024.

- Di Liberto E, Barchitta M, Maugeri A, La Mastra C, Ojeda Granados C, Campisi E, Cappuccio G, Favara G, La Rosa MC, Scandurra M, Magnano San Lio R, Ettore C, Pappalardo E, Galvani F, Ettore G, Agodi A. Utilizzo dei dispositivi elettronici nei primi anni di vita: analisi dei determinanti e del contesto nella coorte MAMI-MED. The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024. - La Rosa MC, Barchitta M, Campisi E, Favara G, Di Liberto E, Ojeda Granados C, La Mastra C, Magnano San Lio R, Maugeri A, Agodi A. Impatto degli impianti di trattamento delle acque reflue sulla diffusione di batteri e geni di resistenza agli antibiotici. The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024.

- Magnano San Lio R, Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Urso M, Gholamiarjenaki N, Scandurra A, Mirabella S, Agodi A. Analisi del resistoma e sviluppo di un biosensore per la determinazione di geni di resistenza agli antimicrobici in campioni di acque reflue. The National Public Health Conference SItI, Palermo, Italy, 23-26 Oct 2024.

- Favara G, Barchitta M, Maugeri A, Magnano San Lio R, Campisi E, Di Liberto E, Ojeda Granados C, La Mastra C, La Rosa MC, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Uso di antibiotici nei primi mille giorni e outcome materno-infantili: risultati della coorte MAMI-MED. XXXII CONGRESSO INTERREGIONALE SICULO- CALABRO "Le strategie per la ripartenza della sanità pubblica e per la salvaguardia della salute collettiva in Sicilia ed in Calabria: un nuovo inizio?" Enna, 23-25 May 2024.

- Campisi E, Barchitta M, Maugeri A, Di Liberto E, Branciforte G, Cavallaro ME, Favara G, Ojeda Granados C, La Mastra C, La Rosa MC, Magnano San Lio R, Agodi A. Sorveglianza nazionale del Consumo di Soluzione IdroAlcolica (CSIA) per l'igiene delle mani: risultati preliminari della Regione Sicilia riferiti all'anno 2023. XXXII CONGRESSO INTERREGIONALE SICULO- CALABRO "Le strategie per la ripartenza della sanità pubblica e per la salvaguardia della salute collettiva in Sicilia ed in Calabria: un nuovo inizio?" Enna, 23-25 May 2024.

- La Rosa MC, Barchitta M, Campisi E, Cappuccio G, Favara G, Di Liberto E, Ojeda Granados C, La Mastra C, Magnano San Lio R, Scandura M, Maugeri M, Agodi A. "Monitoraggio di SARS-CoV-2 e dell'antimicrobico-resistenza tramite le acque reflue". XXXII CONGRESSO INTERREGIONALE SICULO- CALABRO "Le strategie per la ripartenza della sanità pubblica e per la salvaguardia della salute collettiva in Sicilia ed in Calabria: un nuovo inizio?" Enna, 23-25 May 2024.

- Magnano San Lio R, Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Favara G, Ferrante G, Galvani F, Ettore G, Agodi A. Vaccination choices among pregnant women: findings from the MAMI-MED cohort. The 16th European Public Health Conference, Dublin, Ireland, 9 – 11 November 2023.

- La Mastra C, Gholizade Atani Y, Barchitta M, Maugeri A, Magnano San Lio R, Favara G, La Rosa MC, Gallo G, Agodi A. Applicazione di modelli causali per la valutazione della relazione tra aumento di peso gestazionale e lunghezza dei telomeri: risultati della coorte mamma e bambino. Oral presentation at the National Public Health Conference Extraordinary SItI, Cernobbio, Como, Italy, 12- 14 October 2023.

- Maugeri A, Barchitta M, La Rosa MC, La Mastra C, Magnano San Lio R, Favara G, Galvani F, PappalardoE, EttoreC, EttoreG, AgodiA. Effetti dell'incremento di peso gestazionale sugli outcome neonatali: risultati della coorte MAMI-MED. The National Public Health Conference Extraordinary SItI, Cernobbio, Como, Italy, 12- 14 October 2023.

- Favara G, Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Magnano San Lio R, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Dieta e profili nutrizionali nelle donne in gravidanza: risultati della coorte MAMI-MED. Oral presentation at the National Public Health Conference Extraordinary SItI, Cernobbio, Como, Italy, 12- 14 October 2023.

- Magnano San Lio R, Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Favara G, Ferrante G, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. The effect of gestational weight gain on delivery and neonatal characteristics: exploring sex differences in the "MAMI-MED" cohort. 17th World Congress on Public Health, Rome, Italy, May 2023.

- Magnano San Lio R, Barchitta M, Maugeri A, La Rosa MC, La Mastra C, Favara G, Ferrante G, Galvani F, Pappalardo E, Ettore C, Ettore G, Agodi A. Coperture vaccinali nelle donne in gravidanza: risultati della coorte MAMI-MED. The National Public Health Conference SItI, Taormina, Messina, Italy, May 2023.

- Campisi E, Barchitta M, Maugeri A, Manoli M, La Mastra C, La Rosa MC, Favara G, Magnano San Lio R, Ojeda Granados C, Agodi A. Coperture vaccinali tra gli operatori sanitari e i pazienti dei presidi ospedalieri siciliani: risultati della "Sorveglianza europea mediante prevalenza puntuale delle infezioni correlate all'assistenza e sull'uso di antibiotici negli ospedali per acuti". The National Public Health Conference SItI, Taormina, Messina, Italy, May 2023.

- La Mastra C, Barchitta M, Maugeri A, La Rosa MC, Magnano San Lio R, Campisi E, Favara G, Ferrante G, Galvani F, Pappalardo E, Ettore G, Agodi A. Valutazione dei profili nutrizionali in una coorte di donne in gravidanza: risultati dello studio MAMI-MED. The 55th National Congress SItI, Padova, Italy, 28 September - 1 October 2022.

Bibliography

[1] F. S. Collins and H. Varmus. A new initiative on precision medicine. *N Engl J Med*, 372:793–795, 2015.

[2] N. Naithani, S. Sinha, P. Misra, B. Vasudevan, and R. Sahu. Precision medicine: Concept and tools. *Med J Armed Forces India*, 77(3):249–257, Jul 2021.

[3] A. D'Errico, S. Maritano, C. Moccia, E. Isaevska, C. Pizzi, G. Moirano, and M. Popovic. Esposoma: dalla definizione alle sfide future. *Recenti Prog Med*, 114(6):349–354, Jun 2023. Italian.

[4] S. J. Virolainen, A. VonHandorf, K. C. M. F. Viel, M. T. Weirauch, and L. C. Kottyan. Gene-environment interactions and their impact on human health. *Genes Immun*, 24(1):1–11, Feb 2023.

[5] M. L. Wright, A. R. Starkweather, and T. P. York. Mechanisms of the maternal exposome and implications for health outcomes. *ANS Adv Nurs Sci*, 39(2):E17–30, Apr–Jun 2016.

[6] S. M. Rappaport and M. T. Smith. The exposome: A new approach to understanding the health effects of environmental exposures. *Environ Health Perspect*, 118(8):1160–1168, 2010.

[7] D. J. P. Barker. In utero programming of chronic disease. *Clin Sci*, 95(2):115–128, 1998.

[8] R. A. Waterland and R. L. Jirtle. Transgenerational epigenetic inheritance. *Nat Rev Genet*, 4(6):371–378, 2003.

[9] S. S. Araujo et al. The effects of prenatal exposure to air pollution on fetal development. *Environmental Research*, 187:109648, 2020.

[10] E. H. Blackburn. Telomere states and cell fates. *Nature*, 436(7048):149–157, 2005.

[11] R. Tenchov, J. M. Sasso, X. Wang, and Q. A. Zhou. Aging hallmarks and progression and age-related diseases: A landscape view of research advancement. *ACS Chemical Neuroscience*, 15(1):1–30, 2024. Epub 2023 Dec 14, PMID: 38095562, PMCID: PMC10767750.

[12] D. A. Forero and V. Chand. Methods in molecular biology and genetics: looking to the future. *BMC Research Notes*, 16(1):26, 2023. PMID: 36864454, PMCID: PMC9980850.

[13] D. Jungwirth and D. Haluza. Artificial intelligence and public health: An exploratory study. *International Journal of Environmental Research and Public Health*, 20(5):4541, 2023. PMID: 36901550, PMCID: PMC10002031.

[14] J. P. Hamilton. Epigenetics: principles and practice. *Digestive Diseases*, 29(2):130–135, 2011. Epub 2011 Jul 5, PMID: 21734376, PMCID: PMC3134032.

[15] R. Holliday. Epigenetics: a historical overview. *Epigenetics*, 1(1):76–80, 2006.

[16] S. Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.

[17] S. Daios, A. Anogeianaki, G. Kaiafa, A. Kontana, S. Veneti, C. Gogou, E. Karlafti, D. Pilalas, I. Kanellos, and C. Savopoulos. Length as a marker of biological aging: A critical review of recent literature. *Current Medicinal Chemistry*, 29(34):5478–5495, 2022. PMID: 35838223.

[18] C. B. Harley et al. Telomeres shorten during aging of human fibroblasts. *Nature*, 345(6274):458–460, 1990.

[19] N. Ishikawa, K. Nakamura, N. Izumiyama-Shimomura, J. Aida, Y. Matsuda, T. Arai, and K. Takubo. Changes of telomere status with aging: An update. *Geriatrics Gerontology International*, 16(Suppl 1):30–42, 2016. PMID: 27018281.

[20] M. P. Razgonova, A. M. Zakharenko, K. S. Golokhvast, M. Thanasoula, E. Sarandi, K. Nikolouzakis, P. Fragkiadaki, D. Tsoukalas, D. A. Spandidos, and A. Tsatsakis. Telomerase and telomeres in aging theory and chronographic aging theory (review). *Molecular Medicine Reports*, 22(3):1679–1694, 2020. Epub 2020 Jun 25, PMID: 32705188, PMCID: PMC7411297.

[21] I. Shalev et al. Telomere length and the risk of premature birth: a cohort study. *Pediatrics*, 132(6):e1568–e1575, 2013.

[22] J. J. Luxton, M. J. McKenna, A. M. Lewis, L. E. Taylor, S. G. Jhavar, G. P. Swanson, and S. M. Bailey. Telomere length dynamics and chromosomal instability for predicting individual radiosensitivity and risk via machine learning. *J Pers Med*, 11(3):188, 2021.

[23] J. M. Wong and K. Collins. Telomere maintenance and disease. *Lancet*, 362(9388):983–988, 2003.

[24] L. Mirabello, K. Yu, P. Kraft, I. De Vivo, D. J. Hunter, J. Prescott, J. Y. Wong, N. Chatterjee, R. B. Hayes, and S. A. Savage. The association of telomere length and genetic variation in telomere biology genes. *Hum Mutat*, 31(9):1050–1058, 2010.

[25] K. Demanelis, F. Jasmine, L. S. Chen, M. Chernoff, L. Tong, D. Delgado, C. Zhang, J. Shinkle, M. Sabarinathan, H. Lin, E. Ramirez, M. Oliva, S. Kim-Hellmuth, B. E. Stranger, T. P. Lai, A. Aviv, K. G. Ardlie, F. Aguet, H. Ahsan, GTEx Consortium, J. A. Doherty, M. G. Kibriya, and B. L. Pierce. Determinants of telomere length across human tissues. *Science*, 369(6509):eaaz6876, 2020.

[26] Y. Lee, D. Sun, A. P. S. Ori, A. T. Lu, A. Seeboth, S. E. Harris, I. J. Deary, R. E. Marioni, M. Soerensen, J. Mengel-From, J. Hjelmborg, K. Christensen, J. G. Wilson, D. Levy, A. P. Reiner, W. Chen, S. Li, J. R. Harris, P. Magnus, A. Aviv, A. Jugessur, and S. Horvath. Epigenome-wide association study of leukocyte telomere length. *Aging (Albany NY)*, 11(16):5876–5894, 2019.

[27] World Health Organization. Global status report on alcohol and health. World Health Organization, Geneva, Switzerland, 2018.

[28] G. Shen, J. Y. Huang, Y. Q. Huang, and Y. Q. Feng. The relationship between telomere length and cancer mortality: Data from the 1999–2002 national healthy and nutrition examination survey (nhanes). *J Nutr Health Aging*, 24:9–15, 2020.

[29] W. Pan, J. Du, M. Shi, G. Jin, and M. Yang. Short leukocyte telomere length, alone and in combination with smoking, contributes to increased risk of gastric cancer or esophageal squamous cell carcinoma. *Carcinogenesis*, 38(1):12–18, 2017.

[30] H. Peng, Y. Zhu, F. Yeh, S. A. Cole, L. G. Best, J. Lin, E. Blackburn, R. B. Devereux, M. J. Roman, E. T. Lee, et al. Impact of biological aging on arterial aging in american indians: Findings from the strong heart family study. *Aging*, 8:1583–1592, 2016.

[31] D. Révész, J. E. Verhoeven, Y. Milaneschi, and B. W. Penninx. Depressive and anxiety disorders and short leukocyte telomere length: mediating effects of metabolic stress and lifestyle factors. *Psychol Med*, 46(11):2337–2349, 2016.

[32] V. Gorenjak, S. Akbar, M. G. Stathopoulou, and S. Visvikis-Siest. The future of telomere length in personalized medicine. *Front Biosci (Landmark Ed)*, 23(9):1628–1654, 2018.

[33] B. Liu, L. Song, L. Zhang, M. Wu, L. Wang, Z. Cao, C. Xiong, B. Zhang, Y. Li, W. Xia, et al. Prenatal second-hand smoke exposure and newborn telomere length. *Pediatr Res*, 87:1081–1085, 2020.

[34] H. M. Salihu, L. M. King, C. Nwoga, A. Paothong, A. Pradhan, P. J. Marty, R. Daas, and V. E. Whiteman. Association between maternal-perceived psychological stress and fetal telomere length. *South Med J*, 109(12):767–772, 2016.

[35] H. M. Salihu, A. Pradhan, L. King, A. Paothong, C. Nwoga, P. J. Marty, and V. Whiteman. Impact of intrauterine tobacco exposure on fetal telomere length. *Am J Obstet Gynecol*, 212:205.e1–205.e8, 2015.

[36] D. S. Martens, B. Cox, B. G. Janssen, D. B. P. Clemente, A. Gasparrini, C. Vanpoucke, W. Lefebvre, H. A. Roels, M. Plusquin, and T. S. Nawrot. Prenatal air pollution and newborns' predisposition to accelerated biological aging. *JAMA Pediatr*, 171(12):1160–1167, 2017.

[37] H. Zhu, X. Wang, B. Gutin, C. L. Davis, D. Keeton, J. Thomas, I. Stallmann-Jorgensen, G. Mooken, V. Bundy, H. Snieder, P. van der Harst, and Y. Dong. Leukocyte telomere length in healthy caucasian and african-american adolescents: relationships with race, sex, adiposity, adipokines, and physical activity. *J Pediatr*, 158(2):215–220, 2011.

[38] A. L. Fitzpatrick, R. A. Kronmal, M. Kimura, J. P. Gardner, B. M. Psaty, N. S. Jenny, R. P. Tracy, S. Hardikar, and A. Aviv. Leukocyte telomere length and mortality in the cardiovascular health study. *J Gerontol A Biol Sci Med Sci*, 66(4):421–429, 2011.

[39] Nature. How science can put the sustainable development goals back on track, 2021.

[40] K. R. Warren. A review of the history of attitudes toward drinking in pregnancy. *Alcohol Clin Exp Res*, 39(7):1110–1117, 2015.

[41] Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, Spitzer AI, and Ramkumar PN. Machine learning and artificial intelligence: Definitions, applications, and future directions. *Curr Rev Musculoskelet Med*, 13(1):69–76, 2020.

[42] Chen S, Yu J, Chamouni S, et al. Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions. *BMC Med*, 22:354, 2024.

[43] Krauss C, Do XA, and Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702, 2017.

[44] Zhavoronkov A, Mamoshina P, Vanhaelen Q, Scheibye-Knudsen M, Moskalev A, and Aliper A. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Res Rev*, 49:49–66, 2019.

[45] Nersisyan L. Integration of telomere length dynamics into systems biology framework: A review. *Gene Regul Syst Bio*, 10:35–42, 2016.

[46] Gavia-García G, Rosado-Pérez J, Arista-Ugalde TL, Aguiñiga-Sánchez I, Santiago-Osorio E, and Mendoza-Núñez VM. Telomere length and oxidative stress and its relation with metabolic syndrome components in the aging. *Biology (Basel)*, 10(4):253, 2021.

[47] Horvath S. Dna methylation age of human tissues and cell types. *Genome Biol*, 14(10):R115, 2013.

[48] Wang K, Liu H, Hu Q, et al. Epigenetic regulation of aging: implications for interventions of aging and diseases. *Sig Transduct Target Ther*, 7:374, 2022.

[49] Brasil S, Neves CJ, Rijoff T, Falcão M, Valadão G, Videira PA, and Dos Reis Ferreira V. Artificial intelligence in epigenetic studies: Shedding light on rare diseases. *Front Mol Biosci*, 8:648012, 2021.

[50] Baechle JJ, Chen N, Makhijani P, Winer S, Furman D, and Winer DA. Chronic inflammation and the hallmarks of aging. *Mol Metab*, 74:101755, 2023.

[51] Hu S, Cai J, Chen S, Wang Y, and Ren L. Identification of novel biomarkers and immune infiltration characteristics of ischemic stroke based on comprehensive bioinformatic analysis and machine learning. *Biochem Biophys Rep*, 37:101595, 2023.

[52] Prelaj A, Miskovic V, Zanitti M, Trovo F, Genova C, Viscardi G, Rebuzzi SE, Mazzeo L, Provenzano L, Kosta S, Favali M, Spagnoletti A, Castelo-Branco L, Dolezal J, Pearson AT, Lo Russo G, Proto C, Ganzinelli M, Giani C, Ambrosini E, Turajlic S, Au L, Koopman M, Delaloge S, Kather JN, de Braud F, Garassino MC, Pentheroudakis

G, Spencer C, and Pedrocchi ALG. Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. *Ann Oncol*, 35(1):29–65, 2024.

[53] Tahir M, Norouzi M, Khan SS, Davie JR, Yamanaka S, and Ashraf A. Artificial intelligence and deep learning algorithms for epigenetic sequence analysis: A review for epigeneticists and ai experts. *Comput Biol Med*, 183:109302, 2024.

[54] Song AY, Feinberg JI, Bakulski KM, Croen LA, Fallin MD, Newschaffer CJ, Hertz-Picciotto I, Schmidt RJ, Ladd-Acosta C, and Volk HE. Prenatal exposure to ambient air pollution and epigenetic aging at birth in newborns. *Front Genet*, 13:929416, 2022.

[55] Mennickent D, Rodríguez A, Opazo MC, Riedel CA, Castro E, Eriz-Salinas A, Appel-Rubio J, Aguayo C, Damiano AE, Guzmán-Gutiérrez E, and Araya J. Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications. *Front Endocrinol (Lausanne)*, 14:1130139, 2023.

[56] Appleton AA, Lin B, Holdsworth EA, Feingold BJ, and Schell LM. Prenatal exposure to favorable social and environmental neighborhood conditions is associated with healthy pregnancy and infant outcomes. *Int J Environ Res Public Health*, 18(11):6161, 2021.

[57] Barakat C and Konstantinidis T. A review of the relationship between socioeconomic status change and health. *Int J Environ Res Public Health*, 20(13):6249, 2023.

[58] Morniroli D, Tiraferri V, Maiocco G, De Rose DU, Cresi F, Coscia A, Mosca F, and Giannì ML. Beyond survival: the lasting effects of premature birth. *Front Pediatr*, 11:1213243, 2023.

[59] Gómez-Roig MD, Pascal R, Cahuana MJ, García-Algar O, Sebastiani G, Andreu-Fernández V, Martínez L, Rodríguez G, Iglesia I, Ortiz-Arrabal O, Mesa MD, Cabero MJ, Guerra L, Llurba E, Domínguez C, Zanini MJ, Foraster M, Larqué E, Cabañas F, Lopez-Azorín M, Pérez A, Loureiro B, Pallás-Alonso CR, Escuder-Vieco D, and Vento M. Environmental exposure during pregnancy: Influence on prenatal development and early life: A comprehensive review. *Fetal Diagn Ther*, 48(4):245–257, 2021.

[60] Jiaying F, Qingmei L, Baozhuo A, Meijun L, Weidong L, Saijun H, Hong Y, Yin Y, Hualiang L, Jing W, Xi S, and Zilong Z. Associations between maternal exposure to air pollution during pregnancy and trajectories of infant growth: A birth cohort study. *Ecotoxicology and Environmental Safety*, 269:115792, 2024.

[61] Salgado M, Madureira J, Mendes AS, Torres A, Teixeira JP, and Oliveira MD. Environmental determinants of population health in urban settings. a systematic review. *BMC Public Health*, 20(1):853, 2020.

[62] Katzke VA, Kaaks R, and Kühn T. Lifestyle and cancer risk. *Cancer J*, 21(2):104–110, 2015.

[63] Rahmani AM, Azhir E, Ali S, Mohammadi M, Ahmed OH, Yassin Ghafour M, Hasan Ahmed S, and Hosseinzadeh M. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. *PeerJ Comput Sci*, 7:e488, 2021.

[64] Juhn YJ, Ryu E, Wi CI, King KS, Malik M, Romero-Brufau S, Weng C, Sohn S, Sharp RR, and Halamka JD. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the houses index. *J Am Med Inform Assoc*, 29(7):1142–1151, 2022.

[65] Zezhi P, Bin Z, Diwei W, Xinyi N, Jian S, Hongmei X, Junji C, and Zhenxing S. Application of machine learning in atmospheric pollution research: A state-of-art review. *Science of The Total Environment*, 910:168588, 2024.

[66] Padilla CM, Kihal-Talantikit W, Perez S, and Deguen S. Use of geographic indicators of healthcare, environment and socioeconomic factors to characterize environmental health disparities. *Environ Health*, 15(1):79, 2016.

[67] Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, Felländer A, Langhans SD, Tegmark M, and Fuso Nerini F. The role of artificial intelligence in achieving the sustainable development goals. *Nat Commun*, 11(1):233, 2020.

[68] Miele MJ, Souza RT, Calderon IM, Feitosa FE, Leite DF, Rocha Filho EA, Vettorazzi J, Mayrink J, Fernandes KG, Vieira MC, Pacagnella RC, and Cecatti JG. Maternal nutrition status associated with pregnancy-related adverse outcomes. *Nutrients*, 13(7):2398, 2021.

[69] Cordeiro JV. Digital technologies and data science as health enablers: An outline of appealing promises and compelling ethical, legal, and social challenges. *Front Med (Lausanne)*, 8:647897, 2021.

[70] Jiao L, Wang Y, Liu X, Li L, Liu F, Ma W, Guo Y, Chen P, Yang S, and Hou B. Causal inference meets deep learning: A comprehensive survey. *Research (Wash D C)*, 7:0467, 2024.

[71] Lee S and Lim H. Review of statistical methods for survival analysis using genomic data. *Genomics Inform*, 17(4):e41, 2019.

[72] Yu X, Tang L, Long L, et al. Comparison of deep and conventional machine learning models for prediction of one supply chain management distribution cost. *Sci Rep*, 14:24195, 2024.

[73] Ahsan MM, Luna SA, and Siddique Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare (Basel)*, 10(3):541, 2022.

[74] Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, and Kohane IS. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit Med*, 4(1):62, 2021.

[75] Moe SJ, Carriger JF, and Glendell M. Increased use of bayesian network models has improved environmental risk assessments. *Integr Environ Assess Manag*, 17(1):53–61, 2021.

[76] Tan YV and Roy J. Bayesian additive regression trees and the general bart model. *Stat Med*, 38(25):5048–5069, 2019.

[77] Yang TL, Lee KY, Zhang K, et al. Functional linear non-gaussian acyclic model for causal discovery. *Behaviormetrika*, 51:567–588, 2024.

[78] Byeon S and Lee W. Directed acyclic graphs for clinical research: a tutorial. *J Minim Invasive Surg*, 26(3):97–107, 2023.

[79] A. Chatton, F. Le Borgne, C. Leyrat, F. Gillaizeau, C. Rousseau, L. Barbin, D. Laplaud, M. Léger, B. Giraudeau, and Y. Foucher. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Sci Rep*, 10(1):9219, 2020.

[80] C. R. MacIntyre, X. Chen, M. Kunasekaran, A. Quigley, S. Lim, H. Stone, H. Y. Paik, L. Yao, D. Heslop, W. Wei, I. Sarmiento, and D. Gurdasani. Artificial intelligence in public health: the potential of epidemic early warning systems. *J Int Med Res*, 51(3):3000605231159335, 2023.

[81] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler. What is machine learning? a primer for the epidemiologist. *Am J Epidemiol*, 188(12):2222–2239, 2019.

[82] A. Agodi, M. Barchitta, G. Valenti, R. Marzagalli, V. Frontini, and A. E. Marchese. Increase in the prevalence of the mthfr 677 tt polymorphism in women born since 1959: potential implications for folate requirements. *Eur J Clin Nutr*, 65:1302–1308, 2011.

[83] C. J. Piyathilake, M. Macaluso, R. D. Alvarez, M. Chen, S. Badiga, N. R. Siddiqui, J. C. Edberg, E. E. Partridge, and J. L. Johanning. A higher degree of line-1 methylation in peripheral blood mononuclear cells, a one-carbon nutrient related epigenetic alteration, is associated with a lower risk of developing cervical intraepithelial neoplasia. *Nutrition*, 27:513–519, 2011.

[84] P. Haggarty, G. Hoad, D. M. Campbell, G. W. Horgan, C. Piyathilake, and G. McNeill. Folate in pregnancy and imprinted gene and repeat element methylation in the offspring. *Am J Clin Nutr*, 97(1):94–9, 2013.

[85] A. Izzotti, C. Cartiglia, V. E. Steele, and S. De Flora. Micrornas as targets for dietary and pharmacological inhibitors of mutagenesis and carcinogenesis. *Mutat Res*, 751(2):287–303, 2012.

[86] C. W. Chang, H. K. Hsu, C. C. Kao, J. Y. Huang, and P. L. Kuo. Prenatal diagnosis of prader–willi syndrome and angelman syndrome for fetuses with suspicious deletion of chromosomal region 15q11-q13. *Int J Gynaecol Obstet*, 2014.

[87] World Health Organization. *Physical Status: The Use and Interpretation of Anthropometry. Report of a WHO Expert Committee*. Number 854 in Technical Report Series. World Health Organization, Geneva, 1995.

[88] A. Agodi, M. Barchitta, A. Quattrocchi, A. Maugeri, C. Canto, A. E. Marchese, and M. Vinciguerra. Low fruit consumption and folate deficiency are associated with line-1 hypomethylation in women of a cancer-free population. *Genes Nutr*, 10(5):480, 2015.

[89] M. Barchitta, A. Maugeri, A. Quattrocchi, G. Barone, P. Mazzoleni, A. Catalfo, G. De Guidi, M. G. Iemmolo, N. Crimi, and A. Agodi. Mediterranean diet and particulate matter exposure are associated with line-1 methylation: results from a cross-sectional study in women. *Genet*, 2018.

[90] M. Barchitta, A. Maugeri, A. Quattrocchi, O. Agrifoglio, A. Scalisi, and A. Agodi. The association of dietary patterns with high-risk human papillomavirus infection and cervical cancer: a cross-sectional study in italy. *Nutrients*, 10(4):469, 2018.

[91] S. Franceschi, F. Barbone, E. Negri, and et al. Reproducibility of an italian food frequency questionnaire for cancer studies. results for specific nutrients. *Ann Epidemiol*, 5:69–75, 1995.

[92] S. Salvini, M. Parpinel, P. Gnagnarella, and et al. Banca dati di composizione degli alimenti per studi epidemiologici in italia, 1998. Milan: Istituto Europeo di Oncologia.

[93] L. E. Kelemen. Gi epidemiology: nutritional epidemiology. *Aliment Pharmacol Ther*, 25:401–407, 2007.

[94] E. Couto, P. Boffetta, P. Lagiou, P. Ferrari, G. Buckland, K. Overvad, C. C. Dahm, A. Tjønneland, A. Olsen, F. Clavel-Chapelon, M. C. Boutron-Ruault, V. Cottet, D. Trichopoulos, A. Naska, V. Benetou, R. Kaaks, S. Rohrmann, H. Boeing, A. von Ruesten, S. Panico, V. Pala, P. Vineis, D. Palli, R. Tumino, A. May, P. H. Peeters, H. B. Bueno-de Mesquita, F. L. Büchner, E. Lund, G. Skeie, D. Engeset, C. A. Gonzalez, C. Navarro, L. Rodríguez, M. J. Sánchez, P. Amiano, A. Barricarte, G. Hallmans, I. Johansson, J. Manjer, E. Wirfärt, N. E. Allen, F. Crowe, K. T. Khaw, N. Wareham, A. Moskal, N. Slimani, M. Jenab, D. Romaguera, T. Mouw, T. Norat, E. Riboli, and A. Trichopoulou. Mediterranean dietary pattern and cancer risk in the epic cohort. *Br J Cancer*, 104(9):1493–1499, 2011.

[95] A. Trichopoulou, A. Kouris-Blazos, M. L. Wahlqvist, C. Gnardellis, P. Lagiou, E. Polychronopoulos, T. Vassilakou, L. Lipworth, and D. Trichopoulos. Diet and overall survival in elderly people. *BMJ*, 311(7018):1457–1460, 1995.

[96] A. Trichopoulou, P. Orfanos, T. Norat, and et al. Modified mediterranean diet and survival: Epic-elderly prospective cohort study. *BMJ*, 330:991, 2005.

[97] M. Barchitta, A. Quattrocchi, V. Adornetto, A. E. Marchese, and A. Agodi. Tumor necrosis factor-alpha -308 g¿a polymorphism, adherence to mediterranean diet, and risk of overweight/obesity in young women. *Biomed Res Int*, 2014:742620, 2014.

[98] U. M. Devlin, B. A. McNulty, A. P. Nugent, and M. J. Gibney. The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy misreporting. *Proc Nutr Soc*, 71:599–609, 2012.

[99] K. L. Tucker. Dietary patterns, approaches, and multicultural perspective. *Appl Physiol Nutr Metab*, 35(2):211–218, 2010.

[100] W. Willett. *Nutritional Epidemiology*, volume 40. Oxford University Press, Oxford, UK, 3rd edition, 2013.

[101] J. Lever, M. Krzywinski, and N. Altman. Principal component analysis. *Nat. Methods*, 14:641–642, 2017.

[102] A. Agodi, A. Maugeri, S. Kunzova, O. Sochor, H. Bauerova, N. Kiacova, M. Barchitta, and M. Vinciguerra. Association of dietary patterns with metabolic syndrome: Results from the kardiovize brno 2030 study. *Nutrients*, 10:898, 2018.

[103] M. Barchitta, A. Maugeri, R. Magnano San Lio, G. Favara, M. C. La Rosa, C. La Mastra, A. Quattrocchi, and A. Agodi. Dietary patterns are associated with leukocyte line-1 methylation in women: A cross-sectional study in southern italy. *Nutrients*, 11:1843, 2019.

[104] S. M. Moeller, J. Reedy, A. E. Millen, L. B. Dixon, P. K. Newby, K. L. Tucker, S. M. Krebs-Smith, and P. M. Guenther. Dietary patterns: Challenges and opportunities in dietary patterns research an experimental biology workshop, april 1, 2006. *J. Am. Diet. Assoc.*, 107:1233–1239, 2007.

[105] P. Newby and K. Tucker. Empirically derived eating patterns using factor or cluster analysis: A review. *Nutr. Rev.*, 62:177–203, 2004.

[106] A. Maugeri, M. Barchitta, G. Favara, M. C. La Rosa, C. La Mastra, R. Magnano San Lio, and A. Agodi. Maternal dietary patterns are associated with pre-pregnancy body mass index and gestational weight gain: Results from the "mamma bambino" cohort. *Nutrients*, 11:1308, 2019.

[107] A. Maugeri, J. Hruskova, J. Jakubik, O. Hlinomaz, J. R. Medina-Inojosa, M. Barchitta, A. Agodi, and M. Vinciguerra. How dietary patterns affect left ventricular structure, function and remodelling: Evidence from the kardiovize brno 2030 study. *Sci. Rep.*, 9:19154, 2019.

[108] M. Barchitta, A. Maugeri, A. Quattrocchi, O. Agrifoglio, A. Scalisi, and A. Agodi. The association of dietary patterns with high-risk human papillomavirus infection and cervical cancer: A cross-sectional study in italy. *Nutrients*, 10:469, 2018.

[109] N. Altman and M. Krzywinski. Clustering. *Nat. Methods*, 14:545–546, 2017.

[110] A. Maugeri, M. Barchitta, G. Favara, C. La Mastra, M. C. La Rosa, R. Magnano San Lio, and A. Agodi. The application of clustering on principal components for nutritional epidemiology: A workflow to derive dietary patterns. *Nutrients*, 15:195, 2022.

[111] M. Barchitta, R. Magnano San Lio, M. C. La Rosa, C. La Mastra, G. Favara, G. Ferrante, F. Galvani, E. Pappalardo, C. Ettore, G. Ettore, A. Agodi, and A. Maugeri. The effect of maternal dietary patterns on birth weight for gestational age: Findings from the mami-med cohort. *Nutrients*, 15(8):1922, 2023.

[112] M. Barchitta, A. Maugeri, C. La Mastra, G. Favara, M. C. La Rosa, R. Magnano San Lio, Y. Gholizade Atani, G. Gallo, and A. Agodi. Pre-pregnancy bmi, gestational weight gain, and telomere length in amniotic fluid: a causal graph analysis. *Sci Rep*, 14(1):23396, 2024.

[113] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[114] K. S. Gill, Judea Pearl, and D. Mackenzie. The book of why: the new science of cause and effect. *AI & Society*, 35:767–768, 2020.

[115] J. Tian and J. Pearl. On the identification of causal effects. 2015.

[116] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modelling*, 7:1393–1512, 1986.

[117] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 2001.

[118] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.

[119] G. Favara, A. Maugeri, M. Barchitta, R. Magnano San Lio, M. C. La Rosa, C. La Mastra, F. Galvani, E. Pappalardo, C. Ettore, and G. Ettore. Social and nutritional profiles of pregnant women: A cluster analysis on the "mami-med" cohort. *Nutrients*, 16:3975, 2024.

[120] R. C. Nethery, F. Mealli, J. D. Sacks, and F. Dominici. Evaluation of the health impacts of the 1990 clean air act amendments using causal inference and machine learning. *Journal of the American Statistical Association*, 116(535):1128–1139, 2020.

[121] V. Dorie, G. Perrett, J. L. Hill, and B. Goodrich. Stan and bart for causal inference: Estimating heterogeneous treatment effects using the power of stan and the flexibility of machine learning. *Entropy (Basel)*, 24(12):1782, 2022.

[122] B. S. Blette, A. Granholm, F. Li, M. Shankar-Hari, T. Lange, M. W. Munch, M. H. Møller, A. Perner, and M. O. Harhay. Causal bayesian machine learning to assess treatment effect heterogeneity by dexamethasone dose for patients with covid-19 and severe hypoxemia. *Sci Rep*, 13(1):6570, 2023.

[123] M. H. Schwartz, A. J. Ries, and A. G. Georgiadis. Short-term causal effects of common treatments in ambulatory children and young adults with cerebral palsy: three machine learning estimates. *Sci Rep*, 12(1):7818, 2022.

[124] W. Suhre, V. O'Reilly-Shah, and W. Van Cleve. Cannabis use is associated with a small increase in the risk of postoperative nausea and vomiting: a retrospective machine-learning causal analysis. *BMC Anesthesiol*, 20(1):115, 2020.

[125] K. M. Steele and M. H. Schwartz. Causal effects of motor control on gait kinematics after orthopedic surgery in cerebral palsy: A machine-learning approach. *Front Hum Neurosci*, 16:846205, 2022.

[126] L. Hu, J. Y. (Joyce) Lin, K. Sigel, and M. Kale. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. *Ann Epidemiol*, 62:36–42, 2021.

[127] Y. Cherkas, J. Ide, and J. van Stekelenborg. Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions. *Drug Saf*, 45(5):571–582, 2022.

[128] P. Fuster-Parra, P. Tauler, M. Bennasar-Veny, A. Ligeza, A. A. López-González, and A. Aguiló. Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. *Comput Methods Programs Biomed*, 126:128–142, 2016.

[129] S. Mani and G. F. Cooper. Causal discovery using a bayesian local causal discovery algorithm. In *Stud Health Technol Inform*, volume 107, pages 731–735, 2004.

[130] M. Ahangaran, M. R. Jahed-Motlagh, and B. Minaei-Bidgoli. A novel method for predicting the progression rate of als disease based on automatic generation of probabilistic causal chains. *Artif Intell Med*, 107:101879, 2020.

[131] M. Jahangoshai Rezaee, M. Sadatpour, N. Ghanbari-Ghoushchi, E. Fathi, and A. Alizadeh. Analysis and decision based on specialist self-assessment for prognosis factors of acute leukemia integrating data-driven bayesian network and fuzzy cognitive map. *Med Biol Eng Comput*, 58(11):2845–2861, 2020.

[132] M. Sieswerda, S. Xie, R. van Rossum, I. Bermejo, G. Geleijnse, K. Aben, F. van Erning, V. Lemmens, A. Dekker, and X. Verbeek. Identifying confounders using bayesian networks and estimating treatment effect in prostate cancer with observational data. *JCO Clin Cancer Inform*, 7:e2200080, 2023.

[133] R. J. McNally, P. Mair, B. L. Mugno, and B. C. Riemann. Co-morbid obsessive-compulsive disorder and depression: a bayesian network approach. *Psychol Med*, 47(7):1204–1214, 2017.

[134] S. Mani and G. F. Cooper. A study in causal discovery from population-based infant birth and death records. In *Proc AMIA Symp*, pages 315–319, 1999.

[135] H. Kianmehr, J. Guo, Y. Lin, J. Luo, W. Cushman, L. Shi, V. Fonseca, and H. Shao. A machine learning approach identifies modulators of heart failure hospitalization prevention among patients with type 2 diabetes: A revisit to the accord trial. *J Diabetes Complications*, 36(9):108287, 2022.

[136] J. A. Edward, K. Josey, G. Bahn, L. Caplan, J. E. B. Reusch, P. Reaven, D. Ghosh, and S. Raghavan. Heterogeneous treatment effects of intensive glycemic control on major adverse cardiovascular events in the accord and vadt trials: a machine-learning analysis. *Cardiovasc Diabetol*, 21(1):58, 2022.

[137] K. Inoue, T. E. Seeman, T. Horwich, M. J. Budoff, and K. E. Watson. Heterogeneity in the association between the presence of coronary artery calcium and cardiovascular events: A machine-learning approach in the mesa study. *Circulation*, 147(2):132–141, 2023.

[138] H. Chen, J. Xing, X. Yang, and K. Zhan. Heterogeneous effects of health insurance on rural children's health in china: A causal machine learning approach. *Int J Environ Res Public Health*, 18(18):9616, 2021.

[139] B. J. Marafino, A. Schuler, V. X. Liu, G. J. Escobar, and M. Baiocchi. Predicting preventable hospital readmissions with causal machine learning. *Health Serv Res*, 55(6):993–1002, 2020.

[140] J. M. Brooks, C. G. Chapman, S. B. Floyd, B. K. Chen, C. A. Thigpen, and M. Kissenberth. Assessing the ability of an instrumental variable causal forest algorithm to personalize treatment evidence using observational data: the case of early surgery for shoulder fracture. *BMC Med Res Methodol*, 22(1):190, 2022.

[141] K. Verstraete, I. Gyselinck, H. Huts, N. Das, M. Topalovic, M. De Vos, and W. Janssens. Estimating individual treatment effects on copd exacerbations by causal machine learning on randomised controlled trials. *Thorax*, 78(10):983–989, Oct 2023.

[142] T. Mizuguchi and S. Sawamura. Machine learning-based causal models for predicting the response of individual patients to dexamethasone treatment as prophylactic antiemetic. *Sci Rep*, 13(1):7549, May 9 2023.

[143] S. R. Hao, S. C. Geng, L. X. Fan, J. J. Chen, Q. Zhang, and L. J. Li. Intelligent diagnosis of jaundice with dynamic uncertain causality graph model. *J Zhejiang Univ Sci B*, 18(5):393–401, May 2017.

[144] Y. Jiao, Z. Zhang, T. Zhang, W. Shi, Y. Zhu, J. Hu, and Q. Zhang. Development of an artificial intelligence diagnostic model based on dynamic uncertain causality graph for the differential diagnosis of dyspnea. *Front Med*, 14(4):488–497, Aug 2020.

[145] Q. Kang, X. Song, X. Xin, B. Chen, Y. Chen, X. Ye, and B. Zhang. Machine learning-aided causal inference framework for environmental data analysis: A covid-19 case study. *Environ Sci Technol*, 55(19):13400–13410, Oct 5 2021.

[146] S. la Bastide-van Gemert, R. P. Stolk, E. R. van den Heuvel, and V. Fidler. Causal inference algorithms can be useful in life course epidemiology. *J Clin Epidemiol*, 67(2):190–198, Feb 2014.

[147] I. M. Aris, A. L. Sarvet, M. J. Stensrud, R. Neugebauer, L. J. Li, M. F. Hivert, E. Oken, and J. G. Young. Separating algorithms from questions and causal inference with unmeasured exposures: An application to birth cohort studies of early body mass index rebound. *Am J Epidemiol*, 190(7):1414–1423, Jul 1 2021.

[148] L. Laffers, H. Farbmacher, M. Huber, H. Langen, and M. Spindler. Causal mediation analysis with double machine learning, 2020.

[149] H. Kern, G. Corani, D. Huber, N. Vermes, M. Zaffalon, M. Varini, C. Wenzel, and A. Fringer. Impact on place of death in cancer patients: a causal exploration in southern switzerland. *BMC Palliat Care*, 19(1):160, Oct 15 2020.

[150] E. Sokolova, A. M. Oerlemans, N. N. Rommelse, P. Groot, C. A. Hartman, J. C. Glennon, T. Claassen, T. Heskes, and J. K. Buitelaar. A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (adhd) and autism spectrum disorder (asd). *J Autism Dev Disord*, 47(6):1595–1604, Jun 2017.

[151] G. H. Schoenmacker, A. P. Groenman, E. Sokolova, J. Oosterlaan, N. Rommelse, H. Roeyers, R. D. Oades, S. V. Faraone, B. Franke, T. Heskes, A. Arias Vasquez, T. Claassen, and J. K. Buitelaar. Role of conduct problems in the relation between attention-deficit hyperactivity disorder, substance use, and gaming. *Eur Neuropsychopharmacol*, 30:102–113, Jan 2020.

[152] C. Belthangady, S. Giampanis, I. Jankovic, W. Stedden, P. Alves, S. Chong, C. Knott, and B. Norgeot. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. *Nat Commun*, 13(1):6921, Nov 14 2022.

[153] A. S. Tai, Y. T. Huang, H. I. Yang, L. V. Lan, and S. H. Lin. G-computation to causal mediation analysis with sequential multiple mediators-investigating the vulnerable time window of hbv activity for the mechanism of hcv induced hepatocellular carcinoma. *Front Public Health*, 9:757942, Jan 7 2022.

[154] M. Bozigar, C. L. Connolly, A. Legler, W. G. Adams, C. W. Milando, D. B. Reynolds, F. Carnes, R. B. Jimenez, K. Peer, K. Vermeer, J. I. Levy, and M. P. Fabian. In-home environmental exposures predicted from geospatial characteristics of the built environment and electronic health records of children with asthma. *Ann Epidemiol*, 73:38–47, Sep 2022.

[155] T. Banerjee, A. Paul, V. Srikanth, and I. Strümke. Causal connections between socioeconomic disparities and covid-19 in the usa. *Sci Rep*, 12(1):15827, Sep 22 2022.

[156] Y. Cherkas, J. Ide, and J. van Stekelenborg. Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions. *Drug Saf*, 45(5):571–582, May 2022.

[157] H. Li, M. Hai, and W. Tang. Prior knowledge-based causal inference algorithms and their applications for china covid-19 analysis. *Mathematics*, 10:3568, 2022.

[158] A. Bari, M. Heymann, R. J. Cohen, R. Zhao, L. Szabo, S. Apas Vasandani, A. Khubchandani, M. DiLorenzo, and M. Coffee. Exploring coronavirus disease 2019

vaccine hesitancy on twitter using sentiment analysis and natural language processing algorithms. *Clin Infect Dis*, 74(Suppl$_3$) : $e4 - -e9, May152022$.

[159] S. Rao, M. Mamouei, G. Salimi-Khorshidi, Y. Li, R. Ramakrishnan, A. Hassaine, D. Canoy, and K. Rahimi. Targeted-behrt: Deep learning for observational causal inference on longitudinal electronic health records. *IEEE Trans Neural Netw Learn Syst*, 35(4):5027–5038, Apr 2024.

[160] C. Ngufor, X. Yao, J. W. Inselman, J. S. Ross, S. S. Dhruva, D. J. Graham, J. Y. Lee, K. C. Siontis, N. R. Desai, E. Polley, N. D. Shah, and P. A. Noseworthy. Identifying treatment heterogeneity in atrial fibrillation using a novel causal machine learning method. *Am Heart J*, 260:124–140, Jun 2023.

[161] S. Akhtar, H. El-Muzaini, and R. Alroughani. Recombinant hepatitis b vaccine uptake and multiple sclerosis risk: A marginal structural modeling approach. *Mult Scler Relat Disord*, 58:103487, Feb 2022. Epub 2022 Jan 3.

[162] A. Decruyenaere, J. Steen, K. Colpaert, D. D. Benoit, J. M. Decruyenaere, and S. Vansteelandt. The obesity paradox in critically ill patients: a causal learning approach to a casual finding. *Critical Care*, 24, 2020. No page numbers available.

[163] M. M. Hedderson, S. E. Badon, N. Pimentel, F. Xu, A. Regenstein, A. Ferrara, and R. Neugebauer. Association of glyburide and subcutaneous insulin with perinatal complications among women with gestational diabetes. *JAMA Netw Open*, 5(3):e225026, Mar 1 2022.

[164] T. Uchida, K. Fujiwara, K. Nishioji, M. Kobayashi, M. Kano, Y. Seko, K. Yamaguchi, Y. Itoh, and H. Kadotani. Medical checkup data analysis method based on lingam and its application to nonalcoholic fatty liver disease. *Artif Intell Med*, 128:102310, Jun 2022. Epub 2022 Apr 22.

[165] J. Kotoku, A. Oyama, K. Kitazumi, H. Toki, A. Haga, R. Yamamoto, M. Shinzawa, M. Yamakawa, S. Fukui, K. Yamamoto, and T. Moriyama. Causal relations of health indices inferred statistically using the directlingam algorithm from big data of osaka prefecture health checkups. *PLoS One*, 15(12):e0243229, Dec 23 2020.

[166] H. Helajärvi, T. Rosenström, K. Pahkala, M. Kähönen, T. Lehtimäki, O. J. Heinonen, M. Oikonen, T. Tammelin, J. S. Viikari, and O. T. Raitakari. Exploring causality between tv viewing and weight change in young and middle-aged adults. the cardiovascular risk in young finns study. *PLoS One*, 9(7):e101860, Jul 16 2014.

[167] J. J. Anker, E. Kummerfeld, A. Rix, S. J. Burwell, and M. G. Kushner. Causal network modeling of the determinants of drinking behavior in comorbid alcohol use and anxiety disorder. *Alcohol Clin Exp Res*, 43(1):91–97, Jan 2019.

[168] I. Kyparissidis Kokkinidis, E. Logaras, E. S. Rigas, I. Tsakiridis, T. Dagklis, A. Billis, and P. D. Bamidis. Towards an explainable ai-based tool to predict preterm birth. *Stud Health Technol Inform*, 302:571–575, May 18 2023.

[169] M. Akazawa and K. Hashimoto. Prediction of preterm birth using artificial intelligence: a systematic review. *J Obstet Gynaecol*, 42(6):1662–1668, Aug 2022. Epub 2022 Jun 1.