Tracking of Cardiovascular Risk Factors in the INMA-Asturias Cohort

Statistical and Methodological considerations

Universidad de Oviedo

Doctoral Programme

in Biomedicine and Molecular Oncology

PhD Dissertation

Rocío Fernández Iglesias

supervised by

Adonina Tardón García

Ana Fernández Somoano

Seguimiento de Factores de Riesgo Cardiovascular en la Cohorte INMA-Asturias

Consideraciones Estadísticas y Metodológicas

Universidad de Oviedo

Programa Doctoral

en Biomedicina y Oncología Molecular

Tesis Doctoral

Rocío Fernández Iglesias

dirigida por

Adonina Tardón García

Ana Fernández Somoano

Universidad de Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Seguimiento de Factores de Riesgo Cardiovascular en la Cohorte INMA-Asturias. Consideraciones Estadísticas y Metodológicas.	Inglés: Tracking of Cardiovascular Risk Factors in the INMA-Asturias Cohort. Statistical and Methodological considerations.

2.- Autor	
Nombre: Rocío Fernández Iglesias	
Programa de Doctorado: Biomedicina y Oncología Molecular	
Órgano responsable: Centro Internacional de Postgrado	

RESUMEN (en español)

Las enfermedades cardiovasculares no son una causa común de mortalidad infantil; sin embargo, constituyen la principal causa de muerte y discapacidad en adultos en países desarrollados. Su principal desencadenante es la aterosclerosis, un proceso inflamatorio que daña y obstruye las arterias mediante la formación de placas de ateroma a lo largo de sus paredes. Este proceso puede iniciarse en la infancia, incluso durante la gestación, y generalmente progresa gradualmente hasta la edad adulta. A medida que las placas se expanden y se endurecen con el tiempo, pueden llegar a obstruir por completo las arterias, resultando en trombosis arterial y, en última instancia, en eventos cardiovasculares como la cardiopatía isquémica o accidentes cerebrovasculares, entre otros.

Intervenir en los principales factores de riesgo cardiovascular puede retrasar el desarrollo de la aterosclerosis y reducir el riesgo de enfermedades cardiovasculares futura. Los factores de riesgo metabólicos, que incluyen la obesidad central, la resistencia a la insulina, la hipertensión y la dislipidemia, son especialmente relevantes tanto en la edad adulta como en la juventud. Hasta la fecha, no son muchos los estudios que han podido evaluar la asociación entre estos factores metabólicos en edades pediátricas y el riesgo posterior de enfermedades cardiovasculares en la edad adulta, por lo que no está demostrado que estos factores se puedan considerar de riesgo a estas edades. No obstante, dada la naturaleza acumulativa de la aterosclerosis y su inicio temprano en la vida, es crucial iniciar la prevención cardiovascular en la infancia.

En este sentido, los estudios de cohortes longitudinales representan una valiosa herramienta para analizar la progresión de dichos factores de riesgo cardiovascular desde las primeras etapas de la vida. Estos estudios permiten registrar variables indicadoras de riesgo cardiovascular en múltiples momentos a lo largo del período de estudio, lo que posibilita una comprensión más completa y detallada de cómo estos factores se desarrollan y cambian con el tiempo.

Teniendo esto en cuenta, el principal objetivo de este estudio ha sido examinar la evolución o seguimiento de los factores de riesgo metabólicos mencionados anteriormente en edades pediátricas. Se buscó identificar cuáles de estos factores persisten durante la infancia y, por lo tanto, pueden requerir de atención preventiva. Para ello, se utilizaron datos de niños participantes en la cohorte INMA-Asturias, con edades comprendidas entre los 4 y los 8 años.

Por otro lado, la realización de cualquier estudio epidemiológico requiere una toma de decisiones metodológicas y estadísticas fundamentales para abordar adecuadamente las preguntas de investigación. Este trabajo se enfrentó a dos desafíos principales: cómo

abordar la presencia de datos ausentes y qué modelos de regresión eran más apropiados para analizar el seguimiento de los factores de riesgo cardiovascular en edades tan tempranas. Se valoraron las técnicas estadísticas más adecuadas para enfrentar estos desafíos, siendo seleccionadas: la imputación múltiple para manejar los datos ausentes y los modelos de regresión cuantil para realizar el seguimiento de los factores de riesgo metabólicos.

La presente memoria recoge el trabajo anteriormente descrito y está organizada de la siguiente manera. En el Capítulo 1, titulado Introducción, se incluye una breve revisión histórica sobre la etiología de las enfermedades cardiovasculares y la identificación de sus factores de riesgo. Además, se presenta la cohorte INMA-Asturias, y se introduce formalmente el concepto de seguimiento. Finalmente, se abordan las principales dificultades metodológicas que surgieron durante el estudio y se explican los enfoques estadísticos seleccionados como los más apropiados. En el Capítulo 2 se recogen los objetivos específicos propuestos en este trabajo. El Capítulo 3, titulado Métodos y resultados, presenta los tres artículos que respaldan este documento y que recogen los métodos utilizados y los resultados obtenidos. En el Capítulo 4 se lleva a cabo una discusión general de la investigación realizada, presentándola de forma integrada y cohesionada. Finalmente, en el Capítulo 5 se presentan de forma ordenada las conclusiones derivadas de este trabajo.

RESUMEN (en Inglés)

Cardiovascular diseases are not a common cause of mortality in childhood; however, in the developed countries, they constitute the leading cause of death and disability in adults. Their main trigger is atherosclerosis, an inflammatory process that damages and obstructs arteries by forming atheromatous plaques along their walls. This process can initiate in childhood, even during gestation, and typically progresses gradually into adulthood. As plaques expand and harden over time, they can completely block arteries, potentially leading to arterial thrombosis and, ultimately, cardiovascular events such as ischemic heart disease or ischemic stroke, among others.

Intervening in key cardiovascular risk factors can delay the development of atherosclerosis and reduce the risk of cardiovascular diseases in the future. Metabolic risk factors, including central obesity, insulin resistance, hypertension, and dyslipidemia, are particularly relevant both in adulthood and youth. To date, there has been limited research assessing the association between these metabolic factors in pediatric ages and the subsequent risk of cardiovascular diseases in adulthood, so it is not yet proven whether these factors can be considered risks in childhood. Nevertheless, given the cumulative nature of atherosclerosis and its early onset in life, it is crucial to initiate cardiovascular prevention during childhood.

In this sense, longitudinal cohort studies constitute a valuable tool for examining the development and progression of cardiovascular risk factors from early stages of life. These studies enable the collection of cardiovascular-related variables at multiple time points throughout the study period, providing a more comprehensive and deeper understanding of how these factors progress and change over time.

Taking this into account, the primary objective of this study has been to investigate the persistence of the mentioned metabolic risk factors during pediatric ages, a phenomenon known as 'tracking". The aim was to identify which of these factors persist during childhood and, therefore, may require preventive attention. Data from children participating in the INMA-Asturias cohort, aged 4 to 8 years, were used for this purpose.

On the other hand, conducting any epidemiological study involves making several methodological and statistical decisions to adequately address the research question. This work deals with two main challenges: how to address the missing data problem, and which regression models are most appropriate for analyzing tracking of cardiovascular risk factors

Universidad de Oviedo

at these early ages. Suitable statistical techniques were considered to face these challenges, leading to the selection of multiple imputation for handling missing data and quantile regression models for tracking metabolic risk factors.

This thesis presents the previously described work and is organized as follows. Chapter 1, entitled Introduction, includes a brief historical review of the etiology of cardiovascular diseases and the identification of their risk factors. Furthermore, we provide an overview of the INMA-Asturias cohort, the population in which this research was conducted, and we formalize the notion of tracking. Lastly, we delve into the main methodological challenges faced during the study and offer insights into our chosen statistical methodologies. Chapter 2 contains the specific objectives proposed in this dissertation. In Chapter 3, entitled Methods and results, the three articles supporting this PhD memory are provided, including the methods and the results derived from this work. Chapter 4 provides a general discussion of the obtained results, presenting them in an integrated and cohesive form. Lastly, conclusions derived from this work are presented in Chapter 5.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN BIOMEDICINA Y ONCOLOGÍA MOLECULAR

Agradecimientos

Con estas líneas quisiera mostrar mi agradecimiento a quienes me han acompañado durante estos años en la realización de este trabajo. Ha sido un camino largo y en ocasiones difícil, pero no lo he recorrido sola. Con su ayuda y su apoyo ha sido posible llegar hasta aquí.

Quiero comenzar expresando mi agradecimiento a mis directoras de tesis. A la Dra. Adonina Tardón, por brindarme esta oportunidad que tanto me ha aportado a nivel profesional y personal, y por haberme enseñado tantas lecciones valiosas a lo largo de esta etapa académica. A la Dra. Ana Fernández Somoano, gracias por confiar en mi trabajo, por haberme acompañado y animado a seguir adelante, y por tener siempre su puerta abierta para mí.

También quisiera agradecer a la Dra. Isolina Riaño sus aportaciones y el tiempo dedicado a los trabajos recogidos en en esta memoria, así como su amabilidad y cercanía. Extiendo este agradecimiento a las madres, padres y niños que de manera desinteresada participan en el proyecto INMA-Asturias, así como a la Dra. Cristina Rodríguez Dehli, a todo el personal médico, técnicos de laboratorio, de investigación... que, de una manera u otra, han contribuido a este proyecto.

A mis compañeros de trabajo, la mejor parte del día a día durante estos años. Gracias especialmente a Miguel, por su apoyo y confianza constantes, y por ofrecerme siempre su ayuda, antes incluso de que me diese tiempo a pedírsela. Muchas gracias a Isa, a Cris y a Vero por todos los cafés compartidos, por sus palabras de ánimo y por su cariño.

En estos años de formación he tenido la suerte de poder realizar dos estancias predoctorales que han contribuido enormemente a este trabajo, así como a mi aprendizaje académico y personal. Gracias al Dr. Jesús Vioque por darme la oportunidad de realizar mi estancia en la Unidad de Epidemiología de la Nutrición de la Universidad Miguel Hernández de Elche, y al Dr. Antonio Signes Pastor por compartir conmigo sus conocimientos, así como

su entusiasmo. A la Dra. Margaret Karagas, por su cálida acogida en el Departamento de Epidemiología de la Geisel School of Medicine en Dartmouth. Su amabilidad y su ayuda han hecho de esta estancia una experiencia enriquecedora. Gracias de todo corazón al Dr. Pablo Martínez Camblor, por ser tan generoso conmigo e implicarse como lo ha hecho para que consiguiese llegar a buen puerto. Gracias por ejercer como mentor no sólo a nivel científico, sino también personal. A Susana y a Unai, por compartir su tiempo conmigo. Guardo un recuerdo muy especial de esta estancia en Estados Unidos, y me siento muy agradecida de haber seguido disfrutando de tiempo con vosotros a este lado del charco. Gracias por acogerme.

A mis amigos, por su apoyo y comprensión durante este periodo. En especial, gracias a Sonia por sus buenos consejos, por ayudarme con absolutamente todo, y por haberme escuchado y comprendido en los momentos difíciles. A Miriam, que siempre ha estado presente alegrándose por mí, a pesar del tiempo que pasásemos sin vernos, y por ser ejemplo en tantas cosas. A Sheila, por creer de verdad en ese "tú puedes con todo" que hace que hasta yo misma, en ocasiones, me lo crea.

Y finalmente, a mi familia. Gracias a mis padres, por preocuparse siempre de darme todas las oportunidades posibles. Por animarme y apoyarme incondicionalmente en mis decisiones. Vuestro amor y confianza me han permitido llegar a lugares que ni siquiera imaginaba. Gracias a Abel, por cuidarme y ser como un hermano para mí. A Cristina, Alba y Sergio, por todos los buenos ratos compartidos, que tanto me han ayudado a desconectar y coger fuerzas. A Mari y Pepe, por su cariño y apoyo. Y por último, quiero dar las gracias a Nacho. Por ayudarme a superar cualquier reto, por acompañarme en los momentos de incertidumbre y cansancio, pero sobre todo por tu amor y cariño diarios. Gracias por hacer equipo conmigo.

Traslado también mi agradecimiento al Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP). Su financiación me ha permitido realizar el trabajo recogido en la presente memoria al amparo de un contrato laboral, así como la realización de una estancia de investigación de tres meses en el Departamento de Epidemiología de la Geisel School of Medicine (Estados Unidos) y una estancia de un mes en la Unidad de Epidemiología de la Nutrición de la Universidad Miguel Hernández de Elche (Alicante).

También quisiera mostrar mi gratitud a la Comisión Académica del Programa de Doctorado de Biomedicina y Oncología Molecular, a las expertas internacionales que han revisado esta memoria, y a todos los miembros del tribunal que evaluará la defensa de esta tesis doctoral.

Acknowledgements

With these lines, I would like to express my gratitude to those who have supported me during these years in the realization of this work. It has been a long and challenging journey, but I have not walked it alone. With their help and support, getting this far has been possible.

I want to begin by expressing my gratitude to my thesis supervisors. To Dr. Adonina Tardón, for giving me this opportunity that has contributed so much to my professional and personal growth, and for teaching me valuable lessons throughout this academic stage. To Dr. Ana Fernández Somoano, thank you for your confidence in my work, for accompanying and encouraging me at every step, and for always having your door open for me.

I would also like to thank Dr. Isolina Riaño for her contributions and time dedicated to the work presented in this document, as well as her kindness and approachability. I extend this gratitude to the mothers, fathers, and children who selflessly participate in the INMA-Asturias project, as well as to Dr. Cristina Rodríguez Dehli, all the medical staff, laboratory technicians, researchers, and others who, in one way or another, have contributed to this project.

To my colleagues, the best part of day-to-day life during these years. Thanks especially to Miguel for his constant support and trust, and for always offering his help even before I had the chance to ask. Many thanks to Isa, Cris and Vero for all the shared coffees, their words of encouragement, and their affection.

During these years of training, I have been fortunate to undertake two predoctoral stays that have greatly contributed to this work, as well as my academic and personal learning. Thanks to Dr. Jesús Vioque for giving me the opportunity to conduct my stay at the Nutrition Epidemiology Unit of the Miguel Hernández University of Elche, and to Dr. Antonio Signes Pastor for sharing his knowledge and enthusiasm with me. To

Dr. Margaret Karagas, for her warm welcome to the Department of Epidemiology at the Geisel School of Medicine at Dartmouth. Her kindness and assistance made this stay an enriching experience. Heartfelt thanks to Dr. Pablo Martínez Camblor for being so generous with me and for getting involved to ensure the success of this project. Thank you for serving as a mentor not only scientifically but also personally. To Susana and Unai, for spending time with me. I treasure special memories of this stay in the United States, and I am very grateful to have continued enjoying time with you on this side of the pond. Thank you for welcoming me.

To my friends, for their support and understanding during this period. In particular, thanks to Sonia for her good advice, for helping me with absolutely everything, and for listening and understanding me in difficult times. To Miriam, who has always been present, rejoicing in my achievements, despite the time we spent apart, and for being an example in many ways. To Sheila, for truly believing in that "you can do it all" that sometimes even makes me believe it myself.

Last but not least, to my family. Thanks to my parents for always ensuring I have all possible opportunities. For encouraging and supporting me unconditionally in my decisions. Your love and trust have allowed me to reach places I never even imagined. Thank you to Abel for taking care of me and being like a brother to me. To Cristina, Alba, and Sergio, for all the good times shared, which have helped me disconnect and gather strength. To Mari and Pepe, for their love and support. And finally, I want to thank Nacho. For helping me to overcome any challenge, for supporting me in moments of uncertainty and tiredness, but above all, for his daily love and devotion. Thank you for being a team with me.

I also extend my gratitude to the Consortium for Research on Epidemiology and Public Health (CIBERESP). Their funding has allowed me to carry out the work presented in this dissertation under an employment contract, as well as undertake a three-month research stay in the Department of Epidemiology at the Geisel School of Medicine (United States) and a one-month stay in the Nutrition Epidemiology Unit at the Miguel Hernández University of Elche (Alicante).

I would also like to express my gratitude to the Academic Committee of the Biomedicine and Molecular Oncology Doctoral Program, to the international experts who have reviewed this dissertation, and to all the members of the tribunal who will evaluate the defense of this doctoral thesis.

Abstract

Cardiovascular diseases are not a common cause of mortality in childhood; however, in developed countries, they constitute the leading cause of death and disability in adults. Their main trigger is atherosclerosis, an inflammatory process that damages and obstructs arteries by forming atheromatous plaques along their walls. This process can initiate in childhood, even during gestation, and typically progresses gradually into adulthood. As plaques expand and harden over time, they can completely block arteries, potentially leading to arterial thrombosis and, ultimately, cardiovascular events such as ischemic heart disease or ischemic stroke, among others.

Intervening key cardiovascular risk factors can delay the development of atherosclerosis and reduce the risk of cardiovascular diseases in the future. Metabolic risk factors, including central obesity, insulin resistance, hypertension, and dyslipidemia, are particularly relevant in both adulthood and youth. To date, there has been limited research assessing the association between these metabolic factors at pediatric ages and the subsequent risk of cardiovascular diseases in adulthood, therefore it is not yet proven whether these factors can be considered risks in childhood. Nevertheless, given the cumulative nature of atherosclerosis and its early onset in life, it is crucial to initiate cardiovascular prevention during childhood.

In this sense, longitudinal cohort studies constitute a valuable tool for examining the development and progression of cardiovascular risk factors from early stages of life. These studies enable the collection of cardiovascular-related variables at multiple time points throughout the study period, providing a more comprehensive and deeper understanding of how these factors progress and change over time.

Taking this into account, the primary objective of this study has been to investigate the persistence of the mentioned metabolic risk factors during pediatric ages, a phenomenon known as 'tracking". The aim was to identify which of these factors persist during childhood and, therefore, may require preventive attention. Data from children participating in the INMA-Asturias cohort, aged 4 to 8 years, were used for this purpose.

On the other hand, conducting any epidemiological study involves making several methodological and statistical decisions to adequately address the research question. This work deals with two main challenges: how to address the missing data problem, and which regression models are most appropriate for analyzing tracking of cardiovascular risk factors at these early ages. Suitable statistical techniques were considered to face these challenges, leading to the selection of multiple imputation for handling missing data and quantile regression models for tracking metabolic risk factors.

This thesis presents the previously described work and is organized as follows: in **Chapter 1**, entitled **Introduction**, a contextualization of the scientific interest behind studying the persistence of cardiovascular risk factors in childhood is provided. This chapter includes a brief historical review of the etiology of cardiovascular diseases and the identification of their risk factors. Major epidemiological contributions to knowledge in this field that have focused on childhood serve as a guiding thread. Furthermore, we provide an overview of the INMA-Asturias cohort, in which this research was conducted, and we formalize the notion of tracking. Finally, we delve into the main methodological challenges faced during the study and offer insights into our chosen statistical methodologies.

Chapter 2 contains the specific **objectives** proposed in this dissertation.

In **Chapter 3**, entitled **Methods and results**, the three articles supporting this PhD memory are provided, including the methods and the results derived from this work.

Chapter 4 provides a general discussion of the obtained results, presenting them in an integrated and cohesive form. Reflection is made on both their potential implications for child health, and the potential of the chosen statistical methodologies for addressing inherent research limitations.

Lastly, **conclusions** derived from this work are presented in **Chapter 5**. In addition, the **supplementary material** corresponding to the articles included in this document is provided in the section designated with the same name.

Resumen

Las enfermedades cardiovasculares no son una causa común de mortalidad infantil; sin embargo, constituyen la principal causa de muerte y discapacidad en adultos en países desarrollados. Su principal desencadenante es la aterosclerosis, un proceso inflamatorio que daña y obstruye las arterias mediante la formación de placas de ateroma a lo largo de sus paredes. Este proceso puede iniciarse en la infancia, incluso durante la gestación, y generalmente progresa gradualmente hasta la edad adulta. A medida que las placas se expanden y se endurecen con el tiempo, pueden llegar a obstruir por completo las arterias, resultando en trombosis arterial y, en última instancia, en eventos cardiovasculares como la cardiopatía isquémica o accidentes cerebrovasculares, entre otros.

Intervenir en los principales factores de riesgo cardiovascular puede retrasar el desarrollo de la aterosclerosis y reducir el riesgo de enfermedades cardiovasculares futura. Los factores de riesgo metabólicos, que incluyen la obesidad central, la resistencia a la insulina, la hipertensión y la dislipidemia, son especialmente relevantes tanto en la edad adulta como en la juventud. Hasta la fecha, no son muchos los estudios que han podido evaluar la asociación entre estos factores metabólicos en edades pediátricas y el riesgo posterior de enfermedades cardiovasculares en la edad adulta, por lo que no está demostrado que estos factores se puedan considerar como de riesgo ya a estas edades. No obstante, dada la naturaleza acumulativa de la aterosclerosis y su inicio temprano en la vida, es crucial iniciar la prevención cardiovascular en la infancia.

En este sentido, los estudios de cohortes longitudinales representan una valiosa herramienta para analizar la progresión de dichos factores de riesgo cardiovascular desde las primeras etapas de la vida. Estos estudios permiten registrar variables indicadoras de riesgo cardiovascular en múltiples momentos a lo largo del período de estudio, lo que posibilita

una comprensión más completa y detallada de cómo estos factores se desarrollan y cambian con el tiempo.

Teniendo esto en cuenta, el principal objetivo de este estudio ha sido examinar la evolución o seguimiento de los factores de riesgo metabólicos mencionados anteriormente en edades pediátricas. Se buscó identificar cuáles de estos factores persisten durante la infancia y, por lo tanto, pueden requerir de atención preventiva. Para ello, se utilizaron datos de niños participantes en la cohorte INMA-Asturias, con edades comprendidas entre los 4 y los 8 años.

Por otro lado, la realización de cualquier estudio epidemiológico requiere una toma de decisiones metodológicas y estadísticas fundamentales para abordar adecuadamente las preguntas de investigación. Este trabajo se enfrentó a dos desafíos principales: cómo abordar la presencia de datos ausentes y qué modelos de regresión eran más apropiados para analizar el seguimiento de los factores de riesgo cardiovascular en edades tan tempranas. Se valoraron las técnicas estadísticas más adecuadas para enfrentar estos desafíos, siendo seleccionadas: la imputación múltiple para manejar los datos ausentes y los modelos de regresión cuantil para realizar el seguimiento de los factores de riesgo metabólicos.

La presente memoria recoge el trabajo anteriormente descrito y está organizada de la siguiente manera. En el **Capítulo 1**, titulado **Introducción**, se proporciona una contextualización del interés científico detrás del estudio del seguimiento de los factores de riesgo cardiovascular en la infancia. Este capítulo incluye una breve revisión histórica sobre la etiología de las enfermedades cardiovasculares y la identificación de sus factores de riesgo. Se utiliza como hilo conductor los principales estudios epidemiológicos que a lo largo de la historia han contribuido al conocimiento en este campo y han puesto el foco en la infancia. Además, se presenta la cohorte INMA-Asturias, en la que se enmarca este trabajo, y se introduce formalmente el concepto de seguimiento. Finalmente, se abordan las principales dificultades metodológicas que surgieron durante el estudio y se explican los enfoques estadísticos seleccionados como los más apropiados.

En el **Capítulo 2** se recogen los **objetivos** específicos propuestos en este trabajo.

El **Capítulo 3**, titulado **Métodos y resultados**, presenta los tres artículos que respaldan este documento y que recogen los métodos utilizados con el fin de llevar a cabo la investigación, así como los resultados obtenidos.

En el **Capítulo 4** se lleva a cabo una **discusión** general de la investigación realizada, presentándola de forma integrada y cohesionada. Se reflexiona tanto sobre sus posibles implicaciones en términos de salud infantil, como en el potencial de las metodologías estadísticas escogidas para abordar las limitaciones inherentes a la investigación.

Finalmente, en el **Capítulo 5** se presentan de forma ordenada las **conclusiones** derivadas de este trabajo. Además, se incluye el **material suplementario** correspondiente a los artículos que conforman esta tesis, en la sección designada con el mismo nombre.

Abbreviations

AC Atherogenic coefficient

BMI Body mass index

CVDs Cardiovascular diseases

FIML Full information maximum likelihood method

GBD Global Burden of Disease

HDL-c High density lipoprotein cholesterol

INMA INfancia y Medio Ambiente

MAR Missing at random

MCAR Missing completely at random

MI Multiple Imputation

MNAR Missing not at random

QRMs Quantile Regression Models

USA United States of America

WC/Height Waist circunference to height

Contents

CHAPTER 1

Introduction

1.1. Cardiovascular disease

Cardiovascular diseases (CVDs) constitute the major cause of death in both developed and developing countries. In 2019, an estimated 17.9 million people died from CVDs, representing 32% of all global deaths. In addition, of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases, 38% were attributed to CVDs (WHO, 2023). In Spain, in 2022, CVDs continued to hold the first place in the mortality statistics, contributing to 26.0% of all deaths, closely followed by tumors at 24.8% (Instituto Nacional de Estadística, 2022).

The shift in mortality patterns in developed countries began between 1900 and 1930, when infectious diseases stopped being the main cause of death, giving way to chronic diseases. This process is commonly referred to as the "epidemiological transition". During this phase the incidence and mortality rates of CVDs began to show a marked increase, shaping the direction of epidemiological research and interventions needed to address this new scenario (Celentano and Moyses, 2014). Since then, health authorities have dedicated substantial efforts to manage the burden of these diseases, as evidenced by the consistent decline in global mortality rates since the 1990s, including Spain (Amini et al., 2021; Flores-Mateo et al., 2011). However, ischemic heart disease and ischemic stroke remain the leading causes of death. These two conditions also have the highest mortality rates among all CVDs, as illustrated in Table 1. Although mortality rates

have presented this clearly decreasing trend, the reduction in CVDs incidence has not been as pronounced. It is essential not to underestimate the relevance of CVD incidence, as even when a cardiovascular event does not result in death, it frequently leads to varying degrees of disability. This results in a decrease in overall health and quality of life of the population, along with a substantial economic burden on the healthcare system. Although records of CVD incidence and morbidity are not as precise as those of mortality, in Spain, they constitute the primary cause of hospitalization. In 2021, there were 582,446 hospital stays due to CVDs, which represents 12.9% of the total annual admissions (Instituto Nacional de Estadística, 2023). Furthermore, in 2017, CVDs were the leading contributors to the loss of healthy years due to disability (disability-adjusted life years), both globally (Feigin et al., 2021) and in Spain (Soriano et al., 2018).

Table 1: Cardiovascular deaths in all ages (high/middle income countries worldwide, and Spain) - Estimated 2019 data from the Global Burden of Disease (GBD).

	High/middle income countries		**Spain**	
Cardiovascular cause of death	**n**	**%**	**n**	**%**
Ischemic heart disease	2,658,294	51.6	53,633	40.8
Ischemic stroke	1,126,009	21.9	23,256	17.7
Intracerebral hemorrhage	589,148	11.4	11,172	8.5
Hypertensive heart disease	252,277	4.9	8,728	6.6
Cardiomyopathy and myocarditis	150,010	2.9	6,443	4.9
Atrial fibrillation and flutter	83,845	1.6	7,379	5.6
Subarachnoid hemorrhage	79,181	1.5	2,666	2.0
Other cardiovascular/circulatory diseases	58,118	1.1	3,479	2.6
Aortic aneurysm	45,029	0.9	2,402	1.8
Rheumatic heart disease	36,961	0.7	2,493	1.9
Non-rheumatic valvular heart disease	32,893	0.6	6,178	4.7
Peripheral artery disease	26,244	0.5	1,890	1.4
Endocarditis	13,259	0.3	1,782	1.4
Total	**5,151,261**	**100**	**131,501**	**100**

Source: https://vizhub.healthdata.org/gbd-results/

However, CVDs are rare in pediatric ages, with a very low incidence compared with adult populations. In developed countries, including Spain, approximately 10% of deaths in children aged 0 to 14 years are attributed to cardiovascular events. While congenital heart defects are the most frequent, mortality due to chronic CVDs like ischemic heart disease or ischemic stroke is extremely rare during childhood, as shown in Table 2.

Table 2: Cardiovascular deaths in children aged 0 -14 (high/middle income countries worldwide, and Spain) - Estimated 2019 data from the GBD.

	High/middle income countries		**Spain**	
Cardiovascular cause of death	**n**	**%**	**n**	**%**
Congenital heart anomalies	16,278	86.6	125	78.6
Cardiomyopathy and myocarditis	1,000	5.3	11	6.9
Other cardiovascular/circulatory diseases	647	3.4	10	6.3
Intracerebral hemorrhage	300	1.6	5	3.1
Subarachnoid hemorrhage	226	1.2	4	2.5
Rheumatic heart disease	136	0.7	1	0.6
Endocarditis	117	0.6	2	1.3
Ischemic stroke	83	0.4	1	0.6
Total	**18,787**	**100**	**159**	**100**

Source: https://vizhub.healthdata.org/gbd-results/

1.2. Atheroesclerosis

The anatomopathological basis of the most incident CVDs is **arteriosclerosis**, a term originally introduced by Lobstein (1829). Arteriosclerosis is characterized by progressive hardening and narrowing of arteries, independent of vessel size and organ sites (Strasses, 1972). It is important to distinguish the term "arteriosclerosis" from "atherosclerosis". Arteriosclerosis encompasses three different morphological features: Mönckeberg's arteriosclerosis, arteriolosclerosis, and atherosclerosis. Among these, atherosclerosis is the most incident form and a major cause of cardiovascular death.

Atherosclerosis, term first introduced by Marchand (1904), is an inflammatory chronic disease characterized by lesions in the arterial intima and media layers, resulting in narrowing and hardening. These lesions occur because of the accumulation of **atheromatous plaques**, consisting of lipids, fibrous tissue, and inflammatory cells (Ross, 1999). In fact, the word *"athere"* –prefix of atherosclerosis– means mush, gruel, or porridge in Greek to indicate lipid deposition in the arterial wall, whereas the suffix *"sclerosis"* refers to hardening. Atherosclerosis primarily affects large and medium arteries, including the coronary arteries supplying the heart; carotid, vertebral, and cerebral arteries supplying the brain; and iliac and femoral arteries supplying the lower extremities. It is a localized and progressive condition, as not all arteries are susceptible to atherosclerosis, and lesions tend to occur at specific sites such as artery branches and orifices. The disease progresses from early lesions, such as diffuse intimal thickening and fatty streaks, to advanced or complicated ones in the form of atheromatous plaques. The process begins with the deposition of small amounts of fat between the thin layers of the arteries (fatty streak) and slowly progresses with age, and exposure to certain risk factors. This exposure to specific factors, which we will examine further, triggers highly complex cellular and biochemical mechanisms and processes that lead to the growth of the fatty streak by attracting certain types of cells, finally forming the atheromatous plaques. A cascade of inflammatory reactions, along with other mechanical factors, can lead to clinical symptoms such as stenosis, calcification, hemorrhage, ulceration, or rupture of the plaque. If this happens, blood platelets come to the site, aggregate, and result in what is known as thrombosis, which can partially or completely obstruct the arterial lumen, preventing the circulation of blood and, consequently, the supply of oxygen required for the tissues. The consequence is cellular death or tissue necrosis in the areas supplied by the occluded artery, ultimately resulting in ischemia (Fan and Watanabe, 2022). The histopathological classification of atherosclerotic lesions, as published by Stary et al. (1995), categorizes the lesions into six or eight categories based on their progression and clinical relevance. It may take several decades for advanced lesions to develop. Therefore, it could remain clinically silent for many years until advanced lesions occur. However, it is also a disease whose progression can be influenced by medical interventions and lifestyle modifications, and in some cases may remit. (Libby, 2021).

The clinical manifestations of atherosclerosis depend on the affected arteries, degree of arterial occlusion, and speed of progression. Ischemic heart disease is the main manifestation in the coronary arteries, whereas ischemic stroke occurs in the cerebral

arteries, and peripheral arterial disease occurs in the iliac and femoral arteries (Virmani et al., 2006).

Due to the increase in the mortality rate in the early decades of the 1900s, as previously mentioned in the Section 1.1, studies based on autopsies in subjects of different ages and populations were initiated around 1950. These studies revealed early atherosclerotic lesions in infant and young populations. One of the most renowned studies in this regard was conducted by Enos et al. (1955). They analyzed a series of cases involving 300 young American soldiers who died in the Korean War, with a mean age of 22.1 years. An unexpectedly high prevalence of atherosclerotic lesions was observed, ranging from initial to more advanced stages, irrespective of age. However, during the same period, case series were also analyzed in Japanese populations of all age ranges, where the prevalence of atherosclerotic lesions was very low, 1.7% compared to the over 65% observed in the study by Enos et al. (1955). Despite this inconsistency, the study served to draw the attention of the medical community to atherosclerosis in childhood and youth.

Subsequently, various studies were conducted, such as those led by the International Atherosclerosis Project, which analyzed autopsies of child populations in New Orleans (United States of America, USA) and later in other countries, including Japan, Guatemala, Costa Rica, and different regions in South Africa. These studies observed fatty streaks in the aorta in children aged 3, and more frequently in the coronary artery in children from age 10 (Strong and McGill, 1969). The Pathobiological Determinants of Atherosclerosis in Youth study (Sternby et al., 1999) further investigated subjects aged between 5 and 34 years from 15 developed and developing countries, representing five regions of the World Health Organization. This study encompassed different economic, sociocultural, and nutritional patterns. The conclusion drawn was that atherosclerotic lesions begin to appear in the early stages of life, regardless of sex, geographic origin, or socioeconomic level. The rate of appearance of fatty streaks was higher between the ages of 15 and 25, whereas fibrous plaque lesions started to develop during the second decade of life and progressed at an increasing rate during the third and fourth decades (Figure 1). Autopsy studies have continued to demonstrate that initial alterations in the arterial intima can be detected even during prenatal and infant periods (Milei et al., 2008).

Although clinical manifestations of atherosclerosis rarely occur in childhood, its precursor states clearly begin to develop during this period. It is uncertain whether fatty streaks or initial lesions necessarily progress into advanced lesions and subsequently lead to

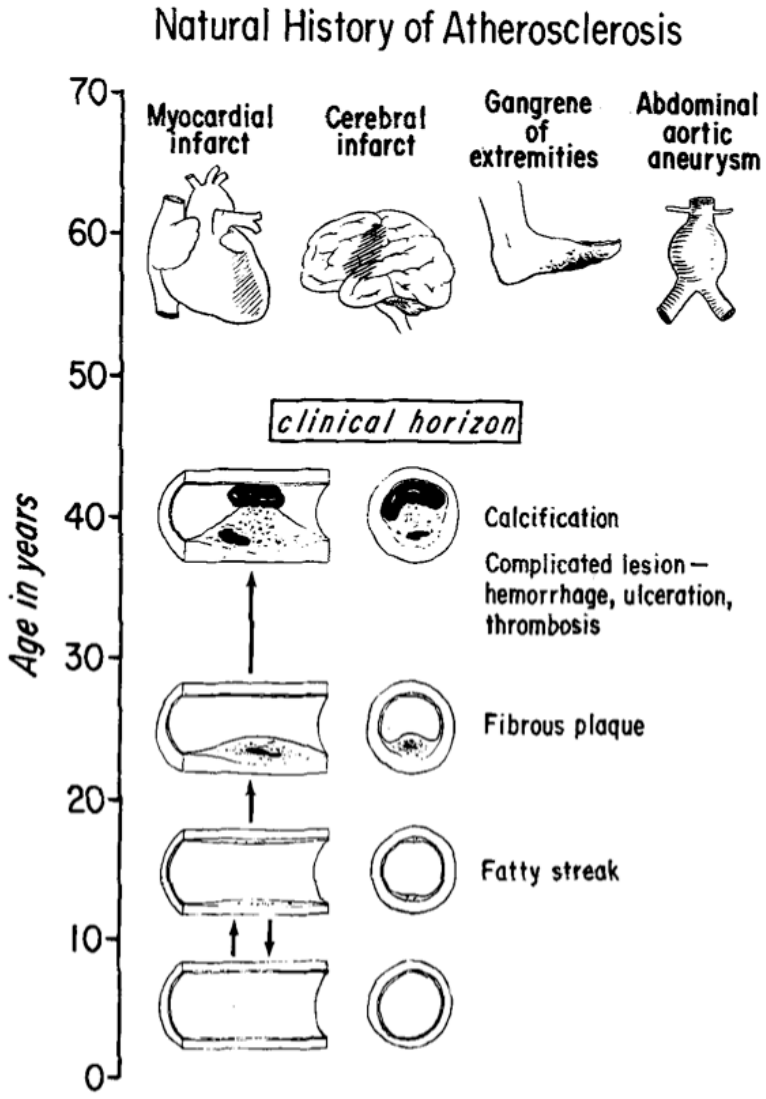


Figure 1: Natural history of atherosclerosis. Source: McGili et al., Natural history of human atherosclerotic lesions. In: Saddlerandg and Bourne (Eds.), Atherosclerosis and its origin, Academic Press, New York, 1963

clinical complications in adulthood, but under certain conditions or in certain anatomical locations, they could do so (Mcgill et al., 2000). Hence, early interventions could have a crucial impact on the future course of CVD.

1.3. Cardiovascular risk factors

The detection and control of cardiovascular risk factors is an essential preventive strategy. At the same time that initial case series autopsy studies were conducted to begin investigating atherosclerosis, around 1950, longitudinal observational studies also began to emerge to understand the determinants of CVD development. In 1948, the landmark Framingham Study, the largest cardiovascular epidemiology study to date, was initiated. Men and women aged 28 to 62 years from Framingham, a small semi-urban industrial town with a relatively stable population of about 28,000 inhabitants located approximately 18 miles west of Boston (Massachusetts, USA), were recruited.

An initial cohort of 5,209 individuals was selected, and a smaller cohort consisting of the offspring of the first participants was later added. This study is currently (October 2023) ongoing with the third generation of participants and it has contributed to the establishment of the multifactorial risk profile for CVDs (Mahmood et al., 2014). Other notable longitudinal studies followed the Framingham Study, further contributing to the same cause. Noteworthy among these is the Seven Countries Study, the first major study to investigate diet and lifestyle along with other risk factors, across contrasting countries and cultures and over an extended period of time (Keys et al., 1984); the SCORE project in Europe, which introduced a risk score system tailored to the European population for assessing cardiovascular risk (Conroy et al., 2003), and that has been adapted specifically for the Spanish population (Sans et al., 2007); or the REGICOR study conducted in Spain, which deserves special recognition for its extensive work in conducting population-based research on ischemic heart disease, its risk factors, and overall cardiovascular prevention efforts (Bardají, 2013). Other relevant studies that focus on the examination of specific risk factors are the ENRICA study, which investigates lifestyle-related factors, and the PREDIMED study, which is linked to the preventive effect of the Mediterranean diet on CVDs. The findings from these observational studies were justified and histopathologically supported by studies related to the onset and development of atherosclerotic lesions mentioned in the Section 1.2.

Figure 2 presents the ranking of the top 12 risk factors contributing to the burden CVD according to the GBD, along with their variation between 1990 and 2019 (Roth et al., 2020). It is noteworthy that the only change observed in the ranking is a decrease in the position of tobacco consumption, while the others remain stable.

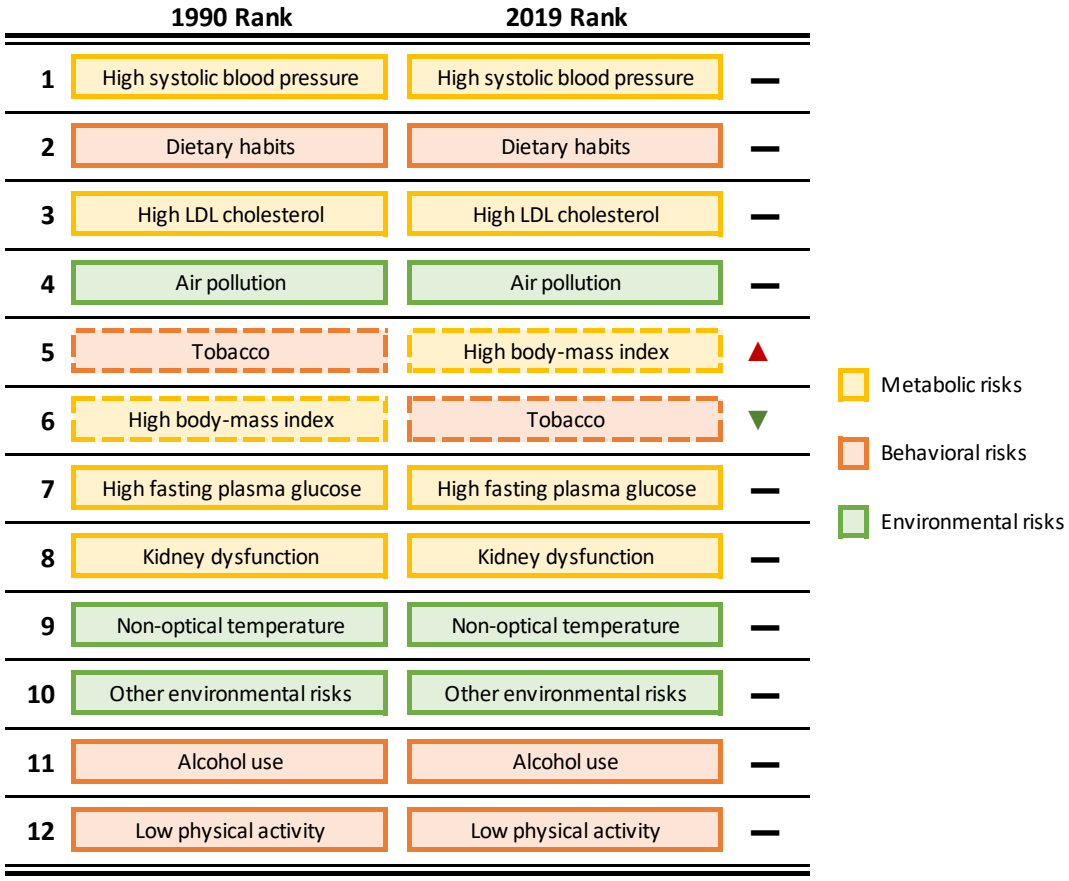


Figure 2: Ranking of cardiovascular risk factors. Adapted from G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Beaton, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. Journal of the American College of Cardiology, 76(25):2982–3021, 2020.

The comprehensive list of recognized risk factors is extensive and includes others that are not shown in Figure 2 as genetic factors (Ho et al., 2020), or family history of CVDs (Chacko et al., 2020). In general, these risk factors are considered "traditional", since their involvement in CVD development has been well-established. However, as early as 1984, Heller et al. (1984) observed a high percentage of CVD patients who did not exhibit any of the traditional risk factors. This observation suggested unexplored areas in the etiopathogenesis of atherosclerosis and the possible existence of other CVD risk factors. Indeed, our understanding of these emerging risk factors has expanded over time, including altered values of variables such as homocysteine (Guieu et al., 2022),

lipoprotein(a) (Duarte Lau and Giugliano, 2022), and C-reactive protein (Denegri and Boriani, 2021). Many of these interact with each other and are intercorrelated in the causal pathway of CVDs, potentially acting as risk factors between them. For instance, behavioral risk factors like dietary habits or physical activity can act as risk factors for others such as cholesterol or blood pressure. Therefore, elucidating the independent attributable risk of each factor constitutes an epidemiological challenge.

Metabolic risk factors are of particular relevance. They are considered to play a central role in the burden of ischemic heart disease (Wang et al., 2021), which as mentioned in the Section 1.1, has the highest mortality rate among CVDs. Additionally, these primary metabolic risk factors –including **central obesity**, **insulin resistance**, **hypertension**, and **dyslipidemia**– tend to cluster together, further increasing the risk of CVD beyond the sum of their individual effects. This clustering phenomenon is defined as metabolic syndrome (Reisinger et al., 2020). In this study, we will focus on these four cardiovascular risk factors.

1.3.1. Cardiovascular risk factors in children

The aforementioned epidemiological studies helped to establish these as risk factors in adulthood. However, considering that atherosclerosis begins to develop in childhood and progresses gradually over subsequent decades, the natural question arose as to whether these risk factors were also relevant in childhood. In other words, from what age would preventive measures targeting these factors reduce the future disease development? Attention was drawn to this issue starting in 1970. Several important longitudinal observational cohort studies emerged in USA with the primary objective of assessing the prevalence and progression of these risk factors during childhood and adolescence. Similar studies were also conducted in Europe or Australia. Table 3 lists some of the most noteworthy, including a Spanish study that, despite not being longitudinal, stands out in the investigation of cardiovascular risk factors in childhood.

Moreover, since the 1990s a new developmental model of diseases has gained interest, because of the findings by Barker (1990). This model suggests that events occurring during critical developmental periods, such as the prenatal period and early childhood, can influence the structure and function of the body, setting the stage for future health issues. Current studies support this theory, also regarding to CVD (Alexander et al., 2015). This underscores the importance of early-life interventions and preventive measures

Table 3: Remarkable observational studies about risk factors during childhood and adolescence.

Study	**Reference**	**Country**
Bogalusa Heart Study	Croft et al. (1984)	USA
Muscatine Study	Lauer et al. (1975)	USA
Princeton Lipid Research Clinics Study	Morrison et al. (2007)	USA
Minneapolis Children's Blood Pressure Study	Gillum et al. (1983)	USA
National Heart, Lung, and Blood Institute Growth and Health Study	Stone (1985)	USA
Australian Schools Health and Fitness Survey	Dwyer and Gibbons (1994)	Australia
European Youth Heart Study	Poortvliet et al. (2003)	Several european countries
Cardiovascular Risk in Young Finns Study	Akerblom et al. (1991)	Finland
4 Provinces Study	Rodríguez-Artalejo et al. (1999)	Spain

to mitigate the risk of chronic diseases, particularly atherosclerosis, adding more interest in the assessment of atherosclerosis risk factors during early childhood.

Despite all this evidences, only recently has it been possible to establish a direct relationship between the presence of adult risk factors in childhood and adolescence and cardiovascular events in adulthood. The cohorts mentioned in Table 3 were still relatively young, needing an extended follow-up period to accumulate a large enough number of CVD events to address the question. Currently (October 2023), the participants in these cohorts have reached an age at which cardiovascular events typically begin to manifest (around 40-50 years old for men and 60-70 for women), although they are still too young to observe a relevant number of events. Therefore, the four largest cohorts (the Cardiovascular Risk in Young Finns, the Australian Schools Health and Fitness Survey, the Bogalusa, and the Muscatine studies) have been consolidated into a consortium known as "The International Childhood Cardiovascular Cohort (i3C) Consortium" (Dwyer et al., 2013). This collaboration increases statistical power, thus reducing the

follow-up period required to determine an adequate number of CVD events. A recent systematic review (Pool et al., 2021), and the results that i3C has begun to produce (Jacobs et al., 2022), already provide consistent evidence of the relationship between the presence of cardiovascular risk factors in childhood and adolescence, and the future development of CVD later in life.

1.4. The INMA-Asturias cohort

Design is a crucial aspect of a research, which determines the final execution of the work, limits the conclusions which can be derived, and suggests the statistical procedures to be employed. It must optimize the available resources in order to get the research objectives. Given its relevance, we will outline the optimal design for investigating the tracking of cardiovascular risk factors during childhood.

Cohort studies

Although not explicitly highlighted thus far, a pattern can be observed in the type of epidemiological studies mentioned in the Section 1.3 as the primary contributors to the study of cardiovascular of risk factors in children and adolescents: they are longitudinal cohort studies.

Cohort studies involve selecting a group of people and observing them over time to assess their exposure level and the incidence of a specific outcome of interest (e.g., a disease), and to investigate the relationship between the exposure and the outcome. These are **observational** designs, meaning that the researcher acts as an observer of the study phenomenon, and does not manipulate any variable of interest. By definition, all cohort studies are **follow-up** studies, as the selected group of people is monitored over time. However, not all cohort studies are longitudinal, although those terms (follow-up and longitudinal) are often used interchangeably. **Longitudinal** studies are defined as those in which exposure and/or outcome variables (or other variables of interest) are repeatedly measured at different time points to assess the changes occurring in them.

One of the main advantages of cohort studies is their utility in establishing causal relationships, as they ensure that exposure temporally precedes the outcome, and facilitate the control of other potentially confounding or interacting variables. In addition, particularly in longitudinal designs, they facilitate the study of the natural history of the disease

or the evolution of its determinant factors. They also have disadvantages, such as high cost if follow-up must be carried out over a long period of time, and the validity of their results may be affected by dropouts of the participating subjects.

In the Section 1.3.1, we showed that prenatal development is one of the most critical windows during which adverse conditions and exposures may influence the future development of a disease, as CVD. In this context, **birth cohort studies**, based on the prospective recruitment and active follow-up of mothers and their children since pregnancy, are the most appropriate design to determine the causal relationship between potential risk factors during the prenatal or postnatal period and the health status of the newborn up to childhood and potentially adulthood (Canova and Cantarutti, 2020).

Considering all the characteristics of the types of studies mentioned in this section, longitudinal birth cohort studies are well-suited for investigating the tracking of cardiovascular risk factors, as they systematically measure variables of interest at different time points during fetal life and early infancy.

The INMA-Asturias cohort

To introduce the INMA-Asturias cohort, we will briefly digress about the topic of cardiovascular risk factors and CVD in general.

In 2003 **the INfancia y Medio Ambiente (INMA) Project** [Environment and Childhood Project] (https://www.proyectoinma.org) was established as a network of prospective birth cohorts in Spain, with the primary objective of investigating the influence of environmental pollutants present in air, water, and diet during pregnancy and early childhood on child growth and development (Guxens et al., 2012).

In particular, the specific objectives of the project were as follows:

INMA Project objectives

- To describe the degree of individual pre-natal exposure to environmental pollutants and the internal dose of these chemicals during pregnancy, at birth, and during childhood.
- To evaluate the impact of exposure to different contaminants on fetal and infant growth, health, and development.
- To evaluate the interaction between pollutants, nutrients, and genetic variants on fetal and infant growth, health, and development.

This network comprises diverse research groups with extensive expertise in environmental pollution and epidemiology. It consists of six cohorts (see Figure 3): three of them, namely Granada, Menorca, and Ribera d'Ebre, were already in existence at the beginning of the network, and their accumulated experience served as the foundation for the project (Granada, Menorca, and Ribera d'Ebre). In addition, four cohorts (Asturias, Sabadell, Valencia and Gipuzkoa) were subsequently established.

A total of 3,944 pregnant women in their first trimester of pregnancy were recruited from the general population residing in each of the specific areas, based on the following inclusion criteria: to be at least 16 years old, to have a singleton pregnancy, not to have followed any of assisted reproduction, to wish to deliver in the hospital of reference, and to have no communication handicaps.

Both mothers and children were followed up at several time points, with some variations among cohorts, but common follow-up periods included during pregnancy, at birth, and at the ages of 4, 8, and 12 years. Trained personnel collected data from different sources during each visit, including questionnaires, medical records, biological and environmental samples, and anthropometric measurements.

This dissertation is set within the framework of the **INMA-Asturias cohort**, which was established in 2004. The cohort is located in a 483 km^2 area in northern Spain, with the San Agustín University Hospital (Avilés, Asturias) serving as hospital of reference (see Figure 4). The economy of this region historically relied on industries characterized by remarkable environmental pollution (Fernández-Somoano and Tardon, 2014). Originally, in 2004, the area included a population of 165,201 inhabitants (reduced to 143,810 in

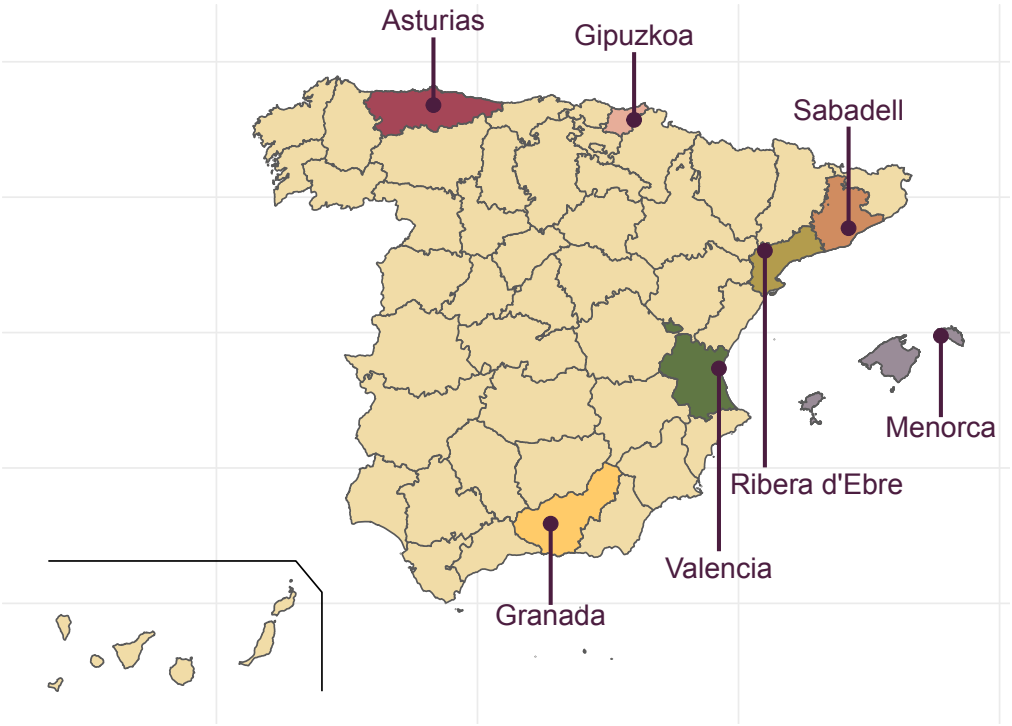


Figure 3: Network of cohorts from the INMA Project. Self-crafted figure.

2022), and the reference hospital was a public health center with 436 beds, providing primary care, as well as central, medical, and surgical services to this population.

From May 2004 to June 2007, pregnant women attending their first prenatal visit at the obstetrics service of San Agustín University Hospital or Las Vegas health center (Corvera, Avilés) were consecutively selected if they met the aforementioned inclusion criteria. Follow-up visits occurred during the first and third trimesters of pregnancy, at birth, and at the ages of 18 months, 4, 8, and 12 years.

To re-approach the topic under consideration here, it is noteworthy that the duration of cohort studies often extends beyond the initially established timeline due to its capacity to exploit the extensive data systematically collected. This valuable information not only supports the initially posited hypotheses, but also facilitates the exploration of new, future-relevant hypotheses or questions.

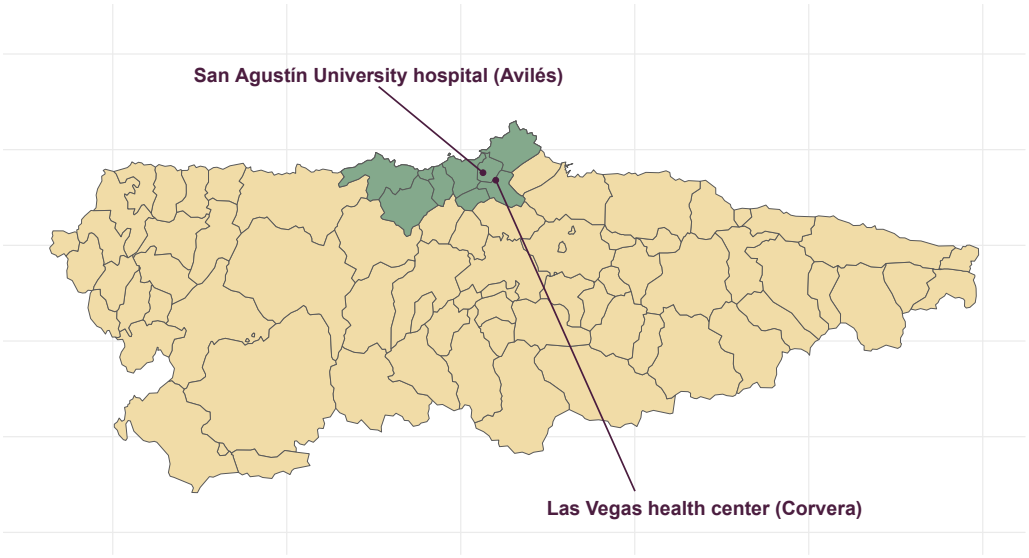


Figure 4: INMA-Asturias cohort recruitment area. Self-crafted figure.

This scenario is exemplified by the INMA study, particularly, by the INMA-Asturias cohort. Although the objectives of the project are linked to environmental exposure, the child population in Asturias currently faces relevant health challenges related to cardiovascular aspects. Asturias has a very high rate of childhood overweight and obesity, estimated at 33.2% in children aged 0 to 14 years (Domínguez Aurrecoechea et al., 2015). These high rates have also been observed in the INMA-Asturias cohort in previous studies, with a percentage of overweight or obese children at 4 years of age of 23.7% (Riaño-Galán et al., 2017). It is well known that obesity increases the risk of other metabolic risk factors (insulin resistance, hypertension, and dyslipidemia), besides increasing the risk of metabolic syndrome (Drozdz et al., 2021). Given this landscape, it is of interest to harness the information provided by the INMA-Asturias cohort to address this epidemiological problem. The availability of blood samples and anthropometric measurements have provided cardiovascular-related variables that reflect well-established CVD risk factors in adulthood, such as lipids, glucose, insulin, blood pressure, and waist circumference, among others. This has expanded the scope of research beyond the initial objectives to those explored in this memory.

1.5. Tracking of cardiovascular risk factors

Given that atherosclerosis is an accumulative process, which has its anatomical basis in the accumulation of atheromatous plaques, it is particularly appropriate to consider not only the presence of cardiovascular risk factors during childhood, but also their **stability** or **maintenance** over time. This is relevant because it can be hypothesized that the longer these risk factors persist, the greater the risk of progression in the natural history of CVD.

In epidemiology, the temporal stability of a biological variable (or specifically, of risk factors for chronic diseases) within a specific population is referred to as **tracking**. While this concept is challenging to translate into an exact and precise definition, its initial use in the late 1970s and subsequent applications have given rise to two definitions:

Definition of *tracking*

1. The ability to predict subsequent observations from earlier observations (Rosner et al., 1977). In a cohort of n children, where we measure their heights k times, resulting in sample values $y_{i,t}$ for $1 \leq i \leq n$ and $1 \leq t \leq k$, the concept of *tracking* refers to the ability to predict $y_{i,t}$ based on the earlier observations $y_{i,1}, \ldots, y_{i,t-1}$, for any $2 \leq t \leq k$.

2. The maintenance of a relative position within a distribution of values in the observed population through time (Berenson et al., 1978; Clarke et al., 1978). In the context of the height of children, this concept raises the question of whether children who are at higher percentiles at time $t-1$ will also be at higher percentiles at time t, for any $2 \leq t \leq k$.

In this work, we are going to use the second definition of tracking. Figure 5 graphically depicts its underlying meaning. The distribution of height in the children participating in the INMA-Asturias cohort at 4 and 8 years of age is presented. We would consider tracking if, for example, those subjects situated at the 95th percentile at 4 years of age remain within the same percentile at 8 years, despite their absolute values change.

Several studies listed in Table 3, such as the Bogalusa Heart Study, the Muscatine Study, or the Cardiovascular Risk in Young Finns Study, have published various tracking analyses primarily focusing on lipids, blood pressure, and obesity. Subsequent studies

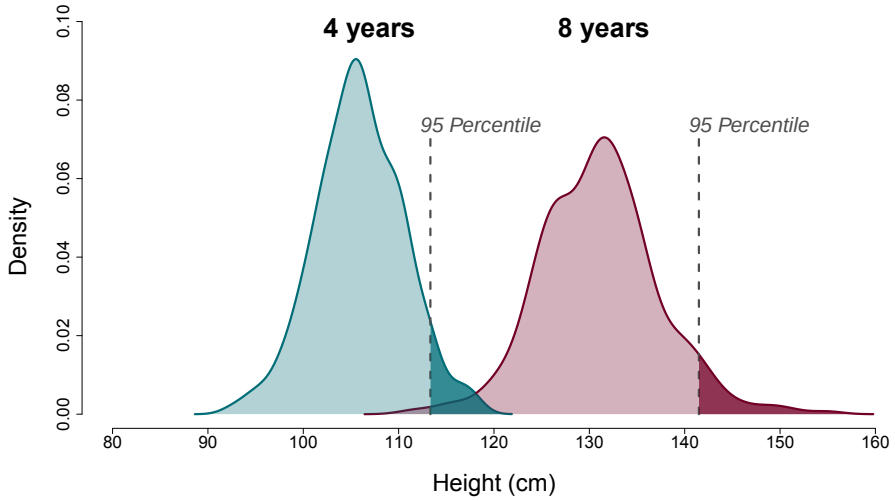


Figure 5: Concept of *tracking*. Self-crafted figure.

continued to investigate the tracking of these risk factors from childhood/adolescence to adulthood, although with a less pronounced focus on early childhood tracking. While numerous studies generally observe tracking, particularly in lipids and obesity, there is considerable variability in their findings. Variations are evident in participant age ranges, time intervals between assessments, and methodological approaches, making it challenging to draw comparisons or to establish conclusions.

1.6. From theory to practice

Statistics plays a crucial role in the field of epidemiology. It provides the tools needed to address the scientific question at hand. Nonetheless, statistical analysis frequently has to cope with limitations and complexities derived from the inherent characteristics of the design or the problem under study. This constitutes additional challenges when translating theoretical research hypotheses into practice.

This is the case for the study of tracking of cardiovascular risk factors, which suffers two main challenges, 1) the difficulty in analyzing tracking when there are not clear threshold definitions for categorizing variables into risk groups, and 2) how to deal with the high number of missing data commonly presented in longitudinal studies.

1.6.1. The controversy over thresholds definitions

While the concept of tracking began to solidify in the 1970s, efforts to establish the methodology and statistical tools required for its measurement and estimation were primarily driven by three articles published in Volume 27, Issue 3 in the specialized journal of Biometrics (https://www.jstor.org/stable/i323058) (Ware and Wu, 1981; Foulkes and Davis, 1981; Mcmahan, 1981), and continue during the next decades. But as the epidemiology notion is complex to define, as we have seen in the Section 1.5, it is also challenging in statistical terms.

One of the biggest problems in tracking analysis is the question of how to evaluate its magnitude. This is evident in the number of statistical techniques employed to estimate it. To review these methods, it is important to recall that metabolic risk factors are defined using biological variables as indicators, with its values serving as criteria to determine whether an individual exhibits the given risk factor or not. For instance, it is considered that an adult has hypertension when their systolic blood pressure equals or exceeds 140 mmHg, or their diastolic blood pressure equals or exceeds 90 mmHg. These thresholds (140 mmHg and 90 mmHg) have been derived by estimating the values of the variable (blood pressure in this example) that effectively discriminate or categorize individuals based on their risk of developing a future disease. In summary, a metabolic risk factor variable is computed by categorizing a continuous cardiovascular-related variable, using a clinically meaningful threshold.

Taking this into account, we can classify statistical techniques for estimating tracking into two types:

Statistical techniques for estimating tracking

1. Those based on the risk factor variable itself, which is a **categorical variable** with two or more categories indicating the presence and extent of the risk factor.
   Let Y represent the variable of interest, and let $y_{i,t}$ be a random sample with $1 \leq i \leq n$, and $1 \leq t \leq k$. Let $\lambda \in \mathbb{R}$ be the threshold used to categorize the risk factor variable, our objective is to estimate the conditional probability that an individual presents a value of Y above λ at time t given that he/she did so at the previous time, $t-1$. If this probability is large, the variable Y is said to *tracked*. This is usually implemented by calculating the proportion of subjects remaining in the risk group between different time periods.

2. Those based on the **continuous variable** that serves as an underlying indicator of this risk factor.
   The most common are correlation coefficients or classical linear regression models. If the correlation between $y_{i,t-1}$ and $y_{i,t}$ is found to be positive, the variable Y is said to *tracked*.

To employ the methods mentioned in the item 1, appropriate thresholds are necessary to categorize the variables. In adult populations, these thresholds are well-established, but in pediatric populations, particularly in healthy cohorts, it copes with unclear or controversial threshold definitions. As mentioned previously, there are not yet many studies that have followed a pediatric population long enough to calculate values that would differentiate the risk of future cardiovascular events. Consequently, thresholds for children are determined based on specific percentiles of the variable of interest, often assuming a normal distribution.

While this approach is widely used in epidemiology research, tracking should not be calculated based on arbitrary or generic percentiles; it should be assessed using clinically meaningful thresholds. Otherwise, it may result in reduced statistical power, less precise estimates, difficulties in comparing results across studies when thresholds are sample-dependent, and challenges in interpreting the obtained results (Bennette and Vickers, 2012). Especially, in the early ages considered in this work, where clear thresholds for defining hypertension, central obesity, insulin resistance, and dyslipidemia are lacking, it is advisable to employ a methodology that allows the use of continuous variables.

However, the continuous approach (item 2) has relevant limitations. It is primarily focused on evaluating the impact within the central part of the probability distribution of the variable. Nevertheless, when considering variables indicative of risk factors, a shift in the mean of the distribution of the variable often lacks clinical or health-related significance. Instead, the consequences observed at the extreme ends of the distribution are the most relevant. Using hypertension as an example again, our primary concern is with high blood pressure values. However, we have little interest in the average blood pressure values. Consequently, the insights generated by these methods may not yield substantial valuable knowledge.

To address this challenge, we require a methodology that allows us to: 1) analyze the tracking of risk factors without relying on thresholds, and 2) maintain our focus on the extreme parts of the distribution. Our proposal is the utilization of **Quantile Regression Models** (QRMs).

QRMs were introduced by Koenker and Bassett (1978). They offer a natural extension of the classical linear regression models in which, instead of specifying the change in the conditional mean of the distribution of the dependent variable associated with a change in the independent variables, the change in any conditional quantile of the distribution is specified.

This is very useful since the effect of a change in the independent variable on the distribution of the dependent variable could be the same in all parts of the distribution, as shown in Figure 6 A), but it could also be different, as shown in Figure 6 B), where the effect is greater at the upper part of the distribution (higher quantiles) in comparison to the lower or central part. Thus, by modeling only the conditional mean, as is done in the classical linear regression, important aspects of the association between the dependent and independent variables can be missed.

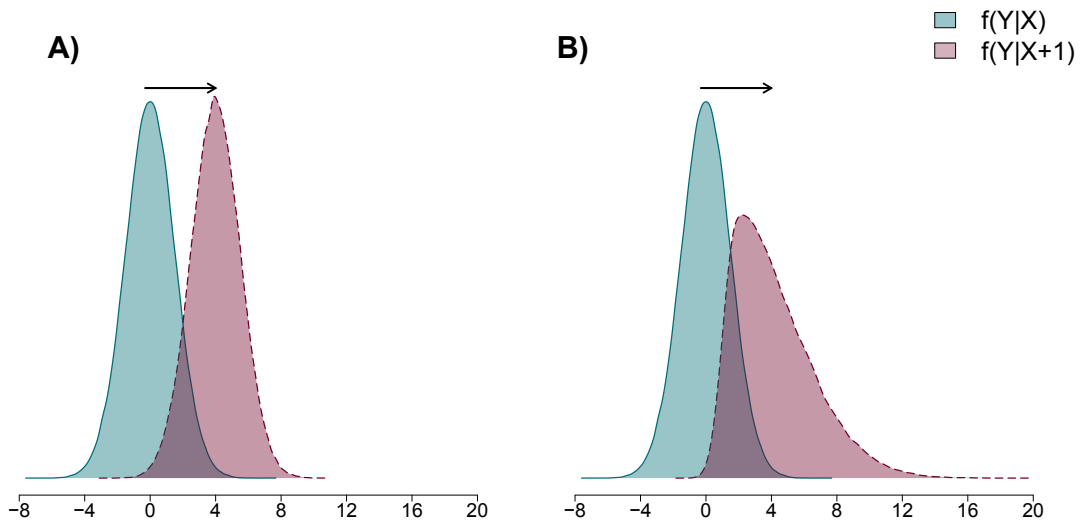


Figure 6: Example of the possible effect of a change in the independent variable X on the distribution of the dependent variable Y. Self-crafted figure.

Mathematically, let be Y the univariate dependent random variable, and $\boldsymbol{X}$ the k-dimensional random vector modeling the predictive components. Let be $q_\tau(Y|\boldsymbol{X})$ the τ-th quantile of the distribution of Y conditioning by $\boldsymbol{X}$, with $\tau \in (0,1)$. Then, the QRM assumes:

$$Y = \beta_\tau \cdot \ \boldsymbol{X} + \epsilon_\tau,$$

where the residuals verify that $\mathcal{P}(\epsilon_\tau \leq 0|\boldsymbol{X}) = \tau$. That is, its conditional τ-th quantile, $q_\tau(\epsilon_\tau|\boldsymbol{X})$, is zero. Therefore,

$$q_\tau(Y|\boldsymbol{X}) = \beta_\tau \cdot \ X$$

.

Let be $\{Y_n, \boldsymbol{X}_n\} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ a random sample from $\{Y, \boldsymbol{X}\}$, and the quantile regression model:

$$q_\tau(y_i|\boldsymbol{x}_i) = \beta_\tau \cdot \ x_i \quad \text{with} \quad 1 \leq i \leq n,$$

and the residuals:

$$\epsilon_{\tau_i} = y_i - q_\tau(y_i|\boldsymbol{x}_i)$$

.

Then, the estimator $\hat{\beta}_\tau$ is obtained by minimizing expression (1), a sum of the weighted absolute residuals that gives asymmetric penalties depending on whether the values of the dependent variable are being overestimated or underestimated:

$$\tau \cdot \sum_{\epsilon_{\tau_i} \geq 0} |\epsilon_{\tau_i}| + (1-\tau) \cdot \sum_{\epsilon_{\tau_i} < 0} |\epsilon_{\tau_i}|. \tag{1}$$

This means that the proportion of data points below the τ-th estimating regression line $\widehat{q_\tau(y_i|\boldsymbol{x}_i)} = \hat{\beta}_\tau \cdot x_i$ $(1 \leq i \leq n)$ is τ and the proportion lying above is $1-\tau$. Expression (1) can be minimized using different algorithms based on linear programming (Koenker, 2005).

The interpretation of the estimated coefficients is analogous to those in classical linear regression, except that instead of referring to the expected effect on the conditional mean of the dependent variable, we refer to the conditional quantile. That is, each $\hat{\beta}_\tau$ can be interpreted as the increment of the τ-th quantile of the dependent variable per unit of change in the value of the corresponding independent variable, while the rest of the independent variables are fixed.

Figure 7 shows an example of quantile regression for a discrete independent variable X which may take values in 1,2,3,4. It is observed the distribution of the dependent variable Y for each value of the independent variable X. This figure clearly illustrates that as the variable X increases by 1-unit, not only the mean of the distribution of variable Y changes, but also its spread and shape. In this example, quantile regression is used to estimate the quantiles $\tau \in \{0.10, 0.5, 0.90\}$, resulting in three distinct regression lines, one for each estimated quantile. It is worth noting that neither the slope nor the intercept is the same in any of the three lines due to the presence of heteroscedasticity.

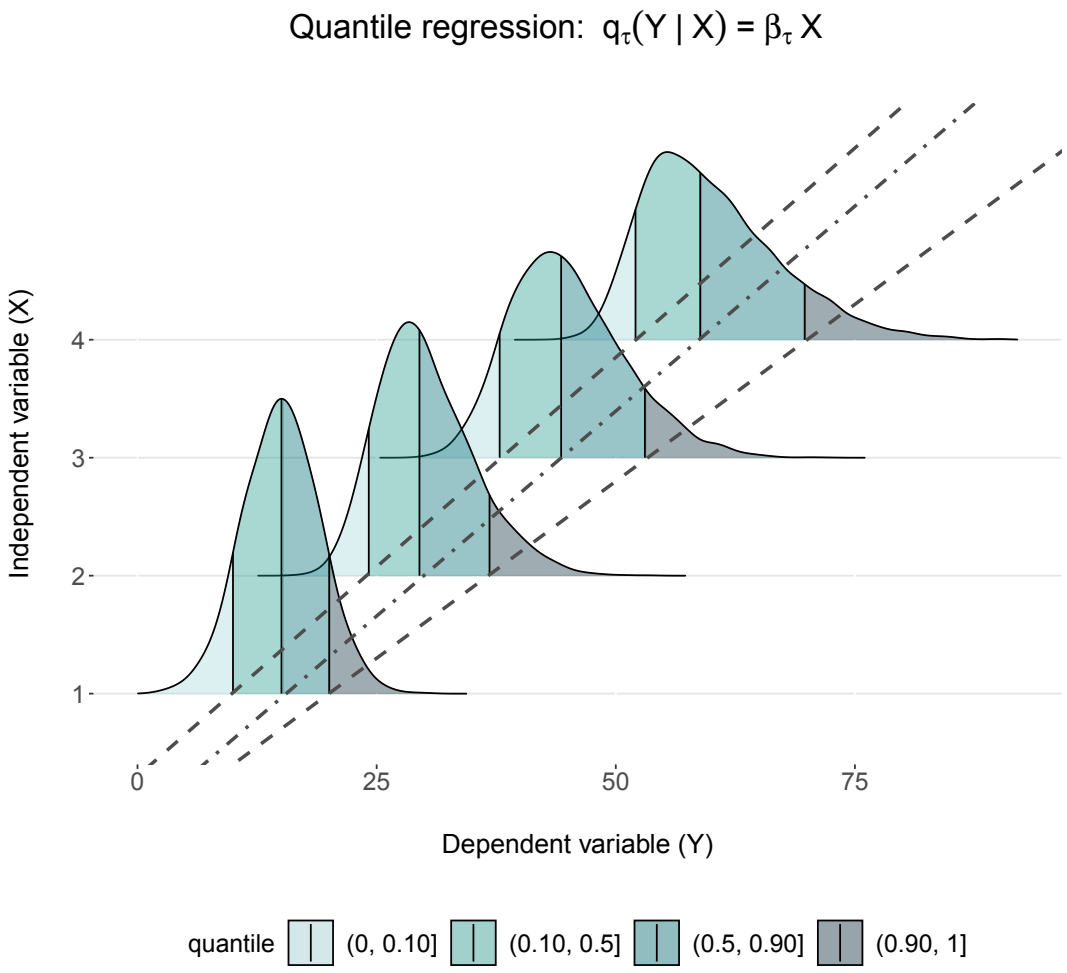


Figure 7: Example of quantile regression. Self-crafted figure.

QRMs overcome some limitations of classical linear regression tools, so they could be more appropriate in some situations (Waldmann, 2018), as the following:

Situations in which QRMs are useful

1. **In presence of outliers**. It handles outliers more effectively as it relies on estimating position measures like quantiles.
2. **In case of heteroscedasticity**. If the variance depends on the independent variables, quantile regression can capture this effect.
3. **When distributional assumptions are not satisfied**. QRMs do not necessary make assumptions about the distribution of errors, so it can be used when the conditions for applying other regression models are not satisfied.
4. **When the focus is on the extremes of the distribution**. Sometimes the real interest of the research question lies in what is happening in the tails of the distribution. QRM allows to answer this question by estimating the extreme quantiles.
5. **When there is no known threshold defining the population at risk**. Since the model can be estimated for any quantile, it allows assessing the impact of independent variables on a specific distribution section without the need to choose a specific point.

Items 4 and 5 are what justify their usefulness in estimating the tracking of metabolic risk factors. So to address the issue of tracking without relying on thresholds, we suggest employing cardiovascular-related variables that serve as underlying indicators of the risk factors themselves, and examining the extremes of their distributions using quantile regression models.

While quantile regression is often considered to be a more flexible alternative to classical linear regression, it may not always be the superior choice. Before concluding this section, the following are some of the disadvantages that QRMs exhibit in comparison to classical linear regression models (Koenker, 2005; Kocherginsky et al., 2005):

Disadvantages of QRMs in comparison to classical linear regression

1. **Less efficiency for estimating the mean**. When our goal is to estimate the effect on the center of the distribution of the dependent variable, we can estimate the mean using a linear regression model or the median (the 0.5 quantile) using a QRM. While QRM makes fewer assumptions, if all assumptions are satisfied, quantile regression may be less efficient than classical linear regression – meaning that, for a given sample size, estimates produced by quantile regression are generally less precise than those produced by classical linear regression.

2. **Larger sample sizes may be required**. Quantile regression might require larger sample sizes to provide accurate estimates, especially when estimating quantiles at the extreme ends of the distribution.

3. **Complex parameter estimation**. Estimating parameters in quantile regression is more complex and it requires more advanced numerical methods. While contemporary computational advancements mitigate this issue, it can still be relatively time-consuming.

4. **Resampling techniques for standard errors**. There are several procedures for computing standard errors in QRMs. Under certain conditions, the usual coefficient estimators are asymptotically normally distributed, but asymptotic standard errors are complex, and resampling approaches are frequently employed. This, once again, can be time-consuming for datasets of moderate to large size.

5. **Interpretational complexity**. Quantile regression tends to be less straightforward in terms of interpreting results. For instance, with a single independent variable, interpreting a single coefficient, as obtained from linear regression, may seem more intuitive than interpreting multiple quantile regression coefficients. Moreover, quantile regression coefficients can lead to different conclusions for different quantiles, adding complexity to interpreting the results. This can make the results less intuitive and more challenging to communicate.

1.6.2. Missing data. A common pitfall in longitudinal cohort studies

The second challenge is how to deal with missing data, a recurrent problem in statistics. When employing multivariate statistical techniques, if a subject is missing one of the variables required in the analysis, it should be excluded. This can lead to a loss of statistical power due to reduced sample size and, perhaps more problematic, the potential introduction of biases.

This problem is particularly pronounced in longitudinal cohort studies, where numerous variables are collected and repeatedly measured at multiple time points. Consequently, the likelihood that not all subjects answer all the questionnaires items or attend for all the samples collections or measurements visits increases. For example, if there are 10 variables involved in the analysis, and each variable independently has a 10% chance of being missing at the first follow-up, then the expected proportion of complete units is $0.9^{10} = 0.35$. And if a second measurement of the same variables is added to the analysis in a second follow-up, and each variable independently of the others and of its first measurement has again a 10% chance of being missing, then the expected proportion of complete units is $0.9^{20} = 0.12$. Although subjects with missing values in one variable are usually the same than those lacking data in other variables, and those with missing values in one follow-up are generally connected to missing values in other follow-ups, discrepancies can still arise. In our study, multiple variables (at least one per metabolic risk factor) are needed at two time points (the 4- and 8-year follow-ups). Consequently, excluding every child with a missing value in at least one variable or follow-up substantially reduces the sample size and can potentially introduce bias.

Until the 1970s, the standard approach to deal with missing data was to delete them, in the so-called **complete-case approach**. Rubin (1976) started to develop a framework of inference from incomplete data that remains in use today, including a missing data classification based on the underlying loss mechanism.

Let $\{Y, \boldsymbol{X}\}$ be a $(k+1)$-dimensional random vector. For the sake of simplicity, we will assume univariate missing data, that is, Y is the only variable containing missing values. Let R be the response indicator matrix, that is $R = 1$ if Y is observed and $R = 0$ otherwise. Then:

Types of missing data

- **Missing completely at random (MCAR) model satisfies**:

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R\},$$

that is, the probability of being missing does not depend either on Y or $\boldsymbol{X}$. This means that there are no systematic differences between the missing and the observed values. For example, serum lipids measurements may be missing because some samples have been lost in transit to the laboratory.

- **Missing at random (MAR) model satisfies**:

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R|\boldsymbol{X}\},$$

that is, the probability of being missing depends on the observed data. For example, serum lipids measures may be more likely to be missing in young people, as they tend to be less concerned and do not attend visits for blood collection.

- **Missing not at random (MNAR) model satisfies**:

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R|Y\},$$

that is, the probability of being missing depends on the missing values itself or on unobserved information. For example, in a study to assess the effect of an hypertensive treatment, hypertensive subjects may present greater collaboration that results in a lower number of missingness.

This distinction is important for understanding why some methods will work or not. Simple methods to deal with missing data are the following, based on single imputations (Schafer and Graham, 2002; van Buuren, 2012):

- **Complete-case analysis**: involves the exclusion of subjects with any missing data from the analysis. It results in a reduction in sample size and consequently an increase in standard errors. It can can lead to biased estimates if the missing data is not MCAR, because the excluded cases may differ systematically from those included in the analysis, which can introduce bias into the results.

- **Available-case analysis**: utilizes all available information for all subjects, using different sets of sample units for different parameter estimates. Because parameters are estimated from different sets, it is difficult to compute standard errors or other measures of uncertainty, and it requires sophisticated optimization techniques and specialized formulas for variance calculation. While this approach can address the issue of sample size reduction observed in complete-case analysis, it may still yield biased estimates when the missing data does not follow a MCAR pattern.

- **Mean imputation**: replaces missing data with the mean of the observed values. Let $Y = \{Y_{obs}, Y_{miss}\}$ a random variable containing missing values with Y_{obs} the observed values and Y_{miss} the missing values, then $\hat{Y}_{miss} = \overline{Y_{obs}}$. This approach is also known as *unconditional imputation* because it does not depend on other variables. While this method preserves the average of the distribution of the variable, its variability is artificially reduced, as it replaces all the missing data by the same value. Again this approach would result in unbiased estimates only under MCAR pattern.

- **Regression imputation**: incorporates knowledge of other related variables $X = (X_1, \ldots, X_k)$ with the idea of producing better imputations using regression models. Missing values Y_{miss} are replaced by its estimates $\hat{Y}_{miss}$ given $X = (X_1, ..., X_p)$, so this approach is also known as *conditional imputation*. For example, if we consider a linear regression model $Y = \beta \cdot X + \epsilon$, then $\hat{Y}_{miss} = E(Y_{obs}|X_{miss}) = \hat{\beta} \cdot X_{miss}$.

- **Stochastic regression imputation**: is a refinement of regression imputation that adds noise to the predictions. In this case each missing value is replaced not by a regression estimate, but by a random draw from the conditional distribution of Y given X. If we consider again the linear regression model $Y = \beta \cdot X + \epsilon$, then $\hat{Y}_{miss} = E(Y_{obs}|X_{miss}) + \epsilon = \hat{\beta} \cdot X_{miss} + \epsilon$, where ϵ is randomly generated from a zero-mean distribution.This method and the regression imputation can provide unbiased estimates under MAR mechanism, but they still tend to underestimate the variance because they ignore the fact that the imputed values are predictions.

Table 4 summarizes these simple methods for dealing with missing data, as well as under which loss mechanism these approaches can yield to unbiased parameter estimates, and their limitations to estimate parameter standard errors.

Table 4: Overview of the most used single imputation methods for dealing with missing data.

Method	**Loss mechanism**	**Standard Error**
Complete-case (listwise deletion)	MCAR	Too large
Available-case analysis (pairwise deletion)	MCAR	Complicated
Mean imputation	MCAR	Too small
Regression imputation	MAR	Too small
Stochastic regression imputation	MAR	Too small

One of the goals of statistics is to provide valid quantifications on both the estimations and the uncertainty associated with those estimations. That allows to produce efficient extrapolations from the studied sample to the target populations. As we have just seen, the single imputation methods are not able to reflect the missing data uncertainty, so more complex methods have been developed to overcome this limitation. The **Multiple Imputation** (MI) method, proposed by Rubin (1987), does not focus on imputing the "closest" possible values to the actual missing values, but rather to make valid and efficient inferences about the parameters of interest and their associated uncertainty. This procedure generates multiple reasonable estimates for the missing values, ensuring that the variation among these estimates accurately represents the true uncertainty around their actual values. Figure 8 illustrates the process: MI uses the distribution of the observed data to estimate a set of plausible values for each missing value, through an imputation model. Random components are incorporated into these estimated values to reflect their uncertainty, resulting in the different values for each estimate. Multiple datasets are created – as many as different estimates for each missing value – and then analyzed individually, generating different parameter estimates and standard errors. Finally, the individual estimatione are combined using specific rules created by Rubin (the so-called Rubin rules) to obtain the overall estimates, their associated standard errors, and appropriate confidence intervals, reflecting the actual uncertainty around the estimation due to the missing values (van Buuren, 2012).

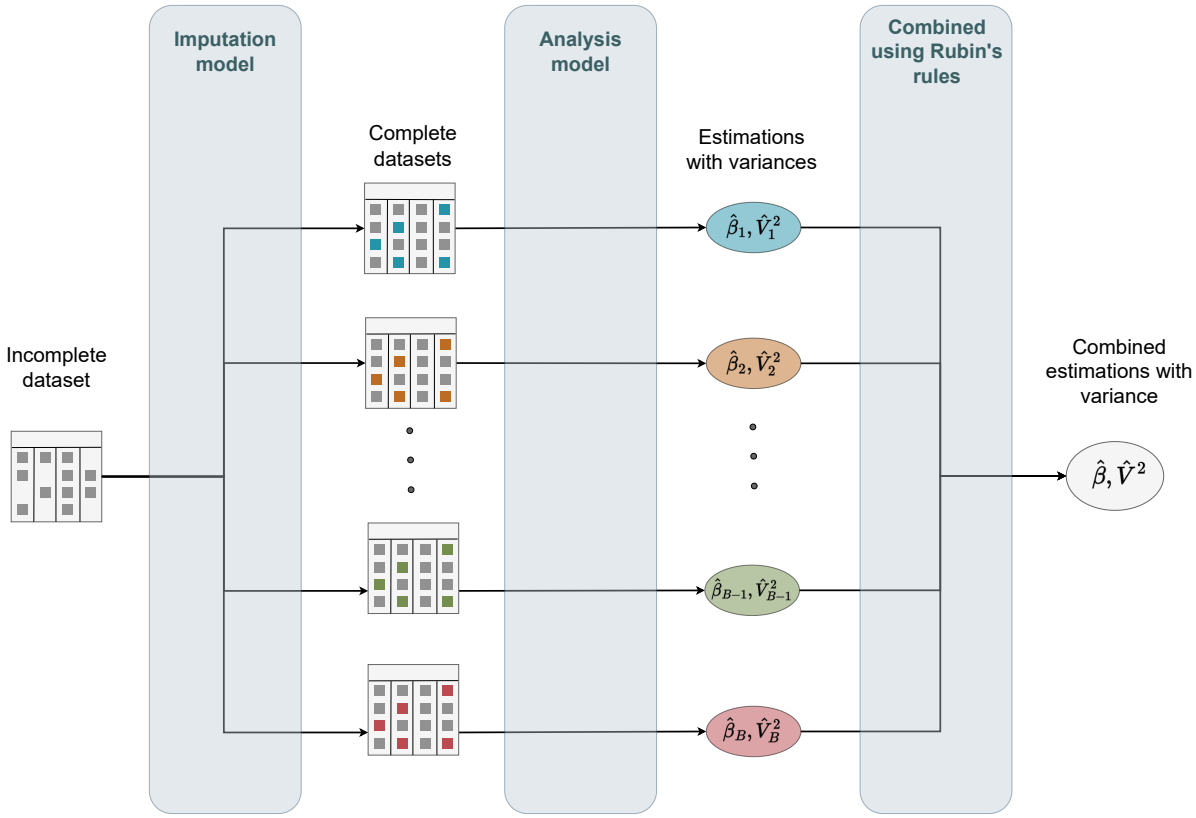


Figure 8: Flowchart of the steps of multiple imputation. Self-crafted figure.

More formally, MI procedures consider the MAR model and the relationship

$$Y = g(\boldsymbol{X}) + \epsilon, \tag{2}$$

where $g(\cdot)$ and ϵ are the link function, and a randomly genereted value from a zero-mean distribution, respectively. And the steps are the following:

Steps of multiple imputation method

Let $\{Y_n, \boldsymbol{X}_n\}$ be a random sample drawn from the random vector $\{Y, \boldsymbol{X}\}$, and let β be the target parameter. We assume that the values $y_{i_1}, \ldots, y_{i_m}$ $(1 \leq i_1 \leq \cdots \leq i_m \leq n)$ are missing.

- **Step 1**. From the non-missing values, we compute the function $\hat{g}(\cdot)$ which estimates $g(\cdot)$ (expression (2)). For each missing value, y_{i_j} $(1 \leq i_1 \leq \cdots \leq i_m \leq n)$ we generate a pseudo-value $\hat{y}_{i_j} = \hat{g}(\boldsymbol{x}_{i_j,n}) + \epsilon_{i_j}$, where ϵ_{i_j} is randomly generated. With this dataset, we estimate the target parameter, $\hat{\beta}$, and its variance, $\hat{V}^2$.
- **Step 2**. We repeat the Step 1 B times (B a large enough number) and get a vector of estimations $\{\hat{\beta}_1, \ldots, \hat{\beta}_B\}$, and another with their respective variabilities $\{\hat{V}_1^2, \ldots, \hat{V}_B^2\}$. Notice that, in each repetition, the error (ϵ) is randomly generated. Therefore, each repetition provides a different dataset.
- **Step 3**. We use the Rubin's rules to combine the vectors obtained in Step 2 in a single estimation with its variability. This estimation reflects both the uncertainty due to the sample variation, and the uncertainty due to the missing data. The B parameter estimates and their respective $\widehat{SE}_k$ standard errors are combined using Rubin's rules, to produce an overall estimate and standard error that reflect both the uncertainty due to the sample variation, and the uncertainty due to the missing data.

By adopting this approach, we obtain the parameter estimates of interest and their corresponding confidence intervals that take advantage of all available information while accounting for the variability and uncertainty introduced by the lack of knowledge regarding the true values of missing data.

Although we focus here on MI, the field of handling missing data is extensive, with other widely studied techniques such as weighting procedures and likelihood-based approaches (see the Expectation-Maximization algorithm or the Full Information Maximum Likelihood (FIML) approach as an example (van Buuren, 2012)).

1.7. Justification

The evidence that atherosclerosis has its origins in childhood underscores the importance of identifying risk factors associated with the development of its clinical manifestations in later stages of life. By addressing these modifiable conditions, it becomes possible to prevent the future burden of the disease.

Central obesity, insulin resistance, hypertension, and dyslipidemia are well-established cardiovascular risk factors among young-adults and adults. The hypothesis that these factors also play the same role in early childhood carries relevant weight. However, the limited evidence supporting this hypothesis is mainly due to the challenges associated with conducting long-term cohort studies in pediatric populations, which require extended observation periods to capture cardiovascular events.

One plausible mechanism that could explain this hypothesis is the tracking of these risk factors from childhood to young-adulthood or adulthood. If extreme values of these factors persist from childhood to the ages where we have conclusive evidence of their role as risk factors, it would substantiate their relevance as early-life determinants of risk. This is also relevant due to the critical period that pediatric ages represent, given the maturation and developmental processes in children during this stage. Furthermore, considering the cumulative nature of atherosclerosis, there is potential for an increased risk of CVD when cardiovascular risk factors persist over time.

Therefore, monitoring and studying the tracking of extreme values of variables that represent adult markers of hypertension, central obesity, insulin resistance, and dyslipidemia holds great potential for early prevention of CVD. This is especially relevant in the studied population, the INMA-Asturias cohort, where the elevated rates of overweight and obesity could position it as particularly susceptible to future development of CVDs.

CHAPTER 2

Objectives

Research question

Building upon the identified issues and gaps highlighted in the previous sections, the following research question arises: *Is the tracking of cardiovascular risk factors, including central obesity, insulin resistance, hypertension, and dyslipidemia, observed during early childhood?*

Objectives

To address this question, the **main objective** of this dissertation is to evaluate the presence of tracking, specifically between the ages of 4 and 8 years, in the INMA-Asturias cohort, with respect to the cardiovascular risk factors of central obesity, insulin resistance, hypertension, and dyslipidemia.

The following **specific objectives** have been established to accomplish this:

1. To evaluate the simultaneous presence of the cardiovascular risk factors under study among the children of the INMA-Asturias cohort, both at 4 and 8 years of age, and to analyze the tracking of these aggregations between these ages.
2. To examine tracking of each of the cardiovascular risk factors under study, between 4 and 8 years of age in the INMA-Asturias cohort.

3. To mitigate the negative effects of missing data on research analysis, and also to explore and compare different methodologies for quantifying tracking, taking into consideration the absence of standard thresholds for the age range of the study population.

Objectives 1, *2*, and *3* are addressed respectively in the documents: Article I, Article II, and Article III, included in the Chapter 3 of this memory.

CHAPTER 3

Methods and results

3.1. Article I: Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort

To cite this article:
R. Fernández-Iglesias, A. Fernández-Somoano, C. Rodríguez-Dehli, R. Venta-Obaya, I. Riaño-Galán, and A. Tardón. Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort. *PLoS ONE*, 18(4), 2023. DOI: 10.1371/journal.pone.0283799.

To link to this article:
https://doi.org/10.1371/journal.pone.0283799

PLOS ONE

RESEARCH ARTICLE

Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort

Rocío Fernández-Iglesias[1,2,3], **Ana Fernández-Somoano**[1,2,3]*, **Cristina Rodríguez-Dehli**[3,4], **Rafael Venta-Obaya**[5,6], **Isolina Riaño-Galán**[1,3,7☯], **Adonina Tardón**[1,2,3☯]

1 Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Madrid, Spain, **2** Unit of Molecular Cancer Epidemiology, University Institute of Oncology of the Principality of Asturias (IUOPA), Department of Medicine, University of Oviedo, Oviedo, Asturias, Spain, **3** Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Oviedo, Asturias, Spain, **4** Servicio de Pediatría, Hospital San Agustín, Avilés, Asturias, Spain, **5** Servicio de Bioquímica, Hospital San Agustín, Avilés, Asturias, Spain, **6** Departamento de Bioquímica y Biología Molecular, University of Oviedo, Oviedo, Asturias, Spain, **7** Servicio de Pediatría, Endocrinología pediátrica, HUCA, Oviedo, Asturias, Spain

☯ These authors contributed equally to this work.
* fernandezsana@uniovi.es

Citation: Fernández-Iglesias R, Fernández-Somoano A, Rodríguez-Dehli C, Venta-Obaya R, Riaño-Galán I, Tardón A (2023) Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort. PLoS ONE 18(4): e0283799. https://doi.org/10.1371/journal.pone.0283799

Editor: Antoine Fakhry AbdelMassih, Cairo University Kasr Alainy Faculty of Medicine, EGYPT

Received: October 18, 2022

Accepted: March 18, 2023

Published: April 12, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0283799

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Abstract

Aim

This study aimed to investigate whether there are subgroups of children with different clusters of cardiovascular disease (CVD) risk factors at 4 and 8 years of age, and their patterns of change between these two time points.

Methods

The analysis was conducted in 332 children who participated in the INMA-Asturias cohort (Spain) at 4 and at 8 years of age. The CVD risk factors were central obesity, dyslipidaemia, hyperglycaemia, and hypertension. Latent transition analysis was used to identify the different clusters and their probabilities of change.

Results

At 4 years, three subgroups were identified: no disorders (prevalence of 55.9%); some disorders (21.2%), and central obesity (22.9%). Three distinct subgroups were identified at 8 years: no disorders (59.8%); hypertension (17.9%), and central obesity (22.3%). Central obesity at 4 years tends to appear simultaneously with dyslipidaemia, while at 8 years it tends to appear simultaneously with dyslipidaemia and/or hypertension. Children aged 4 years with no disorders had a 93.7% probability of remaining in the same status at 8 years of age. Children aged 4 who had some disorders had a 67.7% of probability of having only hypertension and a 32.3% of probability of having central obesity. Children aged 4 in the central obesity subgroup had a 32.4% of probability of having no disorders at 8 years of age, while 67.6% still had central obesity.

Funding: This study was supported by grants from CIBERESP (PhD-employment-contract), ISCIII: PI04/2018, PI09/02311, PI13/02429, PI18/00909 co-funded by FEDER, "A way to make Europe"/ "Investing in your future", Fundación Cajastur and Universidad de Oviedo. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

These exploratory findings suggest that children who do not present any disorder at 4 years of age tend to remain in that state at 8 years of age. And also that central obesity may play a major role in the development of other disorders, as the number of disorders with which it concomitantly occurs increases between the ages of 4 and 8 years.

Introduction

Cardiovascular disease (CVD) is the leading cause of death and chronic disability throughout the world, almost duplicating the cancer mortality rate in 2017, and it increased by nearly 50% from 1990 to 2019 [1, 2]. Disorders, such as hypertension, dyslipidaemia, central obesity, and hyperglycaemia, are the main risk factors for CVD [2–4], and in the last decade, the prevalence of CVD has been increasing [5]. Moreover, these factors tend to appear simultaneously more frequently than would be expected randomly, and this increases the risk of type 2 diabetes, atherosclerosis, and other CVDs in adults. The clustering of at least three of these CVD risk factors has been defined as a complex disorder called metabolic syndrome (MetS) [6].

The accumulation of CVD risk factors has also been observed in children [7, 8]. However, MetS in childhood is controversial for several reasons. One reason is that there is no consensus on the definition of MetS. Although the same four components (hypertension, dyslipidaemia, central obesity, and hyperglycaemia) are usually considered, the cut-off points used to discriminate normal from abnormal values of this components vary between definitions. This causes difficulty in comparing the prevalence of MetS (range: 0.3%–26.4%) among studies worldwide [8–12]. Additionally, the literature on childhood MetS is scarce and shows inconsistent results on whether the joint presence of components implies a higher risk of disease in adulthood than the simple presence of individual components [13–15]. Therefore, more large-scale longitudinal studies are required to validate a causal relationship. Consequently, most studies conducted to date focussed on all the MetS components separately as CVD risk factors and all of their possible aggregation patterns [10, 16]. Especially at an early age, as the International Diabetes Federation suggests that MetS should not be diagnosed younger than 10 years [17].

Furthermore, the temporal stability of these components and their clustering in children are unclear. The phenomenon of longitudinal stability of a variable is known as tracking [18]. Several studies have examined the tracking of clusters of CVD risk factors, but most focussed on adolescence and the transition of patterns to young adulthood or adulthood [19–25]. To date, there have not been many studies that evaluated the short-term tracking of CVD risk factor clustering in children [26]. One of the main difficulties in assessing the stability of these factors over time is the large number of different observable patterns.

Latent transition analysis (LTA) is a statistical method that allows to identify groups of subjects who are distinguished from each other according to the response to observed categorical variables and estimates the probabilities of transition between groups over time [27–29]. This technique is relatively new in the field of epidemiology [27] but is useful for representing the variety and complexity of the relations between MetS components and their changes over time, taking into account within-person variability [26].

There are future implications of the early onset of CVD risk factors and their clustering [8, 9], and there is uncertainty regarding its pathogenic mechanism [10, 16]. Therefore, identifying the different CVD risk factor clusters in children and their patterns of change over time is important. This study aimed to evaluate the following using LTA: (i) the presence of subgroups

of children with different clustering patterns of hypertension, dyslipidaemia, central obesity, and hyperglycaemia at 4 and 8 years of age, (ii) the prevalence of each subgroup, and (iii) the probability of changing between groups between 4 and 8 years of age.

Materials and methods

Study population

The population considered in this study was the Infancia y Medio Ambiente (INMA [Environment and Childhood]) Asturias cohort (northern Spain), which has been described in previous studies [30–32]. In brief, pregnant women in their first trimester of pregnancy were recruited between May 2004 and June 2007. The inclusion criteria were as follows: age $\geq$ 16 years, singleton pregnancy, no assisted conception, delivery scheduled at the San Agustón Hospital (Avilés, Spain), and no communication handicap. After recruitment, data were collected in several phases of follow-up as follows: in the first and third trimesters of pregnancy, at birth, and when children were 18 months, 4 years, 8 years, and 11 years of age. This study analysed data from follow-ups at 4 (T0) and 8 years (T1) of age.

The initial sample was composed of 494 eligible women who agreed to participate. A total of 453 children were followed up at 4 years and 416 children were followed up at 8 years of age. Finally, all 332 children who had available data for each variable used to determine the CVD risk factors (waist circumference [WC], systolic blood pressure [SBP], diastolic blood pressure [DBP], blood glucose, triglycerides [TGs], and high-density lipoprotein cholesterol [HDL-C]) at 4 or at 8 years of age were included in the study (Fig 1).

The study protocol was approved by the Asturias Regional Ethics Committee, and written informed consent was obtained from every participating woman and, in such case, her partner. The research conformed to the principles of the Declaration of Helsinki.

Cardiovascular disease risk factor definitions

CVD risk factors considered in this study were central obesity, hyperglycaemia, dyslipidaemia, and hypertension. To determine whether a child had these disorders, we considered the following variables. WC was a marker for central obesity, blood glucose concentrations were a marker for hyperglycaemia, TG and HDL-C concentrations were markers for dyslipidaemia, and SBP and DBP were markers for hypertension [33, 34]. Predefined cut-off points were used to distinguish between normal or abnormal levels of these variables.

In this study, reference cut-off points provided by the IDEFICS study were applied. These reference values were derived from a large population-based sample of healthy children from a heterogeneous European population (16,228 children from Sweden, Germany, Hungary, Italy, Cyprus, Spain, Belgium and Estonia). This previous study provided age- and sex-specific cut-off points for each variable (also a height-specific cut-off for blood pressure) using two different levels. One cut-off indicated children who required close observation, called the monitoring level (values exceeding the 90th percentile in their sample). The other cut-off indicated children who required an intervention to ameliorate their risk profile, called the action level (values exceeding the 95th percentile in their sample).

In the present analysis, the cut-off points of the monitoring levels were considered as follows. Children were classified as having central obesity if they had a WC above the age- and sex-specific monitoring IDEFICS cut-off point [35]. Children were classified as having hyperglycaemia if they had blood glucose concentrations above the age- and sex-specific monitoring IDEFICS cut-off point [36]. Children were classified as having dyslipidaemia if they had TG concentrations above the age- and sex-specific monitoring IDEFICS cut-off point or HDL-C concentrations below the age- and sex-specific monitoring IDEFICS cut-off point [37].

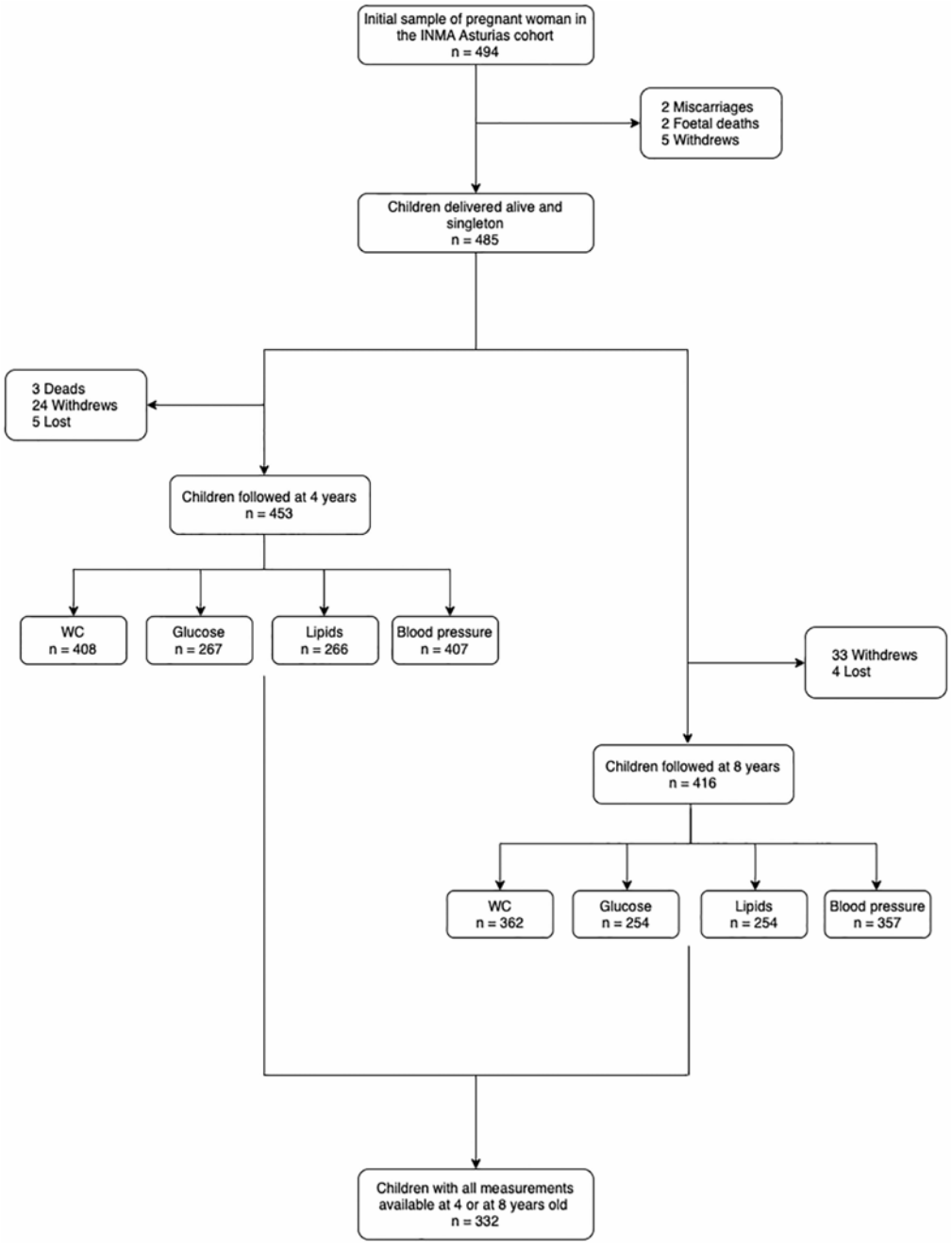


Fig 1. Flowchart of the study population. WC, waist circumference.

https://doi.org/10.1371/journal.pone.0283799.g001

Children were classified as having hypertension if they had SBP or DBP above the age-, sex-, and height-specific monitoring IDEFICS cut-off point [38].

Cardiovascular disease risk factor measurement

WC was measured using a non-stretch nylon tape measure, with the subject in a standing position, at the midpoint between the caudal edge of the last rib and the iliac crest at the level of the umbilicus and with the tape in a horizontal plane. SBP and DBP were obtained, after 5 min rest, with an automated oscillometric system (OMRON®) with the patient in a seated position and the right arm at rest at the level of the heart, measured at the level of the right wrist. Between 2 and 3 measurements were taken and the mean was calculated. All somatometric and blood pressure determinations were performed by experienced and trained healthcare staff. Nonfasting blood samples were collected by laboratory nurses at the Hospital San Agustín de Avilés and determined by laboratory staff using a Roche analyser (Modular Analytics Serum Work Area, Mannheim, Germany).

Statistical analysis

The characteristics of the study population are shown using descriptive statistics. Continuous variables are shown by median and interquartile range, while categorical variables are shown by absolute and relative frequencies.

As the result of the definitions indicated in the 'Cardiovascular disease risk factor definitions' section, four binary categorical variables were obtained that allowed us to distinguish children who had normal or abnormal (monitoring) levels in relation to the four disorders. LTA was used to identify groups of children with different aggregation patterns of CVD risk factors. Each identified group is interpreted as a 'latent status', that represents the different response patterns in the data to the observed variables [29], which are in this case the four CVD risk factor markers considered. LTA was also used to estimate the prevalence of each latent status and the probabilities of change from one latent status to anotherbetween these two time points (S1 Figure in S1 File). In this process, three types of parameters were estimated as follows. Class membership probabilities, which indicate the proportion of the population expected to be classified in a particular latent status. Item-response probabilities, which indicate the probability of being in a particular category of an observed variable (i.e., in the normal or monitoring level category of any of the disorders), conditional to the latent class membership. They also provide the basis for interpretation and labelling the latent status. And finally, transition probabilities, which indicate the probability of membership in one latent status at T1 given the membership at T0. The name given or assigned to the latent status is based on the researchers interpretation of the item-response probabilities. LTA model specification, model estimation, and finally, model selection and interpretation were conducted as reported by Collins and Lanza [29], with an adaptation of the procedure proposed by Ryoo et al. [39]. Further details of the steps followed during this procedure are given in S1 File.

LTA models allow for missing data on the measured outcomes using the full-information maximum likelihood method. In this method, the data for children who have information at only one time point do not contribute to the estimation of the transition parameters. However, the data do contribute to estimating the time-specific parameters, and thus help to produce better results, allowing an increase in the sample size and thus the statistical power [29, 40]. Therefore, only children without any of the measurement components at any time point were eliminated from the study. Consequently, children were included in the study if they met the following criteria: they had a measurement of each of the CVD risk factors considered in at least one of the time points (i.e., those who have measure of WC at T0 or T1, and measure of

blood glucose at T0 or T1, and measure of TG at T0 or T1, and measure of HDL-C at T0 or T1, and measure of SBP at T0 or T1, and measure of DBP measurement at T0 or T1). This resulted in a final study sample of 332 children, which allowed us to reach a sample size of > 300 (for lower sample sizes, the use of LTA models is not recommended [41]). To estimate transition probabilities, we used children who had measurements of all variables at the two time points (n = 154).

Statistical analyses were performed using the R software (version 4.0.5) [42] and the statistical modelling programme Mplus (version 7) [43], which is specially designed for latent variable models. LTA was carried out using Mplus and the R package MplusAutomation [44].

Results

Descriptive statistics of sex, age, cardiometabolic variables, and risk factors of the study population are shown in Table 1. The risk factors with the highest prevalence at T0 at the monitoring level were blood lipids and WC, with a prevalence of 33.7% and 27.1%, respectively. The risk factors with the highest prevalence at T1 were blood pressure and lipids, with a prevalence of 35.8% and 25.0%, respectively.

The distribution of the number of risk factors at the monitoring level that a child presented at the same time (Fig 2) at T0 was as follows. A total of 29.0% of the children had no disorder at this level, 41.6% had one, 21.2% had two, 8.0% had three, and 0.4% had all four. At T1, a higher percentage of children presented with none or one disorder at that level (36.9% and 42.1% respectively), while a lower percentage of children had two or three disorders at that

Table 1. Descriptive statistics of sex, age and cardiometabolic parameters at 4 and 8 years of age.

	N = 327	T0 (4 years)	N = 300	T1 (8 years)
Sex, %				
Male	176	(53.8)	162	(54.0)
Female	151	(46.2)	138	(46.0)
Age, median (IQR[a])	327	4.4 (4.3; 4.5)	300	8.3 (8.1; 8.4)
Waist circunference (cm), median (IQR)	325	53.5 (50.5; 56.0)	300	63.2 (58.5; 69.0)
Glucose (mg/dL), median (IQR)	266	86.0 (81.2; 91.0)	253	86.0 (82.0; 90.0)
Sistolic blood pressure (mmHg), median (IQR)	324	99.0 (90.0; 105.0)	296	107.0 (100.0; 114.0)
Diastolic blood pressure (mmHg), median (IQR)	324	60.0 (54.0; 66.0)	296	67.0 (60.8; 72.0)
Triglycerides (mg/dL), median (IQR)	265	71.0 (55.0; 94.0)	253	64.0 (47.0; 87.0)
HDL-C (mg/dL), median (IQR)	265	57.0 (48.0; 63.0)	251	68.0 (60.0; 80.0)
Waist circunference level, %				
Normal	237	(72.9)	230	(76.7)
Monitorization	88	(27.1)	70	(23.3)
Glucose level, %				
Normal	210	(79.2)	237	(93.7)
Monitorization	55	(20.8)	16	(6.3)
Blood pressure level, %				
Normal	238	(73.5)	190	(64.2)
Monitorization	86	(26.5)	106	(35.8)
Lipids level, %				
Normal	175	(66.3)	189	(75.0)
Monitorization	89	(33.7)	63	(25.0)

[a]IQR, interquartile range.

https://doi.org/10.1371/journal.pone.0283799.t001

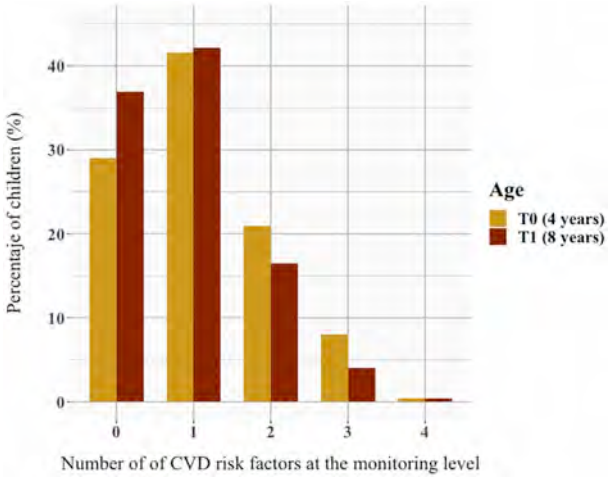


Fig 2. Distribution of the number CVD of risk factors at the monitoring level in the study population.

https://doi.org/10.1371/journal.pone.0283799.g002

level (16.4% and 4.0% respectively). The percentage of children with the four risk factors at that level was the same at both time points. Consequently, 71.2% of children had at least one disorder at the monitoring level at 4 years and 62.9% of children did so at 8 years.

S1 Table in S1 File also contains the distribution of the number of risk factors at the monitoring level at the same time, disaggregated by each risk factor. It shows that at 4 years WC and blood lipids are the variable that appeared at the monitoring level most frequently in combination with others. And that at 8 years WC continued to be the variable that appeared at the monitoring level most frequently in combination with others, followed by blood pressure.

The results of the LTA are described below. On the basis of the criteria indicated in the S1 File, the three latent status model was selected. Details about the information on the relative model fit for selecting the number of latent statuses is shown in Supplementary Table S2 in S1 File.

Table 2 shows the item-response probabilities of being at a normal level for each disorder, in the three latent statuses detected at each time point. It also shows the estimated prevalence of each latent status. On the basis of these item-response probabilities, the latent status detected by LTA was labelled as follows at T0. Children in latent status one were characterised as having a high probability of being in the normal range for all of the disorders (70.9%–100%), so it was labelled as '*no disorders*'. Children in latent status two had the probability of being at normal levels for all of the disorders (48.9 and 79.5%). But all confidence intervals contained the probability of 50%, and therefore, individuals in this group could present with one or more disorders, but were not characterised by any specific one. Consequently, this status was labelled as *'some disorders'*. In latent status three, which was labelled as *'central obesity'*, children had a zero probability of having a normal WC, a high probability of having a normal blood pressure (81.3%), and a high probability of having normal glucose concentrations (76.9%), while they were half as likely to have normal lipid concentrations. The second latent status (some disorders) had a poor homogeneity because all of the confidence intervals contained a 50% probability. That means that no clearly characteristic pattern can be identified in this latent status and that is difficult to interpret it in a meaningful way. The latent status with the highest prevalence at T0 was no disorders (55.9%), followed by central obesity (22.9%) and some disorders (21.2%).

Table 2. Item-response probabilities and confidence intervals of the three latent status model selected, and the prevalence of the latent statuses at each time point.

	T0 (4 years)		
	Latent status 1: No disorders	Latent status 2: Some disorders	Latent status 3: Central obesity
	N = 186 (55.9%)	N = 70 (21.2%)	N = 76 (22.9%)
Waist circunference normal level (%)	**100.0 (100.0; 100.0)**	79.5 (47.9; 100.0)	**0.0 (0.0; 0.0)**
Glucose normal level (%)	**87.2 (79.8; 94.6)**	60.5 (37.1; 83.9)	**76.9 (65.2; 88.6)**
Blood pressure normal level (%)	**79.4 (72.2; 85.5)**	48.9 (10.5; 87.4)	**81.3 (66.7; 93.5)**
Lipids normal level (%)	**70.9 (60.0; 81.9)**	69.6 (39.9; 99.2)	53.0 (38.5; 67.5)
	T1 (8 years)		
	Latent status 1: No disorders	Latent status 2: Hypertension	Latent status 3: Central obesity
	N = 199 (59.8%)	N = 59 (17.9%)	N = 74 (22.3%)
Waist circunference normal level (%)	**92.5 (85.7; 99.2)**	**88.0 (70.6; 100.0)**	24.0 (0.0; 51.7)
Glucose normal level (%)	**93.7 (89.1; 98.2)**	**92.7 (83.0; 100.0)**	**94.5 (86.5; 100.0)**
Blood pressure normal level (%)	**87.3 (71.2; 100.0)**	**0.0 (0.0; 0.0)**	53.5 (34.2; 72.7)
Lipids normal level (%)	**75.5 (67.6; 83.5)**	**99.1 (85.6; 100.0)**	54.2 (36.6; 71.8)

The item-response probabilities shown correspond to the 'normal' category of each disorder. Item-response probabilities corresponding to the 'monitoring' category are the complements of those corresponding to the 'normal' category; therefore, they are not reported here. Statistically significant estimates are shown in bold. Data are expressed as a percentage.

https://doi.org/10.1371/journal.pone.0283799.t002

At T1, latent status one was characterised by the equivalent of the first latent status at T0 and was also labelled as no disorders. Latent status two at T1 was characterised by a zero probability of having normal blood pressure and a high probability of having normal levels in the rest of disorders, so it was labelled as '*hypertension*'. Children in latent status three were characterised as having a low probability of having a normal WC (24.0%), and therefore, this latent status was labelled as "*central obesity*". The difference between this status and the identically named status at T0 is that, in this case, only glucose concentrations had a high probability of being in the normal range. Therefore, subjects in the status of central obesity at T0 could present central obesity alone or accompanied by dyslipidaemia, while at T1, they could suffer from central obesity alone or accompanied by hypertension and/or dyslipidaemia. The latent status with the highest prevalence at T1 was the no disorders status (59.8%), followed by the central obesity status (22.3%) and the hypertension status (17.9%). WC and blood pressure at T1 were strongly related to the children's latent status because these variables clearly differentiated in which latent status a child would be classified.

Transition probabilities of change from one latent status to another between T0 and T1 are shown in Fig 3. The no disorders status (latent status one) showed high stability, with a 93.7% probability that a child without disorders at T0 would remain with no disorders (latent status one) at T1. Among children with an altered latent status, the subgroup with some disorders at T0 (latent status two) had a 67.7% probability of being in the hypertension status (latent status two) at T1 and a 32.3% of probability of being in the central obesity status (latent status three) at T1. The subgroup of central obesity (latent status three) at T0 had a probability of 67.6% of being in the central obesity status (latent status three) at T1 and a probability of 32.4% of being in the no disorders status (latent status one) at T1.

Discussion

In the present study, three latent statuses were identified at 4 and at 8 years of age, and they were different between these two periods. At both years of age, there was a predominant pattern defined by the presence of no disorders, which was highly stable. At 4 years of age, we also

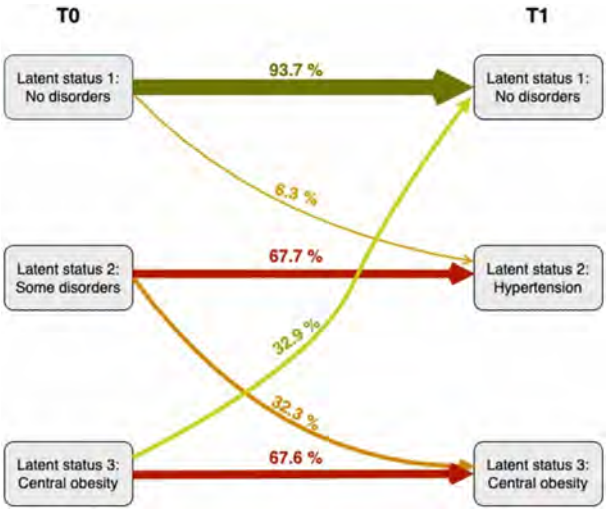


Fig 3. Probability of transitioning to each latent status at T1 (8 years), conditioned by the latent status at T0 (4 years). There were 154 subjects who had measurement of all variables performed at the two time points. Data are expressed as a percentage.

https://doi.org/10.1371/journal.pone.0283799.g003

identified a subgroup characterised by children who had some disorders and by a subgroup that presented with central obesity and could be accompanied by dyslipidaemia. Over time, children transitioned to one subgroup defined by only hypertension and to another subgroup defined by central obesity, which could be accompanied by dyslipidaemia and/or hypertension (in addition to the no disorder subgroup already mentioned). These results support the recommendation of not defining MetS at paediatric ages [17] because no patterns with aggregations of three or more of the disorders of hypertension, central obesity, hyperglycaemia, and dyslipidaemia were clearly identified.

The lack of a standard criterion at paediatric ages to define cut-off points that discriminate between normal or abnormal levels of the variables used in the present study (WC, glucose concentrations, lipid concentrations, and blood pressure), which enable definition of the above-mentioned disorders, causes difficulty in comparing the prevalence between studies [12]. The reason for this difficulty is the high variability depending on the definition used [7, 9, 45]. In this study, the definition provided by the IDEFICS study was used because, although it is not the most widely used, it is applicable to children aged 2–10.9 years. In the IDEFICS study, cut-off points were calculated from a large and healthy European population that covered southern European countries, such as Spain, in which the sample of this study was based.

Taking into account that the cut-off values used to discriminate between normal and abnormal (monitoring levels) for each of the CVD risk factors derived from IDEFICS study were based on the 90th percentile, we expected that 10% of children would have abnormal levels of these disorders. However, in our sample, we observed that the prevalence ranged from 20.8% to 33.7% at 4 years of age, and from 6.3% to 35.8% at 8 years of age. These prevalences are higher than those observed in the reference population [8] and in other European populations in which the same definitions were applied [26]. An exception was glucose concentrations at 8 years of age in our study, which had a lower prevalence (6.3%). In the reference population on which the cut-off points were calculated those who were overweight or obese were excluded, and it has been observed that the prevalence of some of these disorders is positively associated

with overweight and obesity [7, 8, 14, 45–48]. The fact that the INMA-Asturias cohort is characterised by a high prevalence of overweight and obesity [32] could explain the high prevalence of these CVD risk factors in monitoring levels compared with the reference population.

Regarding the number of disorders at the monitoring level, 62.9% of the children at 4 years and 71.2% at 8 years of age presented with at least one component at abnormal levels. These data are consistent with those reported in other studies, in which approximately two thirds of the study population had at least one altered disorder [48, 49]. However, because the definitions and cut-off points used are not the same, comparisons should be considered with caution. Importantly, an improvement in the number of disorders present in the same child was observed between 4 and 8 years of age because the percentage of children who did not present with any CVD increased, that of children who presented with only one CVD risk factor remained the same, and that of children who presented with more than one CVD risk factor simultaneously decreased.

Three latent statuses were identified at 4 years, namely no disorders, some disorders, and central obesity. The no disorders status had a prevalence of 55.9%, and the other two latent statuses had a prevalence of approximately 20%. The central obesity latent status was clearly defined by the presence of central obesity, but the probability of having normal lipid concentrations was approximately 50%. This finding indicated that central obesity could appear alone or be accompanied mainly by dyslipidaemia. The latent status labelled as some disorders has a difficult interpretation due to its poor homogeneity. As all the variables have around 50% of probability of being in the normal category, there is a high degree of uncertainty about what is being captured in that group and no specific pattern can be identify. But that also give us some information. As no other latent status (apart from the central obesity) characterised by a single disorder was found at 4 years, despite the fact that 41.6% of the children had only one disorder, the existence of this some disorders status suggests there is no one disorder that presents a higher prevalence than the others in isolation.

Three latent statuses were identified at 8 years, namely no disorders, hypertension, and central obesity. The no disorders status was the most prevalent, with almost two thirds (59.1%) of the children in that group. At this age, two latent statuses were characterised by the single disorder of hypertension, with a prevalence of 17.9%, and central obesity, with a prevalence of 22.3%. In the case of central obesity, the confidence intervals of the item-response probabilities of normal blood pressure and normal lipid concentrations (confidence interval: 34.2%, 72.7%; and confidence interval: 36.6%, 71.8%, respectively) indicated that central obesity could be partially accompanied by hypertension and dyslipidaemia.

The latent statuses identified at 4 and 8 years of age were distinct (except in the no disorders status), suggested a lack of stability of these disorders between these ages, either in isolation or in clusters. The stability of CVD risk factors associated with MetS at older ages has been discussed in several articles, although mostly from childhood and adolescence to young adulthood and adulthood, instead of between different ages within childhood [13, 19–26]. While the results of these studies are not consistent, most of them are consistent with the lack of stability that we observed [13, 19–23, 26]. The main difference between the time points is that, at 4 years of age, there was no pattern characterised by an isolated disorder, and central obesity could be accompanied by dyslipidaemia. However, at 8 years of age, the detected patterns changed, a latent status composed only of hypertension appeared to be relevant and the latent status of central obesity could be accompanied by dyslipidaemia and/or hypertension. Therefore, the number of disorders that appear at the same time as central obesity increases between 4 and 8 years of age. These findings support the hypothesis that central obesity is a CVD risk factor that triggers other disorders, which usually emerge as comorbidities of central obesity [11, 14, 17, 33]. Regarding to hypertension, the increasing trend of elevated blood pressure

levels in childhood and adolescence in terms of primary hypertension has been mainly related to the also increasing trend of childhood obesity, but also to other factors such as prenatal and postnatal exposures, genetic factors, birth characteristics, dietary habits, or lifestyle factors [50]. Some of them have been shown to be related to primary hypertension independently of obesity, as evidenced for example by Rosner et al. [51] in the case of high sodium intake. Specifically, it has also been found that the Spanish pediatric population exceeds the recommended sodium intake at ages 9 to 12 years regardless of their nutritional status [52], and that it is positive associated with an elevated diastolic blood pressure between 5 and 16 years old, also regardless of their nutritional status [53]. So the latent status characterized only by hypertension is evidenced the group of children with high blood pressure values related to causes independent of obesity, since those who present elevated blood pressure values accompanied by obesity would be classified in the latent status of central obesity.

There are two major similarities between the latent status at 4 and 8 years of age. The no disorders status was observed in slightly more than half of the sample, and the other status, which was defined by central obesity accompanied by other disorders, represented approximately one quarter of the sample at the two time points.

Transition probabilities showed that the no disorders status in children at 4 years of age remained mostly in the same status at 8 years of age, because only a small proportion of children transitioned to other statuses at 8 years (6.3% change to the hypertension status). This observation is in line with other studies, which reported that the group of children who were in a healthy or lower metabolic risk status at the start of the follow-up were more likely to remain in that group [25, 26].

Among the children who suffered from some disorders at 4 years, 67.7% transitioned to have only hypertension at 8 years, while 32.3% transitioned to have central obesity, which could be accompanied by hypertension and/or by dyslipidaemia. The finding that hyperglycaemia disappeared at 8 years of age, because all of the latent statuses at 8 years had a high probability of normal glucose concentrations, could be due to the fact that, at 4 years, the stress caused in such young children by blood collection may produce an unrealistic elevation in glucose concentrations. Additionally, this disappearance could also have been affected by the importance that parents gave to this warning sign, taking measures to reverse it.

Children with central obesity at 4 years of age maintained central obesity at 8 years of age, with a 67.6% probability. However, 32.4% of these children reversed their status by moving to the no disorders status.

Comparing the results of this study with those in other studies with the same or similar objectives is difficult because of the wide variety of variables considered, cut-off points used, diversity in methodology applied, and follow-up periods of the populations studied. We only found one related study in childhood with the use LTA by Bornhorst et al. [26] who analysed the latent status and its transition at 6, 8, and 12 years of age. When we focussed on the results of this previous study between 6 and 8 years of age, because they are the most comparable with our sample, the latent statuses detected were different regarding their pattern and number. However, this previous study also observed a similar prevalence and a higher stability of the no disorders status. Furthermore, central obesity appeared to be the most likely to be accompanied by other disorders, as in our study. The differences detected between the other latent statuses in the Bornhorst et al.'s study and ours may be due to the fact that, in our study, no latent statuses were identified with a prevalence of $< 10\%$. Additionally, the smaller sample size of our study may not have allowed us to distinguish underlying patterns. Moreover, in Bornhorst et al.'s study [26], item-response probabilities were restricted to be equal across time, which meant that the identified latent statuses were forced to be equal at the different time points, unlike in our study.

The present study has several limitations. A larger sample size for this study would have been preferable because LTA is a methodology that uses multiple combinations of categorical variables. There may have been different patterns or combinations of responses with a low or even null frequency, which could have affected the results of the estimates and the width of the confidence intervals. The sample size did not allow us to carry out a stratified analysis by sex, which would have been desirable, because sensitivity analysis due to sex-based differences has been observed in some studies [8, 9, 20]. Additionally, the sample size did not enable the introduction of explanatory variables to estimate their effect on the transition probabilities. With regard to the method of defining the disorders at normal levels, in this study, the variables were dichotomised. This could have resulted in a loss of information. Several authors have suggested that CVD risk factors should be treated as continuous [8, 9, 54, 55]. Another limitation is that, because of the lack of longitudinal studies from childhood to adulthood, there are no cut-off points defined on the basis of a relevant and quantifiable increase in the CVD risk in the future or on the basis of biological evidence [8]. Moreover, in this study, the application of cut-off values from a pre-existing reference population, which was different to the population from which the study sample was drawn, could have overestimated or underestimated the prevalence of the disorders. This possibility suggests the necessity of cut-off points that are ethnicity-specific in each country [56]. Finally, blood samples were not collected after 12 hours of fasting. Recent studies have shown that lipid profiles only minimally change in response to normal food intake in individuals in the general population [57], but they could have a greater effect on glucose measurements. Therefore, the results concerning hyperglycaemia should be treated with caution [58].

There are some strengths of this study. Despite using a different reference population, the chosen population included subjects from Spain and other southern European countries, as well as sex- and age- and even height-specific cut-off points. Therefore, this takes into account the physiological changes in childhood, which is a period in which several modifications that affect cardiometabolic parameters occur [34]. Moreover, this study provides knowledge on the evolution and stability of aggregations of CVD risk factors at the paediatric age through a longitudinal study. There have not been many studies that have conducted this analysis during childhood [19, 25, 26, 59, 60], with most studies from childhood or adolescence to young adulthood or adulthood. Therefore, taking into account the existing gaps in knowledge regarding the interrelations between CVD risk factors and their joint appearance, the information provided by the current study could help to provide further understanding on this topic.

Conclusions

The cluster patterns of different CVDs risk factors are not maintained between the ages of 4 and 8 years, except for those in children who have no disorders. At 8 years of age, the prevalence of hypertension is high and occurs in isolation. Central obesity is found at these two ages. Therefore, the early detection of central obesity has importance for the correct control of its evolution. In addition, central obesity should play a major role in the prevention of the development of other CVD disorders because it is accompanied by other disorders, and the number of disorders that may accompany it increases from the age of 4 to 8 years. The next steps regarding this issue should be focussed on attempting to understand what underlying factors could explain the changes in the latent status throughout childhood.

Supporting information

S1 Data.
(TXT)

S1 File.
(DOCX)

Acknowledgments

The authors would particularly like to thank all of the participants and the families for their generous participation in the study. The authors are grateful to the medical board, the Departments of Gynaecology and Paediatrics in Hospital San Agustín de Avilés, and the Health Centre of Las Vegas in Corvera de Asturias for their disinterested involvement in the project. We thank Ellen Knapp, PhD, from Edanz (https://edanz.com/ac) for editing a draft of this manuscript.

Author Contributions

Conceptualization: Ana Fernández-Somoano, Isolina Riaño-Galán, Adonina Tardón.

Data curation: Rocío Fernández-Iglesias.

Formal analysis: Rocío Fernández-Iglesias.

Funding acquisition: Adonina Tardón.

Investigation: Rocío Fernández-Iglesias, Ana Fernández-Somoano, Isolina Riaño-Galán.

Methodology: Rocío Fernández-Iglesias, Ana Fernández-Somoano.

Project administration: Adonina Tardón.

Resources: Rafael Venta-Obaya.

Supervision: Ana Fernández-Somoano, Adonina Tardón.

Visualization: Rocío Fernández-Iglesias.

Writing – original draft: Rocío Fernández-Iglesias.

Writing – review & editing: Rocío Fernández-Iglesias, Ana Fernández-Somoano, Cristina Rodríguez-Dehli, Rafael Venta-Obaya, Isolina Riaño-Galán, Adonina Tardón.

References

1. Li Z, Lin L, Wu H, Yan L, Wang H, Yang H, et al. Global, Regional, and National Death, and Disability-Adjusted Life-Years (DALYs) for Cardiovascular Disease in 2017 and Trends and Risk Analysis From 1990 to 2017 Using the Global Burden of Disease Study and Implications for Prevention. Front Public Heal. 2021; 9(October). https://doi.org/10.3389/fpubh.2021.559751 PMID: 34778156
2. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. J Am Coll Cardiol. 2020; 76(25):2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010 PMID: 33309175
3. Wang W, Hu M, Liu H, Zhang X, Li H, Zhou F, et al. Global Burden of Disease Study 2019 suggests that metabolic risk factors are the leading drivers of the burden of ischemic heart disease. Cell Metab [Internet]. 2021; 33(10):1943–1956.e2. Available from: https://doi.org/10.1016/j.cmet.2021.08.005 PMID: 34478633
4. O'Donnell CJ, Elosua R. Cardiovascular risk factors. Insights from framingham heart study. Rev Esp Cardiol. 2008; 61(3):299–310.
5. Abbafati C, Abbas KM, Abbasi-Kangevari M, Abd-Allah F, Abdelalim A, Abdollahi M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020; 396(10258):1223–49. https://doi.org/10.1016/S0140-6736(20)30752-2 PMID: 33069327
6. Mottillo S, Filion KB, Genest J, Joseph L, Pilote L, Poirier P, et al. The metabolic syndrome and cardiovascular risk: A systematic review and meta-analysis. J Am Coll Cardiol. 2010 Sep 28; 56(14):1113–32. https://doi.org/10.1016/j.jacc.2010.05.034 PMID: 20863953

7. Kelishadi R. Childhood Overweight, Obesity, and the Metabolic Syndrome in Developing Countries. Epidemiol Rev [Internet]. 2007 Jan 1 [cited 2021 Jul 9]; 29(1):62–76. Available from: https://academic.oup.com/epirev/article/29/1/62/436110 PMID: 17478440

8. Ahrens W, Moreno L, Mårild S, Molnár D, Siani A, De Henauw S, et al. Metabolic syndrome in young children: Definitions and results of the IDEFICS study. Int J Obes [Internet]. 2014 Sep 1 [cited 2021 Feb 10]; 38(2):S4–14. Available from: www.nature.com/ijo PMID: 25376220

9. Reisinger C, Nkeh-Chungag BN, Fredriksen PM, Goswami N. The prevalence of pediatric metabolic syndrome—a critical look on the discrepancies between definitions and its clinical importance. Int J Obes [Internet]. 2020 Nov 18 [cited 2021 Jul 9]; 45(1):12–24. Available from: https://www.nature.com/articles/s41366-020-00713-1 PMID: 33208861

10. Kassi E, Pervanidou P, Kaltsas G, Chrousos G. Metabolic syndrome: Definitions and controversies. Lancet [Internet]. 2005 [cited 2022 Jan 13]; 366(9491):1059–62. Available from: http://www.biomedcentral.com/1741-7015/9/48

11. Al-Hamad D, Raman V. Metabolic syndrome in children and adolescents. Transl Pediatr [Internet]. 2017 Oct 1 [cited 2022 Jan 12]; 6(4):397. https://doi.org/10.21037/tp.2017.10.02 PMID: 29184820

12. Ford ES, Li C. Defining the Metabolic Syndrome in Children and Adolescents: Will the Real Definition Please Stand Up? J Pediatr. 2008; 152(2). https://doi.org/10.1016/j.jpeds.2007.07.056 PMID: 18206681

13. Magnussen CG, Koskinen J, Chen W, Thomson R, Schmidt MD, Srinivasan SR, et al. Pediatric Metabolic Syndrome Predicts Adulthood Metabolic Syndrome, Subclinical Atherosclerosis, and Type 2 Diabetes Mellitus but Is No Better Than Body Mass Index Alone The Bogalusa Heart Study and the Cardiovascular Risk in Young Finns Study. Pediatr Cardiol [Internet]. 2010 Oct 19 [cited 2021 Feb 10]; 122(16):1604–11. Available from: http://circ.ahajournals.org PMID: 20921439

14. Morrison JA, Friedman LA, Gray-Mcguire C. Metabolic Syndrome in Childhood Predicts Adult Cardiovascular Disease 25 Years Later: The Princeton Lipid Research Clinics Follow-up Study. 2007 [cited 2021 May 4]; www.pediatrics.org/cgi/doi/10.1542/

15. Koskinen J, Magnussen CG, Sinaiko A, Woo J, Urbina E, Jacobs DR, et al. Childhood Age and Associations Between Childhood Metabolic Syndrome and Adult Risk for Metabolic Syndrome, Type 2 Diabetes Mellitus and Carotid Intima Media Thickness: The International Childhood Cardiovascular Cohort Consortium. J Am Heart Assoc [Internet]. 2017 Aug 1 [cited 2022 Jan 13]; 6(8). Available from: https://pubmed.ncbi.nlm.nih.gov/28862940/

16. Kahn R, Buse J, Ferrannini E, Stern M. The Metabolic Syndrome: Time for a Critical Appraisal Joint statement from the American Diabetes Association and the European Association for the Study of Diabetes. 2005 [cited 2022 Feb 4]; http://diabetesjournals.org/care/article-pdf/28/9/2289/659731/zdc00905002289.pdf

17. George MMAK, Zimmet P, George Alberti KM, Kaufman F, Tajima N, Silink M, et al. The metabolic syndrome in children and adolescents—an IDF consensus report. Vol. 8, Pediatric Diabetes. 2007.

18. Foulkes MA, Davis CE. An Index of Tracking for Longitudinal Data. Biometrics. 1981; 37(3):439.

19. Gustafson JK, Yanoff LB, Easter BD, Brady SM, Keil MF, Roberts MD, et al. The Stability of Metabolic Syndrome in Children and Adolescents. J Clin Endocrinol Metab [Internet]. 2009 Dec 1 [cited 2022 Feb 18]; 94(12):4828–34. Available from: https://academic.oup.com/jcem/article/94/12/4828/2596689 PMID: 19837941

20. Thangiah N, Chinna K, Su TT, Jalaludin MY, Al-Sadat N, Majid HA. Clustering and Tracking the Stability of Biological CVD Risk Factors in Adolescents: The Malaysian Health and Adolescents Longitudinal Research Team Study (MyHeARTs). Front Public Heal. 2020 Mar 17; 8:69.

21. Asghari G, Eftekharzadeh A, Hosseinpanah F, Ghareh S, Mirmiran P, Azizi F. Instability of different adolescent metabolic syndrome definitions tracked into early adulthood metabolic syndrome: Tehran Lipid and Glucose Study (TLGS). Pediatr Diabetes [Internet]. 2017 Feb 1 [cited 2022 Feb 18]; 18(1):59–66. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/pedi.12349 PMID: 26825860

22. Goodman E, Daniels SR, Meigs JB, Dolan LM. Instability in the diagnosis of metabolic syndrome in adolescents. Circulation [Internet]. 2007 May 1 [cited 2022 Feb 18]; 115(17). Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.106.669994 PMID: 17420347

23. Goodman E, Li C, Tu YK, Ford E, Sun SS, Huang TTK. Stability of the Factor Structure of the Metabolic Syndrome across Pubertal Development: Confirmatory Factor Analyses of Three Alternative Models. J Pediatr [Internet]. 2009 Sep 1 [cited 2022 Feb 18]; 155(3):S5.e1–S5.e8. Available from: https://doi.org/10.1016/j.jpeds.2009.04.045 PMID: 19732562

24. Katzmarzyk PT, Pérusse L, Malina RM, Bergeron J, Després JP, Bouchard C. Stability of indicators of the metabolic syndrome from childhood and adolescence to young adulthood: The Quebec Family Study. J Clin Epidemiol. 2001 Feb 1; 54(2):190–5.

25. Li C, Ford ES, T-K Huang T, Sun SS, Goodman E, Sun S. Patterns of Change in Cardiometabolic Risk Factors Associated with the Metabolic Syndrome among Children and Adolescents: The Fels Longitudinal Study NIH Public Access. J Pediatr. 2009; 155(3).

26. Börnhorst C, Russo P, Veidebaum T, Tornaritis M, Molnár D, Lissner L, et al. Metabolic status in children and its transitions during childhood and adolescence—the IDEFICS/I.Family study. Int J Epidemiol [Internet]. 2019 Oct 1 [cited 2021 Feb 10]; 48(5):1673–83. Available from: https://academic.oup.com/ije/article/48/5/1673/5490631 PMID: 31098634

27. Nguena Nguefack HL, Pagé MG, Katz J, Choinière M, Vanasse A, Dorais M, et al. Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. Clin Epidemiol [Internet]. 2020 [cited 2021 Dec 3]; 12:1205. https://doi.org/10.2147/CLEP.S265287 PMID: 33154677

28. Sorgente A, Lanz M, Serido J, Tagliabue S, Shim S. Latent transition analysis: Guidelines and an application to emerging adults' social development. TPM—Testing, Psychom Methodol Appl Psychol. 2019; 26(1):39–72.

29. Lanza ST, Flaherty BP, Collins LM. Latent Class and Latent Transition Analysis. Handbook of Psychology. 2003.

30. Fernández-Somoano A, Estarlich M, Ballester F, Fernández-Patier R, Aguirre-Alfaro A, Herce-Garraleta MD, et al. Outdoor NO2 and benzene exposure in the INMA (Environment and Childhood) Asturias cohort (Spain). Atmos Environ [Internet]. 2011 Sep 1 [cited 2022 Feb 28]; 45(29):5240–6. Available from: https://doi.org/10.1016/j.atmosenv.2011.02.010

31. Fernández-Somoano A, Tardon A. Socioeconomic status and exposure to outdoor NO2 and benzene in the Asturias INMA birth cohort, Spain. J Epidemiol Community Health [Internet]. 2014 [cited 2022 Feb 28]; 68(1):29. https://doi.org/10.1136/jech-2013-202722 PMID: 23999377

32. Riaño-Galán I, Fernández-Somoano A, Rodríguez-Dehli C, Valvi D, Vrijheid M, Tardón A. Proatherogenic Lipid Profile in Early Childhood: Association with Weight Status at 4 Years and Parental Obesity. J Pediatr [Internet]. 2017 Aug 1 [cited 2022 Feb 28]; 187:153–157.e2. Available from: https://doi.org/10.1016/j.jpeds.2017.04.042 PMID: 28529017

33. Alberti KGMMM, Zimmet P, Shaw J. The metabolic syndrome—a new worldwide definition. Lancet [Internet]. 2005 Sep 24 [cited 2022 Feb 17]; 366(9491):1059–62. Available from: https://www.sciencedirect.com/science/article/pii/S0140673605674028 PMID: 16182882

34. Mameli C, Zuccotti GV, Carnovale C, Galli E, Nannini P, Cervia D, et al. An update on the assessment and management of metabolic syndrome, a growing medical emergency in paediatric populations. Pharmacol Res [Internet]. 2017 May 1 [cited 2022 Apr 26]; 119:99–117. Available from: https://doi.org/10.1016/j.phrs.2017.01.017 PMID: 28111263

35. Nagy P, Kovacs E, Moreno LA, Veidebaum T, Tornaritis M, Kourides Y, et al. Percentile reference values for anthropometric body composition indices in European children from the IDEFICS study. Int J Obes [Internet]. 2014 [cited 2021 May 5]; 38:15–25. Available from: www.nature.com/ijo PMID: 25219408

36. Peplies J, Jiménez-Pavón D, Savva SC, Buck C, Günther K, Fraterman A, et al. Percentiles of fasting serum insulin, glucose, HbA1c and HOMA-IR in pre-pubertal normal weight European children from the IDEFICS cohort. Int J Obes [Internet]. 2014 [cited 2021 May 5]; 38:39–47. Available from: www.nature.com/ijo PMID: 25376219

37. De Henauw S, Michels N, Vyncke K, Hebestreit A, Russo P, Intemann T, et al. Blood lipids among young children in Europe: results from the European IDEFICS study. 2014 [cited 2021 May 5]; http://www.cholestech.com/products/ldx_overview.

38. Barba G, Buck C, Bammann K, Hadjigeorgiou C, Hebestreit A, Mårild S, et al. Blood pressure reference values for European non-overweight school children: The IDEFICS study. Int J Obes [Internet]. 2014 Sep 15 [cited 2021 May 5]; 38(S2):S48–56. Available from: www.nature.com/ijo PMID: 25219411

39. Ryoo JH, Wang C, Swearer SM, Hull M, Shi D. Longitudinal model building using latent transition analysis: An example using school bullying data. Front Psychol. 2018; 9(MAY). https://doi.org/10.3389/fpsyg.2018.00675 PMID: 29867652

40. Stanton WJ, Etzel MJ, Walker BJ. Latent Transition Analysis: Modeling Extensions and an Application to Peer Victimization. 2007;634.

41. Collins LM, Lanza ST, Schafer JL, Flaherty BP. WinLTA USER ' S GUIDE. Simulation. 2002;(May):1–45.

42. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2021. https://www.r-project.org/

43. Muthén LK, Muthén BO. Mplus User's Guide [Internet]. Seventh Ed. Los Angeles: Author. Los Angeles: Muthén & Muthén; 2012. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Mplus+user+guide#8

44. Hallquist MN, Wiley JF. MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. Struct Equ Model [Internet]. 2018 Jul 4 [cited 2022 Jan 21]; 25(4):1–18. Available from: https://www.tandfonline.com/doi/full/10.1080/10705511.2017.1402334

45. Ramírez-Vélez R, Anzola A, Martinez-Torres J, Vivas A, Tordecilla-Sanders A, Prieto-Benavides D, et al. Metabolic Syndrome and Associated Factors in a Population-Based Sample of Schoolchildren in Colombia: The FUPRECOL Study. Metab Syndr Relat Disord. 2016 Nov; 14(9):455–62. https://doi.org/10.1089/met.2016.0058 PMID: 27508490

46. Cook S, Weitzman M, Auinger P, Nguyen M, Dietz WH. Prevalence of a Metabolic Syndrome Phenotype in Adolescents Findings From the Third National Health and Nutrition Examination Survey, 1988–1994. Arch Pediatr Adolesc Med [Internet]. 2003 Aug 1 [cited 2021 Dec 22]; 157(8):821–7. Available from: https://jamanetwork.com/journals/jamapediatrics/fullarticle/481403 PMID: 12912790

47. Galera-Martínez R, García-García E, Vázquez-López MÁ, Ortiz-Pérez M, Ruiz-Sánchez AM, Martín-González M, et al. Prevalencia de síndrome metabólico en la población general adolescente de una ciudad del área Mediterránea: Comparación de dos definiciones. Nutr Hosp. 2015 Aug 4; 32(2):627–33.

48. Costa Dias Pitangueira J, Rodrigues Silva L, Leila Portela de Santana M, da Conceição Monteiro da Silva M, Ribas de Farias Costa P, Marlúcia de Oliveira Assis A. Metabolic syndrome and associated factors in children and adolescents of a Brazilian municipality. Nutr Hosp. 2014; 29(4):865–72. https://doi.org/10.3305/nh.2014.29.4.7206 PMID: 24679029

49. Miller JM, Kaylor MB, Johannsson M, Bay C, Churilla JR. Prevalence of Metabolic Syndrome and Individual Criterion in US Adolescents: 2001–2010 National Health and Nutrition Examination Survey. Metab Syndr Relat Disord [Internet]. 2014; 12(10):527–32. Available from: https://doi.org/10.1089/met.2014.0055 PMID: 25247821

50. S. Machado IB, Tofanelli MR, Saldanha da Silva AA, Simões e Silva AC. Factors Associated with Primary Hypertension in Pediatric Patients: An Up-to-Date. Curr Pediatr Rev. 2021 Apr 5; 17(1):15–37. https://doi.org/10.2174/1573396317999210111200222 PMID: 33430749

51. Rosner B, Cook NR, Daniels S, Falkner B. Childhood Blood Pressure Trends and Risk Factors for High Blood Pressure: The NHANES experience 1988–2008. Hypertension [Internet]. 2013 Aug [cited 2023 Feb 20]; 62(2):247. https://doi.org/10.1161/HYPERTENSIONAHA.111.00831 PMID: 23856492

52. Partearroyo T, Samaniego-Vaesken M de L, Ruiz E, Aranceta-Bartrina J, Gil Á, González-Gross M, et al. Sodium Intake from Foods Exceeds Recommended Limits in the Spanish Population: The ANIBES Study. Nutr 2019, Vol 11, Page 2451 [Internet]. 2019 Oct 14 [cited 2023 Feb 11]; 11(10):2451. Available from: https://www.mdpi.com/2072-6643/11/10/2451/htm PMID: 31615065

53. Pérez-Gimeno G, Rupérez AI, Vázquez-Cobela R, Herráiz-Gastesi G, Gil-Campos M, Aguilera CM, et al. Energy dense salty food consumption frequency is associated with diastolic hypertension in Spanish children. Nutrients. 2020 Apr 1; 12(4). https://doi.org/10.3390/nu12041027 PMID: 32283662

54. Eisenmann JC. On the use of a continuous metabolic syndrome score in pediatric research. Cardiovasc Diabetol [Internet]. 2008 Jun 5 [cited 2021 Feb 10]; 7(17):1–6. Available from: http://www.cardiab.com/content/7/1/17 PMID: 18534019

55. Wijndaele K, Beunen G, Duvigneaud N, Matton L, Duquet W, Thomis M, et al. A Continuous Metabolic Syndrome Risk ScoreUtility for epidemiological analyses. Diabetes Care [Internet]. 2006 Oct 1 [cited 2022 Jan 11]; 29(10):2329–2329. https://doi.org/10.2337/dc06-1341 PMID: 17003322

56. Agirbasli M, Tanrikulu AM, Berenson GS. Metabolic Syndrome: Bridging the Gap from Childhood to Adulthood. Cardiovasc Ther [Internet]. 2016 Feb 1 [cited 2022 Jan 13]; 34(1):30–6. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1755-5922.12165 PMID: 26588351

57. Langsted A, Freiberg JJ, Nordestgaard BG. Fasting and nonfasting lipid levels influence of normal food intake on lipids, lipoproteins, apolipoproteins, and cardiovascular risk prediction. Circulation. 2008; 118(20):2047–56. https://doi.org/10.1161/CIRCULATIONAHA.108.804146 PMID: 18955664

58. Moebus S, Göres L, Lösch C, Jöckel KH. Impact of time since last caloric intake on blood glucose levels. Eur J Epidemiol. 2011; 26(9):719–28. https://doi.org/10.1007/s10654-011-9608-z PMID: 21822717

59. Bugge A, El-Naaman B, Mcmurray RG, Froberg K, Andersen LB. Tracking of clustered cardiovascular disease risk factors from childhood to adolescence. Pediatr Res 2013 732 [Internet]. 2012 Nov 19 [cited 2022 Feb 18]; 73(2):245–9. Available from: https://www.nature.com/articles/pr2012158 PMID: 23165452

60. Bahar A, Esfahani FH, Jafarabadi MA, Mehrabi Y, Azizi F. The structure of metabolic syndrome components across follow-up survey from childhood to adolescence. Int J Endocrinol Metab. 2013; 11(1):16–22. https://doi.org/10.5812/ijem.4477 PMID: 23853615

3.2. Article II: Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA-Asturias cohort

To cite this article:
R. Fernández-Iglesias, P. Martinez-Camblor, A. Fernández-Somoano, C. Rodríguez- Dehli, R. Venta-Obaya, M. R. Karagas, A. Tardón, and I. Riaño-Galán. Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA- Asturias cohort. *European Journal of Pediatrics*, 2023. DOI: 10.1007/S00431-023-05051-8.

To link to this article:
https://doi.org/10.1007/S00431-023-05051-8

European Journal of Pediatrics
https://doi.org/10.1007/s00431-023-05051-8

RESEARCH

Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA-Asturias cohort

Rocío Fernández-Iglesias[1,2,3] · Pablo Martinez-Camblor[4,5] · Ana Fernández-Somoano[1,2,3] · Cristina Rodríguez-Dehli[3,6] · Rafael Venta-Obaya[7,8] · Margaret R. Karagas[9] · Adonina Tardón[1,2,3] · Isolina Riaño-Galán[1,3,10]

Received: 24 February 2023 / Revised: 25 May 2023 / Accepted: 31 May 2023

Abstract

Identifying cardiovascular-related measures that track from early childhood into later ages may help inform early prevention targets for cardiovascular disease. In this study, the tracking of triglycerides (TG), high-density cholesterol (HDL-c), atherogenic coefficient (AC), waist circumference to height ratio (WC/Height), mean arterial pressure (MAP), and homeostatic model assessment of insulin resistance (HOMA-IR) was examined in the INMA-Asturias cohort between 4 and 8 years of age. The analysis was conducted in 307 children who participated in the INMA-Asturias cohort (Spain) at 4 and at 8 years of age. Quantile regression models were used to evaluate tracking between measures at both ages, with each measure at 8 years as the dependent variable and the rank transformation of the same measure at 4 years as the independent variable. We found a positive association between HDL-c rank at 4 years and higher quantiles of the HDL-c distribution at 8 years, with an increase of 2.93 mg/dL (95% CI: 1.98, 3.87) per decile in the 0.9 quantile. A positive association was also found for WC/Height, with an increase of 0.008 (95% CI: 0.004, 0.012) per decile in the 0.9 quantile. We observed that tracking for AC increased in the higher quantiles of the distribution at 8 years, with an increase of 0.11 (95% CI: 0.09, 0.14) in the 0.6 quantile compared to an effect of 0.15 (95% CI: 0.09, 0.21) in the 0.9 quantile.

Conclusions: Adult markers of dyslipidemia and central obesity tracked between ages 4 and 8 years. For AC, tracking increased in the higher quantiles of the distribution.

What is Known:

- *Atherosclerosis begins in early life, so preventive efforts that start in childhood may delay progression to clinical disease. Determine what cardiovascular risk factors track into time since childhood bring the opportunity to identified those subjects at risk for later cardiovascular disease.*
- *The study of risk factors in health populations and, particularly in children, copes with not clear and/or controversial thresholds definition. This makes it challenging to study tracking in pediatric ages.*

What is New:

- *Quantile regression is a useful tool for assessing the tracking of risk factors for which there are no clinically meaningful thresholds. The increasing trend observed in the tracking of dyslipidemia suggests the possible difficulty that children with abnormal values at 4 years of age might have in normalizing them in future years.*
- *The findings of this article may help to determine which cardiovascular-related measures could be screened and followed-up in children.*

Keywords Cardiovascular risk · Childhood · Dyslipidemia · Hyperglycemia · Hypertension · Obesity · Quantile regression · Tracking

Background

Abnormal values of cardiovascular-related measures are frequently detected in adulthood [1] but may also be present in childhood [2]. This does not increase the risk of cardiovascular diseases (CVDs) in childhood itself; children rarely experience cardiovascular diseases and these occurrences

Communicated by Peter de Winter

✉ Ana Fernández-Somoano
fernandezsana@uniovi.es

Extended author information available on the last page of the article

Published online: 20 June 2023

are mainly caused by congenital heart problems or genetic syndromes [3]. Atherosclerosis, one of the main CVD triggers [4] in adults, is an accumulative process that can begin in childhood and youth [5–7]. Therefore, researchers have been trying to answer whether those subjects exposed to specific metabolic alterations in childhood will have higher risk of developing CVDs — or early CVDs — in adulthood [8].

The study of the association between underlying cardiovascular disease indicators in childhood and CVDs in adulthood has been challenging due to the difficulty of following a young sample the time required to observe in this population CVD events. Several studies have shown evidence that CVDs are associated with childhood metabolic alterations [9]. For instance, a study of 38,589 participants aged 3 to 19 years from the USA, Finland, and Australia found an association between body mass index (BMI), systolic blood pressure, triglycerides, and cholesterol with cardiovascular events in midlife [10], and strongest associations with these factors in aggregate. These findings can be explained by a risk accumulation model, in which risk factors present at each life stage further increase risk in adulthood; a risk chain model, in which risk in childhood is mediated by risk in adulthood; or a sensitivity period model, in which exposure at a particular time in life course confers more risk compared with other stages [9, 11]. Under either scenario, identifying metabolic alterations that are more likely to track from childhood into future years will help inform targets of early prevention.

One of the main difficulties in tracking metabolic disorders in children is the disorder definition itself. The lack of adequate studies linking cardiovascular risk factors in childhood to disease in adulthood leaves pediatric definitions of metabolic disorders based on the distribution of cardiovascular measures in generally healthy children [12]. Therefore, thresholds are controversial. One approach is to model measures continuously, and rely on categorization only for clinical diagnosis [13, 14]. Tracking studies — defined as the maintenance over time of a relative position in the distribution of a variable [15] — have been analyzed mainly using thresholds to categorize the variables of interest, and stratify subjects into risk groups [16]. To avoid that, we propose to study the tracking of the rank values instead of the values themselves. That is, we aim to study whether the subjects with higher values at 4 years still have higher values at 8 years in terms of the variable distribution. With this goal, we consider data from the Infancia y Medio Ambiente (INMA)-Asturias cohort [17] and use quantile regression models [18]. This methodology allows to estimate the effect of an explanatory variable on any quantile of the outcome distribution, permitting the analysis of extreme values of the outcome without setting arbitrary thresholds [19].

For this reason, we aimed to apply this approach to assess whether having extreme values in the cardiovascular-related measures at 4 years is associated with having extreme values in the same cardiovascular-related measures at 8 years. We consider the following measures: triglycerides (TG), high-density cholesterol (HDL-c), atherogenic coefficient (AC), waist circumference to height ratio (WC/Height ratio), mean arterial pressure (MAP), and the homeostatic model assessment of insulin resistance (HOMA-IR).

Materials and methods

Study design

Study subjects were children participating in the INMA (Infancia y Medio Ambiente [Environment and Childhood]) Asturias cohort (north of Spain). Details can be found in previous studies [20, 21]. Briefly, between May 2004 and June 2007, pregnant women in their first trimester of pregnancy were recruited at the San Agustín University Hospital (Avilés) following a common protocol [17]. This hospital is a public health center with 436 beds which provides primary care and central, medical, and surgical services to a population of 144,875 inhabitants according to 2021 census [22]. The inclusion criteria were maternal age $\geq$ 16 years, singleton pregnancy, delivery scheduled at the referenced hospital, no assisted conception, and no communication handicap. Data were collected by trained professionals in several phases of follow-up: at first and third trimester of pregnancy, at birth, and at children's ages 18 months, 4, and 8 years. Information was collected by medical registries, interview-based questionnaires with mothers, blood sample collection, and physical examinations of the children conducted by trained staff.

Cardiovascular-related measurements

For this study we focused on cardiovascular-related measures that reflect well-established CVD risk factors in adulthood: central obesity, insulin resistance, dyslipidemia, and hypertension. These included WC/Height ratio for central obesity [23]; MAP for hypertension [24]; TG, HDL-c, and AC for dyslipidemia [25]; and HOMA-IR for insulin resistance [26].

Lipids

Lipids were measured at 4 and 8 years collecting non-fasting blood samples, obtained by antecubital venipuncture. Serum total cholesterol (T-c), TG, HDL-c, and low-density cholesterol (LDL-c) levels were determined using a Roche

analyzer (Modular Analytics Serum Work Area, Mannheim, Germany). AC was calculated as the difference between T-c and HDL-c, divided by HDL-c. Lipids values are presented in milligrams per deciliter (mg/dL).

Anthropometry

At 4 and 8 years, trained staff measured children height and WC. Height was measured twice to the nearest 0.1 cm using a wall-mounted stadiometer after the participant removed their shoes. Waist circumference was measured to the nearest 0.1 cm at the children midpoint between the right lower rib and the iliac crest at the level of the umbilicus, using an inelastic nylon tape in a horizontal plane, and with children in a standing position. WC/Height ratio was calculated as waist circumference in cm divided by height in cm.

Blood pressure

Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were measured using an automated oscillometric system (OMRON®) at children 4 and 8 years. After a 5-min rest period, between two and three consecutive measurements were taken, with children in a seated position and their right arm at rest at the heart level. The SBP and DBP averaged paired values were used and MAP was calculated as DBP + 1/3(SBP − DBP) [27]. Values are presented in millimeter of mercury (mmHg).

Blood glucose and insulin

Blood glucose and insulin levels were determined using the same Roche analyzer at children 4 and 8 years through collecting non-fasting blood samples, obtained by antecubital venipuncture. Glucose values are presented in milligrams per deciliter (mg/dL) and insulin values in microunits per milliliter (μU/mL). HOMA-IR was calculated as glucose multiplied by insulin and divided by 405.

Potential confounding factors

The following parental characteristics were selected as potential confounders: maternal age at enrollment, maternal pre-pregnancy (BMI), paternal BMI, maternal educational level, maternal social class, maternal smoking during pregnancy, and parental CVD antecedents (neither parent has antecedents/one parent has at least one antecedent/both parents have at least one antecedent). Regarding pre-pregnancy BMI, the maternal height and pre-pregnancy weight were self-reported, both at the first-trimester visit. These values were used to calculate the pre-pregnancy BMI (in kg/m^2). Paternal weight and height were reported by the mother at the first-trimester visit and were used to calculate paternal BMI. Questionnaires administered during the first and third trimester of pregnancy obtained information on maternal and paternal age and education, maternal country of birth, maternal and paternal occupation, and maternal smoking during pregnancy. Social class was defined according to the occupation during pregnancy of the mother or father, using a widely used Spanish adaptation of the International Standard Classification of Occupations coding system [28]. Parental CVD antecedent's variable was reported by the mother in the first trimester of pregnancy. She was asked whether she or the father had been diagnosed with diabetes, heart disease, coagulation disorders, hypertension, or hypercholesterolemia and the responses were combined to create a categorical variable according to whether neither parent had any of them, whether one parent had at least one of them, or whether both parents had at least one of them. Children characteristics selected as potential confounders were age, height, weekly out-of-school physical activity time, and the mean of the daily energy intake. All of them were collected at the 4- and 8-year follow-ups. Week of gestation at delivery, birth weight, predominant breastfeeding duration, and sex were also considered. These information were collected from medical records, except for data on predominant breastfeeding duration, which were collected when the children were approximately 6 and 14 months old through questionnaires. Weekly out-of-school physical activity time was self-reported by mothers. The mean of the children daily energy intake was calculated based on validated food frequency questionnaires (FFQs) about children's diet that were administered twice to the parents or care-givers of children over a 9-month period at 4 years, and over a 9–12-month period at 8 years. The FFQs were composed by 105 items at 4 years and by 46 items at 8 years. To explore the reproducibility of the FFQs, the nutrient and food group intake collected from the both FFQs at each age were compared, while validity was examined by contrasting the nutrient values from the FFQs and the average of three 24-h dietary recalls taken in this period, and also with the concentration of several vitamins in the blood (carotenoids, vitamin D, and α-tocopherol) [29, 30]. Nutrient values and total energy intake were calculated based on the US Department of Agriculture's food composition tables and other published national sources. All questionnaires were conducted face-to-face by trained interviewers. The selection of these variables as potential confoundings was based on previous studies.

Study population

Initially, 494 eligible women agreed to participate and, at birth time, 485 children were part of the study. At 4 years, 453 children continued in the follow-up and 91.4% of them attended to this follow-up visit. At 8 years, 416 children continued in the follow-up and 87.0% of them assisted to the

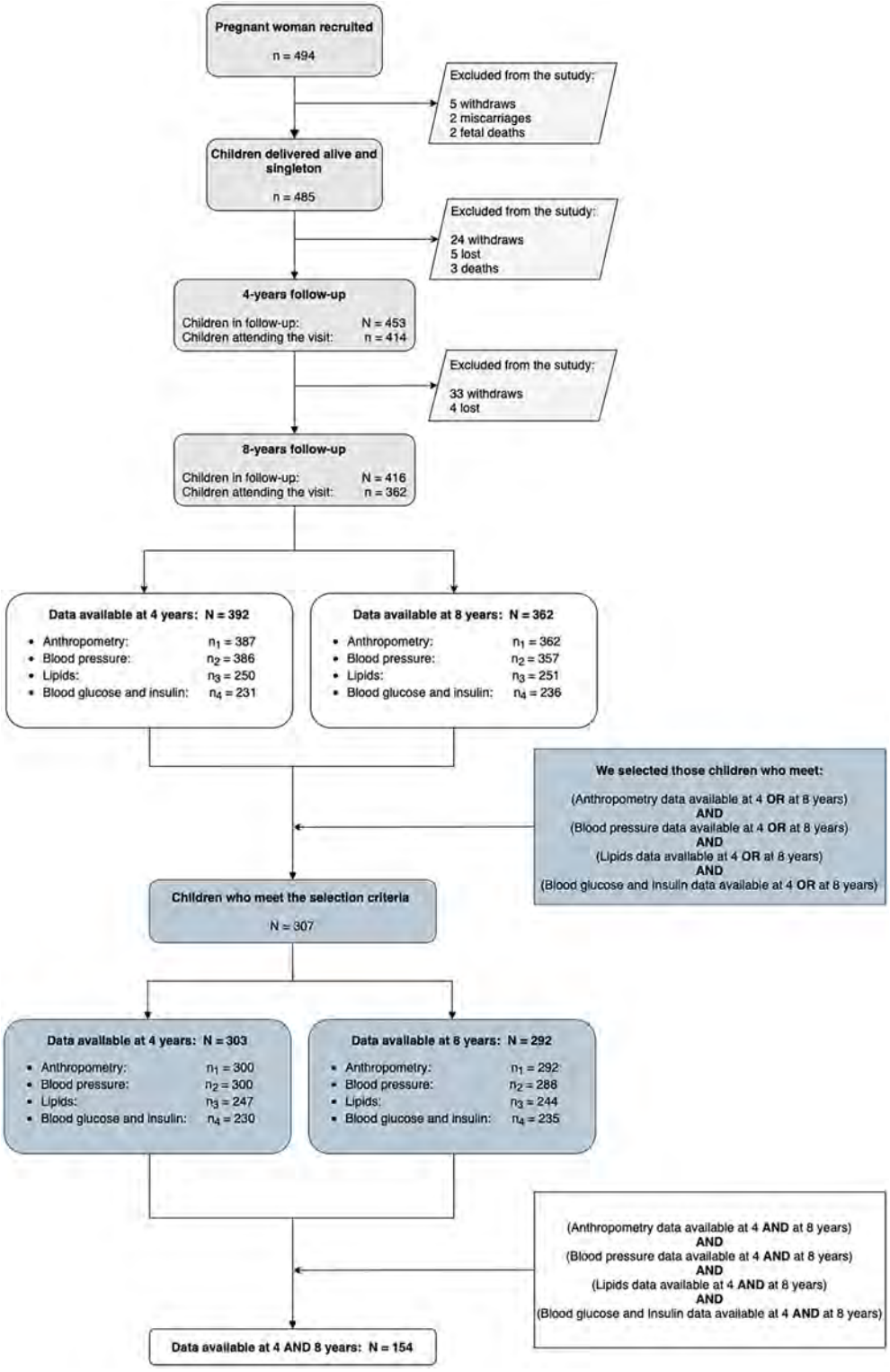


Fig. 1 Flowchart of the study sample

follow-up visit. We limited the study to those 416 children who continued in the study at the 8-year follow-up. Of these 416 children, there were 392 with some measure of anthropometry, blood pressure, lipids, or glucose/insulin at 4 years and 362 at 8 years. Only 154 children had measurements of all variables involved in the study at 4 and 8 years. To optimize the use of the available information, we only excluded from the final sample those children who, for at least one of the cardiovascular-related measures involved in the study, had no data at neither 4 nor 8 years. The final sample was composed by 307 children. Figure 1 shows the flowchart of the study sample and the resulting sample size.

Statistical analysis

Continuous variables were summarized by medians and interquartile ranges, and categorical variables by absolute and relative frequencies.

Crude and adjusted quantile regression models were performed to evaluate the association between the distribution of each cardiovascular-related measure at 8 years as the dependent variable, and the rank transformation — replacing the data by their corresponding ranks — of the same cardiovascular-related measure at 4 years as the independent variable. Quantile regression is a method used to evaluate the effects of exposures on the distribution of a continuous outcome [31]. It allows to assess whether the association between them differs for high-risk subjects (i.e., those at highest quantiles of the outcome) than for average subjects in the outcome distribution. To describe the effect of the independent variable on the cardiovascular-related measure distribution at 8 years, quantile sequence was estimated from 0.1 to 0.9. The models were fully adjusted with potentially confounding variables described in the "Potential confounding factors" section. Models were computed using the *quantreg* R package (version 5.94) [32] and standard errors were estimated using the "*xy*-pair" bootstrap method. To facilitate the interpretation of the regression parameters, the variables resulting from the rank transformation were expressed as percentiles. The reported parameters represent the effect on each quantile of the dependent variable of a 10-unit increase (1-decile) in the independent variable. The analysis was repeated with each cardiovascular-related measure at 8 years as the dependent variable, and the rank transformation of the same cardiovascular as the independent variable, but also including as independent variables the rank variables of other five remaining cardiovascular-related measures, adjusted by covariates (referred as the *complete model* in following sections).

All the analyses were performed after missing value multiple imputation in all the cardiovascular-related measures and the adjustment variables [33]. Under the missing at random (MAR) assumption, that our data suggest that this may be plausible, we applied multivariate imputation by chained equation (MICE) method with fully conditional specification using the mice R package (version 3.14.0) [34]. The results were pooled using Rubin's combination rules [35].

The criteria used to select the final sample resulted in groups of children at 4 and 8 years with a complex structure between them, combining both independent and related measures. To test hypotheses about difference in means or proportions between these groups maintaining the original structure of the data and their relations, we use the general bootstrap algorithm (gBA) for hypothesis testing [36].

All the analyses were conducted using the R statistical software, version 4.2.1 (R Project for Statistical Computing). Statistical significance was considered at p-value < 0.05.

Results

Following the inclusion criteria mentioned in the "Study population" section, a total of 303 children with data on at least one of the cardiovascular-related measures at 4 years were included in the analysis, and a total of 292 children who meet the same criterion were included at 8 years. The merging of these two subsamples results in total sample of 307 children (Table 1). There were 288 children present in both subsamples at the same time, 15 children had only in the 4-year but not in the 8-year subsample, and 4 children had only 8-year data but not 4 years. Due to the small difference between the subjects in each sample, there are no relevant differences in not age-related characteristics. Maternal median age at delivery was 32.9 and 33.1 years at each subsample, respectively, and more than 96% of the mothers were from Spain. Overall, 33.0% of mothers and 66.8% of fathers in the 4-year subsamples, and 32.9% of mothers and 66.2% of fathers in the 8-year subsamples, were overweight or obese (BMI $\geq$ 25 kg/m^2). The average daily energy consumed increased from age 4 to age 8 (a median of 1618 and 1753 cal, respectively; p-value = 0.001), and the number of weekly hours of physical activity outside school was considerably reduced from 4 to 8 years (a median of 11.5 and 3.00 h, respectively; p-value < 0.001). This decrement is explained because at age 4, parents reported an average of 8.3 h per week of playing at home or in playground, and this activity is no longer reported at 8 years. Table 2 contains the summary of anthropometric, serum lipids, blood pressure, and glucose and insulin variables in the 4- and 8-year subsamples.

Additional file 1 shows the percentage of imputed data for each variable over the overall sample. Lipids, glucose, and

Table 1 Characteristics of the study sample, before multiple imputation

	4 years ($n = 303$)	8 years ($n = 292$)	p-Value[a]
Parental characteristics			
Maternal age at delivery (years)	32.9 [30.4, 36.0]	33.1 [30.5, 36.0]	0.815
Maternal origin country			0.999
Spain	292 (96.4%)	282 (96.6%)	
Latin-American	6 (2.0%)	5 (1.7%)	
Europe	4 (1.3%)	4 (1.4%)	
Other	1 (0.3%)	1 (0.3%)	
Maternal level of education			0.996
Primary	47 (15.5%)	44 (15.1%)	
Secondary	140 (46.2%)	134 (45.9%)	
University	116 (38.3%)	114 (39.0%)	
Maternal social class			0.984
Upper I + II	72 (23.8%)	70 (24.0%)	
Middle III	63 (20.9%)	64 (22.0%)	
Low IV + V	167 (55.3%)	157 (54.0%)	
Maternal smoking during pregnancy			0.951
No	244 (84.4%)	233 (84.1%)	
Yes	45 (15.6%)	44 (15.9%)	
Maternal pre-pregnancy BMI (kg/m^2)	23.9 [21.7, 28.1]	23.9 [21.6, 28.1]	0.803
Categorical maternal pre-pregnancy BMI[b]			0.946
Normal	203 (67.0%)	196 (67.1%)	
Overweight	72 (23.8%)	70 (24.0%)	
Obese	28 (9.2%)	26 (8.9%)	
Paternal BMI (kg/m^2)	27.0 [24.6, 30.2]	27.0 [24.6, 30.2]	0.733
Categorical paternal BMI[b]			0.968
Normal	97 (33.2%)	95 (33.8%)	
Overweight	146 (50.0%)	142 (50.5%)	
Obese	46 (16.8%)	44 (15.7%)	
Parental cardiovascular antecedents			
Neither parent has antecedents	264 (87.1%)	253 (86.6%)	0.945
One parent has at least one antecedent	39 (12.9%)	39 (13.4%)	
Both parents have at least one antecedent	0 (0%)	0 (0%)	
Child characteristics			
Sex			0.979
Female	138 (45.5%)	134 (45.9%)	
Male	165 (54.5%)	158 (54.1%)	
Age (years)	4.40 [4.33, 4.53]	8.26 [8.08, 8.38]	< 0.001
Mean daily energy intake (calories)	1618 [1429, 1876]	1753 [1441, 2104]	0.001
Weekly out-of-school physical activity time (h)	11.5 [8.00, 16.0]	3.00 [2.00, 4.75]	< 0.001
Week of gestation at delivery	39.6 [38.6, 40.6]	39.6 [38.5, 40.6]	0.821
Predominant breastfeeding duration (weeks)	10.8 [0.00, 21.6]	10.7 [0.00, 21.6]	0.734
Birth weight (g)	3300 [3010, 3600]	3290 [3000, 3570]	0.879

Characteristics of the 303 children who have data on at least one of the cardiovascular measures involved in the analysis at 4 years of age and of the 292 children who meet the same criteria at 8 years of age. Continuous variables are summarized by medians and interquartile ranges, and categorical variables are summarized by absolute and relative frequencies

[a]The p-values were calculated using the general bootstrap algorithm for hypothesis testing (gBA) mentioned in the "Statistical analysis" section

[b]BMI was categorized according to WHO criteria: normal, BMI < 25 kg/m^2; overweight, 25 kg/m^2 ≤ BMI < 30 kg/m^2; obese, BMI ≥ 30 kg/m^2

Table 2 Cardiovascular-related measures of the study population, before multiple imputation

Measures	4 years ($n=303$)	8 years ($n=292$)
Weight (kg)	18.0 [16.7, 20.0]	29.6 [26.2, 34.3]
Height (cm)	106.0 [103.0, 109.0]	131.0 [127.0, 135.0]
BMI (kg/m^2)	16.0 [15.3, 17.2]	17.3 [15.8, 19.4]
Categorical BMI[a]		
Normal	236 (77.9%)	196 (67.1%)
Overweight	41 (13.5%)	68 (23.3%)
Obese	26 (8.58%)	28 (9.59%)
Waist circumference (cm)	53.5 [50.5, 56.0]	63.5 [58.9, 69.0]
Waist circumference/Height ratio	0.50 [0.48, 0.53]	0.48 [0.46, 0.52]
Triponderal index (kg/m^3)	15.2 [14.4, 16.3]	13.3 [12.1, 14.5]
Systolic blood pressure (mmHg)	99.0 [90.0, 105.0]	107.0 [100.0, 114.0]
Diastolic blood pressure (mmHg)	60.0 [54.0, 66.0]	67.0 [60.8, 72.0]
Categorical blood pressure[b]		
Normal level	218 (72.7%)	104 (63.9%)
Monitoring level	35 (11.7%)	64 (22.2%)
Intervention level	47 (15.7%)	40 (13.9%)
Mean arterial pressure (mmHg)	73.0 [66.7, 77.7]	80.0 [74.2, 84.7]
Total cholesterol (mg/dL)	164.0 [147.0, 178.0]	165.0 [150.0, 183.0]
High-density lipoprotein cholesterol (mg/dL)	57.0 [47.5, 63.0]	68.0 [59.8, 80.0]
Low-density lipoprotein cholesterol (mg/dL)	92.0 [74.0, 106.0]	81.5 [65.0, 98.0]
Triglycerides (mg/dL)	71.0 [55.0, 94.0]	64.0 [47.2, 87.8]
Atherogenic coefficient	1.92 [1.50, 2.46]	1.39 [1.02, 1.78]
Glucose (mg/dL)	86.0 [81.8, 91.0]	86.0 [82.0, 90.0]
Insulin (µU/mL)	5.80 [3.10, 10.60]	7.75 [5.10, 13.6]
HOMA-IR	1.17 [0.64, 2.37]	1.69 [1.05, 2.88]

[a]BMI was categorized according to IOTF criteria

[b]Categorical blood pressure was categorized using age-, sex-, and height-specific thresholds provided by the IDEFICS study. Children were classified in the monitoring or the intervention level if they had SBP or DBP above the age-, sex-, and height-specific corresponding threshold

insulin measurements had the highest percentage of missing data (ranging from 19.2 to 25.1% at 4 years and from 19.9 to 23.1% at 8 years). Missing data in anthropometric and blood pressure measurements ranged from 2.0 to 6.2%.

Triglycerides

Figure 2 shows the estimated quantile regression parameters for each rank cardiovascular-related measure at 4 years, on the distribution of the same cardiovascular-related measure at 8 years, for all quantiles. We observe a positive association between TG rank at 4 years and TG distribution at 8 years above 0.5 quantile. The magnitude of the association was stronger in the upper part of the distribution: 1-decile increase in child's rank at 4 years related to an increase of 2.28 mg/dL (95% CI: 0.13, 4.43) in the 0.6-TG quantile at 8 years compared to an increase of 5.82 mg/dL (95% CI: 1.00, 10.65) in the 0.9-TG quantile at 8 years (Additional file 2). In the complete model (Fig. 3), 1-decile increase at 4 years has an increase effect of 1.77 mg/dL (95% CI: − 0.68, 4.23) in the 0.6 quantile compared to a 2.47-mg/dL (95% CI: − 0.88, 5.83) increase effect in the 0.75 quantile (Additional file 3).

High-density lipoprotein cholesterol

The association between HDL-c rank at 4 years and HDL-c distribution at 8 years was positive across all quantiles (Fig. 2). A gradual increase was observed in the upper part of the distribution with an increase of 2.68 mg/dL (95% CI: 1.75, 3.62) for the 0.6 HDL-c quantile at 8 years and of 2.93 mg/dL (95% CI: 1.98, 3.87) for the 0.9 HDL-c quantile at 8 years (Additional file 2). Associations in the complete model (Fig. 3) were lower than in the individual model but the overall trends were similar.

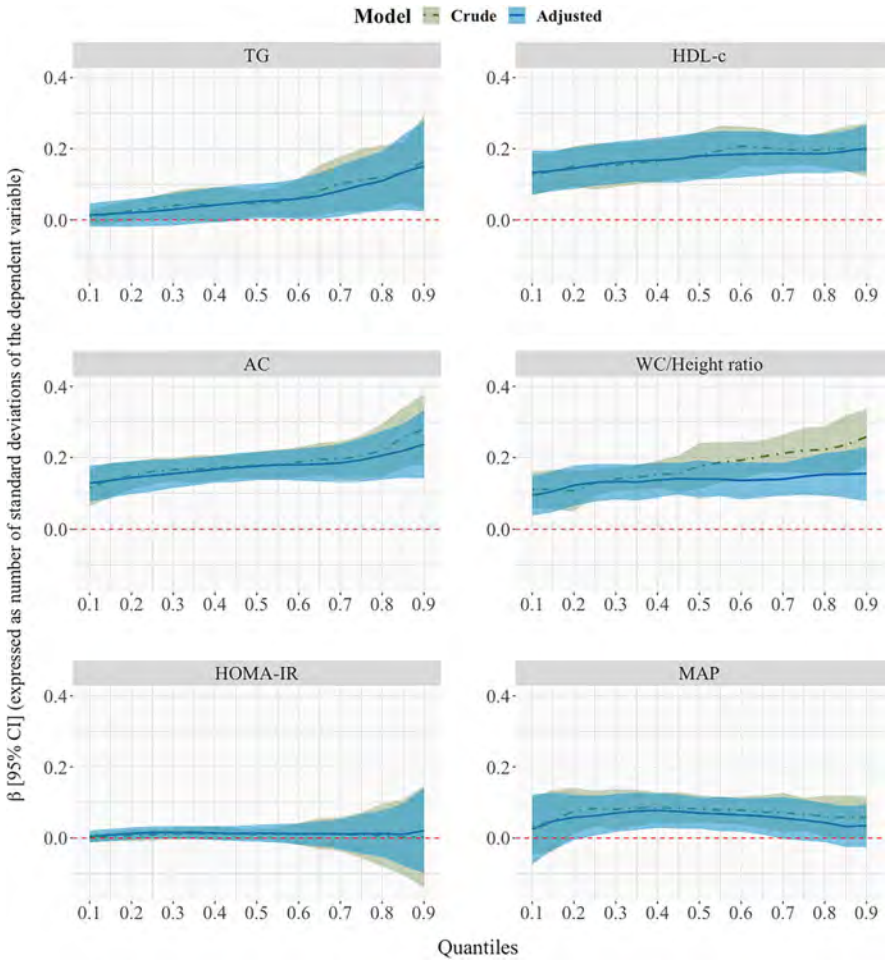


Fig. 2 Quantile regression models with cardiovascular-related measure at 8 years as dependent variable and the rank variable of the corresponding cardiovascular-related measure at 4 years as the independent variable, for the quantiles between 0.1 to 0.9, with increments of 0.05, adjusted for maternal age at delivery, maternal level of education, maternal social class, maternal smoking during pregnancy, maternal pre-pregnancy body mass index, paternal body mass index, parental cardiovascular antecedents, child sex, child mean daily energy intake at 4 and 8 years, child weekly out-of-school physical activity time at 4 and 8 years, week of gestation at delivery, weeks of predominant breastfeeding, and child height at 4 and 8 years. Coefficient estimated are calculated with the independent variables in terms of percentiles and they represent the effect on the dependent variable for each 1-decile increase in the independent variable. They are expressed in terms of number of standard deviations of the dependent variable to homogenize the Y-axis scales

Atherogenic coefficient

The association between the AC rank at 4 years and AC distribution at 8 years also was positive (Fig. 2). The size of increase was greater at the highest part of the AC distribution (an increase of 0.11; 95% CI: 0.09, 0.14) in the 0.6 quantile vs. an effect of 0.15 (95% CI: 0.09, 0.21) in the 0.9 quantile (Additional file 2). Results were similar in the complete model (Fig. 3; Additional file 3).

Waist circumference to height ratio

A positive association was observed between WC/Height rank at 4 years and WC/Height distribution at 8 years (Fig. 2). The crude model shows a positive trend in the effect size as the quantile increases, evidenced by the positive slope of the plot in Fig. 2 (effect of 0.010 (95% CI: 0.006, 0.013) on the 0.6 WC/Height quantile at 8 years vs. an effect of 0.014 (95% CI: 0.010, 0.018) on the 0.9 WC/Height quantile at 8 years; Additional file 2). In the adjusted model, the effect sizes were smaller and generally constant in all quantiles (increase of 0.007 (95% CI: 0.005, 0.010) in the 0.6 WC/Height quantile vs. an increase of 0.008 (95% CI: 0.004, 0.012) in the 0.9 WC/Height quantile at 8 years; Additional file 2). This difference in the trend of the crude and adjusted model is mainly due to the adjustment for maternal BMI and educational level. The complete model (Fig. 3) produced similar results.

HOMA-IR

No clear association was found between HOMA-IR rank at 4 years and HOMA-IR distribution at 8 years (Fig. 2) (effect size of 0.037 (95% CI: − 0.058, 0.131) in the 0.6 HOMA-IR quantile vs. 0.067 (95% CI: − 0.312, 0.445) in the 0.9 HOMA-IR quantile at 8 years; Additional file 2). This also was seen in the complete model (Fig. 3).

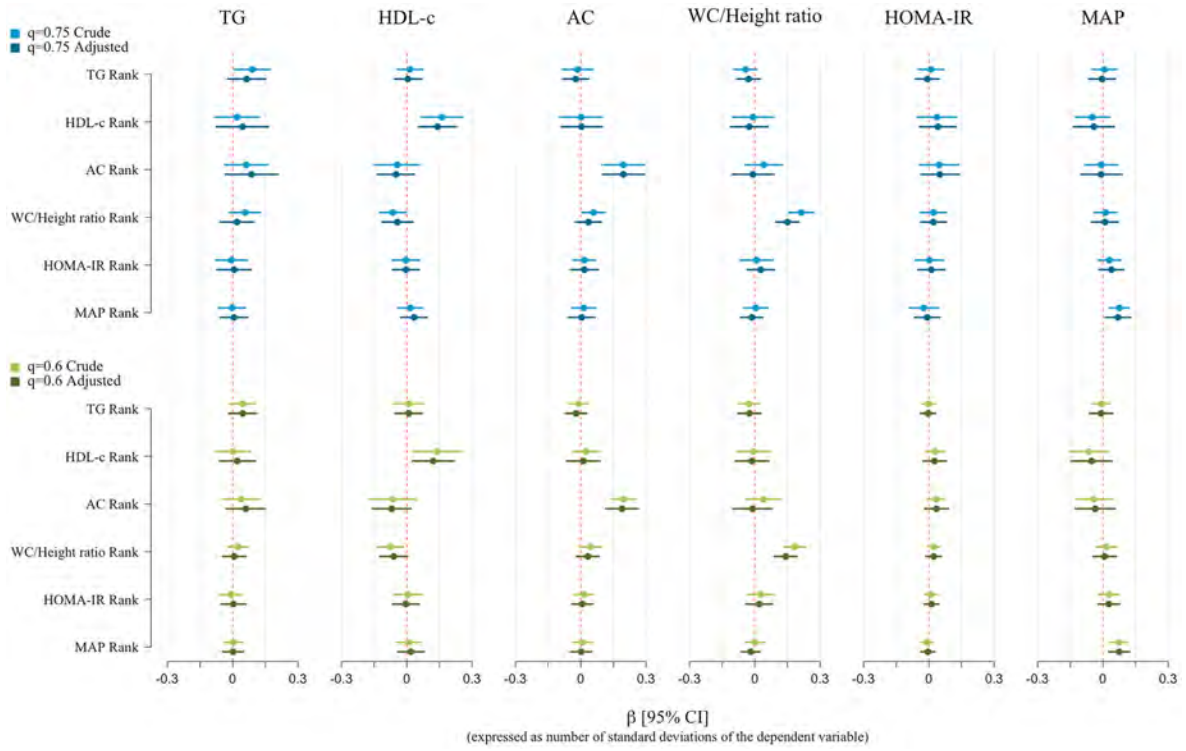


Fig. 3 Quantile regression models with each cardiovascular-related measure at 8 years as dependent variable and the rank of all the cardiovascular-related measures at 4 years as the independent variables, for the quantiles 0.60 and 0.75, adjusted for maternal age at delivery, maternal level of education, maternal social class, maternal smoking during pregnancy, maternal prepregnancy body mass index, paternal body mass index, parental cardiovascular antecedents, child sex, child mean daily energy intake at 4 and 8 years, child weekly out-of-school physical activity time at 4 and 8 years, week of gestation at delivery, weeks of predominant breastfeeding, and child height at 4 and 8 years. Coefficient estimated are calculated with the independent variables in terms of percentiles and they represent the effect on the dependent variable for each 1-decile increase in the independent variable. They are expressed in terms of number of standard deviations of the dependent variable to homogenize the X-axis scales

Mean arterial pressure

The association of MAP rank at 4 years on MAP distribution at 8 years was positive, but only statistically significant between 0.3 and 0.6 quantiles (Fig. 2). Similar results were observed in the complete model (Fig. 3).

Discussion

This study found a positive association between the relative position of children at 4 years in the HDL-c, AC, and WC/Height distributions and all the quantiles of the same variable at 8 years. For TG distribution, it was found a positive association between the relative position at 4 years and the quantiles above 0.5 at 8 years, but which is not observed in the model adjusted for the rest of the cardiovascular-related measures. The stronger associations in the upper parts of the distribution in terms of standard deviations of each variable were found for HDL-c and AC outcomes. For AC, the more extreme the children's values at 8 years, the greater the effect of the association. This trend is also observed for TG, although the effect is not statistically significant. No conclusive association was found for either HOMA-IR or MAP outcomes.

Our findings for HDL-c and AC suggest serum lipid track among children between the ages of 4 and 8 years. These results are in line with those found in children from two different Japan rural areas: one area reported tracking of serum lipids (specifically in T-c, HDL-c, and AC) between 8 and 12 years [37] and the other reported strong T-c tracking in children aged 6–7 after 9 years of follow-up [38]. Previously, The Muscatine Study observed T-c and TG tracking in children between 5 and 12 years after follow-ups of 2, 4, and 6 years [39]. The Bogalusa Heart Study showed tracking of serum lipids in 5-year-old children after a follow-up of 9 years [40]. These studies categorized serum lipids

and evaluated which percentage of children remained in the highest category after the follow-up, which does not allow to observe differences in tracking within the highest-risk category itself. Using our quantile regression approach, we were able to observe that for AC the magnitude of tracking was stronger as the relative position at 8 years of age increases, showing the possible difficulty that children with abnormal values at 4 might have in normalizing them in future years. This is remarkable because AC (the ratio of non-HDL-c to HDL [25]) is clearly related to higher risk of CVDs in adulthood [41, 42]. Whereas higher HDL-c had similar tracking levels in all quantiles and higher HDL-c has unclear association with CVD risk [43].

Tracking of measures related to obesity have been widely studied in childhood and adolescence. The majority of studies use BMI as a marker of obesity [37, 44–48]. Increasingly studies are focusing on other anthropometric measures such as WC, WC/Height ratio, or skinfolds [49–51] and report the presence of tracking, consistent with our findings, in a variety range of ages but mostly between childhood and adolescence. Some of these studies find tracking among the same age ranges that were considered in this study [46, 48, 51], although the evaluation approach makes difficult to compare effect magnitudes. Hayes et al. [46] evaluated tracking of BMI in seven follow-up visits between 2–3 and 16–17 years and reported that the tracking magnitude was lower between 2 and 7 years than at later ages. A meta-regression analysis reported stronger BMI tracking after the age of 7 [52]. This suggests steeper tracking of central obesity than observed in our study for children of older ages. In the crude model, we observed higher tracking in children with a high-risk position in the distribution at 8 years. But when we adjusted for maternal pre-pregnancy BMI, and maternal educational level, the association became very similar across all the quantiles of the distribution: higher maternal BMI is associated with a tracking increasing effect in the highest quantiles, and higher maternal educational level is associated with a decreasing effect in the highest quantiles. Other studies reported similar findings with maternal pre-pregnancy BMI and measures related with obesity, and with blood pressure as well [53]. We only have observed this effect in measures related with obesity. Several studies have reported the influence of socioeconomical inequalities in BMI tracking, using parental educational level [54], parental socioeconomic position [46], or a combination of parental educational level, household income, and occupation [48]. We have observed this effect with maternal educational level but not with socioeconomic status.

Neither the MAP nor HOMA-IR showed a relevant association between the relative position of children at age 4 years and the relative position at age 8 years in the upper part of the distribution. As with serum lipids, there are few studies that analyzed the tracking of these measures in childhood and adolescence, rather than in adulthood. In our study, MAP is used as a blood pressure index to try to capture the effect of systolic and diastolic pressure using a single measurement. However, we also examined SBP and DBP separately (data not shown), yielding identical findings and conclusions as those obtained using MAP. Existing studies predominantly employ SBP and DBP and most of them report weak or poor blood pressure tracking [37, 39, 55], in line with the results we obtained here. One exception is the study by Sánchez-Bayle et al. [56] that reported considerably higher level of blood pressure tracking in a school-aged population. For measures related to insulin resistance, two studies examined tracking between 8 and 21 years [57], and between 10 and 17 years [58], with disparate results. Joshi et al. [57] reported moderate tracking of the HOMA-IR but no tracking of fasting insulin or glucose measures, while Wang et al. [58] reported no tracking in the HOMA-IR but tracking of fasting glucose. It should be pointed that in longitudinal studies a change in the behavior of the subjects can arise based on the knowledge of the results and the recommendations given in this regard, influenciating the tracking effect. Despite the limitation this may imply, these recommendations make it possible to reverse trends that would be more difficult in adulthood.

Numerous studies have reported interrelations between cardiovascular risk measures including markers of obesity, blood pressure, insulin resistance, and lipids, although these relations and the pathways explaining them are not clear yet [59, 60]. Therefore, in our study, the analyses were repeated including the cardiovascular-related measures as independent variables altogether in our models to observe whether any were acting as confounding factors for each other. The results of the individual analysis were remained, although with a general attenuation of the magnitude of the effects.

To our knowledge, this is among the first studies to address the analysis of cardiovascular measure tracking in children using quantile regression. Only one prior study conducted the analysis of BMI tracking between childhood and adulthood using this statistical technique [54]. This approach allows the introduction of several adjustment variables and exploration of the effects of different variables at the same time, as well as avoids using thresholds, always controversial in pediatric ages. Among other strengths of this study is that by using quantile regression and the independent variables in terms of their ranges allows to control for age-dependent variation in the effects observed for measures as HDL-c or TG.

This study has also some limitations. The sample size is moderate/small, with the consequent loss of power in the analysis. This is an exploratory study, in which numerous hypotheses are tested (different quantiles and different results), so that multi-testing problems arise, making problematic to calculate the statistical power of the study. Some

of the variables were self-reported by the children's parents. Other variables that have shown to influence the associations evaluated here have not been included such as children diet quality [61], maternal diet quality during pregnancy and breastfeeding [62], or pre-eclampsia [63]. It should also be noted that blood samples were collected under non-fasting conditions. It is unclear to what extent glucose and insulin levels might be biased due to prior caloric intake [64]. However, in non-diabetic subjects, the HOMA-IR would not be expected to show large variations. If blood glucose is higher due to previous intake, insulin also raises its levels, and therefore, the ratio between them will be similar. No blood glucose levels suggestive of diabetes have been detected in our sample, so we expect the HOMA-IR to be similar to that under fasting conditions. Regarding to lipids, the use of non-fasting samples is already recommended, except in unusual cases that do not apply to our sample [65–67]. On the other hand, most studies evaluate tracking between longer periods, and extreme values of cardiovascular-related measurements in childhood have shown to have an age-dependent impact on adult cardiovascular health, being predictive of subclinical atherosclerosis from the age of 9 [68]. Yet, it is still relevant to know the tracking of these measures throughout early childhood and adolescence more as a continuum, as long-term effect of childhood exposures on adult health is likely cumulative [11].

Conclusions

Our study found tracking between 4 and 8 years of age at the highest quantiles of the distribution of cardiovascular-related measures established as adult markers of dyslipidemia and central obesity (specifically for HDL-c and AC for dyslipidemia and WC/Height ratio for central obesity). The results indicated that for AC distribution tracking appears to be stronger at higher quantiles, suggesting the difficulty of normalizing their extreme values. These findings can help determine what cardiovascular-related measures could be the targets of screening and monitoring in children.

Abbreviations AC: Atherogenic coefficient; BMI: Body mass index; CVDs: Cardiovascular diseases; DBP: Diastolic blood pressure; gBA: General bootstrap algorithm; HDL-c: High-density cholesterol; HOMA-IR: Homeostatic model assessment of insulin resistance; LDL-c: Low-density cholesterol; MAP: Mean arterial pressure; MAR: Missing at random; MICE: Multivariate imputation by chained equations; SBP: Systolic blood pressure; T-c: Total cholesterol; TG: Triglycerides; WC: Waist circumference; WC/Height: Waist circumference to height ratio

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00431-023-05051-8.

Acknowledgements The authors would particularly like to thank all the participants and the families for their valuable collaboration. They also like to thank the medical board, the gynecology and pediatrics services of Hospital San Agustín (Avilés, Asturias), and the health center of Las Vegas (Corvera, Asturias) for their generous implication in the project.

Author contributions Rocío Fernández-Iglesias: formal analysis, data curation, investigation, methodology, visualization, writing—original draft, writing—review and editing. Pablo Martínez-Camblor: formal analysis, conceptualization, investigation, methodology, supervision, visualization, writing—review and editing. Ana Fernández-Somoano: conceptualization, investigation, methodology, supervision, writing—review and editing. Cristina Rodríguez-Dehli: data collection, writing—review and editing. Rafael Venta-Obaya: resources, writing—review and editing. Margaret R. Karagas: conceptualization, writing—review and editing. Ana Fernández-Somoano: conceptualization, investigation, methodology, supervision, writing—review and editing. Adonina Tardón: conceptualization, supervision, project administration, funding acquisition, writing—review and editing. Isolina Riaño-Galán: conceptualization, data collection, writing—review and editing. All authors have read an agreed to the publisher version of the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This study was supported by grants from CIBERESP (PhD-employment-contract and fellowship for short stays abroad-2022), ISCIII: PI04/2018, PI09/02311, PI13/02429, and PI18/00909 co-funded by the European Regional Development Fund (FEDER), "A way to make Europe"/ "Investing in your future"; Fundación Cajastur; and Universidad de Oviedo.

Availability of data and materials The data and computing code are available for replication from the corresponding author on reasonable request.

Additional information The CHAMP statement and The STROBE guideline were used to ensure the reporting of this observational study [69, 70].

Declarations

Ethics approval and consent to participate The study was conducted to conform to the principles of the Declaration of Helsinki and its protocol was approved by the Asturias Regional Ethics Committee. Informed consent was obtained from every participant woman and, in such case, her partner.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Abbafati C, Abbas KM, Abbasi-Kangevari M, et al (2020) Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 396:1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2
2. Genovesi S, Giussani M, Orlando A, et al (2019) Prevention of cardiovascular diseases in children and adolescents. High Blood Press Cardiovasc Prev 26:191–197. https://doi.org/10.1007/s40292-019-00316-6
3. Vetter VL, Covington TM, Dugan NP, et al (2015) Cardiovascular deaths in children: General overview from the National Center for the Review and Prevention of Child Deaths. Am Heart J 169:426-437.e23. https://doi.org/10.1016/j.ahj.2014.11.014
4. Zanchetti A, Bond MG, Hennig M, et al (1998) Risk factors associated with alterations in carotid intima—media thickness in hypertension: baseline data from the European Lacidipine Study on Atherosclerosis. J Hypertens 16:949–61. https://doi.org/10.1097/00004872-199816070-00008
5. Milei J, Ottaviani G, Lavezzi AM, et al (2008) Perinatal and infant early atherosclerotic coronary lesions. Can J Cardiol 24:137–41. https://doi.org/10.1016/s0828-282x(08)70570-1
6. Mcgill HC, Mcmahan ; C Alex, Herderick EE, et al (2000) Origin of atherosclerosis in childhood and adolescence. Am J Clin Nutr 72:1307S-1315S. https://doi.org/10.1093/ajcn/72.5.1307s
7. Erald G, Erenson SB, Rinivasan ARS, et al (1998) Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. N Engl J Med 338:1650–1656. https://doi.org/10.1056/NEJM199806043382302
8. De Ferranti SD, Steinberger J, Ameduri R, et al (2019) Cardiovascular risk reduction in high-risk pediatric patients: A scientific statement from the American Heart Sssociation. Circulation 139:E603–E634. https://doi.org/10.1161/CIR.0000000000000618
9. Pool LR, Aguayo L, Brzezinski M, et al (2021) Childhood risk factors and adulthood cardiovascular disease: A systematic review. J Pediatr 232:118-126.e23. https://doi.org/10.1016/j.jpeds.2021.01.053
10. Jacobs DR, Woo JG, Sinaiko AR, et al (2022) Childhood cardiovascular risk factors and adult cardiovascular events. N Engl J Med 386:1877–1888. https://doi.org/10.1056/NEJMoa2109191
11. Ben-Shlomo Y, Mishra G, Kuh D (2014) Life course epidemiology. Handb Epidemiol Second Ed 1521–1549. https://doi.org/10.1007/978-0-387-09834-0_56
12. Lurbe E, Agabiti-Rosei E, Cruickshank JK, et al (2016) 2016 European Society of Hypertension guidelines for the management of high blood pressure in children and adolescents. J Hypertens 34:1887–1920. https://doi.org/10.1097/HJH.0000000000001039
13. Reisinger C, Nkeh-Chungag BN, Fredriksen PM, Goswami N (2020) The prevalence of pediatric metabolic syndrome—a critical look on the discrepancies between definitions and its clinical importance. Int J Obes 45:12–24. https://doi.org/10.1038/s41366-020-00713-1
14. Ahrens W, Moreno L, Mårild S, et al (2014) Metabolic syndrome in young children: Definitions and results of the IDEFICS study. Int J Obes 38:S4–S14. https://doi.org/10.1038/ijo.2014.130
15. Foulkes MA, Davis CE (1981) An index of rracking for longitudinal data. Biometrics 37:439-446. https://doi.org/10.2307/2530557
16. Twisk JWR (2003) The problem of evaluating the magnitude of tracking coefficients. Eur J Epidemiol 18:1025–1026. https://doi.org/10.1023/a:1026161919170
17. Guxens M, Ballester F, Espada M, et al (2012) Cohort profile: The INMA-INfancia y Medio Ambiente-(environment and childhood) project. Int J Epidemiol 41:930–940. https://doi.org/10.1093/ije/dyr054
18. Koenker R (2005) Quantile regression. Cambridge University Press, Quantile Regres. Cambridge
19. Waldmann E (2018) Quantile regression: A short story on how and why. Stat Modelling 18:203–218. https://doi.org/10.1177/1471082X18759142
20. Fernández-Somoano A, Estarlich M, Ballester F, et al (2011) Outdoor NO2 and benzene exposure in the INMA (Environment and Childhood) Asturias cohort (Spain). Atmos Environ 45:5240–5246. https://doi.org/10.1016/j.atmosenv.2011.02.010
21. Fernández-Somoano A, Tardon A (2014) Socioeconomic status and exposure to outdoor NO2 and benzene in the Asturias INMA birth cohort, Spain. J Epidemiol Community Health 68:29-36. https://doi.org/10.1136/JECH-2013-202722
22. Sociedad Asturiana de Estudios Económicos e Industriales (2021) Actualización demográfica del Mapa Sanitario de Asturias. Available from: https://www.astursalud.es/documents/35439/37169/Mapa_Padron_2021_Libro.pdf/0c9cf850-f5d4-9292-3fb1-9fa26d449276
23. Mokha JS, Srinivasan SR, DasMahapatra P, et al (2010) Utility of waist-to-height ratio in assessing the status of central obesity and related cardiometabolic risk profile among normal weight and overweight/obese children: The Bogalusa Heart Study. BMC Pediatr 10:1–7. https://doi.org/10.1186/1471-2431-10-73
24. Sesso HD, Stampfer MJ, Rosner B, et al (2000) Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in Men. Hypertens (Dallas, Tex 1979) 36:801–807. https://doi.org/10.1161/01.HYP.36.5.801
25. Çelik E, Çora AR, Karadem KB (2021) The effect of untraditional lipid parameters in the development of coronary artery disease: atherogenic index of plasma, atherogenic coefficient and lipoprotein combined index. J Saudi Hear Assoc 33:244–250. https://doi.org/10.37616/2212-5043.1266
26. Yajnik CS, Katre PA, Joshi SM, et al (2015) Higher glucose, insulin and insulin resistance (HOMA-IR) in childhood predict adverse cardiovascular risk in early adulthood: the Pune Children's Study. Diabetologia 58:1626–1636. https://doi.org/10.1007/s00125-015-3602-z
27. DeMers D, Wachs D (2022) Physiology, mean arterial pressure. StatPearls Publishing, StatPearls
28. Domingo-Salvany A, Regidor E, Alonso J, Alvarez-Dardet C (2000) Una propuesta de medida de la clase social. Atención Primaria 25:350–363. https://doi.org/10.1016/S0212-6567(00)78518-0
29. Vioque J, Gimenez-Monzo D, Navarrete-Muñoz EM, et al (2016) Reproducibility and validity of a food frequency questionnaire designed to assess diet in children aged 4-5 years. PLoS One 11:1–17. https://doi.org/10.1371/JOURNAL.PONE.0167338
30. Vioque J, Garcia-De-La-Hera M, Gonzalez-Palacios S, et al (2019) Reproducibility and validity of a short food frequency questionnaire for dietary assessment in children aged 7-9 years in Spain. Nutrients 11:1–15. https://doi.org/10.3390/NU11040933
31. Hao L, Naiman DQ (2007) Quantile regression. Quantile Regres. Thousand Oaks, Calif: Sage Publications
32. Koenker R (2022) Package 'quantreg.' Quantile Regres 1–349
33. Sterne JAC, White IR, Carlin JB, et al (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338:b2393. https://doi.org/10.1136/bmj.b2393
34. van Buuren S, Groothuis-Oudshoorn K (2011) MICE: multivariate imputation by chained equations. J Stat Softw 45:1–67. https://doi.org/10.18637//jss.v045.i03
35. Heymans MW, Eekhout I (2019) Applied Missing Data Analysis with SPSS and (R)Studio. https://bookdown.org/mwheymans/bookmi/. Accessed 6 Nov 2022
36. Martínez-Camblor P, Corral N (2012) A general bootstrap algorithm for hypothesis testing. J Stat Plan Inference 142:589–600. https://doi.org/10.1016/j.jspi.2011.09.003

37. Tan F, Okamoto M, Suyama A, Miyamoto T (2010) Tracking of cardiovascular risk factors and a cohort study on hyperlipidemia in rural schoolchildren in Japan. J Epidemiol 10:255–261. https://doi.org/10.2188/jea.10.255
38. Osawa E, Asakura K, Okamura T, et al (2022) Tracking pattern of total cholesterol levels from childhood to adolescence in Japan. J Atheroscler Thromb 29:38–49. https://doi.org/10.5551/jat.59790
39. Clarke WR, Schrott HG, Leaverton PE, et al (1978) Tracking of blood lipids and blood pressures in school age children: The Muscatine study. Circulation 58:626–634. https://doi.org/10.1161/01.CIR.58.4.626
40. Freedman DS, Shear CL, Srinivasan SR, et al (1985) Tracking of serum lipids and lipoproteins in children over an 8-year period: the Bogalusa heart study. Prev Med (Baltim) 14:203–216. https://doi.org/10.1016/0091-7435(85)90036-2
41. Brunner FJ, Waldeyer C, Ojeda F, et al (2019) Application of non-HDL cholesterol for population-based cardiovascular risk stratification: results from the Multinational Cardiovascular Risk Consortium. Lancet (London, England) 394:2173–2183. https://doi.org/10.1016/S0140-6736(19)32519-X
42. McBride P (2008) Triglycerides and risk for coronary artery disease. Curr Atheroscler Rep 10:386–390. https://doi.org/10.1007/s11883-008-0060-9
43. Casula M, Colpani O, Xie S, et al (2021) HDL in atherosclerotic cardiovascular disease: in search of a role. Cells 10:1–17. https://doi.org/10.3390/cells10081869
44. Freedman DS, Lawman HG, Galuska DA, et al (2018) Tracking and variability in childhood levels of BMI: the Bogalusa Heart Study. Obesity 26:1197–1202. https://doi.org/10.1002/OBY.22199
45. De Wilde JA, Middelkoop BJC, Verkerk PH (2018) Tracking of thinness and overweight in children of Dutch, Turkish, Moroccan and South Asian descent from 3 through 15 years of age: a historical cohort study. Int J Obes 42:1230–1238. https://doi.org/10.1038/s41366-018-0135-9
46. Hayes AJ, Carrello JP, Kelly PJ, et al (2021) Looking backwards and forwards: tracking and persistence of weight status between early childhood and adolescence. Int J Obes 45:870–878. https://doi.org/10.1038/s41366-021-00751-3
47. Toselli S, Brasili P, Di Michele R (2013) Tracking of weight status and body fatness in Italian children. Eat Weight Disord 18:383–388. https://doi.org/10.1007/S40519-013-0074-3
48. Wheaton N, Millar L, Allender S, Nichols M (2015) The stability of weight status through the early to middle childhood years in Australia: a longitudinal study. BMJ Open 5:1–9. https://doi.org/10.1136/BMJOPEN-2014-006963
49. Zvonar M, Štefan L, Kasović M, Piler P (2022) Tracking of anthropometric characteristics from childhood to adolescence: an 8-year follow-up findings from the Czech ELSPAC study. BMC Public Health 22:1–8. https://doi.org/10.1186/s12889-022-13178-w
50. Ochiai H, Shirasawa T, Nishimura R, et al (2020) Changes in overweight/obesity and central obesity status from preadolescence to adolescence: a longitudinal study among schoolchildren in Japan. BMC Public Health 20:1–7. https://doi.org/10.1186/s12889-020-8343-3
51. Ortiz-Marrón H, Ortiz-Pinto MA, Cabañas Pujadas G, et al (2022) Tracking and risk of abdominal and general obesity in children between 4 and 9 years of age. The Longitudinal Childhood Obesity Study (ELOIN). BMC Pediatr 22:1–11. https://doi.org/10.1186/s12887-022-03266-6
52. Bayer O, Krüger H, Von Kries R, Toschke AM (2011) Factors associated with tracking of BMI: a meta-regression analysis on BMI tracking. Obesity 19:1069–1076. https://doi.org/10.1038/oby.2010.250
53. Cox B, Luyten LJ, Dockx Y, et al (2020) Association Between Maternal Prepregnancy Body Mass Index and Anthropometric Parameters, Blood Pressure, and Retinal Microvasculature in Children Age 4 to 6 Years. JAMA Netw open 3:e204662–e204662. https://doi.org/10.1001/jamanetworkopen.2020.4662
54. Norris T, Bann D, Hardy R, Johnson W (2020) Socioeconomic inequalities in childhood-to-adulthood BMI tracking in three British birth cohorts. Int J Obes 44:388–398. https://doi.org/10.1038/s41366-019-0387-z
55. Sarganas G, Schaffrath Rosario A, Niessner C, et al (2018) Tracking of blood pressure in children and adolescents in Germany in the context of risk factors for hypertension. Int J Hypertens 2018:1–10. https://doi.org/10.1155/2018/8429891
56. Sánchez-Bayle M, Muñoz-Fernández MT, Gonzzález-Requejo A (1999) A longitudinal study of blood pressure in Spanish schoolchildren. Res Rep U S Nav Sch Aviat Med 81:169–171. https://doi.org/10.1161/01.cir.26.4.530
57. Joshi SM, Katre PA, Kumaran K, et al (2014) Tracking of cardiovascular risk factors from childhood to young adulthood - The Pune Children's Study. Int J Cardiol 175:176–178. https://doi.org/10.1016/j.ijcard.2014.04.105
58. Wang G, Arguelles L, Liu R, et al (2011) Tracking blood glucose and predicting prediabetes in chinese children and adolescents: a prospective twin study. PLoS One 6:e28573. https://doi.org/10.1371/journal.pone.0028573
59. Alaqil AI, Petushek EJ, Gautam YR, et al (2022) Determining independence and associations among various cardiovascular disease risk factors in 9-12 years old school-children: a cross sectional study. BMC Public Health 22:1–9. https://doi.org/10.1186/s12889-022-14035-6
60. Al-Hamad D, Raman V (2017) Metabolic syndrome in children and adolescents. Transl Pediatr 6:397–407. https://doi.org/10.21037/TP.2017.10.02
61. Siddiqui NZ, Nguyen AN, Santos S, Voortman T (2022) Diet quality and cardiometabolic health in childhood: the Generation R Study. Eur J Nutr 61:729–736. https://doi.org/10.1007/s00394-021-02673-2
62. Castillo P, Kuda O, Kopecky J, et al (2022) Reverting to a healthy diet during lactation normalizes maternal milk lipid content of diet-induced obese rats and prevents early alterations in the plasma lipidome of the offspring. Mol Nutr Food Res 66:1–12. https://doi.org/10.1002/mnfr.202200204
63. Wojczakowski W, Kimber-Trojnar Ż, Dziwisz F, et al (2021) Preeclampsia and cardiovascular risk for offspring. J Clin Med 10:1–17. https://doi.org/10.3390/jcm10143154
64. Moebus S, Göres L, Lösch C, Jöckel KH (2011) Impact of time since last caloric intake on blood glucose levels. Eur J Epidemiol 26:719–728. https://doi.org/10.1007/s10654-011-9608-z
65. Nordestgaard BG, Langsted A, Mora S, et al (2016) Fasting is not routinely required for determination of a lipid profile: clinical and laboratory implications including flagging at desirable concentration cut-points—a joint consensus statement from the European Atherosclerosis Society and European Federa. Eur Heart J 37:1944–1958. https://doi.org/10.1093/EURHEARTJ/EHW152
66. Pearson GJ, Thanassoulis G, Anderson TJ, et al (2021) 2021 Canadian cardiovascular society guidelines for the management of dyslipidemia for the prevention of cardiovascular disease in adults. Can J Cardiol 37:1129–1150. https://doi.org/10.1016/j.cjca.2021.03.016
67. Grundy SM, Stone NJ, Bailey AL, et al (2019) 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 73:3168–3209. https://doi.org/10.1016/j.jacc.2018.11.002
68. Juonala M, Magnussen CG, Venn A, et al (2010) Influence of age on associations between childhood risk factors and carotid

intima-media thickness in adulthood. Circulation 122:2514–2520. https://doi.org/10.1161/CIRCULATIONAHA.110.966465
69. Mansournia MA, Collins GS, Nielsen RO, et al (2021) A Checklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration. Br J Sports Med 55:1009–1017. https://doi.org/10.1136/BJSPORTS-2020-103652
70. Vandenbroucke JP, Von Elm E, Altman DG, et al (2007) Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Med 4:1628–1654. https://doi.org/10.1371/journal.pmed.0040297

Authors and Affiliations

Rocío Fernández-Iglesias[1,2,3] · Pablo Martinez-Camblor[4,5] · Ana Fernández-Somoano[1,2,3] · Cristina Rodríguez-Dehli[3,6] · Rafael Venta-Obaya[7,8] · Margaret R. Karagas[9] · Adonina Tardón[1,2,3] · Isolina Riaño-Galán[1,3,10]

1 Spanish Consortium for Research On Epidemiology and Public Health (CIBERESP), Monforte de Lemos Avenue, 3-5, 28029 Madrid, Spain

2 Unit of Molecular Cancer Epidemiology, University Institute of Oncology of the Principality of Asturias (IUOPA), Department of Medicine, University of Oviedo, Julian Clavería Street S/N, 33006 Oviedo, Asturias, Spain

3 Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Roma Avenue S/N, 33001 Oviedo, Asturias, Spain

4 Biomedical Data Science Department, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

5 Faculty of Health Sciences, Universidad Autonoma de Chile, 7500912 Providencia, Chile

6 Servicio de Pediatría, Hospital San Agustín, Heros Street, 4, 33410 Avilés, Asturias, Spain

7 Servicio de Bioquímica, Hospital San Agustín, Heros Street, 4, 33410 Avilés, Asturias, Spain

8 Departamento de Bioquímica y Biología Molecular, University of Oviedo, Fernando Bongera Street, S/N, 33006 Oviedo, Asturias, Spain

9 Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

10 Endocrinología Pediátrica, Servicio de Pediatría, HUCA, Roma Avenue S/N, 33001 Oviedo, Asturias, Spain

3.3. Article III: Statistical considerations for analyzing data derived from long longitudinal cohort studies

To cite this article:
R. Fernández-Iglesias, P. Martínez-Camblor, A. Tardón, and A. Fernández-Somoano. Statistical considerations for analyzing data derived from long longitudinal cohort studies. *Mathematics*, 11(19), 2023.
DOI: 10.3390/math11194070.

To link to this article:
https://www.mdpi.com/2227-7390/11/19/4070

mathematics

MDPI

Article

Statistical Considerations for Analyzing Data Derived from Long Longitudinal Cohort Studies

Rocío Fernández-Iglesias [1,2,3,*], Pablo Martínez-Camblor [4,5], Adonina Tardón [1,2,3] and Ana Fernández-Somoano [1,2,3]

[1] Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Monforte de Lemos Avenue 3-5, 28029 Madrid, Spain
[2] University Institute of Oncology of the Principality of Asturias (IUOPA)—Department of Medicine, University of Oviedo, Julian Clavería Street s/n, 33006 Oviedo, Asturias, Spain
[3] Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Roma Avenue s/n, 33001 Oviedo, Asturias, Spain
[4] Biomedical Data Science Department, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA
[5] Faculty of Health Sciences, Universidad Autonoma de Chile, Providencia 7500912, Chile
* Correspondence: rocio.fdez.iglesias@gmail.com

Abstract: Modern science is frequently based on the exploitation of large volumes of information storage in datasets and involving complex computational architectures. The statistical analyses of these datasets have to cope with specific challenges and frequently involve making informed but arbitrary decisions. Epidemiological papers have to be concise and focused on the underlying clinical or epidemiological results, not reporting the details behind relevant methodological decisions. In this work, we used an analysis of the cardiovascular-related measures tracked in 4–8-year-old children, using data from the INMA-Asturias cohort for illustrating how the decision-making process was performed and its potential impact on the obtained results. We focused on two particular aspects of the problem: how to deal with missing data and which regression model to use to evaluate tracking when there are no defined thresholds to categorize variables into risk groups. As a spoiler, we analyzed the impact on our results of using multiple imputation and the advantage of using quantile regression models in this context.

Keywords: missing data; quantile regression; tracking; cohort studies; children's health; cardiovascular risk

MSC: 62P10; 92B15; 92D30

check for updates

Citation: Fernández-Iglesias , R.; Martínez-Camblor, P.; Tardón, A.; Fernández-Somoano, A. Statistical Considerations for Analyzing Data Derived from Long Longitudinal Cohort Studies. *Mathematics* **2023**, *11*, 4070. https://doi.org/10.3390/math11194070

Academic Editors: Miguel Ángel Montero-Alonso and Juan De Dios Luna del Castillo

Received: 7 August 2023
Revised: 18 September 2023
Accepted: 21 September 2023
Published: 25 September 2023

1. Introduction

Modern science is frequently based on the exploitation of large volumes of information stored in datasets and involving complex computational architectures [1]. Sometimes, these datasets compromise a huge number of participants. That is the case for those studies based on large registries, which frequently include hundreds of thousands or even millions of patients [2]. In this situation, despite some aspects of the statistical analyses becoming unuseful (i.e., *p* values), the main challenge is the computational capacity for handling the number of subjects. The so-called "omic sciences", including genomics, transcriptomics, proteomics, and metabolomics, among other technologies, represent a clear example of research requiring a high computational capacity but usually involving few subjects. In these studies, the researchers collected a number of variables and had to cope with several specific methodological challenges. Among those, we have examples such as preserving the security of the data, the difficulty of cleaning and checking their consistency, or the presence of missing values. The loss of subjects between follow-ups in the case of longitudinal studies, the data harmonization when the information comes from different records or systems, and apparently trivial aspects such as sometimes being able

to know the information contained in each variable are just a few examples of the issues that researchers have to deal with. Schmitt et al. [3] presented an interesting document in which the authors described the cohort, quality assurance procedures, and results of the Successful Aging after Elective Surgery (SAGES) study, highlighting the relevance of the processes related to data collection for having a successful project.

We consider here studies in which a relevant amount of information is systematically collected with the goal of studying the evolution of the enrolled participants in an undefined future, trying to delineate the associations between exposure to potential risk factors and posterior health status. Cohort designs, such as the landmark Framingham study [4], which was originally aimed to identify the determinants of cardiovascular disease (CVD) and whose collected data have been used with different goals; the European Prospective Intake and Cancer (EPIC) study [5], designed to investigate the relationships between nutritional, lifestyle, and environmental factors and the incidence of different types of cancer and other chronic diseases; or the Environmental Influences On Child Health Outcomes (ECHO) program [6], a network of pediatric cohorts that aims to understand the effects of a broad range of early environmental influences on child health and development, are just few examples. Particularly, there are a number of them that enrolled pregnant women and had active follow-ups with themselves and their children to determine whether pre-, peri-, or post-natal exposures may influence childhood or even adulthood health outcomes. Examples of these so-called birth cohorts are the Infancia y Medio Ambiente (INMA) (Environment and Childhood) project [7], a network concerned with the relation of environmental exposures with growth, health, and development from early fetal life until puberty, or the New Hampshire Birth Cohort (NHBC) study [8], which investigated the effect of several factors such as environment contaminants on the health outcomes of pregnant women and their children.

Usually, related subprojects involve part of the subjects and a limited number of variables. They suffer from the same problems. The use of multivariate statistical techniques implies that even if a subject is only missing one of the required variables, then it should be completely excluded from the analysis. Additionally, in longitudinal studies in which large numbers of variables are collected at different follow-ups, subjects having missing information at one follow-up can differ from those having missing information in another. This can result in a drastic reduction in the available sample size and, perhaps worse, the potential introduction of systematic biases. Aside from that, the study of risk factors in health populations, and particularly in children, copes with unclear or controversial threshold definitions. As a result, children thresholds are chosen as a specific percentile of the variable of interest [9], usually assuming that it is normally distributed with parameters estimated in healthy children; that is, there is not enough knowledge about the targets and clinically meaningful thresholds.

In this work, we aim to provide some statistical insight for longitudinal cohort studies involving controversial threshold definitions. Despite some of the considered techniques being new, we put the focus on their utilization in this particular setting. Dealing with missing data or selecting the adequate regression methodology implies making a number of decisions which could impact the final conclusions. Published documents are overwhelmingly focused on describing the obtained results and, in general, do not present in detail each decision made. Here, we pay more attention to those methodological details, analyzing the impact of the made decisions on the final results and discussing their suitability in relation to the possible alternatives.

2. Materials and Methods

2.1. The INMA-Asturias Cohort

In 2004, the INMA-Asturias cohort [10] was established as a prospective, population-based cohort study. As part of the INMA project [7], its aim is to examine the potential impact of environmental exposures on maternal and child health outcomes, with special emphasis on exposure to environmental pollutants and genetic and nutritional factors.

The cohort is located in a 483 km^2 area in northern Spain, with San Agustín University Hospital (Avilés, Asturias) serving as the reference hospital. The economy of this region historically relied on industries characterized by important environmental pollution. Originally, the area included a population of 165,201 inhabitants (reduced to 144,875 in 2021), and the reference hospital is a public health center with 436 beds, providing primary care as well as central, medical, and surgical services to this population.

From May 2004 to June 2007, pregnant women attending their first prenatal visits at the obstetrics service of San Agustín University Hospital or the Las Vegas health center (Corvera, Avilés) were consecutively selected if they met the following criteria: mother's age $\geq$ 16 years, singleton pregnancy, scheduled delivery at San Agustín University Hospital, no assisted conception, and no communication handicap. Extensive data were collected by trained staff through questionnaires, medical records, biological and environmental samples, and anthropometric measures. Follow-up visits took place at the first and third trimesters of pregnancy, at birth, and when the children's ages were 18 months and 4, 8, and 12 years.

The availability of blood samples enabled the measurement of markers of adult cardiovascular risk factors, including serum lipid, glucose, insulin, blood pressure, and anthropometric measures. These markers have expanded the scope of research beyond the initial objectives, allowing study of the tracking of cardiovascular-related measures. Here, we use the work by Fernández-Iglesias et al. [11] to illustrate the motivation behind specific methodological decisions and their potential impact on the results obtained.

2.2. Tracking of Cardiovascular-Related Measures

In epidemiology, predictability or maintenance of the range of a biological variable (or specifically of risk factors for chronic diseases) within a specific population is referred to as *tracking*. Particularly in children, early studies of growth established that some measures are relatively stable over time periods [12]. This phenomenon has interested both biologists and statisticians since the early 1980s, although there is no widely accepted definition of the term. Attempts to put the underlying concept into practice have resulted in the two main conceptions shown in Box 1.

Box 1. Tracking definitions.

- The ability to predict subsequent observations $(t+1)$ from earlier observations $(1, \ldots, t)$ [13]. If, in a cohort of n children, we measured their heights $y_{i,t}$, with $1 \leq i \leq n$ and $1 \leq t \leq k$, then *tracking* is the ability to predict $y_{i,t+1}$ from $y_{i,1}, \cdots, y_{i,t}$.
- The maintenance of a relative position within a distribution of values in the observed population through time [14,15]. Therefore, in the children's height example, the question is whether children at higher percentiles at time t will also be at higher percentiles at time $t+1$.

Here, we focus on this second conception in an attempt to explore the relationship between longitudinal measurements.

Considering that atherosclerosis is a progressive accumulation process that can begin in childhood and youth [16,17], in Fernández-Iglesias et al. [11], we studied the tracking between 4 and 8 years of the following cardiovascular-related variables that reflect well-established CVD risk factors in adulthood: waist-to-height ratio (WC/Height ratio) for central obesity, mean arterial pressure (MAP) for hypertension, triglycerides (TG), high-density lipoprotein cholesterol (HDL-c), and the atherogenic coefficient (AC) for dyslipidemia, and the homeostatic model assessment of insulin resistance (HOMA-IR) for insulin resistance.

Operationally, tracking is challenging [18], particularly when examining risk factors. The most commonly used statistical techniques in the literature include logistic regression, correlation coefficients, or linear regression models. Logistic regression models require the use of thresholds to categorize risk factors that are inherently continuous, typically

using specific quantiles. This is an extremely common approach in epidemiology research, but it has major limitations. It may lead to a loss of statistical power, to less precise estimates, or to difficulty in comparing results between studies when the thresholds are sample-dependent [19–21]. Choosing them arbitrarily can be a pitfall, especially when studying adult risk factors in generally healthy children. In such cases, it is advisable to use a methodology that allows for the use of continuous measures. However, commonly used continuous approaches, such as correlation coefficients or linear regression models [22–24], also have important limitations. These methods primarily concentrate on assessing the impact within the central part of the variable's distribution. However, in the context of variables denoting risk factors, a shift in the variable's mean often does not imply a meaningful clinical or health-related impact. Instead, it is the consequences observed at the extreme part of the distribution that hold a relevant significance. Consequently, the insights yielded by these techniques may not contribute substantial valuable knowledge. To overcome this challenge, in Section 2.4, we propose the use of quantile regression models to overcome two challenges: (1) to analyze the tracking of risk factors while avoiding the use of thresholds and (2) to maintain the focus on the extreme parts of the distribution.

2.3. Missing Data: Multivariate Imputation

Missing data is a recurrent problem in statistics which is especially impactful on longitudinal studies. Little and Rubin [25] proposed a missing data classification based on the underlying loss mechanism (Box 2).

Box 2. Types of missing data according to missingness mechanisms.

Let $\{\boldsymbol{X}, Y\}$ be a k-dimensional random matrix. For the sake of simplicity, we will assume univariate missing data; that is, Y is the only variable containing missing values. Let R be the response indicator vector; that is, $R = 1$ if Y is observed, and we have $R = 0$ otherwise. Then, the following apply:

- **The missing completely at random (MCAR) model satisfies**

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R\},$$

That is, the probability of being missing does not depend either on Y or $\boldsymbol{X}$. This means that there are no systematic differences between the missing and observed values. For example, serum lipid measurements may be missing because some samples have been lost in transit to the laboratory.

- **The missing at random (MAR) model satisfies**

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R|\boldsymbol{X}\},$$

That is, the probability of being missing depends on the observed data. For example, serum lipid measures may be more likely to be missing in young people, as they tend to be less concerned and do not attend visits for blood collection.

- **The missing not at random (MNAR) model satisfies**

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R|Y\},$$

That is, the probability of being missing depends on the missing values themselves or on unobserved information. For example, in a study to assess the effect of a hypertensive treatment, hypertensive subjects may present greater collaboration that results in a lower number of missingness.

The statistical analysis approach depends on each of these situations. Under the MCAR model, the observed data can be considered a random sample from the original target sample. In such cases, a complete-case analysis does not introduce bias in the estimated parameters but implies a sample size reduction with the associated loss of power. When missing data are not MCAR, as observed, the data do not represent the full population, and the complete-case approach may provide biased results. Multiple imputation (MI) methods

can produce unbiased estimations and preserve the original sample size under the MAR situation [26]. However, under the MNAR model, as long as the missingness depends on unobserved information, MI could fail [27]. Strategies to handle the MNAR model include collecting more information about the causes for the missingness or performing sensitivity analyses to evaluate the results under various scenarios [26].

The MI method, proposed in Rubin [28], does not focus on imputing the "closest" possible values to the actual missing values but rather making valid and efficient inferences about the parameters of interest. The key concept of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing ones. Random components are incorporated into these estimated values to reflect their uncertainty. Multiple datasets are created and then analyzed individually. Finally, the individual estimations are combined to obtain the overall estimates, their standard errors, and adequate confidence intervals.

MI procedures consider the MAR model and the relationship

$$Y = g(\boldsymbol{X}) + \epsilon, \tag{1}$$

where $g(\cdot)$ and ϵ are the link function and random white noise, respectively. Box 3 summarizes the MI algorithm.

Box 3. Steps of the MI method.

Let $\{\boldsymbol{X}_n, Y_n\}$ be a random sample drawn from $\{\boldsymbol{X}, Y\}$, and let β be the target parameter. We assume that the values $y_{i_1}, \cdots, y_{i_m}$ $(1 \leq i \leq n,\ \ m < n)$ are missing.

- **Step 1**. From the non-missing values, we compute the function $\hat{g}(\cdot)$ which estimates $g(\cdot)$ (Equation (1)). For each missing value, y_{i_j} $(1 \leq i \leq n,\ \ 1 \leq j \leq m)$ generates a pseudo-value $\hat{y}_{i_j} = \hat{g}(\boldsymbol{X}_{i_j,n}) + \epsilon_{i_j}$, where ϵ_{i_j} is randomly generated. With this dataset, we estimate the target parameter $\hat{\beta}$ and its variance, $\hat{V}^2$.
- **Step 2**. We repeat Step 1 B times (where B is a large enough number) and obtain a vector of estimations $\{\hat{\beta}_1, \cdots, \hat{\beta}_B\}$ and another with their respective variabilities $\{\hat{V}_1^2, \cdots, \hat{V}_B^2\}$. Notice that in each repetition, the error (ϵ) is randomly generated. Therefore, each repetition provides a different dataset.
- **Step 3**. We use Rubin's rules to combine the vectors obtained in Step 2 into a single estimation with its variability. This estimation reflects both the uncertainty due to the sample variation and the uncertainty due to the missing data. The m $\hat{\beta}^k$ estimates and $\hat{SE}^k$ standard errors are combined using Rubin's rules to produce an overall estimate and standard error that reflect both the uncertainty due to the sample variation and the uncertainty due to the missing data.

Different algorithms have been proposed for estimating Equation (1) [29]. For instance, if we consider the linear model

$$Y = \boldsymbol{\beta} \cdot \boldsymbol{X} + \epsilon, \tag{2}$$

then we have the imputation process

$$\hat{y}_{i_j} = \hat{\boldsymbol{\beta}} \cdot X_{i_j} + \epsilon_{i_j}, \quad (1 \leq i \leq n,\ \ 1 \leq j \leq m)$$

where ϵ_{i_j} is randomly generated.

In many MI algorithms, a Bayesian perspective is often adopted, treating the parameters associated with the link function $g(\cdot)$ as random variables rather than fixed constants. This approach introduces uncertainty about missing values not only by incorporating random noise through the error term ϵ_{i_j}, as noted in Step 1 of Box 3, but also by introducing uncertainty into the link function parameters, whose state of knowledge is represented through a posterior distribution [26]. For instance, if we consider the same linear model (Equation (2)), then we have the imputation process

$$\hat{y}_{i_j} = \hat{\boldsymbol{\beta}} \cdot X_{i_j} + \epsilon_{i_j}, \quad (1 \leq i \leq n,\ \ 1 \leq j \leq m)$$

where ϵ_{i_j} is randomly generated and $\hat{\boldsymbol{\beta}}$ is sampled from its posterior distribution based on the available data.

Rubin's rules [28] combine the results of the B analysis performed to obtain

$$\bar{\hat{\beta}} = \frac{1}{B}\sum_{k=1}^{B}\hat{\beta}_k.$$

The variance of $\bar{\hat{\beta}}$, denoted as V_T^2, is calculated by

$$V_T^2 = V_W^2 + V_B^2 \cdot \left(1 + \frac{1}{B}\right), \tag{3}$$

where V_W^2 is the within-imputation variance and represents the sample variation and V_B^2 is the between-imputations variance and represents the extra variance due to the uncertainty around the imputed data; that is, we have

$$V_W^2 = \frac{1}{B}\sum_{k=1}^{B}\hat{V}_k^2, \quad \text{and}$$
$$V_B^2 = \frac{1}{B-1}\sum_{k=1}^{B}(\hat{\beta}_k - \bar{\hat{\beta}})^2$$

Inflating the between-imputation variance in Equation (3) by the factor $1/B$ reflects the extra variability as a consequence of imputing the missing data using a finite number of imputations instead of an infinite number. For constructing $100 \times (1-\alpha)\%$ confidence intervals, we assume $\bar{\hat{\beta}}$ is normally distributed and use the general formula

$$\bar{\hat{\beta}} \pm z_{\alpha/2}\sqrt{V_T^2},$$

where $z_{\alpha/2}$ is the critical value of the standard normal distribution.

Different indexes have been proposed for measuring the severity of the missing data problem. We consider here the so-called fraction of missing information (FMI), which estimates the proportion of the total variance due to the imputations and is defined by

$$\text{FMI} = \frac{V_B^2(1+1/B)}{V_T^2}.$$

The FMI ranges between 0 and 1. It is equal to zero only if the missing data do not add extra variation to the sample variance, an exceptional situation which implies perfect imputation models. And it is equal to one when the whole variation is caused by the missing data. In practice, this is equally unlikely since it means that there is no variation in the observed information [30]. The higher the value of this indicator, the greater the influence of the imputation model on the final results. Another index is the relative efficiency (RE), which represents the relative efficiency of using B rather than an infinite number of imputations:

$$\text{RE} = \frac{1}{1+\text{FMI}/B}. \tag{4}$$

It ranges from 0.5 to 1, where the higher the value, the less efficiency would be gained by increasing the number of imputed datasets.

The INMA-Asturias Cohort Example

In our study, we had a total of 416 children, but just 154 (37.02%) had all the required information. The missing percentage for cardiovascular-related variables oscillated between 6.97% and 44.47%. In the models, measures at age 4 play the role of the independent variable, and the same measures at age 8 play the role of the dependent variable. We excluded children who lacked data at the 4 and 8 year time points simultaneously. The final

considered sample (307 children) showed missing percentages which ranged from 2.3% to 25.1% (see Table S1).

The first decision related to missing data is the plausibility of the MAR assumption. We can reject data to be MCAR using the Little [31] test or by exploring if there are variables associated with missingness. But there is no way to distinguish whether the data are MAR or MNAR without additional information. In general, the MAR assumption will be more reasonable the more variables are included in the imputation model that are related to the missingness of the data on the variable of interest or to the variable of interest itself. In our case example, assuming the data were MAR, we had extra related auxiliary variables that could be incorporated into the imputation model, suggesting that the imputation methods could perform considerably well.

The second point is to specify the imputation model. To avoid unnecessary complexity, we have represented in this section the multiple imputation theory for univariate missing data. In our case, missing data occurred in more than one variable, and thus we applied a multiple imputation strategy for imputing multivariate missing data. In particular, we applied the *multivariate imputation by chained equations (MICE)* method [32] using the predictive mean matching algorithm. A detailed description and definition of this algorithm, which is based on Bayesian imputations in the MICE package, can be found in the work of Stef van Buuren [33]. Initially, we specified the imputation model with five (=B) imputed datasets [32,34–36]. Regarding to imputation model diagnosis, we assessed the maintenance of the observed relationship between the dependent and independent variables in the complete datasets. Figure 1, for example, shows that the distributions of observed and imputed data for TG were quite similar, as expected under the MAR approach.

The next decision was to determine the final B value. As the rate of missing information was below 0.5, we applied the criteria suggested by White et al. [36], Graham et al. [37], and Bodner [38]. We started with B equal to the maximum percentage of missing data observed ($B = 26$). Then, we applied the corresponding analysis to each generated dataset and combined the results. The FMI was calculated and verified whether $100 \cdot \text{FMI} \leq B$. B should be adjusted to the minimum number that satisfies this criterion otherwise. Of the 81 quantile regression models performed, the FMI median was 0.25 (interquartile range (IR): 0.12; 0.29), but the maximum was 0.46. As the computation time and storage capacity were not a concern, we finally selected $B = 50$.

After that, MI was repeated with the new number of imputations ($B = 50$), the models were estimated for each of the 50 datasets created, and the overall estimates and variances were calculated using Rubin's rules. The influence that the imputation had on these estimates was checked. Table 1 summarizes the corresponding indicators for each of the cardiovascular-related measures. The proportion of the total variance due to the imputation procedure was around 28% in the models involving measures with higher percentages of missing data and around 10% in those measures with low percentages. Note that by using a number of imputations B satisfying $100 \cdot \text{FMI} \leq B$ and taking into account Equation (4), it is expected to obtain REs higher than 99%, as we observed in Table 1. Therefore, minimal variation would occur just by increasing the number of imputations.

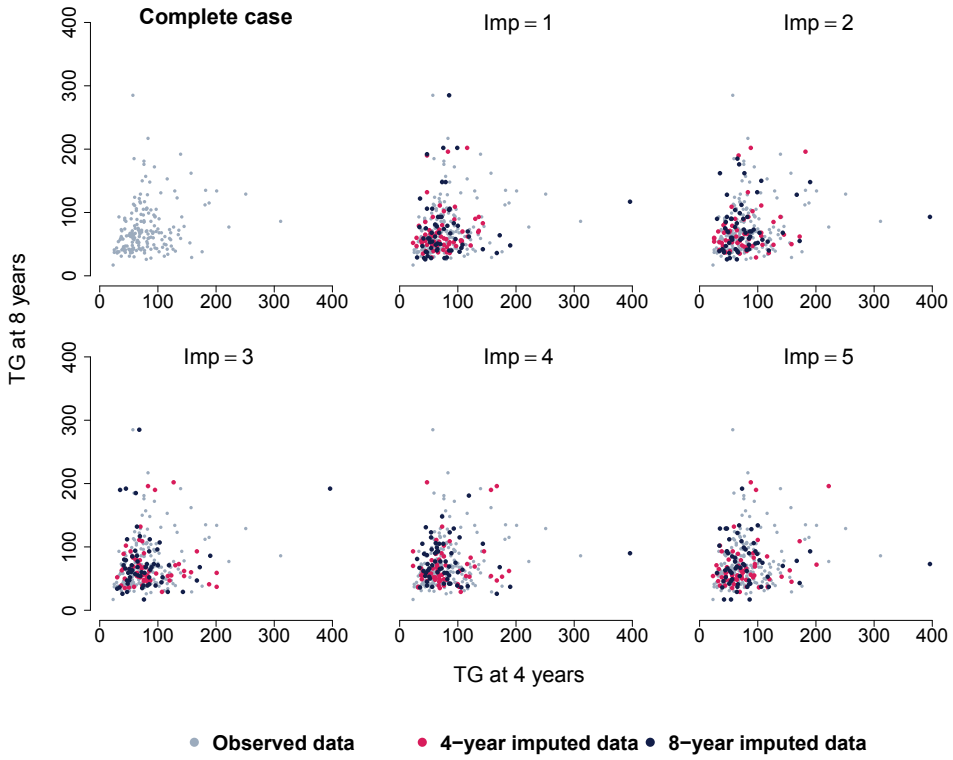


Figure 1. Scatter plots for each of the initial five imputed datasets of TG measure at 4 vs. TG measure at 8 years. TG = triglycerides; Imp = imputation.

Table 1. Median, first, and third quartiles for the indicators of the impact of the missing data, expressed as percentages.

Measure	FMI	RE
TG	30.2 (28.7; 32.6)	99.4 (99.4; 99.4)
HDL-c	25.9 (23.1; 28.9)	99.5 (99.4; 99.5)
AC	28.6 (24.3; 29.7)	99.4 (99.4; 99.5)
WC/Height ratio	11.6 (9.6; 14.6)	99.8 (99.7; 99.8)
MAP	8.9 (7.4; 10.9)	99.8 (99.8; 99.9)
HOMA-IR	29.2 (27.6; 32.5)	99.4 (99.4; 99.5)

FMI = fraction of missing information; RE = relative efficiency; TG = triglycerides; HDL-c = high-density lipoprotein cholesterol; AC = atherogenic coefficient; WC/Height ratio = waist-to-height ratio; MAP = mean arterial pressure; HOMA-IR = homeostatic model assesment of insulin resistance.

2.4. Quantile Regression

Quantile regression models (QRMs) were introduced in 1978 by Koenker and Bassett [39]. They offer a natural extension of the classical linear regression models in which, instead of specifying the change in the conditional mean of the dependent variable's distribution associated with a change in the independent variables, the change in any conditional quantile of the distribution is specified. In longitudinal studies, QRMs have been applied in a wide variety of problems. For instance, Lipsitz et al. [40] used this technique for analyzing the changes in the distribution of CD4 cell counts in patients with human immunodeficiency virus. They are also commonly used for identifying risk factors in particular populations. Fenske et al. [41] applied a QRM for detecting obesity risk factors in childhood.

Mathematically, given the dependent variable Y, the k-dimensional independent variable X, and the τth quantile with $\tau \in (0,1)$, the QRM can be specified as follows:

$$Y = \beta_\tau \cdot X + \epsilon_\tau,$$

where the residuals verify that $\mathcal{P}(\epsilon_\tau \leq 0|X) = \tau$; that is, its conditional τth quantile, $q_\tau(\cdot|\cdot)$ is zero. Therefore, we have

$$q_\tau(Y|X) = \beta_\tau \cdot X + q_\tau(\epsilon_\tau|X) = \beta_\tau \cdot X.$$

Let $\{X_n, Y_n\}$ be a random sample from $\{X, Y\}$ (sample size n). The estimator $\hat{\beta}_\tau$ is obtained by minimizing a sum of weighted absolute residuals that gives asymmetric penalties depending on whether the values of the dependent variable are being overestimated or underestimated:

$$\tau \cdot \sum_{\epsilon_{\tau_i} \geq 0} |\epsilon_{\tau_i}| + (1-\tau) \cdot \sum_{\epsilon_{\tau_i} < 0} |\epsilon_{\tau_i}| \quad (1 \leq i \leq n). \tag{5}$$

This means that the proportion of data points below the τth estimating regression line $\hat{y}_i = \hat{\beta}_\tau \cdot X_i$ $(1 \leq i \leq n)$ is τ and the proportion lying above it is $1-\tau$. Equation (5) can be minimized using different algorithms based on linear programming [42].

The interpretation of the coefficient estimates is analogous to those in classical linear regression, except that instead of referring to the effect on the conditional mean of the dependent variable, we refer to the conditional quantile. Each $\hat{\beta}_\tau$ can be interpreted as the increment of the τth quantile of the dependent variable per unit of change in the value of the corresponding independent variable, while the rest of the independent variables are fixed.

There are several procedures for computing both the standard errors and confidence intervals for the quantile regression coefficients. Under certain conditions, the usual coefficient estimators are asymptotically normally distributed [42]. However, asymptotic standard errors are complex, and resampling approaches are frequently employed [43].

QRMs overcome some limitations of classical linear regression tools, even if the researcher is only interested in a central position and its behavior. Box 4 provides some guidance on the situations for which a QRM may be appropriate. The last two points are the keys to its usefulness in evaluating tracking. But it is worth noting that the last point also makes these models highly suitable for assessing whether the effects of an exposure are the same in all quantiles. And the third point also solves the incredibly common cases where exposures follow skewed distributions.

Box 4. Situations in which quantile regression is useful.

1. **In the presence of outliers**. It is able to cope better with outliers, since it is based on the estimation of a position measure such as the quantile. Outliers only have an influence on the estimation of the quantile close to them.
2. **In case of heteroscedasticity**. If the variance depends on the independent variables, quantile regression can capture this effect.
3. **When distributional assumptions are not satisfied**. QRMs do not make assumptions about the distribution of errors, and thus they can be used when the conditions for applying other regression models are not satisfied.
4. **When the interest is at the extremes of the distribution**. Sometimes the real interest of the research question lies in what happens in the tails of the distribution. The QRM allows one to answer this question by estimating the extreme quantiles.
5. **When there is no known threshold defining the at-risk population**. As the model can be estimated for any quantile, it becomes possible to evaluate the impact of the independent variables on a specific section of the distribution without having to select a particular point.

The INMA-Asturias Cohort Example

Taking the TG measure as an example, we estimated the following QRMs for τ values ranging from 0.1 to 0.9 in 0.05 intervals:

$$q_\tau(TG_8|rank(TG_4)) = \hat{\beta}_{\tau_0} + \hat{\beta}_{\tau_1} \cdot rank(TG_4),$$

where TG_8 represents the TG measure at 8 years and $rank(TG_4)$ is the rank transformation of the TG measure at 4 years. As previously mentioned, the tracking conception is based on the relative positions of subjects within the distribution of the variable of interest. In order to incorporate this relative position within the independent variable, a rank transformation was applied. Here, we use the crude analysis as an example for simplicity, but as in any regression model, adjustment variables can be included.

Our aim is studying the impact on the upper tail of the TG at the 8 year distribution, (i.e., to estimate $\hat{\beta}_{\tau_1}$ for high τ values). However, estimating the effect for quantiles across the whole distribution and plotting $\hat{\beta}_{\tau_1}$ estimates against τ serves as a useful exploratory tool to assess whether the size and nature of the effect remains constant. Figure 2 shows that the association differed for high-risk subjects (those at the highest quantiles of TG at the 8 year distribution) compared with average subjects (those around the 0.5 quantile), reflecting an increasing trend in the the association's effect. This observation would not

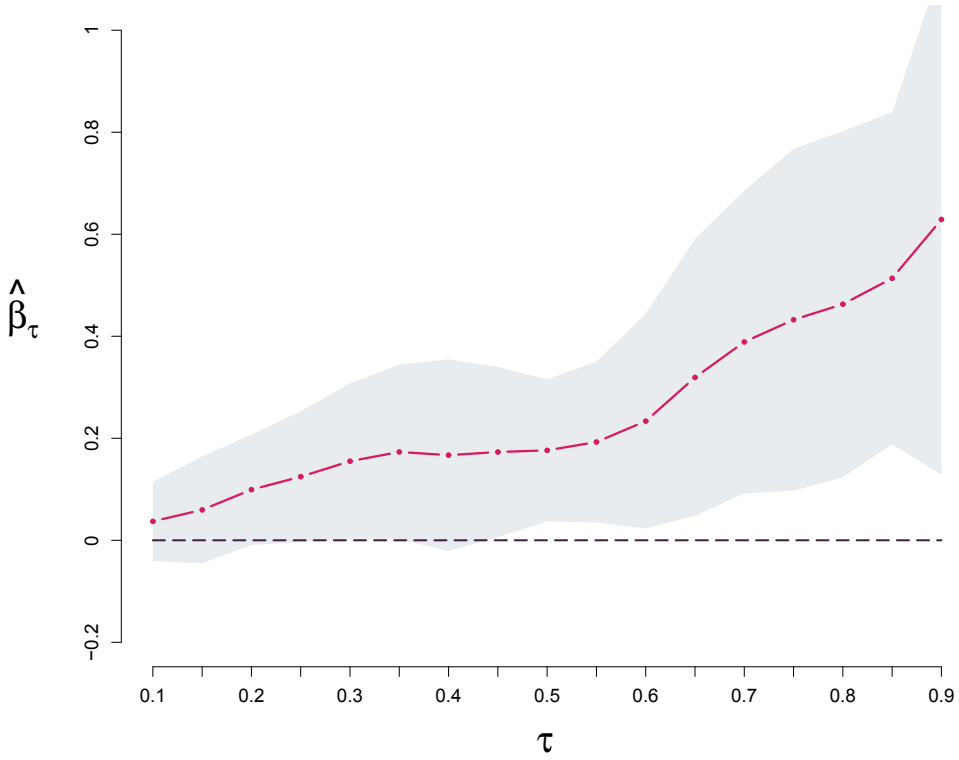


Figure 2. Quantile regression parameters ($\hat{\beta}_\tau$) per quantile (τ) for the effect of the rank transformation of TG at 4 on TG at 8. The red dots and lines represent point estimate of the parameters, while the grey bounds represent the confidence interval estimate. TG = triglycerides.

3. Results

The final results of the analysis may depend on the methodological decisions made. Regarding the missing data, the possible alternative here would be to conduct a complete case analysis. We compared the results between these two approaches and did not observe any systematic differences. However, contrary to what might be expected, not all

confidence intervals were narrower when applying MI in contrast to the complete case analysis. Figure 3 shows an example of the variances in the parameter estimates for several TG at 8 years quantiles by analysis type, and we can observe that the estimates were not more accurate for all parameters when using MI ($\tau = 0.3$ and $\tau = 0.5$). We observed this phenomenon in all the models involving measures with a high percentage of missing data (TG, HDL-c, AC, and HOMA-IR) but not in models involving the WC/Height ratio and MAP, which had less than 10% missing data.

Regarding the statistical model, a binary response model such as logistic regression could have been considered as an alternative to quantile regression. With this approach, we would still focus on the upper tail of the distribution and explore the probability of being in a high-risk TG category at 8 years, depending on the TG values at 4 years. For that purpose, we considered the 0.9 quantile, which is both age- and sex-specific, to calculate the binary variable that divided TG into the normal category (TG $<$ 0.9 quantile) and the risk category (TG $\geq$ 0.9 quantile). In our sample, without imputation, 72.3% of the 4-year-old children had normal TG levels, 8.2% had risk levels, and 19.5% had missing data. At 8 years, 71.6% of the children had normal levels, 8.5% had risk levels, and 19.9% had missing data.

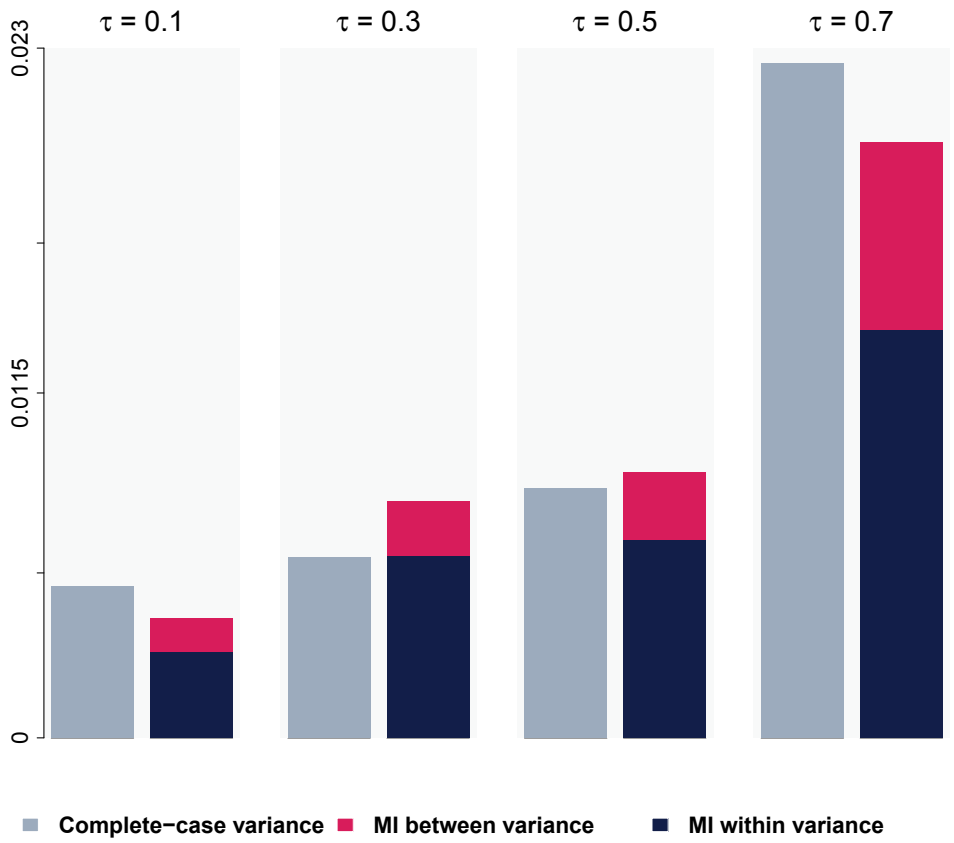


Figure 3. Variance of parameter estimates in quantile regression models for TG at 8 years and TG at 4 years by type of analysis: complete case or MI analysis. TG = triglycerides; MI = multiple imputation; τ = quantile.

Using a QRM, we observed a positive association between the rank of TG at 4 years and the 0.9 quantile of TG at 8 years ($\hat{\beta}_{0.9}$: 0.629, 95%CI: 0.129–1.129). The logistic regression model showed a positive association between the rank of TG at 4 years and the odds of being in the risk category of TG at 8 years (odds ratio (OR): 1.009, 95%CI: 0.995–1.023). While the observed association and overall conclusion were the same, the estimated parameters were not directly comparable. In the QRM, $\hat{\beta}_{0.9}$ represents an additive effect on the dependent

variable, whereas the OR in the logistic regression model represents a multiplicative effect. More specifically, for the same one-unit increase in the rank of TG at 4 years, in the first case, we estimated a 0.629 mg/dL increase in the 0.9 TG quantile at 8 years, while in the second case, we estimated there to be 1.009 times the risk of being in the TG risk category at 8 years. Moreover, the outcome did not represent the same construct. In the QRM, the outcome was a specific point of the TG distribution at 8 years, while in logistic regression, it was a section of the distribution, assuming no variation in the effect within that section. Another option would be to use the binary TG variable for both 4 and 8 years in the logistic regression. In this case, we found that children who were in the risk category at 4 years were 3.287 times (95%CI: 1.173–9.212) more likely to be in the risk category at age 8 than those who were not. Again, the evidence on the nature of the association between variables was the same, as high TG values at 4 years were positively associated with high TG values at 8 years, but the estimated effects were not comparable.

4. Discussion

The epidemiology literature has plenty of statistical analysis. Despite these usually being briefly explained, it is never clear what the impact on the observed results would be if a different decision was made. Lack of space is a common problem in specialized journals, and deep explanations are relegated to Supplementary Materials or directly omitted. Here, we explored the impact of the decisions taken, particularly with regard to missing data and the selection of the most appropriate statistical model for the study of variables involving controversial thresholds.

In recent years, MI has become a quite popular method for dealing with missing data. As we saw in Section 2.3, the most appropriate approach depends on the data missingness mechanism and on the amount of missing data or the role played by the involved variables (dependent or independent variables, adjustment variables, etc.) [44]. Several authors recommend the use of MI procedures regardless of the mechanism of missingness [45]. They argue that, under the MCAR condition, it is preferred against a complete case analysis because it results in more power. Under the MAR condition, it is preferred because, aside from more power, it will give unbiased results, whereas complete case analysis may not. And under the MNAR model, some authors suggest that it will provide less biased results than complete case analysis [46]. However, the decision is not always straightforward, and using MI only to maximize the sample size is a kind of artificial approach, which may not always be successful when it is correctly performed. Here, we presented an example where utilizing MI resulted in a higher sum of within-imputation variance and between-imputation variance and, consequently, total variance for certain quantiles compared with the variance obtained through complete case analysis. This may occur in cases where the proportion of imputed data is large and there are no variables closely related to the missing data or to the variables containing the missing data themselves. The MI model would reflect the high uncertainty around the missing data, and the target parameter estimation would be highly dependent on the generated datasets. This adds extra noise and increases uncertainty when combining the results, and it potentially leads to higher between-imputation variance values that offset the gain in the within-imputation variance resulting from the increase in sample size. Another example in which MI might yield to less precise confidence intervals, despite an increased sample size, is when there is a large proportion of missing data in the explanatory variables, and these are highly correlated with the response variable. In this case, MI can affect the precision of the estimates.

We considered tracking analysis of the cardiovascular-related measures—particularly TG—in healthy children as an example of analysis that requires avoiding the use of arbitrary thresholds while focusing on the extreme parts of the outcome distribution. Correlation coefficients and linear regression models are frequently used to explore tracking while preserving the continuous nature of the variables. These methods would focus on estimating the effect of TG at 4 years on the average TG at 8 years. Nevertheless, this does not provide us with any information on the magnitude or direction of the association

in the upper part of the TG distribution. In contrast, quantile regression addresses this constraint and allows us to assess the impact within the region of interest without relying on quantile-based categorization. We compared our approach with two variations of the classical logistic regression analysis using thresholds. The overall finding was the same: There was a positive association between the high TG values at 4 and 8 years of age. However, quantile regression provided much richer information. If there were a clear threshold enabling the categorization of a cardiovascular-related variable into normal and risk values, then logistic regression would allow us to estimate the effect on the probability of being in the risk category at 8 years associated with an increase in TG values at 4 years, thus providing an estimation of the tracking of the variable between these ages. However, in the specific case of cardiovascular-related variables in children, where consensus on the threshold values is lacking, we are truly estimating the effect on the probability of being in a category that holds no clinical significance. And we are also assuming homogeneity of risk within categories. In other words, the risk is the same for all individuals within the normal category and the same for all individuals within the risk category. On the contrary, quantile regression allows us to estimate this effect across all quantiles, thus covering the entire part of the distribution that may imply potential risk. In our example, using quantile regression, we were able to observe that the effect was not constant across all quantiles of the distribution at age 8. Instead, it increased as the quantile increased. Using logistic regression, we would not be able to see this behavior.

This suggests that the magnitude of tracking increases the more extreme the values are, providing relevant insights. While there is no established risk threshold for TG in pediatric ages, our findings indicate that increasing TG levels at 4 years may lead not only to a higher average at 8 years but also to a longer upper tail of the TG distribution at 8 years. Although it is not the purpose of this article, it should be mentioned that this could imply difficulty in normalizing TG values in the future for those children who present extreme values at 4 years of age and a progressive increase in TG values at 8 years of age. These results have potential implications for children's health, as the consequences associated with such changes in TG levels are not yet known. These findings can also inform the identification of cardiovascular-related measures that should be considered as targets for screening and monitoring in clinical practice, as well as in the development of public health guidelines and recommendations for children [11].

Quantile regression has gained widespread popularity in social science, economics, environmental modeling, public health research [47–50], and in recent years, in the field of environmental pollutant exposure [51–55]. In longitudinal data analysis, which suffers from a high level of complexity due to the intercorrelation among repeatedly measured observations, QRMs have also gained increasing popularity. Most longitudinal modeling methods primarily focus on mean regression, concentrating solely on the average effects of covariates and the mean trajectory of longitudinal outcomes. Consequently, similar to independent data, quantile regression has also been extended and applied to longitudinal data. Quantile regression for longitudinal data possesses the capacity, at both the population and individual levels, to identify heterogeneous covariate effects, elucidate variations in longitudinal changes across different quantiles of the outcome, and offer more robust estimates when heavy-tailed distributions and outliers are present [56]. Despite this, its application in longitudinal cohort studies for tracking purposes has been limited [57]. This work serves as an example of its potential for investigating risk variables without known thresholds or when research interests lie in non-central areas of the distribution, as occurs in tracking studies or also when evaluating the possible effects of exposures. Even in other cases, it can complement traditional analysis methods by estimating a family of conditional quantile functions, providing a more nuanced understanding of variable effects.

5. Conclusions

Details are important in statistical analysis, as they can impact the final results. In our data, the findings seemed to be robust with respect the to main decisions taken but led

to differences in terms of accuracy and richness of information obtained. Here we point out that although multiple imputation methods are generally useful for mitigating biases in estimates, they may not necessarily improve the precision of standard error estimates. Moreover, we illustrate that quantile regression can be a powerful tool in addressing challenges associated with controversial threshold definitions and tracking analyses in cohort studies, providing valuable additional information. Given the strengths of these models, they should be considered in analyses of continuous outcomes, at least as a first step for making future modeling decisions. Finally, it is always unclear what impact different decisions would have on the obtained results, and there are always numerous alternatives to choose from. Therefore, it is essential to describe and report precisely how the analysis was conducted, including its limitations and strengths, even if it has to be included in Supplementary Materials.

6. Computational Considerations

Nowadays, there are many resources that allow a wide range of statistical analyses to be performed, including those that may require a high computational capacity, such as the ones presented here. In this work, we used R statistical software (version 4.2.1; R Foundation for Statistical Computing, Vienna, Austria, www.r-project.org). In particular, we used the package *MICE* [32] developed by van Buuren and Groothuis-Outshoorn, which includes several different imputation model options to perform multivariate imputation with chained equations. The package *quantreg* [58] was used for QRM estimation and inference, which provides several alternative methods to estimate model parameters and to compute standard errors.

Supplementary Materials: The following supporting information can be downloaded at www.mdpi.com/article/10.3390/math11194070/s1. Table S1: Number of participants with missing data for each variable, expressed in absolute and relative frequencies, for the final sample composed by 307 children.

Author Contributions: Conceptualization, R.F.-I., A.F.-S. and P.M.-C.; data curation, R.F.-I.; formal analysis, R.F.-I.; funding acquisition, A.T.; investigation, R.F.-I., A.F.-S. and P.M.-C.; methodology, R.F.-I., A.F.-S. and P.M.-C.; supervision, A.F.-S., P.M.-C. and A.T.; visualization, R.F.-I. and P.M.-C.; writing—original draft, R.F.-I. and P.M.-C.; writing—review and editing, A.F.-S. and P.M.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by grants from CIBERESP (PhD employment contract and fellowship), ISCIII: PI04/2018, PI09/02311, PI13/02429, PI18/00909 co-funded by FEDER, "A way to make Europe"/"Investing in your future", Fundación Cajastur, and Universidad de Oviedo.

Institutional Review Board Statement: The study was conducted while conforming to the principles of the Declaration of Helsinki, and its protocol was approved by the Asturias Regional Ethics Committee.

Informed Consent Statement: Written informed consent was obtained from every participanting woman and, in such cases, her partner.

Data Availability Statement: The data and computing code are available for replication from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAGES	Successful Aging after Elective Surgery
CVD	Cardiovascular disease
EPIC	European Prospective Intake and Cancer
ECHO	Environmental Influences On Child Health Outcomes
INMA	INfancia y Medio Ambiente (Environment and Childhood)

NHBC	New Hampshire Birth Cohort
WC/Height ratio	Waist-to-height ratio
MAP	Mean arterial pressure
TG	Triglycerides
HDL-c	High-density lipoprotein cholesterol
AC	Atherogenic coefficient
HOMA-IR	Homeostatic model assessment of insulin resistance
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
MI	Multiple imputation
FMI	Fraction of missing information
RE	Relative efficiency
MICE	Multivariate imputation by chained equations
QRMs	Quantile regression models

References

1. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.Z. Big Data for Health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1193–1208. [CrossRef] [PubMed]
2. Rodriguez-Martinez, A.; Zhou, B.; Sophiea, M.K.; Bentham, J.; Paciorek, C.J.; Iurilli, M.L.; Carrillo-Larco, R.M.; Bennett, J.E.; Di Cesare, M.; Taddei, C.; et al. Height and body-mass index trajectories of school-aged children and adolescents from 1985 to 2019 in 200 countries and territories: A pooled analysis of 2181 population-based studies with 65 million participants. *Lancet* **2020**, *396*, 1511–1524. [CrossRef] [PubMed]
3. Schmitt, E.; Saczynski, J.; Kosar, C.; Jones, R.; Alsop, D.; Fong, T.; Metzger, E.; Cooper, Z.; Marcantonio, E.R.; Travison, T.; et al. The successful aging after elective surgery study: Cohort description and data quality procedures. *J. Am. Geriatr. Soc.* **2015**, *63*, 2463–2471. [CrossRef] [PubMed]
4. Tsao, C.W.; Vasan, R.S. Cohort Profile: The Framingham Heart Study (FHS): Overview of milestones in cardiovascular epidemiology. *Int. J. Epidemiol.* **2015**, *44*, 1800–1813. [CrossRef] [PubMed]
5. Riboli, E.; Kaaks, R. The EPIC Project: Rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **1997**, *26*, 6–14. [CrossRef]
6. Blaisdell, C.J.; Park, C.; Hanspal, M.; Roary, M.; Arteaga, S.S.; Laessig, S.; Luetkemeier, E.; Gillman, M.W. The NIH ECHO Program: Investigating how early environmental influences affect child health. *Pediatr. Res.* **2022**, *92*, 1215–1216. [CrossRef]
7. Guxens, M.; Ballester, F.; Espada, M.; Fernández, M.F.; Grimalt, J.O.; Ibarluzea, J.; Olea, N.; Rebagliato, M.; Tardon, A.; Torrent, M.; et al. Cohort profile: The INMA–INfancia y Medio Ambiente–(Environment and Childhood) Project. *Int. J. Epidemiol.* **2012**, *41*, 930–940. [CrossRef]
8. Gilbert-Diamond, D.; Cottingham, K.L.; Gruber, J.F.; Punshon, T.; Sayarath, V.; Gandolfi, A.J.; Baker, E.R.; Jackson, B.P.; Folt, C.L.; Kargas, M.R.; et al. Rice consumption contributes to arsenic exposure in US women. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20656–20660. [CrossRef]
9. Lurbe, E.; Agabiti-Rosei, E.; Cruickshank, J.K.; Dominiczak, A.; Erdine, S.; Hirth, A.; Invitti, C.; Litwin, M.; Mancia, G.; Pall, D.; et al. 2016 European Society of Hypertension guidelines for the management of high blood pressure in children and adolescents. *J. Hypertens.* **2016**, *34*, 1887–1920. [CrossRef]
10. Fernández-Somoano, A.; Estarlich, M.; Ballester, F.; Fernández-Patier, R.; Aguirre-Alfaro, A.; Herce-Garraleta, M.D.; Tardón, A. Outdoor NO2 and benzene exposure in the INMA (Environment and Childhood) Asturias cohort (Spain). *Atmos. Environ.* **2011**, *45*, 5240–5246. [CrossRef]
11. Fernández-Iglesias, R.; Martinez-Camblor, P.; Fernández-Somoano, A.; Rodríguez-Dehli, C.; Venta-Obaya, R.; Karagas, M.R.; Tardón, A.; Riaño-Galán, I. Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA-Asturias cohort. *Eur. J. Pediatr.* 2023, *online ahead of print*. [CrossRef] [PubMed]
12. Binkin, N.J.; Yip, R.; Fleshood, L.; Trowbridge, F.L. Birth weight and childhood growth. *Pediatrics* **1988**, *82*, 828–834. [CrossRef]
13. Rosner, B.; Hennekens, C.H.; Kass, E.H.; Miall, W.E. Age-specific correlation analysis of longitudinal blood pressure data. *Am. J. Epidemiol.* **1977**, *106*, 306–313. [CrossRef] [PubMed]
14. Berenson, G.S.; Foster, T.A.; Frank, G.C.; Frerichs, R.R.; Srinivasan, S.R.; Voors, A.W.; Webber, L.S. Cardiovascular disease risk factor variables at the preschool age. The Bogalusa heart study. *Circulation* **1978**, *57*, 603–612. . [CrossRef] [PubMed]
15. Clarke, W.R.; Schrott, H.G.; Leaverton, P.E.; Connor, W.E.; Lauer, R.M. Tracking of blood lipids and blood pressures in school age children: The Muscatine study. *Circulation* **1978**, *58*, 626–634. [CrossRef] [PubMed]
16. Milei, J.; Ottaviani, G.; Lavezzi, A.M.; Grana, D.R.; Stella, I.; Matturri, L. Perinatal and infant early atherosclerotic coronary lesions. *Can. J. Cardiol.* **2008**, *24*, 137–141. [CrossRef]
17. Mcgill, H.C.; Mcmahan, C.A.; Herderick, E.E.; Malcom, G.T.; Tracy, R.E.; Strong, J.P. Origin of atherosclerosis in childhood and adolescence. *Am. J. Clin. Nutr.* **2000**, *72*, 1307S–1315S. [CrossRef]

18. Wang, Y.; Wang, X. How do statistical properties influence findings of tracking (maintenance) in epidemiologic studies? An example of research in tracking of obesity. *Eur. J. Epidemiol.* **2003**, *18*, 1037–1045. [CrossRef]
19. Ragland, D.R. Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* **1992**, *3*, 434–440. [CrossRef]
20. Altman, D.G.; Royston, P. The cost of dichotomising continuous variables. *BMJ* **2006**, *332*, 1080. [CrossRef]
21. Bennette, C.; Vickers, A. Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med. Res. Methodol.* **2012**, *12*, 21. [CrossRef] [PubMed]
22. Sarganas, G.; Schaffrath Rosario, A.; Niessner, C.; Woll, A.; Neuhauser, H.K. Tracking of Blood Pressure in Children and Adolescents in Germany in the Context of Risk Factors for Hypertension. *Int. J. Hypertens.* **2018**, *2018*, 8429891. [CrossRef] [PubMed]
23. Joshi, S.M.; Katre, P.A.; Kumaran, K.; Joglekar, C.; Osmond, C.; Bhat, D.S.; Lubree, H.; Pandit, A.; Yajnik, C.S.; Fall, C.H. Tracking of cardiovascular risk factors from childhood to young adulthood—The Pune Children's Study. *Int. J. Cardiol.* **2014**, *175*, 176–178. [CrossRef] [PubMed]
24. De Wilde, J.A.; Middelkoop, B.J.; Verkerk, P.H. Tracking of thinness and overweight in children of Dutch, Turkish, Moroccan and South Asian descent from 3 through 15 years of age: A historical cohort study. *Int. J. Obes.* **2018**, *42*, 1230–1238. [CrossRef]
25. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2019.
26. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef]
27. Kristman, V.L.; Manno, M.; Côté, P. Methods to account for attrition in longitudinal data: Do they work? A simulation study. *Eur. J. Epidemiol.* **2005**, *20*, 657–662. [CrossRef]
28. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, NY, USA, 1987.
29. Laqueur, H.S.; Shev, A.B.; Kagawa, R.M.C. SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations. *Am. J. Epidemiol.* **2021**, *191*, 516–525. [CrossRef]
30. Van Buuren, S. *Flexible Imputation of Missing Data*, 2nd ed.; Chapman & Hall: London, UK, 2018.
31. Little, R.J. A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* **1988**, *83*, 1198–1202. [CrossRef]
32. Van Buuren, S.; Groothuis-Oudshoorn, K. MICE: Multivariate imputation by chained equations. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
33. Van Buuren, S.. *Flexible Imputation of Missing Data*; Chapman & Hall/CRC Interdisciplinary Statistics: London, UK, 2012; p. 342.
34. Austin, P.C.; White, I.R.; Lee, D.S.; van Buuren, S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can. J. Cardiol.* **2021**, *37*, 1322–1331. [CrossRef]
35. Lee, K.J.; Roberts, G.; Doyle, L.W.; Anderson, P.J.; Carlin, J.B. Multiple imputation for missing data in a longitudinal cohort study: A tutorial based on a detailed case study involving imputation of missing outcome data. *Int. J. Soc. Res. Methodol.* **2016**, *19*, 575–591. [CrossRef]
36. White, I.R.; Royston, P.; Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30*, 377–399. [CrossRef] [PubMed]
37. Graham, J.W.; Olchowski, A.E.; Gilreath, T.D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **2007**, *8*, 206–213. [CrossRef]
38. Bodner, T.E. What improves with increased missing data imputations? *Struct. Equ. Model. A Multidiscip. J.* **2008**, *15*, 651–675. [CrossRef]
39. Koenker, R.; Bassett, G. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
40. Lipsitz, S.R.; Fitzmaurice, G.M.; Molenberghs, G.; Zhao, L.P. Quantile Regression Methods For Longitudinal Data with Drop-Outs: Application to CD4 Cell Counts of Patients Infected with the Human Immunodeficiency Virus. *J. R. Stat. Soc. Ser. C* **1997**, *46*, 463–476. [CrossRef]
41. Fenske, N.; Fahrmeir, L.; Rzehak, P.; Höhle, M. *Detection of Risk Factors for Obesity in Early Childhood with Quantile Regression Methods for Longitudinal Data*; Technical Report; University of Munich: Munich, Germany, 2008. [CrossRef]
42. Koenker, R. *Quantile Regression*; Cambridge University: Cambridge, UK, 2005.
43. Hao, L.; Naiman, D.Q. *Quantile Regression*; Sage Publications: Thousand Oaks, CA, USA, 2007.
44. Enders, C.K. *Applied Missing Data Analysis*; Guilford Press: New York, NY, USA, 2010.
45. Van Ginkel, J.R.; Linting, M.; Rippe, R.C.; van der Voort, A. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *J. Personal. Assess.* **2020**, *102*, 297–308. [CrossRef]
46. Schafer, J.L. *Analysis of Incomplete Multivariate Data*, 1st ed.; Chapman & Hall: London, UK, 1997.
47. Yu, K.; Lu, Z.; Stander, J. Quantile regression: Applications and current research areas. *J. R. Stat. Soc.* **2003**, *52*, 331–350. [CrossRef]
48. Staffa, S.J.; Kohane, D.S.; Zurakowski, D. Quantile Regression and Its Applications: A Primer for Anesthesiologists. *Anesth. Analg.* **2019**, *128*, 820–830. [CrossRef]
49. Oconnor, C. Robust estimates of vulnerability to poverty using quantile models. *Econ. Model.* **2023**, *123*, 106274. [CrossRef]
50. Amjad, M.; Akbar, M. The Association between Fruit and Vegetable Intake and Socioeconomic Factors in the Households of Pakistan Using Quantile Regression Model. *Soc. Work Public Health* **2023**, *38*, 248–258. [CrossRef] [PubMed]
51. Wei, Y.; Kehm, R.D.; Goldberg, M.; Terry, M.B. Applications for Quantile Regression in Epidemiology. *Curr. Epidemiol. Rep.* **2019**, *6*, 191–199. [CrossRef]

52. Peralta, A.A.; Schwartz, J.; Gold, D.R.; Vonk, J.M.; Vermeulen, R.; Gehring, U. Quantile regression to examine the association of air pollution with subclinical atherosclerosis in an adolescent population. *Environ. Int.* **2022**, *164*, 107285. [CrossRef] [PubMed]
53. Strickland, M.; Lin, Y.; Darrow, L.; Warren, J.; Mulholland, J.; Chang, H. Associations Between Ambient Air Pollutant Concentrations and Birth Weight: A Quantile Regression Analysis. *Epidemiology* **2019**, *30*, 624–632. . [CrossRef]
54. Cowell, W.; Jacobson, M.H.; Long, S.E.; Wang, Y.; Khan, L.G.; Ghassabian, A.; Naidu, M.; Torshizi, G.D.; Afanasyeva, Y.; Liu, M.; et al. Maternal urinary bisphenols and phthalates in relation to estimated fetal weight across mid to late pregnancy. *Environ. Int.* **2023**, *174*, 107922. [CrossRef]
55. Kapwata, T.; Wright, C.Y.; Reddy, T.; Street, R.; Kunene, Z.; Mathee, A. Environmental Science and Pollution Research Relations between personal exposure to elevated concentrations of arsenic in water and soil and blood arsenic levels amongst people living in rural areas in Limpopo, South Africa. *Environ. Sci. Pollut. Res.* **2023**, *30* , 65204–65216. [CrossRef]
56. Huang, Q.; Zhang, H.; Chen, J.; He, M. Quantile Regression Models and Their Applications: A Review. *J. Biom. Biostat.* **2017**, *8*, 354. [CrossRef]
57. Norris, T.; Bann, D.; Hardy, R.; Johnson, W. Socioeconomic inequalities in childhood-to-adulthood BMI tracking in three British birth cohorts. *Int. J. Obes.* **2020**, *44*, 388–398. [CrossRef]
58. Koenker, R. *Quantreg: Quantile Regression*, R Package Version 5.94; R Foundation for Statistical Computing: Vienna, Austria, 2017. Available online: https://CRAN.R-project.org/package=quantreg (accessed on 25 August 2023).

CHAPTER 4

Discussion

The methodology and results reported in each of the previously presented articles have already been extensively discussed within their respective contexts. In this section, our aim is to provide a comprehensive and integrated overview of the research, highlighting key points, rather than approaching each article in isolation.

4.1. Tracking of cardiovascular risk factors

This dissertation focuses on the study of tracking of metabolic risk factors, and as discussed in the Section 1.3, in recent decades the phenomenon of metabolic syndrome has gained particular relevance. In a very general sense, metabolic syndrome involves the simultaneous presence of three or more metabolic risk factors in the same subject: central obesity, insulin resistance, hypertension, and dyslipidemia. Various definitions exist, with slight variations, but as it is not a specific part of this work, we will not go into it further. The dangers associated with this clustering of metabolic syndrome components have been demonstrated in adults, where the presence of three or more components notably increases the risk of coronary heart disease, including death or non-fatal myocardial infarction, as well as the onset of diabetes (Sattar et al., 2003).

In the case of children, defining metabolic syndrome faces some challenges (see Section 1.6.1) regarding controversial thresholds, as it relies on risk factor definitions in children, with not recognised clinically meaningful thresholds. Instead, values above the 90th, 95th, or 97th percentiles for sex and age are used to define metabolic syndrome components.

There is not a universal agreement on which threshold use for the metabolic syndrome criteria. Also, the International Diabetes Federation does not even recommend considering metabolic syndrome in children under 6 years of age (Zimmet et al., 2007). Our study focused on children aged 4 and 8 years, so metabolic syndrome was not one of the objectives. Nonetheless, considering its relevance, as a first step in the study of tracking of metabolic risk factors, we examined whether there were patterns of aggregation beyond what might be expected by chance alone, and if such patterns persisted between ages 4 and 8. This is the Objective 1 of this dissertation, which is reported in Article I (Section 3.1).

Conducting this initial exploration of metabolic risk factors in children, we encountered the challenge of determining appropriate thresholds for classifying whether a child exhibited these risk factors. While recommendations and guidelines have been developed by authoritative societies such as the European Society of Hypertension for blood pressure (Lurbe et al., 2016), these mostly rely on adult values adapted for childhood, or on percentiles. In our work, we chose to employ thresholds established within the framework of the IDEFICS study for each of the cardiovascular-related variables under study (Ahrens et al., 2014). The primary reasons for this choice were twofold: firstly, by not calculating thresholds using percentiles derived from our own sample, we aimed to ensure that the same reference values could be applied in potential replications of the study in other populations, thereby avoiding sample-dependence. Secondly, the selected thresholds were applicable to children aged 2 to 10.9 years, making them suitable for 4-8–year-olds, and they were derived from a large, healthy European population, including southern European countries like Spain.

However, when applying the threshold values derived from the IDEFICS study, which were based on the 90th percentile, we expected that approximately 10% of children in the INMA-Asturias cohort sample would exhibit these risk factors. Surprisingly, our sample showed a prevalence ranging from 20.8% to 33.7% at 4 years of age and from 6.3% to 35.8% at 8. Furthermore, we faced notable challenges when comparing our results with those from other studies due to variations in the thresholds used, and there were limited investigations into this topic at such early ages as 4 years. We also found a number of references suggesting that CVD risk factors should be treated as continuous variables in research, and only be relegated to categorical variables in clinical diagnosis. Some overall continuous CVD risk scores have been developed (Eisenmann, 2008; Wijndaele

et al., 2006), potentially offering a solution to the thresholds issue. However, our research aimed to investigate the tracking of each individual risk factor, and collapsing them into a single score would result in a loss of information.

In the study presented in the Article I, our conclusions indicated that clear aggregation patterns were not readily identifiable. However, it was noteworthy that nearly all children without risk factors at 4 years did not exhibit them at 8 years, and central obesity was the one suggesting tracking from 4 to 8 years, as well as the most frequent co-occurring risk factor. The absence of clearly differentiated patterns with concurrent presence of multiple risk factors constitutes a finding itself, which supports the criterion of the International Diabetes Federation that it is not necessary to define the metabolic syndrome at such early ages in healthy populations without specific related pathologies. On the other hand, the pathogenic mechanism of metabolic syndrome remains highly complex. It is believed that central obesity and/or insulin resistance initiate many different pathogenic pathways that increase metabolic risk and end up in the full expression of the metabolic syndrome (Reisinger et al., 2020). In this regard, the fact that central obesity frequently appears alongside other factors in our study lends support to the hypothesis that central obesity could play a central role in the development of other metabolic factors. In our study, a high prevalence of elevated blood pressure values was also identified at the age of 8, occurring independently of the presence of other risk factors. One of the primary contributing factors to hypertension in pediatric populations is obesity. However, our study revealed a group of children with elevated blood pressure values not associated with obesity, suggesting a potential link to other independent causes such as prenatal and postnatal exposures, genetic factors, birth characteristics, dietary habits, or lifestyle factors (S. Machado et al., 2021). Some of these factors have been shown to be related to primary hypertension, independent of obesity, as demonstrated, for example, by Rosner et al. (2013) in the case of high sodium intake. Specifically, it has also been observed that the Spanish pediatric population exceeds the recommended sodium intake between the ages of 9 and 12, irrespective of their nutritional status (Partearroyo et al., 2019). This excess sodium intake is positively associated with elevated diastolic blood pressure levels between 5 and 16 years of age, again irrespective of their nutritional status (Pérez-Gimeno et al., 2020). It is important to note that blood pressure measurements in our study were taken according to a protocol where the mean of three different measurements was recorded, but all measurements were taken at the same moment and not on different days, as ideally recommended (Lurbe et al., 2016).

Since clear patterns of aggregation among various metabolic risk factors at age 4 and 8 were not identified, we decided to investigate individual tracking of these factors. This corresponds to the Objective 2 of this dissertation, which results and conclusions are detailed in the Article II (Section 3.2).

Using quantile regression, we found a positive association between the relative position of children at 4 years in the high density lipoprotein cholesterol distribution (HDL-c) and all the quantiles of the same variable at 8 years. Similar results were found for the atherogenic coefficient (AC), which is defined as the ratio between non-HDL-c to HDL-c, suggesting that serum lipids track among children between the ages of 4 and 8 years. But quantile regression lets us to appreciate that, in the case of the AC, the more extreme the values at 8 years, the greater the effect of the association. While there is no established risk threshold for AC in pediatric ages, our findings indicate that increasing AC levels at 4 years may lead not only to a higher average at 8, but also to a longer upper tail of the AC distribution at 8 years. This could imply difficulty in normalizing AC values in the future for those children who present extreme values at 4 years of age and a progressive increase in AC values at 8. These results have potential implications for the health of children, as the consequences associated with such changes in the distribution of AC levels are not yet known. This is also remarkable because AC is related to higher risk of CVDs in adulthood (Brunner et al., 2019; McBride, 2008). Whereas higher HDL-c had similar tracking levels in all quantiles and higher HDL-c has unclear association with CVD risk (Casula et al., 2021). For triglycerides distribution, it was found a positive association between the relative position at 4 years and the quantiles above 0.5 at 8. This means that higher triglyceride levels at 4 years increase values above the median at 8, but do not affect values below it. Therefore, the effect of an increase in triglycerides at 4 years would be linked to an increase at 8 in those values that are potentially of risk.

This study found a positive association between the relative position of children at 4 years in waist circunference to height (WC/Height) ratio distribution and all the quantiles of the same variable at 8 years. Here, this association has special relevance because in the crude analysis, without covariates, we observed higher tracking in children with a high-risk position in the distribution at 8 years, but when we adjusted for maternal pre-pregnancy body mass index (BMI), and maternal educational level, the association became similar across all the quantiles of the distribution. In other words, higher maternal BMI values are linked to a more pronounced increase in the highest quantiles, while a

higher maternal educational level is associated with a reduction in the highest quantiles. Thus, both maternal BMI and maternal education amplify the tracking of central obesity between ages 4 and 8.

All this yield to the conclusion of a suggesting tracking between 4 and 8 years of those variables that are indicators of dyslipidemia and central obesity, with the greatest effect in the highest parts of the distribution of the dyslipidemia indicator variable.

We would like to highlight that in both the analysis carried out in Article I and in Article II (Section 3.2) the sample sizes were different. In the study of aggregation patterns of cardiovascular risk factors, the sample size for the analysis, after imputing missing data using the FIML method, was 332 children. However, in the tracking study, after performing MI, it was 307 children. The difference arises because, in the first case, we began with the initial pool of 453 children who continued participating in the INMA-Asturias study at 4 years. From this group, we selected all those who had data for at least one of the cardiovascular-related variables at one of the two time points (either the 4-year or 8-year follow-up), as described in the article. However, in the tracking study, we conducted a more detailed analysis of how to handle missing data. Those children who did not continue in the study at 8 years were not considered as missing data; instead, they were regarded as dropouts. And dropouts are usually due to different reasons compared to the children who continue to participate but do not attend some of the visits. Therefore, the decision was made to use the sample size of 416 children who were still participating in the study at 8 years as the starting point for the analysis, and then proceed as described in Article II. This difference in criteria is what led to the different sample sizes used in both analyses.

There is also a difference in the cardiovascular-related variables presented in each article as indicators of cardiovascular risk factors. In Article I, waist circumference was considered as a marker of obesity, while in Article II we decided to include the WC/Height ratio as it is considered a more robust measure. The second article examined various central obesity indicators, including those previously mentioned, as well as BMI and triponderal index. All these indicators were analyzed, leading to consistent results and conclusions. Regarding dyslipidemia, Article II introduced the atherogenic coefficient as an additional indicator, to reflect the relationship between the non-HDL-c and the HDL-c values. With regard to blood pressure, in Article II we added to systolic and diastolic blood pressure variables the mean arterial pressure, and analized all of them. We obtained the same

association and the same tendency regardless of the measure used. As the study carried out in Article II includes a considerable number of cardiovascular-related measures, we decided to present only the mean arterial pressure, which intrinsically includes both systolic and diastolic arterial pressure. Results for systolic and diastolic blood pressure not included in Article II are provided in the appendix 5. In this way, the second study aimed to improve the measurement of metabolic risk factors using underlying indicators compared to the first study.

4.2. Methodological aspects

Beyond the specific findings related to the study's focus – Objectives 1 and 2 –, we also derived important methodological insights. After developing the research reported in Article I, we concluded that, for continuing the research, it would be benefit to study the variables continuously, as indicated by the aforementioned challenges in defining thresholds. Secondly, we recognised the substantial loss of sample size and potential bias associated with complete-case analysis. In the Article I we had already addressed this employing the FIML method (Enders, 2010), as a more sophisticated imputation method than the simple methods discussed in the Section 1.6.2. However, we considered necessary to explore additional imputation approaches that would allow for a better inspection and analysis of the imputed data, improving our ability to evaluate their impact on the obtained results (van Buuren, 2012).

When considering possible options for approaching the study of tracking using continuous variables, we explored alternatives to the more common mean-based analyses, leading us to quantile regression (Wei et al., 2019). Quantile regression has gained widespread popularity not only in social science, economics, environmental modeling, and public health research but also, in recent years, in the field of environmental pollutant exposure (Yu et al., 2003).

Correlation coefficients and linear regression models are commonly employed to examine tracking while maintaining the continuous nature of the variables. However, these approaches do not provide insights into the magnitude or direction of the association within specific segments of the variable's distribution. In contrast, quantile regression addresses this limitation and allows us to assess the impact within a region of interest without the need for quantile-based categorization, making it well-suited for our purposes.

It is noteworthy that even when accepted and clinically meaningful thresholds existed, quantile regression could still be a more suitable alternative to logistic regression, which is one of the most commonly used methods, as it offers much richer information. If a clear threshold existed for categorizing a cardiovascular-related variable into normal and risk values, logistic regression could estimate the effect on the probability of falling into the risk category at 8 years associated with an increase in the same variable at 4 years, providing an estimation of tracking between these ages. However, this approach assumes homogeneity of risk within categories, implying that the risk is the same for all individuals within the normal or risk category. In contrast, quantile regression enables us to estimate this effect across all quantiles, covering the entire distribution range that may imply potential risk. For example, in our study, using quantile regression, we observed that the effect was not constant across all quantiles of the distribution at age 8 for some cardiovascular-related variables. Instead, it increased as the quantile increased. Logistic regression would not capture this behavior. The work presented in the Article II (Section 3.2) serves as an example not only of quantile regression's potential for investigating variables without known thresholds, but also in cases where research interests extend to non-central areas of the distribution, such as when evaluating the potential effects of exposures (Kapwata et al., 2023; Peralta et al., 2022). Moreover, in various scenarios, quantile regression can complement traditional analysis methods by estimating a family of conditional quantile functions, providing a more nuanced understanding of variable effects, and also examining whether the effects of the independent variable on the dependent variable result in other changes, such as changes in the shape or dispersion of the dependent variable.

Regarding the chosen option for handling missing data, the primary reasons for selecting MI were, as mentioned in the Section 1.6.2: Under the MAR mechanism, it can produce unbiased estimations and preserve the original sample size. The difficult here is that the complete true distribution of the variables with missing data is unknown, and we cannot test if missing data depends on the missing data itself (MNAR mechanism), so MAR is only an assumption that cannot being really checked (Schafer and Graham, 2002). Collins et al. (2001) demonstrated that in many realistic cases, an erroneous assumption of MAR may often have only a minor impact on estimates and standard errors. Moreover, MAR assumption will be more reasonable the more observed variables are included in the imputation model (Schafer, 1997). In our case, we have extra auxiliary variables that are related to the missingness of the data on the variable of interest or to the variable of

interest itself and can be incorporated into the imputation model. We also apply Little test (Little, 1988) to reject data to be MCAR and find there are variables associated with missingness. In summary, although any of that provides a proof that the missing data are MAR, suggest that the imputation methods could perform well. MI has become a widely popular method for dealing with missing data, and some authors argue that even under the MCAR condition, it is preferred over complete case analysis because it can result in higher statistical power (van Ginkel et al., 2020). In the work presented in the Article III, we explored the differences between the results obtained using MI and what we would have obtained through complete-case analysis. It was revealing to observe that MI did not always lead to narrower confidence intervals and, therefore, more precise estimations. What we noticed was that in cases where the proportion of imputed data is large, and there are no variables closely related to the missing data or the variables containing the missing data themselves, the MI model introduce additional noise to the imputed data, reflecting appropriately the uncertainty that actually exists around them. Hence, using it as a technique to artificially increase the available sample size, when we lack suitable additional information for imputation, may not be effective.

4.3. Strengths and limitations

Each one of the three academic papers provided in the Chapter 3 already discuss specific strengths and limitations of each work. The most notable ones are highlighted below.

In terms of the applied methodology, it is always unclear the impact different decisions would have on the obtained results, and there are always numerous alternatives to choose from. We have presented and discussed the reasons for choosing the selected options, but other choices were possible, and we have not conducted an exhaustive analysis of all the different possibilities. Additionally, the available sample size, even after imputing missing data, remains moderately small, resulting in reduced statistical power for the analysis. It should also be acknowledged that this study is exploratory in nature, involving the testing of numerous hypotheses, which introduces the issue of multiple testing, thereby complicating the calculation of the statistical power of the study. Moreover, comparing the results obtained in both Article I and II with those from other studies is challenging due to the wide range of thresholds and techniques employed.

In terms of interpreting the results and their potential implications for child health,

there are also some limitations to highlight. One of them is that blood samples were not collected under fasting conditions. However, as argued in the articles, we now know that this is not a problem in the case of lipids. Regarding glucose and insulin, the HOMA-IR index used to measure insulin resistance should also not imply a distortion issue in healthy individuals without diabetes. Article I is affected by the use of thresholds to define risk factors, which can potentially impact the results. It is also essential to be aware that in this study, we are examining the stability of extreme values of variables related to cardiovascular risk in a population of healthy children. Therefore, while these values may be extreme in this population, we cannot conclude that they are indicative of a pathology or even a future increased risk. Furthermore, it is important to consider the fact that maintaining a relatively high value of a risk factor over time may not be as crucial in predicting the development of a disease as a substantial increase in the value of these risk factors (for example a stable high body weight vs. a rapid increase in body weight).

Regarding strengths, the study is being among the first to address the analysis of cardiovascular-related variable tracking in children using quantile regression. This opens the door to the versatile use of this tool in this field, and allows for much more accuracy and richness tracking information. Furthermore, the use of multivariate models, in conjunction with the extensive data collection capabilities of the INMA project, allows for the incorporation of a wide range of information related to potential confounding factors of the relationships under study. This facilitates the control of the effects of numerous potential confounding variables to estimate the effects of primary interest. While it is impossible to isolate all potential confounding effects, cohort studies, combined with multivariate techniques, lend robustness to this aspect of the research. The age of children involved is also a key point, as not many studies addressing cardiovascular risk factors examine children as young as 4 years of age. Moreover, the details and discussion provided in Article III regarding the use of quantile regression and multiple imputation can be useful to other researchers interested in both topics.

4.4. Public health implications

In order to translate the conclusions here presented to real public health considerations we have to consider the following points. It is crucial to emphasize the importance of early identification of cardiovascular risk factors in children. Prevention starting early in life is likely to have a substantial impact on reducing disease incidence and its associated effects

at the personal, economic, and social levels. Our study highlights the fact that children who were free from risk factors at age 4, remained so at age 8. That emphasizes the importance of maintaining a healthy state from early childhood, could been preventive of altered values in variables indicative of cardiovascular risk factors since such an early ages as 4 years old.

The concept of "tracking" of risk factors at an early age suggests that some children may be at higher risk of developing cardiovascular issues as they grow. This underscores the need for continuous monitoring and intervention strategies for these high-risk groups. As a result, the findings presented in this dissertation can help to identify which cardiovascular-related measures should be the focus of screening and monitoring in children. According to our results, central obesity and dyslipidemia emerge as potential areas of concern at early ages. Children with extreme values in these factors at a young age may face greater challenges in normalizing these values. Additionally, concerning dyslipidemia, if values of related variables in early ages, such as 4 years old, continue to increase in future generations, this could lead to even higher values at age 8. With the potential consequences that a substantial increase in extremely high values of dyslipidemia-related variables in childhood could have on the development of future disease in adulthood. The results related to central obesity are of particular relevance due to its association with chronic inflammation, as well as metabolic and vascular alterations that predispose individuals to the development of metabolic syndrome (Lurbe and Ingelfinger, 2021). Moreover, considering the previously mentioned high prevalence of obesity and overweight, which affects both children in Asturias and, in particular, the INMA-Asturias cohort.

This comprehensive analysis provides essential insights into the early identification and monitoring of cardiovascular risk factors in children, offering a foundation for more effective prevention and intervention strategies.

4.5. Future research

This dissertation assesses tracking within the INMA-Asturias cohort between the ages of 4 and 8. Our findings suggest tracking of cardiovascular-related variables that act as indicators for dyslipidemia and central obesity. Furthermore, cardiovascular risk factors in childhood have been shown to have an age-dependent impact on adult cardiovascular health, with predictability for subclinical atherosclerosis as early as age 9 (Juonala et al.,

2010). That suggests compelling reasons to expand our investigations into subsequent age groups, so future research should consider examining tracking in age groups from 8 years old to adolescence and early adulthood. It would also be pertinent to study which modifiable factors are influencing or altering the tracking of cardiovascular risk factors. These studies will offer valuable insights for prevention strategies and public health, and been a focal point for future research and public health policies.

CHAPTER 5

Conclusions

1) No clearly discernible patterns of aggregation in cardiovascular risk factors were identified at 4 and 8 years of age. Despite this, nearly all the children who did not exhibit any risk factors at age 4 did not exhibit them at age 8 either.

2) Tracking was observed between the ages of 4 and 8 years in children from the INMA-Asturias cohort. Particularly in the highest quantiles of the distribution of cardiovascular-related measures that serve as adult markers for dyslipidemia and central obesity. Remarkable, this was evident for the hight density lipoprotein cholesterol and the atherogenic coefficient in the context of dyslipidemia, and for the waist circunference/height ratio for central obesity.

3) The phenomenon of tracking appears to exhibit greater effect at higher quantiles of the atherogenic coefficient distribution, indicating possibles challenges in normalizing extreme values of that cardiovascular-related variable.

4) The complexity of the statistical analysis decisions, and the multitude of available alternatives often obscure the potential impact of the decision-making process. Therefore, it is imperative to precisely document and report the methodology employed, including both its limitations and strengths.

5) In the presence of missing data, multiple imputation methods are generally useful for mitigating biases in estimates. However, it is important to note that they may

not necessarily enhance the precision of standard error estimates. So using MI only to maximize the sample size is not always successful when it is correctly performed.

6) Quantile regression emerges as a potent tool for addressing challenges related to controversial threshold definitions, and tracking analyses in cohort studies. It provides valuable additional insights. Given its robust capabilities, it should be considered in the analysis of continuous outcomes, serving as an initial step for informing future modelling decisions.

Conclusiones

1) No se identificaron patrones claramente discernibles de agregación en los factores de riesgo cardiovascular a las edades de 4 y 8 años. A pesar de esto, prácticamente todos los niños que no presentaban factores de riesgo a los 4 años tampoco los presentaban a los 8 años.

2) Se observó seguimiento entre los 4 y 8 años en niños de la cohorte INMA-Asturias en los cuantiles más altos de la distribución de las medidas que sirven de marcadores de dislipidemia y de obesidad central. En concreto se observó en la distribución del colesterol de lipoproteínas de alta densidad y del coeficiente aterogénico, y de la relación circunferencia de la cintura/estatura.

3) El fenómeno de seguimiento parece mostrar un mayor efecto en los cuantiles más altos de la distribución del coeficiente aterogénico, lo que indica una posible dificultad en la normalización de valores extremos de esa variable.

4) La complejidad de las decisiones llevadas a cabo a la hora de realizar un análisis estadístico, así como la multitud de alternativas disponibles, a menudo dificultan discernir su potencial impacto. Por lo tanto, es de gran relevancia y pertinencia documentar y reportar con precisión la metodología empleada, incluyendo tanto sus limitaciones como sus fortalezas.

5) Los métodos de de imputación mútiple son generalmente útiles para mitigar posibles sesgos en las estimaciones en presencia de datos ausentes. Sin embargo, es importante destacar que no necesariamente mejoran la precisión de las estimaciones de los errores estándar. Por lo que sí su uso se centra en maximizar el tamaño muestral, no siempre resultará exitoso, cuando la imputación se ha realizado correctamente.

6) La regresión cuantil constituye una herramienta de gran valor en la realización de análisis donde se deben definir grupos de riesgo, pero no existen puntos de corte claramente definidos que permitan discriminar dichos grupos. También aporta gran valor en la realización de análisis de seguimiento en estudios de cohortes. Además, teniendo en cuenta sus ventajas sobre otros modelos de regresión tradicionales,

debería considerarse un paso inicial en el análisis de variables continuas, sirviendo para tomar futuras decisiones con respecto al análisis estadístico.

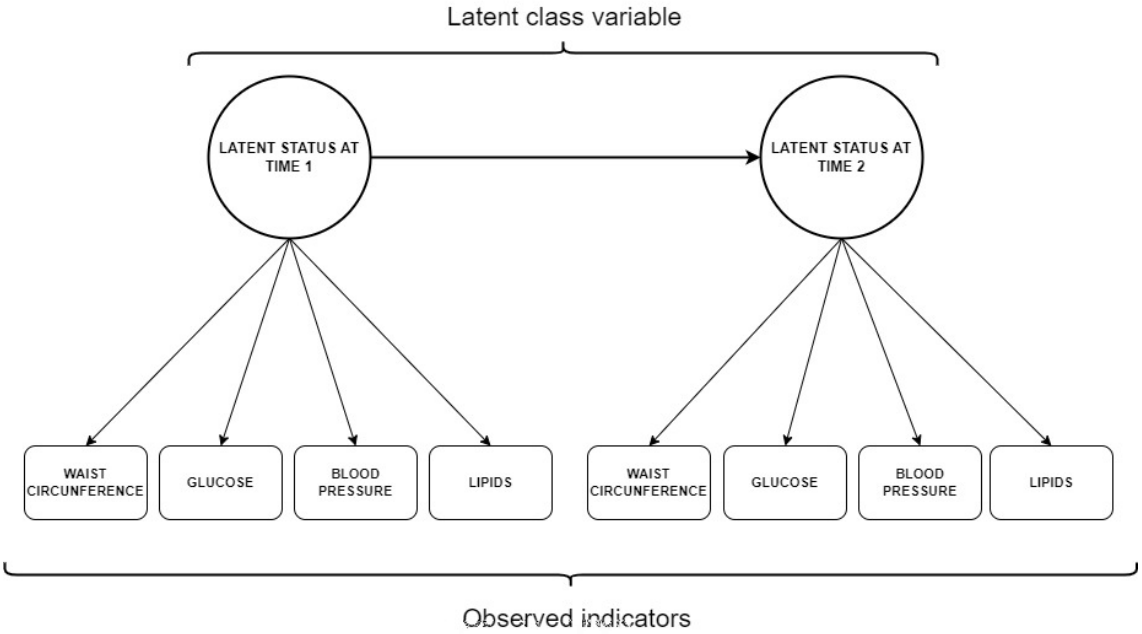
Supplementary material

Article I: Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort

To cite this article:
R. Fernández-Iglesias, A. Fernández-Somoano, C. Rodríguez-Dehli, R. Venta-Obaya, I. Riaño-Galán, and A. Tardón. Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort. *PLoS ONE*, 18(4), 2023. DOI: 10.1371/journal.pone.0283799.

To link to this supplementary material:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283799#sec014

Supplementary material



S1 Figure. Graphical representation of LTA model.

Supplementary material S1: Description of the process for specifying, estimating, and selecting the final LTA model.

In the first step, models from two to five latent statuses were estimated. To be able to interpret the latent status in a meaningful way, good homogeneity and latent class separation were considered to select the number of latent statuses in the model. Entropy was also used to select the best model as a measure of classification uncertainty (this measure can range from 0 to 1, with higher values representing a better fit, and > 0.7 is considered acceptable) (28). Akaike information criterion and Bayesian information criterion values were considered. The prevalence of the different latent statuses was not much lower than 10%, to avoid detecting excessively minority patterns.

The second and third steps involved testing the hypothesis of longitudinal measurement invariance and the hypothesis of change between time points invariance, respectively, using the likelihood ratio test. In the longitudinal measurement invariance hypothesis, the identified latent status is the same at 4 and 8 years of age. If this hypothesis is accepted, the latent status at 4 and 8 years is forced to be the same, which constrains the item-response probabilities to be the same at both time points. In contrast, if this hypothesis is rejected, the latent status can be assumed to be different at 4 and 8 years, and item-response probabilities can be freely estimated. The longitudinal measurement invariance hypothesis should usually be assumed to make

the model easier to interpret because fewer parameters need to be estimated under this assumption. However, this study was exploratory in nature, and we did not assume an a priori hypothesis on the behavior of the latent status in children. Therefore, this hypothesis was tested. In the change between time points invariance hypothesis, children who are in a latent status at 4 years of age will be in the same latent status at 8 years of age, without the possibility of change. If this hypothesis is accepted, the transition probabilities are constrained to be equal to 0, and if it is rejected, the transition probabilities are freely estimated.

After this process, the three latent status model was selected. Although this model did not have the best information criterion values (Akaike information criterion and Bayesian information criterion) and the highest entropy, it had good homogeneity, latent class separation, and interpretability. A summary of information on the relative model fit for selecting the number of latent statuses is shown in Supplementary Table S2.

The hypothesis of longitudinal measurement invariance was rejected (p = 0.009, see S2 Table). Therefore, although constraining the item-response probabilities to be equal across times would make interpreting the model easier, it would not be a reasonable assumption and would not capture the underlying structure of the data. On the basis that the latent status was different at the two time points, the hypothesis of change between time points invariance was also rejected because children who were in a latent status at T0 inevitably changed to another latent status at T1. Therefore, transition probabilities between T0 and T1 were different from zero.

S1 Table. Number and percentage of children with zero, one, two, three or four risk factors at the monitoring level.

Nº of risk factors at the monitoring level	Disorders	T0		T1	
		N	%	N	%
0	WC- G- BP- LIP-	76	29.0%	92	36.9%
1	WC+ G- BP- LIP-	25	9.5%	18	7.2%
1	WC- G+ BP- LIP-	17	6.5%	5	2.0%
1	WC- G- BP+ LIP-	32	12.2%	51	20.5%
1	WC- G- BP- LIP+	35	13.4%	31	12.4%
2	WC+ G+ BP- LIP-	6	2.3%	2	0.8%
2	WC+ G- BP+ LIP-	8	3.1%	16	6.4%
2	WC+ G- BP- LIP+	18	6.9%	10	4.0%
2	WC- G+ BP+ LIP-	7	2.7%	4	1.6%
2	WC- G+ BP- LIP+	8	3.1%	2	0.8%
2	WC- G- BP+ LIP+	8	3.1%	7	2.8%
3	WC+ G+ BP+ LIP-	3	1.1%	0	0.0%
3	WC- G+ BP+ LIP+	3	1.1%	1	0.4%
3	WC+ G- BP+ LIP+	7	2.7%	8	3.2%
3	WC+ G+ BP- LIP+	8	3.1%	1	0.4%
4	WC+ G+ BP+ LIP+	1	0.4%	1	0.4%

WC+: Waist circumference at the monitoring level. WC-: Waist circumference at the normal level.

G+: Blood glucose at the monitoring level. G-: Blood glucose at the normal level.

BP+: Blood pressure at the monitoring level. BP-: Blood pressure at the normal level.

Lip+: Lipid levels at the monitoring level. Lip-: Lipid levels at the normal level.

S2 Table. Summary of information for selecting the number of latent status and fit statistics for test the hypothesis of measurement invariance.

Measurement invariance	Number of latent status	Number of paremeters estimated	G^2	df	AIC	BIC	LL	Entropy	p-value*
Yes	2	11	201.7	241	2477.9	2519.7	-1227.9	0.791	
No	2	19	168.1	234	2457.6	2529.9	-1209.8	0.800	<0.001
Yes	3	20	174.3	233	2466.2	2542.3	-1213.1	0.688	
No	3	32	145.9	221	2460.7	2582.5	-1198.4	0.767	<0.001
Yes	4	31	150.9	222	2465.1	2583.0	-1201.5	0.842	
No	4	47	127.9	206	2470.9	2649.8	-1188.5	0.855	0.009
Yes	5	44	130.6	209	2470.9	2638.3	-1191.5	0.819	
No	5	64	108.5	189	2485.4	2728.9	-1178.7	0.840	0.19

*p-value obtained from the chi-square difference test based on loglikelihood values for testing the hypothesis of measurement invariance across times. G^2, likelihood-ratio statistic. df, degrees of freedom. AIC, Akaike information criterion. BIC, Bayesian information criterion. LL, loglikelohood value.

Article II: Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA-Asturias cohort

To cite this article:
R. Fernández-Iglesias, P. Martinez-Camblor, A. Fernández-Somoano, C. Rodríguez- Dehli, R. Venta-Obaya, M. R. Karagas, A. Tardón, and I. Riaño-Galán. Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA- Asturias cohort. *European Journal of Pediatrics*, 2023. DOI: 10.1007/S00431-023-05051-8.

To link to this supplementary material:
https://link.springer.com/article/10.1007/s00431-023-05051-8#Sec21

Table S1. Number of children with missing data for each variable, expressed in absolute and relative frequencies, for the final analysis sample composed by 307 children.

	n = 307
Parental characteristics	
Mother age at delivery (years)	0 (0.0%)
Mother origin country	0 (0.0%)
Mother level of education	0 (0.0%)
Mother social class	1 (0.3%)
Mother smoking during pregnancy	15 (4.9%)
Mother BMI (kg/m^2)	0 (0.0%)
Father BMI (kg/m^2)	11 (3.6%)
Parental cardiovascular antecedents	0 (0.0%)
Child characteristics	
Sex	0 (0.0%)
Age (years)	5 (1.6%)
Week of gestation at delivery	6 (2.0%)
Predominant breastfeeding duration (weeks)	37 (12.1%)
Birth weight (gr)	6 (2.0%)
At 4 years	
Mean daily energy intake (calories)	8 (2.6%)
Weekly out-of-school physical activity time (hours)	12 (3.9%)
Weight (kg)	6 (2.0%)
Height (cm)	6 (2.0%)
BMI (kg/m^2)	6 (2.0%)
Waist circunference (cm)	7 (2.3%)
Waist circunference/Height ratio	7 (2.3%)
Triponderal index (kg/m^3)	6 (2.0%)
Sistolic blood pressure (mmHg)	7 (2.3%)
Diastolic blood pressure (mmHg)	7 (2.3%)
Mean arterial pressure (mmHg)	7 (2.3%)
Total cholesterol (mg/dL)	60 (19.5%)
High density lipoprotein cholesterol (mg/dL)	60 (19.5%)
Low density lipoprotein cholesterol (mg/dL)	60 (19.5%)
Triglycerides (mg/dL)	60 (19.5%)
Atherogenic index	60 (19.5%)
Glucose (mg/dL)	59 (19.2%)
Insulin (μU/mL)	77 (25.1%)
HOMA index	77 (25.1%)
At 8 years	
Mean daily energy intake (calories)	21 (6.8%)
Weekly out-of-school physical activity time (hours)	19 (6.2%)
Weight (kg)	15 (4.9%)

Height (cm)	15 (4.9%)
BMI (kg/m^2)	15 (4.9%)
Waist circunference (cm)	15 (4.9%)
Waist circunference/Height ratio	15 (4.9%)
Tri-ponderal index (kg/m^3)	15 (4.9%)
Sistolic blood pressure (mmHg)	19 (6.2%)
Diastolic blood pressure (mmHg)	19 (6.2%)
Mean arterial pressure (mmHg)	19 (6.2%)
Total cholesterol (mg/dL)	61 (19.9%)
High density lipoprotein cholesterol (mg/dL)	63 (20.5%)
Low density lipoprotein cholesterol (mg/dL)	63 (20.5%)
Triglycerides (mg/dL)	61 (19.9%)
Atherogenic index	63 (20.5%)
Glucose (mg/dL)	61 (19.9%)
Insulin (μU/mL)	71 (23.1%)
HOMA-IR	72 (23.5%)

BMI, body mass index; HOMA-IR, Homeostatic Model Assessment for Insulin Resistance

Table S2. Association between the distribution of the cardiovascular-related measures at 8 years and the same cardiovascular-related measures rank at 4 years, using the imputed data.

			Crude		Adjusted	
Dependent variable	**Independent variable**	**Quantile**	**β^***	**95% CI**	**β^***	**95% CI**
TG at 8 years	**Rank TG at 4 years**	0.1	0.37	(-0.41, 1.15)	0.53	(-0.71, 1.76)
		0.2	0.99	(-0.09, 2.07)	0.78	(-0.71, 2.26)
		0.3	1.55	(0.02, 3.08)	1.13	(-0.61, 2.87)
		0.4	1.67	(-0.21, 3.55)	1.60	(-0.27, 3.47)
		0.5	1.76	(0.37, 3.16)	2.02	(0.18, 3.87)
		0.6	2.34	(0.23, 4.44)	2.28	(0.13, 4.43)
		0.7	3.89	(0.92, 6.86)	3.16	(0.40, 5.93)
		0.8	4.63	(1.24, 8.02)	4.21	(1.01, 7.41)
		0.9	6.29	(1.29, 11.29)	5.82	(1.00, 10.65)
HDL-c at 8 years	**Rank HDL-c at 4 years**	0.1	1.85	(1.02, 2.68)	1.93	(1.03, 2.84)
		0.2	2.20	(1.38, 3.01)	2.11	(1.27, 2.95)
		0.3	2.24	(1.35, 3.13)	2.34	(1.47, 3.20)
		0.4	2.41	(1.56, 3.25)	2.44	(1.53, 3.35)
		0.5	2.64	(1.69, 3.59)	2.61	(1.66, 3.56)
		0.6	3.02	(2.21, 3.82)	2.68	(1.75, 3.62)
		0.7	2.88	(2.23, 3.54)	2.72	(1.89, 3.54)
		0.8	2.87	(2.16, 3.59)	2.71	(1.91, 3.51)
		0.9	2.84	(1.73, 3.94)	2.93	(1.98, 3.87)
AC at 8 years	**Rank AC at 4 years**	0.1	0.07	(0.04, 0.10)	0.08	(0.05, 0.11)
		0.2	0.09	(0.07, 0.12)	0.09	(0.06, 0.12)
		0.3	0.11	(0.09, 0.13)	0.10	(0.07, 0.12)
		0.4	0.11	(0.09, 0.13)	0.11	(0.08, 0.13)
		0.5	0.11	(0.09, 0.14)	0.11	(0.09, 0.14)
		0.6	0.12	(0.09, 0.15)	0.11	(0.09, 0.14)
		0.7	0.12	(0.09, 0.16)	0.12	(0.08, 0.15)
		0.8	0.14	(0.09, 0.19)	0.13	(0.09, 0.17)
		0.9	0.18	(0.12, 0.24)	0.15	(0.09, 0.21)
WC/Height at 8 years	**Rank WC/Height at 4 years**	0.1	0.006	(0.003, 0.009)	0.005	(0.002, 0.008)
		0.2	0.006	(0.003, 0.009)	0.007	(0.004, 0.010)
		0.3	0.008	(0.006, 0.009)	0.007	(0.005, 0.010)
		0.4	0.008	(0.006, 0.010)	0.007	(0.005, 0.010)
		0.5	0.010	(0.006, 0.013)	0.008	(0.005, 0.010)
		0.6	0.010	(0.008, 0.013)	0.007	(0.005, 0.010)
		0.7	0.011	(0.009, 0.014)	0.008	(0.005, 0.010)
		0.8	0.012	(0.009, 0.016)	0.008	(0.005, 0.012)

		0.9	0.014	(0.010, 0.018)	0.008	(0.004, 0.012)
HOMA-IR at 8 years	**Rank HOMA-IR at 4 years**	0.1	0.004	(-0.038, 0.047)	0.016	(-0.037, 0.068)
		0.2	0.024	(-0.027, 0.074)	0.041	(-0.014, 0.097)
		0.3	0.037	(-0.012, 0.086)	0.050	(-0.008, 0.108)
		0.4	0.042	(-0.016, 0.101)	0.045	(-0.017, 0.107)
		0.5	0.033	(-0.009, 0.076)	0.043	(-0.032, 0.119)
		0.6	0.038	(-0.059, 0.136)	0.037	(-0.058, 0.131)
		0.7	0.034	(-0.118, 0.185)	0.036	(-0.100, 0.172)
		0.8	0.026	(-0.254, 0.306)	0.040	(-0.166, 0.246)
		0.9	0.010	(-0.444, 0.465)	0.067	(-0.312, 0.445)
MAP at 8 years	**Rank MAP at 4 years**	0.1	0.29	(-0.38, 0.96)	0.21	(-0.62, 1.05)
		0.2	0.65	(0.11, 1.20)	0.50	(-0.05, 1.04)
		0.3	0.69	(0.21, 1.18)	0.60	(0.15, 1.05)
		0.4	0.73	(0.37, 1.10)	0.66	(0.25, 1.07)
		0.5	0.71	(0.38, 1.03)	0.60	(0.23, 0.97)
		0.6	0.67	(0.34, 1.00)	0.55	(0.15, 0.96)
		0.7	0.61	(0.14, 1.09)	0.48	(-0.02, 0.99)
		0.8	0.52	(0.03, 1.02)	0.37	(-0.10, 0.84)
		0.9	0.51	(0.02, 1.00)	0.30	(-0.22, 0.81)

Quantile regression models with cardiovascular-related measure at 8 years as dependent variable and the rank variable of the corresponding cardiovascular-related measure at 4 years as the independent variable, for the quantiles between 0.1 to 0.9, with increments of 0.1, adjusted for mother age at delivery, mother level of education, mother social class, mother smoking during pregnancy, mother pre-pregnancy body mass index, father body mass index, parental cardiovascular antecedents, child sex, child mean daily energy intake at 4 and 8 years, child weekly out-of-school physical activity time at 4 and 8 years, week of gestation at delivery, weeks of predominant breastfeeding, and child height at 4 and 8 years. *Coefficient estimated are calculated with the independent variables in terms of percentiles (not quantiles) and they represent the effect on the dependent variable for each 1-decile increase in the independent variable.

Table S3. Association between the distribution of the cardiovascular-related measures at 8 years and all the cardiovascular-related measures rank at 4 years, using the imputed data.

			Crude		Adjusted	
Dependent variable	**Independent variable**	**Quantile**	**β**	**95% CI**	**β**	**95% CI**
TG at 8 years	**Rank TG at 4 years**	0.60	1.76	(-0.55, 4.06)	1.77	(-0.68, 4.23)
		0.75	3.48	(0.30, 6.67)	2.47	(-0.88, 5.83)
	Rank HDL-c at 4 years	0.60	0.07	(-3.05, 3.19)	0.81	(-2.34, 3.96)
		0.75	0.79	(-3.20, 4.78)	1.74	(-2.88, 6.35)
	Rank AC at 4 years	0.60	1.52	(-1.81, 4.86)	2.31	(-1.07, 5.69)
		0.75	2.38	(-1.43, 6.19)	3.35	(-1.31, 8.01)
	Rank WC/Height at 4 years	0.60	0.94	(-0.84, 2.72)	0.27	(-1.79, 2.34)
		0.75	2.21	(-0.55, 4.97)	0.77	(-2.18, 3.71)
	Rank HOMA-IR at 4 years	0.60	-0.32	(-2.24, 1.61)	0.15	(-2.07, 2.37)
		0.75	-0.22	(-3.04, 2.59)	0.25	(-2.78, 3.28)
	Rank MAP at 4 years	0.60	0.16	(-1.46, 1.77)	0.07	(-1.77, 1.91)
		0.75	-0.11	(-2.52, 2.29)	0.24	(-2.20, 2.68)
HDL-c at 8 years	**Rank TG at 4 years**	0.60	0.14	(-0.86, 1.14)	0.14	(-0.76, 1.03)
		0.75	0.21	(-0.64, 1.07)	0.07	(-0.82, 0.97)
	Rank HDL-c at 4 years	0.60	2.02	(0.37, 3.67)	1.76	(0.36, 3.17)
		0.75	2.35	(0.98, 3.71)	2.04	(0.79, 3.30)
	Rank AC at 4 years	0.60	-0.93	(-2.53, 0.67)	-1.00	(-2.27, 0.27)
		0.75	-0.64	(-2.13, 0.86)	-0.71	(-1.94, 0.52)
	Rank WC/Height at 4 years	0.60	-1.09	(-1.96, -0.22)	-0.87	(-1.79, 0.05)
		0.75	-0.93	(-1.81, -0.04)	-0.62	(-1.65, 0.41)
	Rank HOMA-IR at 4 years	0.60	0.09	(-0.87, 1.04)	-0.06	(-0.94, 0.82)
		0.75	-0.06	(-0.98, 0.86)	-0.05	(-0.94, 0.83)
	Rank MAP at 4 years	0.60	0.13	(-0.69, 0.94)	0.27	(-0.60, 1.15)
		0.75	0.24	(-0.58, 1.06)	0.49	(-0.39, 1.36)
AC at 8 years	**Rank TG at 4 years**	0.60	-0.01	(-0.04, 0.02)	-0.01	(-0.05, 0.02)
		0.75	-0.01	(-0.05, 0.03)	-0.01	(-0.05, 0.97)
	Rank HDL-c at 4 years	0.60	0.02	(-0.03, 0.06)	0.01	(-0.04, 0.06)
		0.75	0.00	(-0.06, 0.06)	0.00	(-0.06, 3.30)
	Rank AC at 4 years	0.60	0.13	(0.09, 0.16)	0.12	(0.07, 0.17)
		0.75	0.13	(0.06, 0.19)	0.13	(0.06, 0.52)
	Rank WC/Height at 4 years	0.60	0.03	(0.00, 0.06)	0.02	(-0.01, 0.05)
		0.75	0.04	(0.00, 0.07)	0.02	(-0.02, 0.41)
	Rank HOMA-IR at 4 years	0.60	0.01	(-0.02, 0.04)	0.00	(-0.03, 0.04)
		0.75	0.01	(-0.02, 0.04)	0.01	(-0.03, 0.83)

	Rank MAP at 4 years	0.60	0.01	(-0.02, 0.03)	0.00	(-0.03, 0.03)
		0.75	0.01	(-0.03, 0.04)	0.00	(-0.04, 1.36)
WC/Height at 8 years	**Rank TG at 4 years**	0.60	-0.001	(-0.004, 0.001)	-0.001	(-0.004, 0.001)
		0.75	-0.002	(-0.005, 0.001)	-0.002	(-0.005, 0.001)
	Rank HDL-c at 4 years	0.60	0.000	(-0.005, 0.004)	-0.001	(-0.005, 0.004)
		0.75	0.000	(-0.006, 0.005)	-0.001	(-0.006, 0.003)
	Rank AC at 4 years	0.60	0.002	(-0.002, 0.007)	-0.001	(-0.005, 0.004)
		0.75	0.002	(-0.002, 0.007)	0.000	(-0.006, 0.005)
	Rank WC/Height at 4 years	0.60	0.010	(0.007, 0.013)	0.008	(0.005, 0.011)
		0.75	0.012	(0.009, 0.015)	0.008	(0.005, 0.011)
	Rank HOMA-IR at 4 years	0.60	0.002	(-0.002, 0.005)	0.001	(-0.002, 0.005)
		0.75	0.000	(-0.004, 0.005)	0.001	(-0.002, 0.005)
	Rank MAP at 4 years	0.60	0.000	(-0.002, 0.002)	-0.001	(-0.003, 0.001)
		0.75	0.000	(-0.003, 0.003)	-0.001	(-0.004, 0.002)
HOMA-IR at 8 years	**Rank TG at 4 years**	0.60	-0.005	(-0.100, 0.090)	-0.012	(-0.114, 0.089)
		0.75	0.029	(-0.145, 0.203)	-0.021	(-0.185, 0.142)
	Rank HDL-c at 4 years	0.60	0.086	(-0.042, 0.215)	0.077	(-0.082, 0.237)
		0.75	0.111	(-0.162, 0.385)	0.125	(-0.125, 0.374)
	Rank AC at 4 years	0.60	0.100	(-0.021, 0.221)	0.103	(-0.061, 0.266)
		0.75	0.140	(-0.132, 0.411)	0.151	(-0.110, 0.413)
	Rank WC/Height at 4 years	0.60	0.070	(-0.011, 0.151)	0.065	(-0.041, 0.171)
		0.75	0.063	(-0.119, 0.245)	0.062	(-0.115, 0.238)
	Rank HOMA-IR at 4 years	0.60	0.027	(-0.062, 0.117)	0.036	(-0.065, 0.136)
		0.75	0.007	(-0.185, 0.200)	0.034	(-0.150, 0.219)
	Rank MAP at 4 years	0.60	-0.025	(-0.111, 0.062)	-0.014	(-0.11, 0.081)
		0.75	-0.072	(-0.272, 0.129)	-0.024	(-0.199, 0.151)
MAP at 8 years	**Rank TG at 4 years**	0.60	-0.06	(-0.41, 0.30)	-0.07	(-0.52, 0.38)
		0.75	0.06	(-0.41, 0.54)	-0.04	(-0.56, 0.49)
	Rank HDL-c at 4 years	0.60	-0.56	(-1.32, 0.21)	-0.44	(-1.23, 0.34)
		0.75	-0.42	(-1.09, 0.26)	-0.35	(-1.15, 0.44)
	Rank AC at 4 years	0.60	-0.36	(-1.08, 0.36)	-0.31	(-1.08, 0.47)
		0.75	-0.07	(-0.71, 0.57)	-0.07	(-0.87, 0.74)
	Rank WC/Height at 4 years	0.60	0.14	(-0.26, 0.53)	0.07	(-0.39, 0.52)
		0.75	0.10	(-0.35, 0.55)	0.08	(-0.44, 0.60)
	Rank HOMA-IR at 4 years	0.60	0.24	(-0.15, 0.63)	0.23	(-0.20, 0.66)
		0.75	0.25	(-0.20, 0.70)	0.33	(-0.15, 0.81)
	Rank MAP at 4 years	0.60	0.62	(0.28, 0.97)	0.63	(0.22, 1.05)
		0.75	0.64	(0.25, 1.02)	0.59	(0.08, 1.09)

Quantile regression models with cardiovascular-related measure at 8 years as dependent variable and all the cardiovascular-related measure rank at 4 years as the independent variables, for the quantiles 0.60 and 0.75, adjusted for mother age at delivery, mother level of education, mother social class, mother smoking during pregnancy, mother pre-pregnancy body mass index, father body mass index, parental cardiovascular antecedents, child sex, child mean daily energy intake at 4 and 8 years, child weekly out-of-school physical activity time at 4 and 8 years, week of gestation at delivery, weeks of predominant breastfeeding, and child height at 4 and 8 years. *Coefficient estimated are calculated with the independent variables in terms of percentiles (not quantiles) and they represent the effect on the dependent variable for each 1-decile increase in the independent variable.

Article III: Statistical considerations for analyzing data derived from long longitudinal cohort studies

To cite this article:
R. Fernández-Iglesias, P. Martínez-Camblor, A. Tardón, and A. Fernández-Somoano. Statistical considerations for analyzing data derived from long longitudinal cohort studies. *Mathematics*, 11(19), 2023.
DOI: 10.3390/math11194070.

To link to this article:
https://www.mdpi.com/2227-7390/11/19/4070

Table S1: Number of participants with missing data for each variable, expressed in absolute and relative frequencies, for the final sample composed by 307 children.

	n = 307
Parental characteristics	
Mother age at delivery (years)	0 (0.0%)
Mother origin country	0 (0.0%)
Mother level of education	0 (0.0%)
Mother social class	1 (0.3%)
Mother smoking during pregnancy	15 (4.9%)
Mother BMI (kg/m2)	0 (0.0%)
Father BMI (kg/m2)	11 (3.6%)
Parental cardiovascular antecedents	0 (0.0%)
Child characteristics	
Sex	0 (0.0%)
Age (years)	5 (1.6%)
Week of gestation at delivery	6 (2.0%)
Predominant breastfeeding duration (weeks)	37 (12.1%)
Birth weight (gr)	6 (2.0%)
At 4 years	
Mean daily energy intake (cals)	8 (2.6%)
Weekly out-of-school physical activity time (hours)	12 (3.9%)
Weight (kg)	6 (2.0%)
Height (cm)	6 (2.0%)
BMI (kg/m2)	4 (1.3%)
Waist circunference (cm)	7 (2.3%)
Waist circunference/Height ratio	7 (2.3%)
Triponderal index (kg/m3)	6 (2.0%)
Sistolic blood pressure	7 (2.3%)
Diastolic blood pressure	7 (2.3%)
Mean arterial pressure	7 (2.3%)
Total cholesterol (mg/dL)	60 (19.5%)
cHDL (mg/dL)	60 (19.5%)
cLDL (mg/dL)	60 (19.5%)
Triglycerides (mg/dL)	60 (19.5%)
Atherogenic index	60 (19.5%)
Glucose (mg/dL)	59 (19.2%)
Insulin (μU/mL)	77 (25.1%)
HOMA index	77 (25.1%)
At 8 years	
Mean daily energy intake (cals)	21 (6.8%)
Weekly out-of-school physical activity time (hours)	19 (6.2%)
Weight (kg)	15 (4.9%)
Height (cm)	15 (4.9%)
BMI (kg/m2)	15 (4.9%)
Waist circunference (cm)	15 (4.9%)
Waist circunference/Height ratio	15 (4.9%)
Triponderal index (kg/m3)	15 (4.9%)
Sistolic blood pressure	19 (6.2%)
Diastolic blood pressure	19 (6.2%)

Mean arterial pressure	19 (6.2%)
Total cholesterol (mg/dL)	61 (19.9%)
cHDL (mg/dL)	63 (20.5%)
cLDL (mg/dL)	63 (20.5%)
Triglycerides (mg/dL)	61 (19.9%)
Atherogenic index	63 (20.5%)
Glucose (mg/dL)	61 (19.9%)
Insulin (μU/mL)	71 (23.1%)
HOMA index	72 (23.5%)

List of Figures

List of Tables

Bibliography

W. Ahrens, L. Moreno, S. Mårild, D. Molnár, A. Siani, S. De Henauw, J. Böhmann, K. Günther, C. Hadjigeorgiou, L. Iacoviello, L. Lissner, T. Veidebaum, H. Pohlabeln, and I. Pigeot. Metabolic syndrome in young children: definitions and results of the IDEFICS study. *International Journal of Obesity*, 38(2):S4–S14, 2014. DOI: 10.1038/ijo.2014.130.

H. K. Akerblom, M. Uhari, E. Pesonen, M. Dahl, E. A. Kaprio, E. M. Nuutinen, M. Pietikäinen, M. K. Salo, A. Aromaa, and L. Kannas. Cardiovascular risk in young Finns. *Ann Med*, 23(1):35–39, 1991. DOI: 10.3109/07853899109147928.

B. T. Alexander, J. H. Dasinger, and S. Intapad. Fetal programming and cardiovascular pathology. *Comprehensive Physiology*, 5(2):997–1025, 2015. DOI: 10.1002/CPHY.C140036.

M. Amini, F. Zayeri, and M. Salehi. Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. *BMC Public Health*, 21(1):401, 2021. DOI: 10.1186/s12889-021-10429-0.

A. Bardají. REGICOR: 35 years of excellence in cardiovascular research. *Rev Esp Cardiol*, 66(12):923–925, 2013. DOI: 10.1016/j.rec.2013.07.008.

D. J. Barker. The fetal and infant origins of adult disease. *British Medical Journal*, 301: 1111, 1990. DOI: 10.1136/BMJ.301.6761.1111.

C. Bennette and A. Vickers. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, 12 (1):21, 2012. DOI: 10.1186/1471-2288-12-21.

G. S. Berenson, T. A. Foster, G. C. Frank, R. R. Frerichs, S. R. Srinivasan, A. W. Voors, and L. S. Webber. Cardiovascular disease risk factor variables at the preschool age. The Bogalusa heart study. *Circulation*, 57(3):603–612, 1978. DOI: 10.1161/01.cir.57.3.603.

F. J. Brunner, C. Waldeyer, F. Ojeda, V. Salomaa, F. Kee, S. Sans, B. Thorand, S. Giampaoli, and et al. Application of non-HDL cholesterol for population-based cardiovascular risk stratification: results from the multinational cardiovascular risk consortium. *Lancet*, 394(10215):2173–2183, 2019. DOI: 10.1016/S0140-6736(19)32519-X.

C. Canova and A. Cantarutti. Population-based birth cohort studies in epidemiology. *International Journal of Environmental Research and Public Health*, 17(15):1–6, 2020. DOI: 10.3390/ijerph17155276.

M. Casula, O. Colpani, S. Xie, A. L. Catapano, and A. Baragetti. HDL in atherosclerotic cardiovascular disease: in search of a role. *Cells*, 10(8):1869, 2021. DOI: 10.3390/cells10081869.

D. D. Celentano and S. Moyses. *Gordis Epidemiology*. Elsevier Inc., 6 edition, 2014. ISBN 987-84-9113-633-0.

M. Chacko, P. S. Sarma, S. Harikrishnan, G. Zachariah, and P. Jeemon. Family history of cardiovascular disease and risk of premature coronary heart disease: a matched case-control study. *Wellcome Open Res*, 12(5):70, 2020.

W. R. Clarke, H. G. Schrott, P. E. Leaverton, W. E. Connor, and R. M. Lauer. Tracking of blood lipids and blood pressures in school age children: the Muscatine study. *Circulation*, 58(4):626–634, 1978. DOI: 10.1161/01.cir.58.4.626.

L. M. Collins, J. L. Schafer, and C. M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedure. *Psychological Methods*, 6(3):330–351, 2001. DOI: 10.1037/1082-989X.6.4.330.

R. M. Conroy, K. Pyörälä, A. P. Fitzgerald, S. Sans, A. Menotti, G. De Backer, D. De Bacquer, P. Ducimetière, P. Jousilahti, U. Keil, I. Njølstad, R. G. Oganov, T. Thomsen, H. Tunstall-Pedoe, A. Tverdal, H. Wedel, P. Whincup, L. Wilhelmsen, I. M. Graham, and on behalf of the SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal*, 24 (11):987–1003, 2003. DOI: 10.1016/S0195-668X(03)00114-3.

J. B. Croft, L. S. Webber, F. C. Parker, and G. S. Berenson. Recruitment and participation of children in a long-term study of cardiovascular disease: the Bogalusa heart study, 1973-1982. *Am J Epidemiol*, 120(3):436–448, 1984. DOI: 10.1093/oxfordjournals.aje.a113908.

A. Denegri and G. Boriani. High sensitivity C-reactive protein (hsCRP) and its implications in cardiovascular outcomes. *Curr Pharm Des*, 27(2):263–275, 2021. DOI: 10.2174/1381612826666200717090334.

B. Domínguez Aurrecoechea, M. Sánchez Echenique, M. A. Ordóñez Alonso, J. I. Pérez Candás, and J. Delfrade Osinaga. Nutritional status of children population in Asturias (SNUPI-AS study): thinness, overweight, obesity and stunting. *Pediatria de Atencion Primaria*, 17(65):e21–e31, 2015. DOI: 10.4321/s1139-76322015000100006.

D. Drozdz, J. Alvarez-Pitti, M. Wójcik, C. Borghi, R. Gabbianelli, A. Mazur, V. Herceg-čavrak, B. G. Lopez-Valcarcel, M. Brzeziński, E. Lurbe, and E. Wühl. Obesity and cardiometabolic risk factors: from childhood to adulthood. *Nutrients*, 13(11):4176, 2021. DOI: 10.3390/NU13114176/S1.

F. Duarte Lau and R. P. Giugliano. Lipoprotein(a) and its significance in cardiovascular disease: A review. *JAMA Cardiol*, 7(7):760–769, 2022. DOI: 10.1001/jamacardio.2022.0987.

T. Dwyer and L. E. Gibbons. The australian schools health and fitness survey. Physical fitness related to blood pressure but not lipoproteins. *Circulation*, 89(4):1539–1544, 1994. DOI: 10.1161/01.cir.89.4.1539.

T. Dwyer, C. Sun, C. G. Magnussen, O. T. Raitakari, N. J. Schork, A. Venn, T. L. Burns, M. Juonala, J. Steinberger, A. R. Sinaiko, R. J. Prineas, P. H. Davis, J. G. Woo, J. A. Morrison, S. R. Daniels, W. Chen, S. R. Srinivasan, J. S. Viikari, and G. S. Berenson. Cohort profile: the international childhood cardiovascular cohort (i3C) consortium. *International Journal of Epidemiology*, 42(1):86–96, 2013. DOI: 10.1093/ije/dys004.

J. C. Eisenmann. On the use of a continuous metabolic syndrome score in pediatric research. *Cardiovascular Diabetology*, 7(17):1–6, 2008. DOI: 10.1186/1475-2840-7-17.

C. K. Enders. *Applied missing data analysis*. Guilford Press, New York, NY, US, 2010. ISBN 978-1-60623-639-0.

J. Enos, William F., J. C. Beyer, and R. H. Holmes. Pathogenesis of coronary disease in american soldiers killed in Korea. *Journal of the American Medical Association*, 158 (11):912–914, 1955. DOI: 10.1001/jama.1955.02960110018005.

J. Fan and T. Watanabe. Atherosclerosis: known and unknown. *Pathology international*, 72(3):151–160, 2022. DOI: 10.1111/pin.13202.

V. L. Feigin, B. A. Stark, C. O. Johnson, G. A. Roth, C. Bisignano, G. G. Abady, M. Abbasifard, M. Abbasi-Kangevari, and et al. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Neurology*, 20(10):1–26, 2021. DOI: 10.1016/S1474-4422(21)00252-0.

A. Fernández-Somoano and A. Tardon. Socioeconomic status and exposure to outdoor NO2 and benzene in the Asturias INMA birth cohort, Spain. *Journal of Epidemiology and Community Health*, 68(1):29–36, 2014. DOI: 10.1136/JECH-2013-202722.

G. Flores-Mateo, M. Grau, M. O'Flaherty, R. Ramos, R. Elosua, C. Violan-Fors, M. Quesada, R. Martí, J. Sala, J. Marrugat, and S. Capewell. Análisis de la disminución de la mortalidad por enfermedad coronaria en una población mediterránea: España 1988-2005. *Revista Española de Cardiología*, 64(11):988–996, 2011. DOI: 10.1016/j.recesp.2011.05.033.

M. A. Foulkes and C. E. Davis. An index of tracking for longitudinal data. *Biometrics*, 37(3):439–446, 1981. DOI: 10.2307/2530557.

R. F. Gillum, R. J. Prineas, G. Sopko, Y. Koga, W. Kubicek, N. M. Robitaille, J. Bass, and A. Sinaiko. Elevated blood pressure in school children—prevalence, persistence, and hemodynamics: the Minneapolis children's blood pressure study. *American Heart Journal*, 105(2):316–322, 1983. DOI: https://doi.org/10.1016/0002-8703(83)90533-1.

R. Guieu, J. Ruf, and G. Mottola. Hyperhomocysteinemia and cardiovascular diseases. *Ann Biol Clin (Paris)*, 80(1):7–14, 2022. DOI: 10.1684/abc.2021.1694.

M. Guxens, F. Ballester, M. Espada, M. F. Fernández, J. O. Grimalt, J. Ibarluzea, N. Olea, M. Rebagliato, A. Tardón, M. Torrent, J. Vioque, M. Vrijheid, and J. Sunyer. Cohort profile: the INMA-INfancia y Medio Ambiente-(environment and childhood) project. *International Journal of Epidemiology*, 41:930–940, 2012. DOI: 10.1093/ije/dyr054.

R. F. Heller, S. Chinn, H. D. Pedoe, and G. Rose. How well can we predict coronary heart disease? findings in the United Kingdom heart disease prevention project. *Br Med J (Clin Res Ed)*, 288(6428):1409–1411, 1984. DOI: 10.1136/bmj.288.6428.1409.

W.-M. Ho, Y.-Y. Wu, and Y.-C. Chen. Genetic variants behind cardiovascular diseases and dementia. *Genes (Basel)*, 11(12):1514, 2020. DOI: 10.3390/genes11121514.

Instituto Nacional de Estadística. *Defunciones según la causa de muerte. Año 2022. Datos provisionales*, June 2022.

Instituto Nacional de Estadística. *Encuesta de morbilidad hospitalaria. Año 2021*, March 2023.

D. R. Jacobs, J. G. Woo, A. R. Sinaiko, S. R. Daniels, J. Ikonen, M. Juonala, N. Kartiosuo, T. Lehtimäki, C. G. Magnussen, J. S. Viikari, N. Zhang, L. A. Bazzano, T. L. Burns, R. J. Prineas, J. Steinberger, E. M. Urbina, A. J. Venn, O. T. Raitakari, and T. Dwyer. Childhood cardiovascular risk factors and adult cardiovascular events. *New England Journal of Medicine*, 386(20):1877–1888, 2022. DOI: 0.1056/NEJMoa2109191.

M. Juonala, C. G. Magnussen, A. Venn, T. Dwyer, T. L. Burns, P. H. Davis, W. Chen, S. R. Srinivasan, S. R. Daniels, M. Kähönen, T. Laitinen, L. Taittonen, G. S. Berenson, J. Viikari, and O. T. Raitakari. Influence of age on associations between childhood risk factors and carotid intima-media thickness in adulthood. *Circulation*, 122(24): 2514–2520, 2010. DOI: 10.1161/CIRCULATIONAHA.110.966465.

T. Kapwata, · . Caradee, Y. Wright, T. Reddy, R. Street, Z. Kunene, and A. Mathee. Relations between personal exposure to elevated concentrations of arsenic in water and soil and blood arsenic levels amongst people living in rural areas in Limpopo, South Africa. *Environmental Science and Pollution Research*, 30(24):65204–65216, 2023. DOI: 10.1007/s11356-023-26813-9.

A. Keys, A. Menotti, C. Aravanis, H. Blackburn, B. S. Djordevic, R. Buzina, A. S. Dontas, F. Fidanza, M. J. Karvonen, and N. Kimura. The seven countries study: 2,289 deaths in 15 years. *Prev Med*, 13(2):141–154, 1984. DOI: 10.1016/0091-7435(84)90047-1.

M. Kocherginsky, X. He, and Y. Mu. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14(1):41–55, 2005. DOI: 10.1198/106186005X27563.

R. Koenker. *Quantile regression*. Econometric Society monographs ; 38. Cambridge University Press, Cambridge, 2005. ISBN 1-107-71383-8.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. DOI: 10.2307/1913643.

R. M. Lauer, W. E. Connor, P. E. Leaverton, M. A. Reiter, and W. R. Clarke. Coronary heart disease risk factors in school children: the Muscatine study. *The Journal of Pediatrics*, 86(5):697–706, 1975. DOI: 10.1016/S0022-3476(75)80353-2.

P. Libby. The changing landscape of atherosclerosis. *Nature*, 592(7855):524–533, 2021. DOI: 10.1038/s41586-021-03392-8.

R. J. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988. DOI: 10.1080/01621459.1988.10478722.

J. F. Lobstein. *Traité d'anatomie pathologique*, volume v.1. Paris Levrault, 1829.

E. Lurbe and J. Ingelfinger. Developmental and early life origins of cardiometabolic risk factors. *Hypertension*, 77(2):308–318, 2021. DOI: 10.1161/HYPERTENSIONAHA.120.14592.

E. Lurbe, E. Agabiti-Rosei, J. K. Cruickshank, A. Dominiczak, S. Erdine, A. Hirth, C. Invitti, M. Litwin, G. Mancia, D. Pall, W. Rascher, J. Redon, F. Schaefer, T. Seeman, M. Sinha, S. Stabouli, N. J. Webb, E. Wühl, and A. Zanchetti. 2016 European society of hypertension guidelines for the management of high blood pressure in children and adolescents. *Journal of Hypertension*, 34(10):1887–1920, 2016. DOI: 10.1097/HJH.0000000000001039.

S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang. The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*, 383 (9921):999–1008, 2014. DOI: 10.1016/S0140-6736(13)61752-3.

F. Marchand. Ueber atherosclerosis, 1904.

P. McBride. Triglycerides and risk for coronary artery disease. *Curr Atheroscler Rep*, 10 (5):386–390, 2008. DOI: 10.1007/s11883-008-0060-9.

H. C. Mcgill, . C. A. Mcmahan, E. E. Herderick, G. T. Malcom, R. E. Tracy, and J. P. Strong. Origin of atherosclerosis in childhood and adolescence. *American Journal of Clinical Nutrition*, 72(5 Suppl):1307S–1315S, 2000. DOI: 10.1093/ajcn/72.5.1307s.

C. A. Mcmahan. An index of tracking. *Biometrics*, 37(3):447–455, 1981. DOI: 10.2307/2530558.

J. Milei, G. Ottaviani, A. M. Lavezzi, D. R. Grana, I. Stella, and L. Matturri. Perinatal and infant early atherosclerotic coronary lesions. *Can J Cardiol*, 24(2):137–41, 2008. DOI: 10.1016/s0828-282x(08)70570-1.

J. A. Morrison, L. A. Friedman, and C. Gray-McGuire. Metabolic syndrome in childhood predicts adult cardiovascular disease 25 years later: the Princeton lipid research clinics follow-up study. *Pediatrics*, 120(2):340–345, 2007. DOI: 10.1542/peds.2006-1699.

T. Partearroyo, M. d. L. Samaniego-Vaesken, E. Ruiz, J. Aranceta-Bartrina, Á. Gil, M. González-Gross, R. M. Ortega, L. Serra-Majem, and G. Varela-Moreiras. Sodium intake from foods exceeds recommended limits in the spanish population: The ANIBES study. *Nutrients 2019, Vol. 11, Page 2451*, 11(10):2451, 2019. DOI: 10.3390/NU11102451.

A. A. Peralta, J. Schwartz, D. R. Gold, J. M. Vonk, R. Vermeulen, and U. Gehring. Quantile regression to examine the association of air pollution with subclinical atherosclerosis in an adolescent population. *Environment International*, 164:107285, 2022. DOI: 10.1016/J.ENVINT.2022.107285.

G. Pérez-Gimeno, A. I. Rupérez, R. Vázquez-Cobela, G. Herráiz-Gastesi, M. Gil-Campos, C. M. Aguilera, L. A. Moreno, M. R. L. Trabazo, and G. Bueno-Lozano. Energy dense salty food consumption frequency is associated with diastolic hypertension in spanish children. *Nutrients*, 12(4), 2020. DOI: 10.3390/NU12041027.

L. R. Pool, L. Aguayo, M. Brzezinski, A. M. Perak, M. M. Davis, P. Greenland, L. Hou, B. S. Marino, L. Van Horn, L. Wakschlag, D. Labarthe, D. Lloyd-Jones, and N. B. Allen. Childhood risk factors and adulthood cardiovascular disease: a systematic review. *Journal of Pediatrics*, 232:118–126.e23, 2021. DOI: 10.1016/j.jpeds.2021.01.053.

E. Poortvliet, A. Yngve, U. Ekelund, A. Hurtig-Wennlöf, A. Nilsson, M. Hagströmer, and M. Sjöström. The European Youth Heart Survey (EYHS): an international study that

addresses the multi-dimensional issues of CVD risk factors. *Forum Nutr*, 56:254–256, 2003.

C. Reisinger, B. N. Nkeh-Chungag, P. M. Fredriksen, and N. Goswami. The prevalence of pediatric metabolic syndrome—a critical look on the discrepancies between definitions and its clinical importance. *International Journal of Obesity*, 45(1):12–24, 2020. DOI: 10.1038/s41366-020-00713-1.

I. Riaño-Galán, A. Fernández-Somoano, C. Rodríguez-Dehli, D. Valvi, M. Vrijheid, and A. Tardón. Proatherogenic lipid profile in early childhood: association with weight status at 4 years and parental obesity. *Journal of Pediatrics*, 187:153–157.e2, 2017. DOI: 10.1016/j.jpeds.2017.04.042.

F. Rodríguez-Artalejo, C. Garcés, A. Gil, M. A. Lasunción, J. M. Martín Moreno, L. Gorgojo, and M. de Oya. The 4 provinces study: its principal objectives and design. The researchers of the 4 provinces study. *Rev Esp Cardiol*, 52(5):319–326, 1999. DOI: 10.1016/s0300-8932(99)74922-5.

B. Rosner, C. H. Hennekens, E. H. Kass, and W. E. Miall. Age-specific correlation analysis of longitudinal blood pressure data. *American journal of epidemiology*, 106(4): 306–313, 1977. DOI: 10.1093/oxfordjournals.aje.a112466.

B. Rosner, N. R. Cook, S. Daniels, and B. Falkner. Childhood blood pressure trends and risk factors for high blood pressure: The NHANES experience 1988–2008. *Hypertension*, 62(2):247, 2013. DOI: 10.1161/HYPERTENSIONAHA.111.00831.

R. Ross. Atherosclerosis — an inflammatory disease. *New England Journal of Medicine*, 340(2):115–126, 1999. DOI: 10.1056/NEJM199901143400207.

G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Beaton, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021, 2020. DOI: 10.1016/j.jacc.2020.11.010.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. DOI: 10.2307/2335739.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley, 1987. ISBN 9780471655749.

I. B. S. Machado, M. R. Tofanelli, A. A. Saldanha da Silva, and A. C. Simões e Silva. Factors associated with primary hypertension in pediatric patients: An up-to-date. *Current Pediatric Reviews*, 17(1):15–37, 2021. DOI: 10.2174/1573396317999210111200222.

S. Sans, A. P. Fitzgerald, D. Royo, R. Conroy, and I. Graham. Calibrating the SCORE cardiovascular risk chart for use in Spain. *Revista Española de Cardiología*, 60(5): 476–485, 2007. DOI: 10.1016/S1885-5857(07)60188-1.

N. Sattar, A. Gaw, O. Scherbakova, I. Ford, D. S. J. O'Reilly, S. M. Haffner, C. Isles, P. W. Macfarlane, C. J. Packard, S. M. Cobbe, and J. Shepherd. Metabolic syndrome with and without C-reactive protein as a predictor of coronary heart disease and diabetes in the west of Scotland coronary prevention study. *Circulation*, 108(4):414–419, 2003. DOI: 10.1161/01.CIR.0000080897.52664.94.

J. L. Schafer. *Analysis of incomplete multivariate data*. Monographs on statistics and applied probability. Chapman Hall, 1997. ISBN 0412040611.

J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. DOI: 10.1037/1082-989X.7.2.147.

J. B. Soriano, D. Rojas-Rueda, J. Alonso, J. M. Antó, P.-J. Cardona, E. Fernández, A. L. Garcia-Basteiro, and et al. La carga de enfermedad en España: resultados del estudio de la carga global de las enfermedades 2016. *Medicina Clínica*, 151(5):171–190, 2018. DOI: 10.1016/j.medcli.2018.05.011.

N. H. Sternby, J. E. Fernandez-Britto, and P. Nordet. Pathobiological determinants of atherosclerosis in youth (PBDAY study), 1986-96. *Bulletin of the World Health Organization*, 77(3):250, 1999.

E. J. Stone. School-based health research funded by the national heart, lung, and blood institute. *J Sch Health*, 55(5):168–174, 1985. DOI: 10.1111/j.1746-1561.1985.tb04113.x.

H. Strasses. Aterosclerosis y cardiopatía coronaria: la contribución de la epidemiología. *Crónica de la OMS*, 26(1):7–12, 1972.

J. P. Strong and H. C. McGill. The pediatric aspects of atherosclerosis. *Journal of Atherosclerosis Research*, 9(3):251–265, 1969. DOI: 10.1016/S0368-1319(69)80020-7.

S. van Buuren. *Flexible Imputation of Missing Data*. Interdisciplinary Statistics. Chapman Hall/CRC, 2012. ISBN 1439868247.

J. R. van Ginkel, M. Linting, R. C. Rippe, and A. van der Voort. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 102(3):297–308, 2020. DOI: 10.1080/00223891.2018.1530680.

R. Virmani, A. P. Burke, A. Farb, and F. D. Kolodgie. Pathology of the vulnerable plaque. *Journal of the American College of Cardiology*, 47(8, Supplement):C13–C18, 2006. DOI: https://doi.org/10.1016/j.jacc.2005.10.065.

E. Waldmann. Quantile regression: a short story on how and why. *Statistical Modelling*, 18(3-4):203–218, 2018. DOI: 10.1177/1471082X18759142.

W. Wang, M. Hu, H. Liu, X. Zhang, H. Li, F. Zhou, Y. M. Liu, F. Lei, J. J. Qin, Y. C. Zhao, Z. Chen, W. Liu, X. Song, X. Huang, L. Zhu, Y. X. Ji, P. Zhang, X. J. Zhang, Z. G. She, J. Yang, H. Yang, J. Cai, and H. Li. Global burden of disease study 2019 suggests that metabolic risk factors are the leading drivers of the burden of ischemic heart disease. *Cell Metabolism*, 33(10):1943–1956.e2, 2021. DOI: 10.1016/j.cmet.2021.08.005.

J. H. Ware and M. C. Wu. Tracking: prediction of future values from serial measurements. *Biometrics*, 37(3):427, 1981. DOI: 10.2307/2530556.

Y. Wei, R. D. Kehm, M. Goldberg, and M. B. Terry. Applications for quantile regression in epidemiology. *Current Epidemiology Reports*, 6(2):191–199, 2019. DOI: 10.1007/s40471-019-00204-6.

WHO. Cardiovascular diseases (CVDs), 10 2023. URL `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`.

K. Wijndaele, G. Beunen, N. Duvigneaud, L. Matton, W. Duquet, M. Thomis, J. Lefevre, and R. M. Philippaerts. A continuous metabolic syndrome risk score. Utility for epidemiological analyses. *Diabetes Care*, 29(10):2329–2329, 2006. DOI: 10.2337/DC06-1341.

K. Yu, Z. Lu, and J. Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society Series D: The Statistician*, 52(3):331–350, 2003. DOI: 10.1111/1467-9884.00363.

P. Zimmet, K. G. M. Alberti, F. Kaufman, N. Tajima, M. Silink, S. Arslanian, G. Wong, P. Bennett, J. Shaw, S. Caprio, and I. C. Group. The metabolic syndrome in children and adolescents – an IDF consensus report. *Pediatric Diabetes*, 8(5):299–306, 2007. DOI: https://doi.org/10.1111/j.1399-5448.2007.00271.x.

Appendix

Report on the impact factor of the publications

Article	Bibliometric indicators
Cardiovascular risk factors and its patterns of change between 4 and 8 years of age in the INMA-Asturias cohort. *PLoS ONE*, 18(4), 2023	IF: 3.7 Q2 (26/73)[a]
Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA- Asturias cohort. *European Journal of Pediatrics*, 2023	IF: 3.6 Q1 (23/130)[b]
Statistical considerations for analyzing data derived from long longitudinal cohort studies. *Mathematics*, 11(19), 2023	IF: 2.4 Q1 (23/330)[c]

Bibliometric indicators according to Journal Citation Reports 2022: Impact factor (IF), quartile, and rank. Categories for quartile and rank indicators: [a]Multidisciplinary sciences, [b]Pediatrics, [c]Mathematics.

Results for tracking of cardiovascular-related variables not included in Article II

Figure 1 shows quantile regression models for systolic blood pressure at 8 years as dependent variable and the rank of systolic blood pressure at 4 years as the independent variables, for the quantiles between 0.1 to 0.9, with increments of 0.05, adjusted for maternal age at delivery, maternal level of education, maternal social class, maternal smoking during pregnancy, maternal pre-pregnancy body mass index, paternal body mass index, parental cardiovascular antecedents, child sex, child mean daily energy intake

at 4 and 8 years, child weekly out-of-school physical activity time at 4 and 8 years, week of gestation at delivery, weeks of predominant breastfeeding, and child height at 4 and 8 years. It also shows the same information but using diastolic blood pressure at 8 years as the dependent variable and the rank variable of diastolic blood pressure at 4 years as the independent variable. Coefficient estimated are calculated with the independent variables in terms of percentiles and they represent the effect on the dependent variable for each 1-decile increase in the independent variable. They are expressed in terms of number of standard deviations of the dependent variable to homogenize the Y-axis scales.

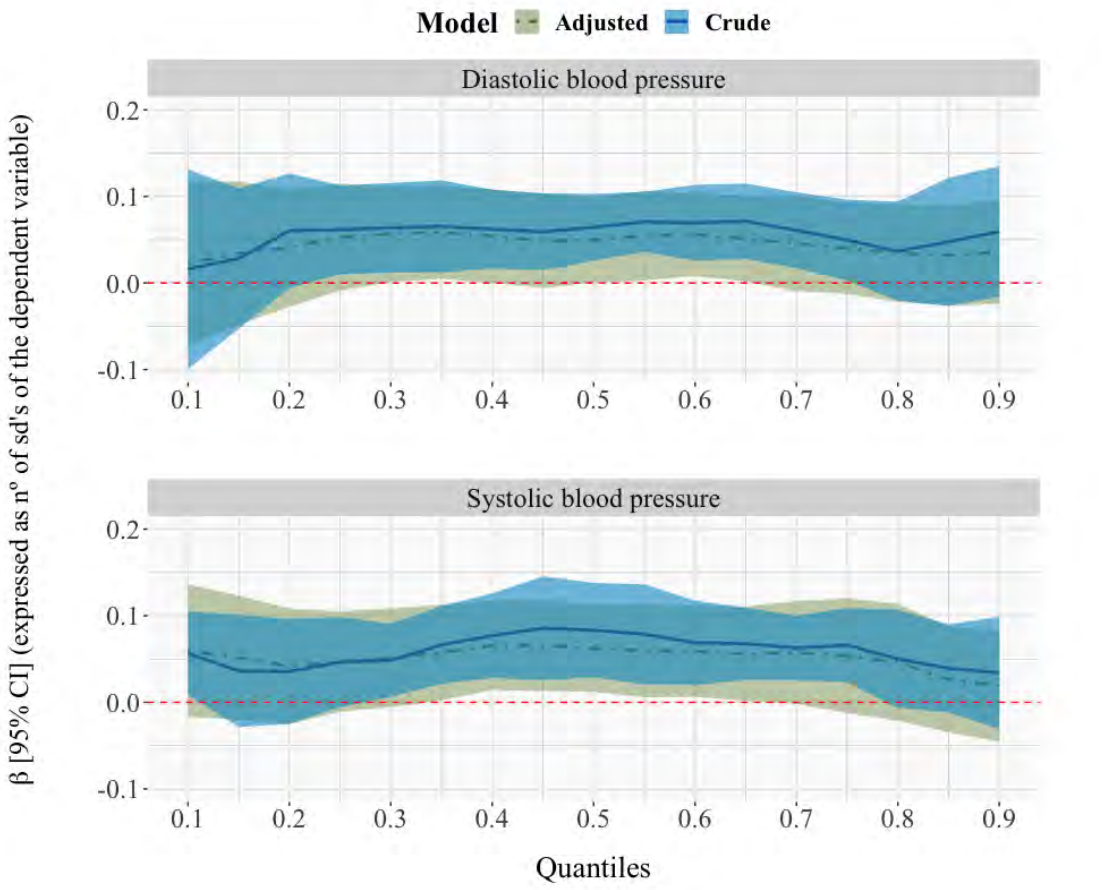


Figure 1: Quantile regression models for systolic and diastolic blood pressure