TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2024 222

Sara Castel Feced

Riesgo cardiovascular en dos cohortes de estudio: análisis con datos de vida real

Director/es
Rabanaque Hernández, María José

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Tesis Doctoral

RIESGO CARDIOVASCULAR EN DOS COHORTES DE ESTUDIO: ANÁLISIS CON DATOS DE VIDA REAL

Autor

Sara Castel Feced

Director/es

Rabanaque Hernández, María José

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Medicina

2024

Universidad Zaragoza
1542

Tesis Doctoral

Riesgo cardiovascular en dos cohortes de estudio: análisis con datos de vida real

Autor

Sara Castel Feced

Directora

Mª José Rabanaque Hernández

Facultad de Medicina
2023

A mi madre,
por su apoyo incondicional,
por educarme en la cultura del esfuerzo y
enseñarme a afrontar las dificultades.

A mi padre,
que siempre me motivó e impulsó a crecer.

AGRADECIMIENTOS

En primer lugar, a mi directora de tesis, Mª José Rabanaque, por su profesionalidad, dedicación y capacidad de trabajo, pero especialmente por su paciencia, accesibilidad, cariño, consejos y apoyo durante estos 4 años de trabajo.

A mis compañeras del Grupo de Investigación GRISSA y del Área de Medicina Preventiva y Salud Pública de la Universidad de Zaragoza. En especial a Sara Malo, Isabel Aguilar y Lina Maldonado, por su colaboración, consejos y disponibilidad para ayudar en todo momento.

A mis compañeros del Departamento de Epidemiología y Salud Global de la Universidad de Umeå, en especial a Miguel San Sebastián, por su absoluta disponibilidad y acogida. Todos ellos hicieron que me sintiera como en casa, incluso estando a miles de kilómetros.

A mis amigos, por su paciencia, escucharme cuando lo he necesitado y darme las palabras de aliento que precisaba escuchar. Porque todo esfuerzo requiere una motivación y empuje, y su apoyo ha sido fundamental en este viaje.

A mi familia, por su inquebrantable apoyo y por el tiempo "robado" de su valiosa compañía durante este período, sobre todo a mis abuelos. A mi padre, aunque falleció meses antes de embarcarme en esta aventura, siempre me motivó para hacer lo que me gustaba. Finalmente, y en especial, a mi madre, pilar fundamental en mi vida, por su preocupación y por inculcarme desde pequeña el amor al trabajo, la constancia y la perseverancia.

RESUMEN

Introducción

La enfermedad cardiovascular supone un desafío para la salud pública debido a su alta incidencia y gravedad. Las guías actuales destacan la importancia del cálculo preciso del riesgo cardiovascular para implementar medidas preventivas y tratamientos efectivos. El análisis longitudinal de factores de riesgo cardiovascular (FRCV), mediante herramientas como el análisis de clusters, proporciona una nueva visión y contribuye a mejorar la estimación del riesgo y al desarrollo de estrategias preventivas más personalizadas. Además, las técnicas de aprendizaje automático ofrecen un enfoque prometedor para mejorar la precisión en las estimaciones, aprovechando datos de vida real (RWD). Las diferencias entre sexos también afectan al campo de la enfermedad cardiovascular y el análisis contrafactual puede ser utilizado para identificar causas de estas diferencias. Este enfoque integral proporciona información valiosa para desarrollar políticas de salud pública dirigidas a reducir el riesgo y la frecuencia de enfermedades cardiovasculares e implementar estrategias preventivas más personalizadas.

Objetivos

Estimar la frecuencia de factores de riesgo cardiovascular en dos poblaciones diferentes de Aragón, cohortes Aragon Workers' Health Study (AWHS) y CArdiovascular Risk factors for hEalth Service research (CARhES), describiendo su evolución en función de perfiles de pacientes; estudiar la adherencia a tratamientos; aplicar diferentes técnicas para estimar el riesgo de enfermedad cardiovascular, y analizar las diferencias entre sexos en la incidencia de evento cardiovascular mayor (MACE), estudiando posibles factores que influyen en las mismas.

Metodología

El trabajo de tesis se ha desarrollado en el contexto de dos cohortes: la cohorte AWHS y la cohorte CARhES. Dentro de los estudios realizados en los hombres incluidos en el AWHS, se implementaron técnicas de machine learning, integrando variables de riesgo cardiovascular y aquellas que miden la adherencia a tratamientos para explorar su impacto en la ocurrencia de eventos cardiovasculares. Además, se han aplicado técnicas de clustering longitudinal para identificar grupos de individuos según la evolución de factores de riesgo cardiovascular.

La segunda parte de la tesis se centra en la cohorte CARhES, que abarca tanto a hombres como mujeres aragonesas con algún factor de riesgo cardiovascular: hipertensión, hipercolesterolemia o diabetes. En esta fase, también se aplicaron técnicas de machine learning para analizar la relación entre factores de riesgo cardiovascular y la incidencia de MACE, aportando información que permite enfoques más personalizados en la prevención. También se utilizaron técnicas contrafactuales para estudiar los factores que influyen en las diferencias en la frecuencia de eventos cardiovasculares entre hombres y mujeres.

Resultados

En el análisis realizado con datos del AWHS, la hipercolesterolemia fue el factor de riesgo cardiovascular más prevalente y la incidencia acumulada de enfermedad cardiovascular resultó del 7,9% en 10 años. Los modelos predictivos, destacaron la edad y la exposición al tratamiento como principales predictores de la incidencia de evento cardiovascular. Finalmente, los varones del AWHS se agregaron en dos clusters atendiendo a la evolución de sus factores de riesgo y nivel de riesgo cardiovascular (utilizando la herramienta SCORE). En comparación al cluster 2, el cluster 1 estuvo formado por trabajadores más jóvenes, con valores medios de índice de masa corporal, perímetro de cintura, glucemia y SCORE más bajos, y valores medios de colesterol HDL más altos.

En la cohorte CARhES, la hipertensión fue el FRCV más común entre los individuos de la cohorte, todos ellos con algún FRCV. La incidencia de MACE fue del 1,1% en 4 años. Los modelos predictivos resaltaron la edad y la adherencia a antidiabéticos como factores clave en la incidencia de MACE. La incidencia de MACE fue mayor en hombres, sujetos con diabetes y jubilados que ganaban menos de 18,000€ al año. Finalmente, diabetes y nivel socioeconómico fueron los mayores factores contribuyentes a las diferencias observadas entre sexos.

Conclusiones

Los modelos predictivos desarrollados tanto en la cohorte AWHS como en la cohorte CARhES destacan la influencia de la edad y la adherencia al tratamiento en el riesgo de sufrir un evento cardiovascular. El análisis por clusters realizado con los trabajadores de la cohorte AWHS evidenció dos grupos de individuos según la evolución de factores de riesgo y nivel de riesgo cardiovascular. Uno con peor evolución de factores de riesgo y mayor aumento de riesgo que el otro. Finalmente, la diabetes y el nivel socioeconómico parecen ser dos factores contribuyentes a las diferencias en la incidencia de MACE entre hombres y mujeres de la cohorte CARhES.

ABSTRACT

Introduction

Cardiovascular disease poses a significant challenge to public health due to its high incidence and severity. Current guidelines emphasize the importance of accurate cardiovascular risk assessment to implement preventive measures and effective treatments. Longitudinal analysis of cardiovascular risk factors (CVRF), using tools such as cluster analysis, provides a new perspective and contributes to improving risk estimation and developing more personalized preventive strategies. Additionally, machine learning techniques offer a promising approach to enhance precision in estimates by leveraging real-world data (RWD). Gender differences also impact the cardiovascular disease field, and counterfactual analysis can be used to identify causes of disparity. This comprehensive approach provides valuable insights for developing public health policies aimed at reducing the risk and frequency of cardiovascular diseases and implementing more personalized preventive strategies.

Objectives

The objectives include estimating the frequency of cardiovascular risk factors in two different populations in Aragon, the cohorts Aragon Workers' Health Study (AWHS) and CArdiovascular Risk factors for hEalth Service research (CARhES), describing their evolution based on patient profiles, studying treatment adherence, applying various techniques to estimate cardiovascular disease risk, and analyzing gender differences in the incidence of major adverse cardiovascular events (MACE), exploring possible influencing factors.

Methodology

The thesis work has been conducted within the context of two cohorts: the Aragon Workers' Health Study (AWHS) cohort and the CARhES cohort. In studies involving men from the AWHS, machine learning techniques were implemented, integrating cardiovascular risk variables and those measuring treatment adherence to explore their impact on cardiovascular event occurrence. Longitudinal clustering techniques were also applied to identify groups of individuals based on the evolution of CVRF.

The second part focuses on the CARhES cohort, encompassing both Aragonese men and women with some cardiovascular risk factor: hypertension, hypercholesterolemia, or

diabetes. Machine learning techniques were again applied to analyze the relationship between cardiovascular risk factors and MACE incidence, providing information for more personalized prevention approaches. Counterfactual techniques were also used to study factors influencing differences in the frequency of cardiovascular events between men and women.

Results

In the AWHS analysis, hypercholesterolemia was the most prevalent CVRF, with a 7.9% incidence of cardiovascular disease over 10 years. Predictive models highlighted age and treatment exposure as key predictors of cardiovascular disease incidence. Finally, male workers in the AWHS were grouped into two clusters based on the evolution of their risk factors and cardiovascular risk (measured through the tool SCORE). Compared to Cluster 2, Cluster 1 consisted of younger workers with lower average values of body mass index, waist circumference, blood glucose, and SCORE, and higher average values of HDL cholesterol.

In the CARhES cohort, hypertension was the most common CVRF among individuals with any risk factor. The incidence of MACE was 1.1% over 4 years. Predictive models emphasized age and adherence to antidiabetic medications as key factors in MACE incidence. MACE incidence was higher in men, individuals with diabetes, and retirees earning less than €18,000 per year. Finally, diabetes and socioeconomic level were the major contributing factors to observed gender differences.

Conclusions

Predictive models developed in both the AWHS and CARhES cohorts highlight the influence of age and treatment adherence on the risk of experiencing a cardiovascular event. Cluster analysis with AWHS workers revealed two groups based on the evolution of risk factors and cardiovascular risk, with one showing worse risk factor evolution and a higher increase in risk than the other. Ultimately, diabetes and socioeconomic level appear to be two contributing factors to differences in MACE incidence between men and women in the CARhES cohort.

ÍNDICE

1. INTRODUCCIÓN 23

Definición de enfermedad cardiovascular 24

Frecuencia e impacto de la enfermedad cardiovascular 25

Factores de riesgo cardiovascular 29

Estrategias de prevención cardiovascular 33

Algunos estudios de cohortes en el estudio de la enfermedad cardiovascular 36

Métodos para la estimación de riesgo cardiovascular 38

Técnicas de machine learning para la estimación del riesgo cardiovascular 39

Justificación del estudio 41

Hipótesis 42

2. OBJETIVOS 44

Objetivo general 45

Objetivos específicos 45

1.- Trabajo con datos de la cohorte AWHS 45

2.- Para el trabajo con datos de la cohorte CARhES 46

3. MATERIAL Y MÉTODOS 47

Estudio en la cohorte AWHS 48

Descripción de la cohorte 48

Selección de sujetos para la presente tesis 49

Variables del estudio y fuentes de información utilizadas para la presente tesis 50

Selección de sujetos y análisis realizados para dar respuesta a los objetivos planteados 58

Estudio en la cohorte CARhES 64
Descripción de la cohorte 64
Comparación cohorte CARhES y población aragonesa 65
Selección de sujetos para la presente tesis 66
Variables de estudio y fuentes de información utilizadas en la presente tesis 67
Selección de sujetos y análisis estadísticos por objetivos 71
Consideraciones éticas 76
Financiación 76
4. RESULTADOS 77
Resultados del estudio en la cohorte AWHS 78
Resultados que dan respuesta al OBJETIVO 1.1.- Análisis descriptivo de factores de riesgo cardiovascular (hipertensión, hipercolesterolemia, diabetes y estado físico), exposición a tratamientos preventivos e incidencia de ECV en la cohorte AWHS. 78
Resultados que dan respuesta al OBJETIVO 1.2.- Describir los valores analíticos y variables médicas relacionadas con FRCV y el SCORE, analizando la evolución de los mismos. 80
Resultados que dan respuesta al OBJETIVO 1.3.- Analizar la capacidad de diferentes métodos de machine learning para predecir la aparición de ECV y describir la influencia de distintos FRCV y la exposición a tratamientos preventivos en la incidencia de evento mediante el análisis de dichos modelos, incluyendo las siguientes variables: edad, hipertensión, hipercolesterolemia, diabetes, estado físico y exposición al tratamiento. 84
Resultados que dan respuesta al OBJETIVO 1.4.- Identificar perfiles de participantes en la cohorte AWHS en función de la evolución de FRCV y del SCORE utilizando la información de tres momentos, aplicando técnicas de cluster longitudinal. 90

Resultados cohorte CARhES 97

Resultados que dan respuesta al objetivo 2.1.- Describir la prevalencia de factores de riesgo en la población de Aragón, así como la frecuencia de FRCV, adherencia a tratamientos e incidencia de MACE en la cohorte CARhES. 97

Resultados que dan respuesta al objetivo 2.2.- Analizar diferencia en la prevalencia de FRCV y nivel socioeconómico y la incidencia de MACE entre hombres y mujeres. 102

Resultados que dan respuesta al objetivo 2.3.- Analizar la capacidad de distintos métodos de machine learning para predecir la incidencia de MACE en la cohorte CARhES de manera separada para hombres y mujeres, analizando la influencia de 4 grupos de variables (edad, FRCV, valores analíticos y mediciones de TA y adherencia a tratamientos antihipertensivos, antidiabéticos e hipolipemiantes) en dicha predicción. 105

Resultados que dan respuesta al objetivo 2. 4.- Estudiar el impacto que tienen las diferencias en la distribución de hipertensión, hipercolesterolemia, diabetes y nivel socioeconómico entre sexos en las diferencias observadas en la incidencia de MACE..... 111

5. DISCUSIÓN 114

Poblaciones de estudio 115

Discusión de los resultados cohorte AWHS 116

Resultados análisis descriptivos cohorte AWHS..... 116

Resultados análisis machine learning en la cohorte AWHS..... 118

Resultados análisis cluster longitudinal en la cohorte AWHS 121

Discusión resultados cohorte CARhES 124

Resultados análisis descriptivos cohorte CARhES..... 124

Resultados análisis de machine learning en la cohorte CARhES..... 126

Resultados análisis contrafactual en la cohorte CARhES..... 129

Comparación resultados en las dos cohortes 131

Limitaciones y fortalezas del estudio..... 132
Limitaciones del estudio..... 132
Fortalezas del estudio..... 135
6. CONCLUSIONES..... 138
7. BIBLIOGRAFÍA 143
8. ANEXOS..... 157
ANEXO I 158
ANEXO II 173
ANEXO III 187
ANEXO IV..... 214

ÍNDICE DE TABLAS

Tabla 1: Bases de datos consultadas e información obtenida de la cohorte AWHS. 51
Tabla 2: Relación variables y objetivos en la cohorte AWHS. 57
Tabla 3: Comparación entre los datos de las distintas variables reales e imputados de la cohorte AWHS. 63
Tabla 4: Distribución por sexo, edad, nivel socioeconómico y zona de residencia de la cohorte CARhES y de la población de Aragón mayor de 16 años. 66
Tabla 5: Bases de datos consultadas e información obtenida en la cohorte CARhES. 68
Tabla 6: Relación variables y objetivos en la cohorte CARhES 71
Tabla 7: Análisis descriptivo de las variables estratificado según la incidencia de evento cardiovascular en la cohorte AWHS. 79
Tabla 8: Análisis descriptivo de las variables de estudio 81
Tabla 9:Resultados del análisis por cuartiles en el AWHS. Los valores representan el porcentaje de individuos que pasaron de un cuartil a otro entre los momentos 1 y 2 y los momentos 2 y 3 para las siguientes variables: índice de masa corporal, glucemia y SCORE de riesgo 83
Tabla 10: Evaluación de la validez y rendimiento utilizando solo factores de riesgo cardiovascular como variables predictivas en la cohorte AWHS. 86
Tabla 11: Evaluación de la validez y rendimiento utilizando solo factores de riesgo cardiovascular y exposición al tratamiento como variables predictivas en la cohorte AWHS. 88
Tabla 12: Resultados del análisis de cuartiles por clusters en la cohorte AWHS. Los valores representan el porcentaje de individuos que pasaron de un cuartil a otro entre los momentos 1 y 2 y los momentos 2 y 3 para las siguientes variables: índice de masa corporal, glucemia y SCORE 94
Tabla 13: Prevalencias de los FRCV en la población de Aragón. 97
Tabla 14: Análisis descriptivo de los sujetos incluidos en este análisis de CARhES. 99

Tabla 15: Análisis descriptivo según la incidencia de MACE, de los sujetos incluidos de CARhES. 101
Tabla 16: Características de la población total y por sexo al inicio del periodo del estudio de los sujetos incluidos de CARhES. 103
Tabla 17: Incidencia de MACE en función de los factores explicativos entre los sujetos incluidos de CARhES 104
Tabla 18: Métricas de rendimiento de los modelos creados con Random Forest con la población incluida que formaba CARhES. 106
Tabla 19: Métricas rendimiento de los modelos creados con XG Boost con la población incluida de CARhES 109

ÍNDICE DE GRÁFICOS

Figura 1. a. Años de vida ajustados por discapacidad (DALYs) estandarizados por edad causados por enfermedades cardiovasculares. b. Proporción de muertes prematuras atribuidas a enfermedades cardiovasculares. 24
Figura 2: Causas de muerte por frecuencia en la Unión Europea (EU) en 2020. . 26
Figura 3: Muertes por enfermedades del sistema circulatorio, cáncer y COVID-19 por países de la Unión Europea (EU)..... 27
Figura 4: Defunciones por causa estratificadas por sexo en Aragón..... 27
Figura 5: Altas hospitalarias por diagnóstico principal en Aragón..... 28
Figura 6: Criterios de selección de sujetos y periodos de estudio de la cohorte AWHS..... 50
Figura 7: Criterios de selección de sujetos y periodos de estudio de la cohorte CARhES..... 67
Figura 8: Capacidad predictiva de las variables incluidas en el estudio según los tres métodos empleados: XG Boost, Random Forest y Naïve Bayes en la cohorte AWHS. 85
Figura 9: Capacidad predictiva de las variables incluidas en el estudio según los tres métodos empleados: XG Boost, Random Forest y Naïve Bayes, en la cohorte AWHS. 87
Figura 10: Curva PR para los modelos desarrollados con Random Forest, considerando factores de riesgo solos o factores de riesgo y exposición al tratamiento como variables predictivas en la cohorte AWHS. 89
Figura 11: Índices de calidad en función del número de conglomerados.. 91
Figura 12: Gráficos de caja que representan los valores medios (puntos) de cada variable por cluster en la cohorte AWHS. 92
Figura 13: Contribuciones relativas de las variables en los modelos Random Forest para hombres y mujeres de CARhES. 107
Figura 14: Contribuciones relativas de las variables en los modelos XG Boost para hombres y mujeres de la cohorte CARhES. 110
Figura 15: RR para las poblaciones natural y contrafactual y porcentaje de contribución de cada factor explicativo en CARhES 111

LISTADO DE ABREVIATURAS

ABC	Ascertaining Barriers for Compliance
ACC	American College of Cardiology
ACV	Accidente cerebrovascular
AHA	American Heart Association
APVP	Años potenciales de vida perdidos
ATC	Código anatómico terapéutico químico
AUC	Área bajo la curva
AUC-PR	Área bajo la curva precision-recall
AWHS	Aragon Workers' Health Study
BDU	Base de Datos de Usuarios del Sistema de Salud de Aragón
CARhES	CArdiovascular Risk factors for hEalth Service research
CIE-9	Clasificación Internacional de Enfermedades, 9ª revisión
CIE-10	Clasificación Internacional de Enfermedades, 10ª revisión
CMBD	Conjunto Mínimo Básico de Datos
CV	Cardiovascular
DALYs	Disability Adjusted Life Year
DDD	Número de dosis diarias definidas
DE	Desviación estándar
DM	Diabetes mellitus
ECV	Evento cardiovascular
ESC	Sociedad Europea de Cardiología
FRCV	Factores de Riesgo Cardiovascular
GMA	Grupos ajustados por morbilidad
HC	Hipercolesterolemia
HTA	Hipertensión
IAM	Infarto agudo de miocardio
IMC	Índice de masa corporal
MACE	Evento cardiovascular mayor
MPR	Medication posesion ratio
NCEP	NAtional Cholesterol Education Program

NB	Naïve Bayes
OMS	Organización Mundial de la Salud
PCE	Pooled cohort equations
PDC	Proporción de días cubiertos
Q	Cuartil
RCV	Riesgo cardiovascular
ROSE	Random Over Sampling Examples
SCORE	Systemic Coronary Risk Estimation
TA	Tensión arterial
TAS	Tensión arterial sistólica
TAD	Tensión arterial diastólica
ULSAM	Uppsala Longitudinal Study of Adult Men
VPP	Valor predictivo positivo
VPN	Valor predictivo negativo
WHF	World Heart Federartion

1. INTRODUCCIÓN

1. INTRODUCCIÓN

Definición de enfermedad cardiovascular

Las enfermedades cardiovasculares (CV) han pasado a ser la primera causa de muerte y morbilidad en todo el mundo. El impacto de estas enfermedades en los países desarrollados es bien conocido desde hace tiempo, mientras que en los países en vías de desarrollo son un problema relativamente reciente. De cada 4 muertes por enfermedad CV, 3 se producen en estos países, donde el acceso a servicios sanitarios y a programas de atención primaria para la detección y tratamiento precoz de factores de riesgo es limitado. Además, las enfermedades no transmisibles en general, y la enfermedad cardiovascular en particular, contribuyen a la pobreza de las familias y del país debido a la mortalidad temprana y morbilidad que provocan (Figura 1. a. y 1. b.) y al gasto sanitario directo e indirecto que suponen [1,2].

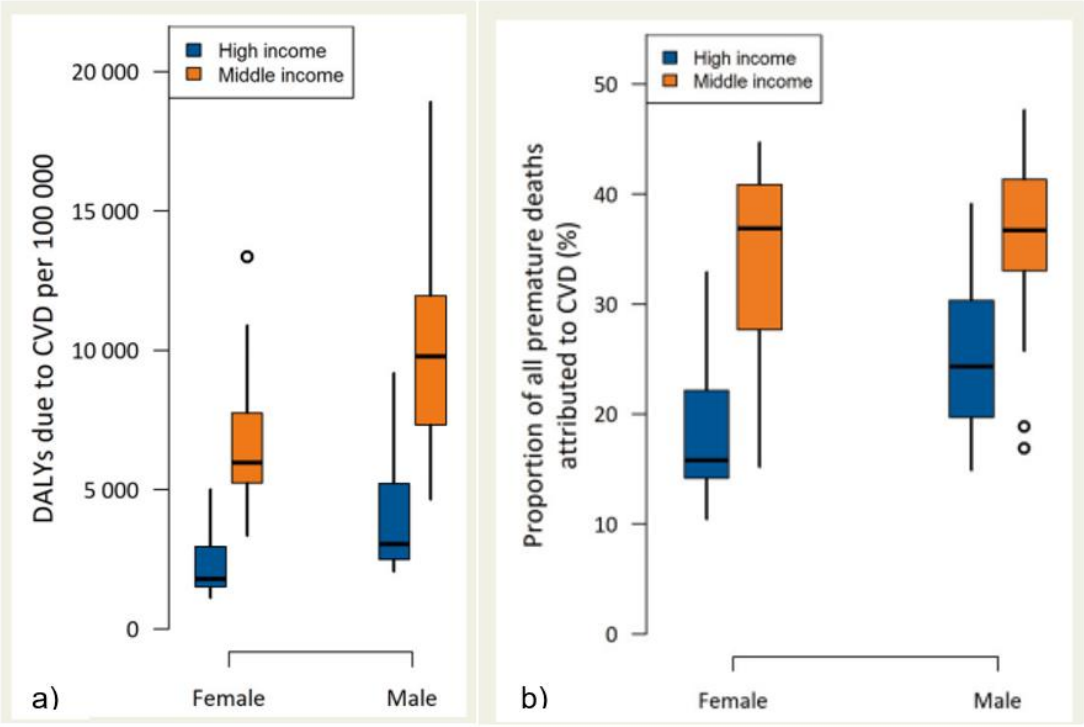


Figura 1. a. Años de vida ajustados por discapacidad (DALYs) estandarizados por edad causados por enfermedades cardiovasculares. b. Proporción de muertes prematuras atribuidas a enfermedades cardiovasculares. Fuente: Roth G. et al [3]

Dentro de las enfermedades cardiovasculares se incluyen un amplio grupo de patologías que afectan al corazón y a los vasos sanguíneos abarcando, entre otras, cardiopatías coronarias, enfermedades cerebrovasculares y cardiopatías reumáticas. Los episodios agudos más comunes dentro de estas enfermedades son el infarto agudo de miocardio (IAM) y los accidentes cerebrovasculares (ACV). Estos procesos suelen estar causados por la obstrucción de los vasos sanguíneos, normalmente por depósitos de grasa en sus paredes, que impiden que la sangre irrigue estos órganos. El ACV puede ocasionarse también por una hemorragia de los vasos cerebrales.

Desde el punto de vista fisiopatológico, la aterosclerosis es la causa de ACV e IAM, entre otras enfermedades, en la que se provocan lesiones ateroscleróticas en los vasos grandes y medianos que llevan a un cambio estructural en sus paredes. Ésta es una enfermedad silenciosa y crónica, sin manifestaciones clínicas, pero con grandes consecuencias. En la aterosclerosis la pared íntima de las arterias se ve afectada por el depósito de grasa y colesterol, volviéndose delgada e irregular. Las placas creadas por el depósito de estas sustancias van creciendo progresivamente y producen el estrechamiento de las arterias, reduciéndose el flujo sanguíneo y aumentando la probabilidad de que se cree un trombo[4].

Frecuencia e impacto de la enfermedad cardiovascular

En el año 2020 las enfermedades cardiovasculares causaron 1,7 millones de muertes en toda Europa [5], lo que supuso el 33% del total de fallecidos en el continente (Figura 2).

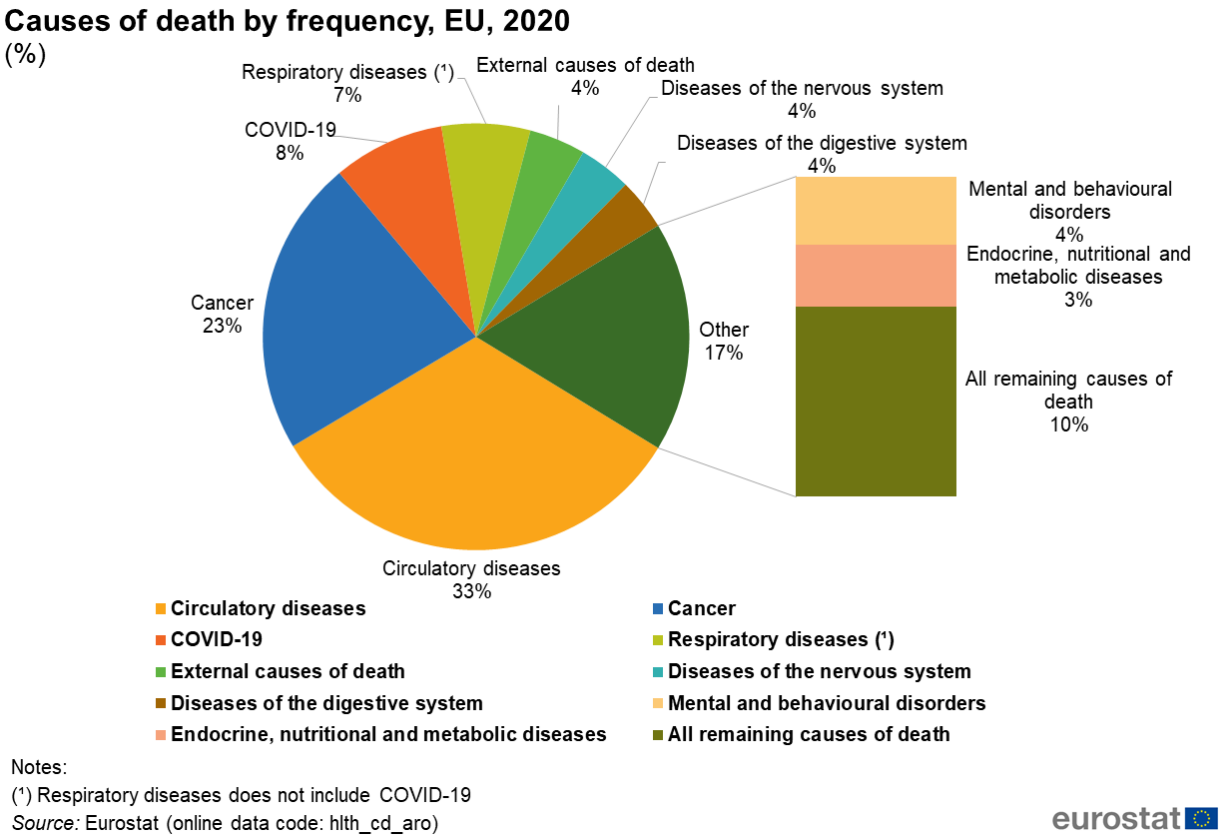


Figura 2: Causas de muerte por frecuencia en la Unión Europea (EU) en 2020. Fuente: Eurostat [5]

En España, aunque estas enfermedades son la primera causa de muerte, su incidencia es de las menores de Europa [5] (Figura 3). En nuestro país, las enfermedades del sistema circulatorio son responsables del 26,4% de los fallecimientos [6]. En Aragón, durante el año 2022, estas afecciones se posicionaron como la principal causa de fallecimiento en mujeres. En el caso de los hombres, ocuparon la segunda posición, siendo superadas únicamente por los tumores, representando el 24,7% del total de defunciones [7]. En contraste, en las mujeres, las enfermedades circulatorias alcanzaron el 30,1% del total de fallecimientos [7] (Figura 4).

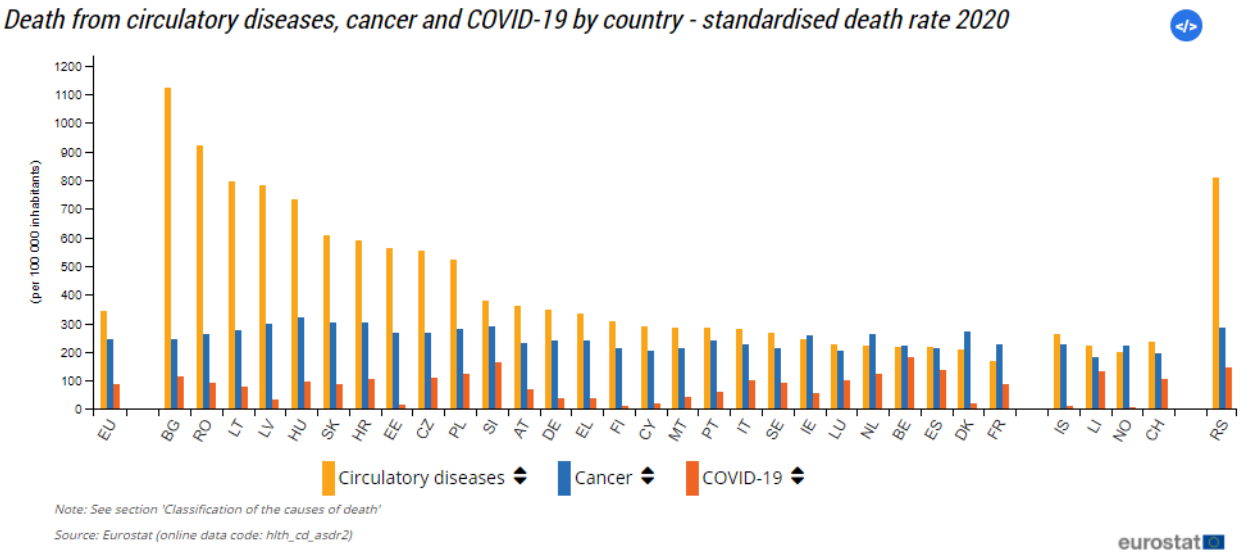


Figura 3: Muertes por enfermedades del sistema circulatorio, cáncer y COVID-19 por países de la Unión Europea (EU). Fuente: Eurostat [5]

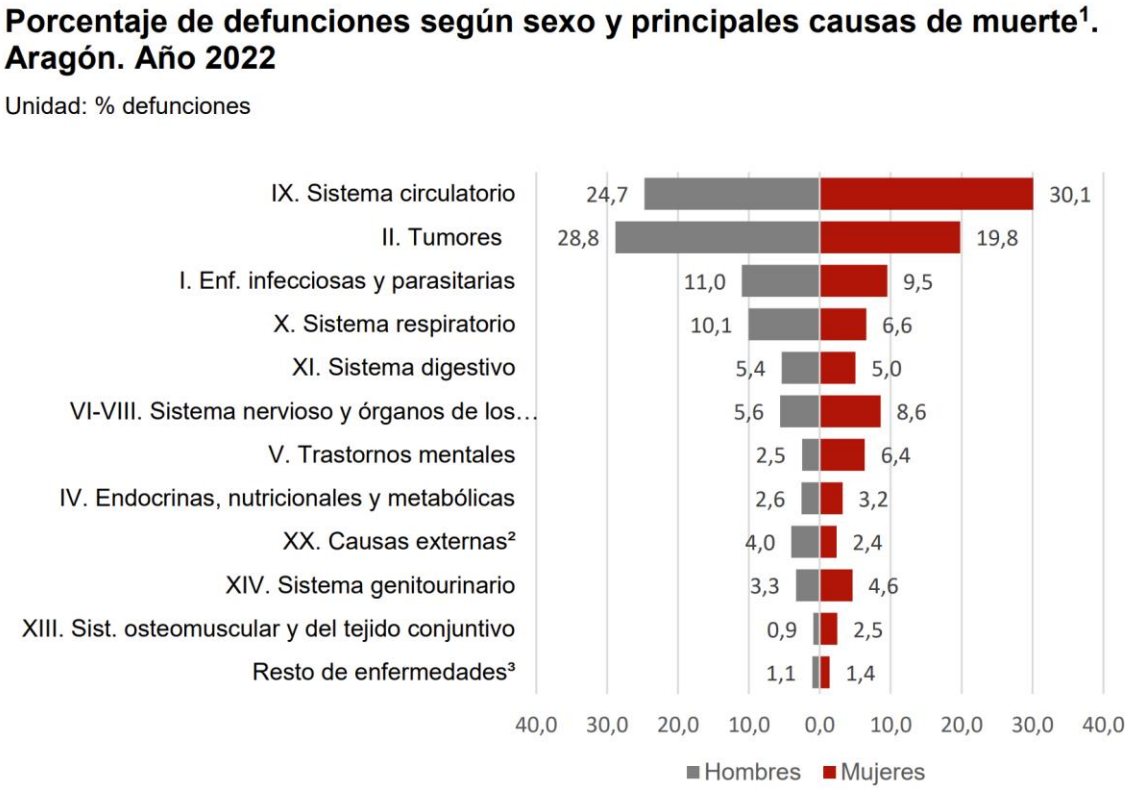


Figura 4: Defunciones por causa estratificadas por sexo en Aragón. Fuente: Instituto Aragonés de Estadística [7].

En España el 9,8% de la población española sufre alguna enfermedad cardiovascular, y de ellos el 52,6% son mujeres y el 47,4% hombres [8]. Además, cada año se diagnostica un nuevo caso por cada 100 habitantes y son la primera causa de ingreso hospitalario, siendo en 2021 las responsables del 10,9% de los ingresos en mujeres y del 15,1% en hombres [8–10]. En Aragón, en el año 2022, la prevalencia de ACV por cada 1000 habitantes fue del 15,1 en hombres y 12,9 en mujeres, mientras que la de cardiopatía isquémica fue 32,5 y 14,5 respectivamente para cada sexo. En dicha comunidad las enfermedades del sistema circulatorio representaron el 12,6% de los ingresos en hospital (Figura 5) y el 16% de años potenciales de vida perdidos (APVP) [11–14].

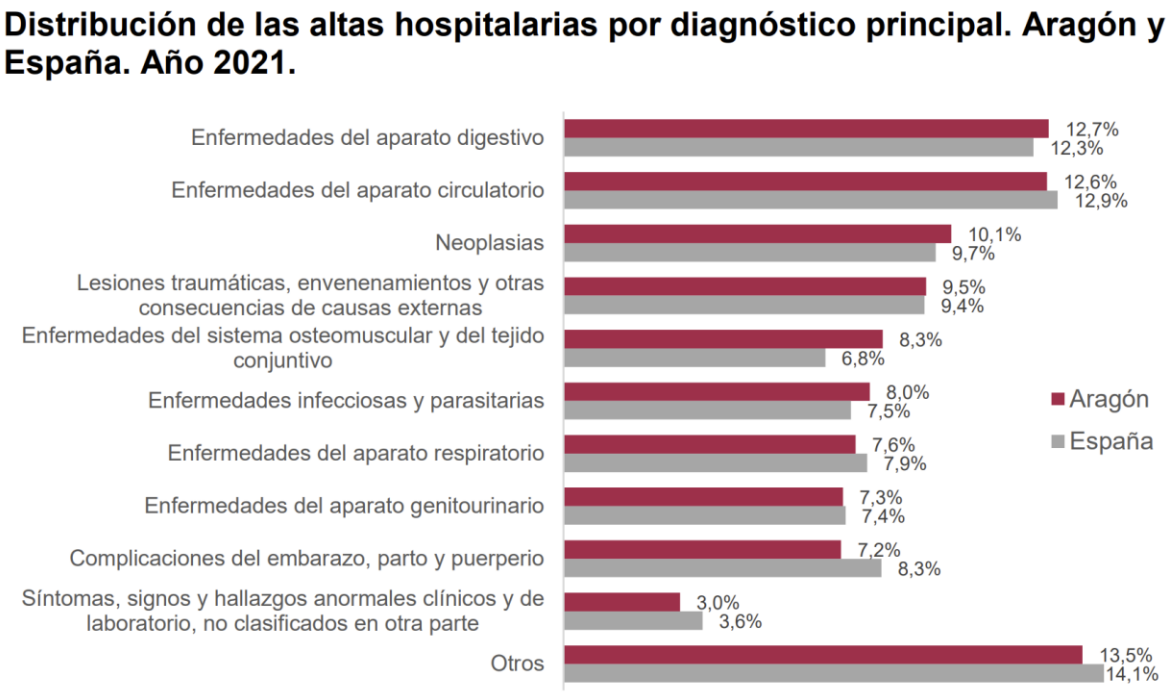


Figura 5: Altas hospitalarias por diagnóstico principal en Aragón. Fuente: Instituto Aragonés de Estadística[7]

Factores de riesgo cardiovascular

Hay algunas características biológicas y hábitos o estilos de vida que aumentan la probabilidad de padecer una enfermedad cardiovascular y que se conocen como factores de riesgo cardiovascular (FRCV). Según la Organización Mundial de la Salud (OMS), las principales causas de enfermedad cardiovascular son hábitos de vida poco saludables como una dieta inadecuada, la falta de ejercicio físico, o el consumo de tabaco y alcohol. Estos hábitos pueden desencadenar la aparición de los FRCV clásicos como son hipertensión arterial (HTA), hipercolesterolemia (HC), diabetes mellitus (DM) y sobrepeso u obesidad [1].

Las dietas poco saludables que se relacionan con la aparición de HTA, HC, diabetes y obesidad se caracterizan por tener altos contenidos en grasas saturadas, azúcares y sal. En España se estima que alrededor del 30% de la población no consume fruta fresca diariamente, y la mitad no ingiere diariamente verduras, ensaladas u hortalizas. La realización de ejercicio físico se asocia a menores niveles de morbilidad y mortalidad, así como una menor incidencia de diabetes mellitus tipo 2. A pesar de sus beneficios, el 32,4% de los hombres y 40,3% de las mujeres en España refieren llevar una vida sedentaria. El tabaquismo es la principal causa de morbimortalidad cardiovascular prematura, se estima que en menores de 50 años el riesgo cardiovascular (RCV) de fumadores es 5 veces mayor que el de los no fumadores y en 2017 causó el 12,9% de las muertes en España. Además, no solo el tabaquismo activo se asocia con un aumento del RCV, también el tabaquismo pasivo aumenta el riesgo de sufrir una enfermedad cardiovascular. Finalmente, aunque el consumo moderado de alcohol ha sido relacionado en algunos estudios con una reducción del RCV, sus efectos nocivos sobre múltiples resultados de salud (tanto cardiovascular como no cardiovascular) superan los posibles beneficios, por lo que se le considera también un factor de riesgo. Además, en España, es el 4º factor de riesgo de pérdida de años de vida ajustados por discapacidad (DALYs) [3,15,16].

Todos estos FRCV relacionados con los estilos de vida se consideran modificables y es común que las personas presenten más de uno al mismo tiempo, interaccionando éstos

entre sí, por lo que el riesgo de una persona con varios FRCV es mayor que la suma del riesgo que suponen individualmente[1,3,15].

Así pues, se hace necesario un abordaje integral de todos los FRCV en el que no solo se tenga en cuenta el individuo sino también su contexto, pudiendo abordar los FRCV tanto por las autoridades sanitarias como educativas, mediante políticas públicas para mejorar las oportunidades individuales y colectivas, así como por la propia persona adoptando estilos de vida saludables[3,8,15].

Como ya se ha mencionado, HTA, HC y DM son tres FRCV bien conocidos. La hipertensión es la primera causa de mortalidad atribuible y una de las principales causas de enfermedad cardiovascular, ya que se ha estimado que provoca 9,4 millones de muertes y el 7% de APVP en todo el mundo. En el año 2020, el 19,3% de los españoles refirió sufrir hipertensión, sin embargo, estudios transversales han estimado una prevalencia superior al 40%, debido a un alto infradiagnóstico, especialmente en personas menores de 35 años. El beneficio de reducir la tensión arterial sistólica (TAS) depende del RCV global evaluado considerando todos los FRCV en conjunto. Por último, el riesgo de muerte aumenta linealmente cuando la TAS se incrementa por encima de 90 mm Hg y la tensión arterial diastólica (TAD) por encima de 75 mm Hg [15]. Se ha estimado que un buen control de la HTA puede disminuir las tasas de IAM en un 24% y la mortalidad por ACV un 42% [8].

El riesgo cardiovascular por hipercolesterolemia se multiplica si se presenta junto a otros FRCV, y una pequeña reducción absoluta del colesterol-LDL puede ser beneficiosa en personas con RCV alto o muy alto [15]. En España, la prevalencia de hipercolesterolemia auto declarada era de 15,5% en 2020 y algunos estudios han mostrado que, en pacientes con síndrome coronario agudo, la prevalencia puede estar entre 40 y 58% y entre 40 y 72% en pacientes con cardiopatía[8].

En cuanto a la diabetes, no sólo se consideran FRCV la diabetes tipo 1 y 2, sino también la prediabetes. La diabetes es la octava causa de DALYs y en España provoca el 4,15% de los APVP y el 2,36% de muertes. En España la prevalencia de diabetes se considera que es del 7,5%, y un 12% de la población está diagnosticada de prediabetes, pero se estima que un 6% de los diabéticos no están diagnosticados. Las personas con diabetes

tipo 2 tienen un RCV alto, están predispuestas a la aparición de enfermedad cardiovascular precoz y se ha demostrado que tiene un fuerte impacto en el riesgo de ictus, especialmente entre las mujeres. Además, los pacientes con diabetes tipo 2 tienen mayor riesgo de padecer otros FRCV (como hipertensión e hipercolesterolemia), lo que aumenta aún más el RCV [15]. Aunque la mortalidad por diabetes se ha reducido en los últimos años, el control de estos pacientes en España es mejorable, ya que una parte significativa de los diabéticos no alcanzan los objetivos recomendados. Además, las mujeres con diabetes suelen tener peor control de otros FRCV en comparación con los hombres [17].

Otros FRCV bien conocidos y que están relacionados con una mayor mortalidad son la obesidad y el sobrepeso. En España se estima que el 60% de las personas tienen sobrepeso u obesidad, porcentaje que ha ido en aumento, especialmente en varones. Estos también afectan a la población infantil, y su prevalencia en este grupo de la población ha incrementado en los últimos años, situándose el porcentaje de niños entre 6 y 9 años con sobrepeso u obesidad en un 40,6%, con las consecuencias que esto conlleva tanto a largo como a corto plazo [17].

Además de los FRCV ya mencionados, muchos de ellos modificables y sobre los que se puede actuar, hay otros factores como los genéticos, el sexo y la edad sobre los cuales no se puede actuar y que predisponen a sufrir enfermedades cardiovasculares. Entre los genéticos se encuentran determinadas características que predisponen a tener hipertensión arterial o hipercolesterolemia familiar, aunque se considera que por sí mismos no justifican el desarrollo de enfermedad cardiovascular si no hay otros factores medioambientales o FRCV modificables [17].

Como se ha mencionado, junto con los factores genéticos, otros factores no modificables que están relacionados con el riesgo de desarrollar enfermedades cardiovasculares son el sexo/género y la edad. La edad se considera el FRCV más importante ya que el RCV aumenta exponencialmente con los años, por lo que se la asocia a un mayor riesgo de enfermedades cardiovasculares [15,18,19]. Se estima que 58% de los diagnosticados con alguna enfermedad cardiovascular tiene más de 65 años, y los resultados de estimación de riesgo muestran que las mujeres menores de 50 y hombres menores de 40 años

tienden a tener un riesgo bajo de sufrir enfermedades cardiovasculares, mientras que las mayores de 75 y 65, respectivamente, tienden a tener riesgo alto [15].

En cuanto al papel del género/sexo, las enfermedades cardiovasculares y los FRCV son diferentes entre hombres y mujeres y, aunque tradicionalmente estas se han considerado una "enfermedad de hombres", son la principal causa de muerte y discapacidad en las mujeres[20–22]. El patrón de enfermedades cardiovasculares y la repercusión de los FRCV difieren entre hombres y mujeres, y se han asociado a condiciones específicas de la mujer, relacionados sobre todo con el embarazo y la menopausia [3,15]. Entre ellas se encuentran la preeclampsia, la hipertensión relacionada con el embarazo, la diabetes gestacional, el síndrome de ovario poliquístico y la menopausia prematura. Se dan independientemente de los FRCV convencionales, pero repercuten en el RCV.

Según los FRCV tradicionales y las diferencias entre sexos, la literatura muestra una mayor asociación de la diabetes con el desarrollo de enfermedad cardiovascular en las mujeres que en los hombres [23–26]. Las diferencias entre hombres y mujeres en el impacto de la hipertensión y la hipercolesterolemia no están claras, aunque algunos estudios han encontrado un mayor impacto de la hipercolesterolemia con el desarrollo de estas enfermedades en hombres que en mujeres[27,28].

Por último, se han descrito disparidades por género en la atención cardiovascular, especialmente en la atención aguda. Las mujeres suelen someterse a menos ecocardiogramas, pruebas de esfuerzo, angiografías coronarias o procedimientos de revascularización. Además, durante mucho tiempo las enfermedades cardiovasculares se han considerado un problema masculino, y en las mujeres se ha producido una infrautilización del cribado de estas. Otros factores basados en el género están interrelacionados con las diferencias mencionadas en las mujeres en comparación con los hombres, como un nivel socioeconómico más bajo, un menor nivel de actividad física y un mayor estrés debido a las responsabilidades familiares [20–22].

Así pues, el desarrollo de FRCV y la salud de las personas están influenciados por las condiciones de vida y los determinantes sociales de la salud, que son las circunstancias en que las personas nacen crecen, trabajan, viven y envejecen, sistemas que influyen sobre las condiciones de la vida cotidiana [29,30]. Estas circunstancias no se distribuyen de

forma homogénea en la población, haciendo que la salud de las personas se pueda ver afectada por el entorno en el que viven. En este sentido, la contaminación ambiental se ha descrito como un nuevo e importante factor de riesgo que influye en la salud de las personas, reduciendo la esperanza de vida y aumentando la mortalidad, tanto global como por enfermedad cardiovascular. Según la Federación Mundial del Corazón (WHF), la Sociedad Europea de Cardiología (ESC), la Asociación Americana del Corazón (AHA) y el Colegio Americano de Cardiología (ACC), hasta el 50% de las muertes por contaminación en 2019 fueron por enfermedades cardiovasculares. Además, según la OMS, la contaminación del aire es responsable del 25% de todas las muertes por enfermedad cardiaca y del 24% de todas las muertes por ictus [31–33].

Estrategias de prevención cardiovascular

Dada la prevalencia e impacto de la enfermedad cardiovascular, su prevención resulta esencial para reducir o minimizar su impacto y la incapacidad asociada. Se estima que con la eliminación de comportamientos de riesgo para la salud sería posible prevenir hasta el 80% de los casos de las enfermedades cardiovasculares [16]. La prevención debe estar dirigida hacia la población general, promoviendo estilos de vida saludables y a nivel individual, en sujetos con un RCV alto o moderado, centrándose en corregir hábitos no saludables, como dietas poco saludables, inactividad física y tabaquismo, y controlando sus FRCV mediante tratamientos farmacológicos[15,16,34]. En estos sujetos objeto de una prevención primaria, es importante que las decisiones sean tomadas entre el médico y el paciente, ya que cuando este se siente implicado es más fácil que adopte las recomendaciones dadas [34].

Con el fin de mejorar las acciones preventivas a nivel individual y colectivo distintas asociaciones como la Sociedad Europea de Cardiología (ESC), la American College of Cardiology (ACC) o la American Heart Association (AHA), entre otras, publican periódicamente guías para dar a conocer las actividades preventivas a realizar en la práctica clínica diaria y establecer unos valores determinados para ciertos marcadores[15,34]. En 2022 se publicó la última guía desarrollada por la ESC y en ella se presentan estrategias para actuar tanto a nivel individual como colectivo.

A nivel individual, esta guía basa la prevención en la identificación de pacientes con RCV más alto, siendo su piedra angular la estimación del RCV. Una vez realizada la estimación del riesgo a través de distintas herramientas en función de las características del paciente, la comunicación del riesgo y de los objetivos esperados con el tratamiento de los FRCV son claves para reducir las probabilidades de sufrir una enfermedad cardiovascular y sus consecuencias. Las medidas tomadas para la reducción del riesgo suelen estar basadas en cambios de estilos de vida y en intervenciones farmacológicas[15].

Las intervenciones recomendadas basadas en la optimización de estilos de vida se centran en la prescripción de actividad física y ejercicio, que pueden mejorar los desenlaces adversos y mejorar los FRCV a cualquier edad y en ambos sexos; consejo para la adquisición de hábitos alimentarios saludables dirigidos a mejorar los valores de lípidos, el peso o la diabetes con una dieta sana y el bajo consumo de bebidas alcohólicas y azucaradas; y, finalmente, la deshabituación tabáquica es una de las medidas más conocidas por su alta efectividad en la reducción del riesgo de enfermedad cardiovascular [15].

A parte de las medidas nombradas, la intervención farmacológica en ocasiones es también necesaria para la reducción del RCV. En esta línea, la adherencia a los tratamientos prescritos adquiere un papel muy relevante y un importante punto de mejora, ya que la adherencia a los tratamientos prescritos varía entre el 50% en pacientes con prevención primaria y el 66% con secundaria. Las causas de la falta de adherencia a tratamientos son múltiples, desde la complejidad del régimen de tratamiento o las dosis, falta de aceptación de la enfermedad, mala relación médico-paciente, creencias sobre los tratamientos y miedo a efectos adversos, o limitaciones físicas, cognitivas o económicas [15].

Para estandarizar la terminología relacionada con la adherencia a las terapias farmacológicas, el proyecto europeo Ascertaining Barriers for Compliance (ABC) [35] propuso una Taxonomía de la Adherencia que consta de tres componentes: iniciación, implementación y discontinuación. El inicio del tratamiento corresponde a la toma de la primera dosis de un medicamento prescrito. El proceso continúa con la implementación del régimen de dosificación, definido como el grado en que la dosificación real del

paciente se corresponde con el régimen de dosificación prescrito, desde el inicio hasta la toma de la última dosis. La interrupción marca el final de la terapia, cuando no se toma la siguiente dosis prescrita y deja de tomar la medicación a partir de entonces. La persistencia es el tiempo transcurrido entre el inicio y la última dosis, que precede inmediatamente a la interrupción.

Hay distintos métodos para estimar la adherencia durante cada una de estas fases. En la fase de iniciación sólo es posible si se tienen datos de la fecha de prescripción y dispensación. Para el cálculo de la fase de implementación la Proporción de Días Cubiertos (PDC) y Medication Posesion Ratio (MPR) son ampliamente utilizados y para el cálculo de la fase de discontinuación, la persistencia es el estimador más utilizado [36].

En cuanto a las medidas a tomar dirigidas a la población, la guía desarrollada por la ESC presenta algunas recomendaciones para el desarrollo de políticas y estrategias de intervención poblacional. Estas políticas y estrategias están enfocadas tanto a la prevención de la enfermedad cardiovascular como a la intervención sobre los FRCV específicos.

El objetivo de las estrategias desarrolladas para la prevención de la enfermedad cardiovascular es producir cambios en los individuos de una población con el objeto de modificar el riesgo atribuible en dicha población. Para ello se centran en medidas preventivas que requieren grandes intervenciones en salud pública enfocadas en los estilos de vida y en promover la vigilancia de las enfermedades cardiovasculares.

Con las intervenciones poblacionales enfocadas a factores de riesgo específico se busca alterar el entorno social y modificar algunos determinantes sociales para proporcionar incentivos que fomenten cambios en el comportamiento individual y la exposición a FRCV, teniendo en consideración la alfabetización en salud. Los FRCV específicos en los que se recomienda centrar estas intervenciones son el control de la tensión arterial (TA), la dieta, el tabaquismo y consumo de alcohol.

Hay que tener en cuenta que las modificaciones de estilos de vida en la población conlleva tiempo y un elevado coste económico, además de que los beneficios pueden

tardar en manifestarse. Sin embargo, su efecto perdura a largo plazo y mejoran la calidad de vida y el bienestar de las personas.

Algunos estudios de cohortes en el estudio de la enfermedad cardiovascular

Dada la importancia de las enfermedades cardiovasculares y la carga que suponen en los sistemas sanitarios, numerosas cohortes se han creado para su estudio. En 1948, el servicio de Salud Pública de Estados Unidos inició la cohorte Framingham con el objetivo de investigar la epidemiología y FRCV [37]. Este estudio sigue recopilando datos de 3 generaciones de participantes sobre factores de riesgo biológicos, estilos de vida y resultados de diferentes tipos de enfermedades (incluidas las cardiovasculares). La cohorte inicial estuvo formada por 5209 participantes con edades entre 28 y 62 años en el momento de reclutamiento, que han sido seguidas cada 2 años. A partir de esta cohorte, en 1971 se inició la cohorte de Hijos, en la que se incluyeron los hijos y cónyuges de la cohorte inicial, y en 2002 se comenzó el seguimiento de la tercera generación. Como consecuencia de las investigaciones llevadas a cabo en esta cohorte, se desarrolló un score para estimar el riesgo de sufrir enfermedad cardiovascular a 10 años ampliamente utilizado en todo el mundo[37].

Desde entonces, otras cohortes en todo el mundo se han creado con el objetivo de estudiar enfermedades cardiovasculares y sus factores de riesgo. A continuación, se comentarán brevemente algunas de ellas.

En 1970, en la ciudad sueca de Uppsala, se creó una cohorte formada por 2322 hombres, llamada Uppsala Longitudinal Study of Adult Men (ULSAM), con el objetivo de investigar los FRCV. Esta cohorte sigue activa y se ha ido adaptando, añadiendo mujeres y recogiendo distintos tipos de datos [38].

En España se han creado diferentes cohortes para el estudio de la enfermedad cardiovascular con múltiples enfoques. El estudio REGICOR [39] tiene como objetivo el estudio de la enfermedad coronaria y estrategias de prevención a nivel poblacional, estudiando la influencia de mecanismos moleculares, celulares y ambientales y se centra en la nutrición, actividad física y obesidad. Dentro de este estudio se adaptó y validó la

calculadora de riesgo desarrollada en el estudio Framingham a la población española. La cohorte Heart Healthy Hoods [40], se creó con el propósito de asociar enfermedades cardiovasculares con el entorno en el que las personas viven y desarrollan su vida diaria. Su objetivo es realizar estudios epidemiológicos con una perspectiva social, realizar un retrato de la salud cardiovascular europea y compararlo con la situación en América. Para ello, analizan la disponibilidad y tipos de alimentos, instalaciones y posibilidades para realizar actividad física, hábitos de consumo de alcohol y de tabaco en varios barrios madrileños, y tratan de asociarlos con distintos registros de atención primaria relacionados con la enfermedad cardiovascular. Por otro lado, el estudio PREDIMED y Predimed-plus [41], analiza el efecto de la dieta mediterránea en la salud cardiovascular. El estudio Predimed [42] comenzó en el año 2002 y su objetivo principal fue valorar los efectos de la Dieta Mediterránea en la prevención primaria de las enfermedades crónicas, en concreto, intentaba averiguar si la Dieta Mediterránea suplementada con aceite de oliva virgen extra y frutos secos evitaba la aparición de enfermedades cardiovasculares en comparación con una dieta baja en grasa. Como resultado de estas investigaciones se han encontrado evidencias de los beneficios de la dieta mediterránea en la prevención de estas y otras enfermedades crónicas como cáncer, HTA y enfermedades neurodegenerativas. Desde el año 2013, este proyecto continuó con el estudio Predimed-Plus [41], que cuenta con 6800 participantes procedentes de 23 puntos repartidos por toda España. Este estudio analiza el efecto de intervenciones intensivas para la pérdida de peso, basadas en el consumo de la dieta mediterránea hipocalórica, promoción de actividad física y cambios de hábitos para la prevención de las enfermedades cardiovasculares.

Finalmente, en Aragón también se han desarrollado dos cohortes con el objetivo de estudiar las enfermedades cardiovasculares. El estudio Aragon Workers' Health Study (AWHS)[43] se basa en una cohorte de 5400 trabajadores de una fábrica automovilística de Zaragoza y de los que se obtienen datos durante las revisiones médicas anuales. Recientemente, se ha creado la cohorte CArdiovascular Risk factors for hEalth Service research (CARhES), que recoge a todos los sujetos con algún FRCV en Aragón y cuyo objetivo es el estudio de la frecuencia de FRCV en la comunidad, del uso de servicios

sanitarios y del tratamiento farmacológico preventivo, tratando de identificar desigualdades en la atención sanitaria y estudiar su impacto en salud.

Métodos para la estimación de riesgo cardiovascular

Las guías de prevención de enfermedad cardiovascular consideran la evaluación del riesgo como la piedra angular para la prevención primaria de estas enfermedades, ya que la presencia simultánea de distintos FRCV que interaccionan entre sí aumenta el riesgo de sufrir alguna de estas enfermedades. A partir de este cálculo es posible personalizar la intensidad de las intervenciones preventivas según el riesgo absoluto del paciente. Así personas con mayor riesgo necesitarán intervenciones y tratamientos más intensos que aquellas con menor riesgo, maximizando los beneficios y minimizando los posibles daños derivados de un posible tratamiento excesivo [34].

Diferentes calculadoras son propuestas por las guías de prevención de enfermedad cardiovascular para el cálculo del riesgo de sufrir un evento cardiovascular (ECV) en 10 años. Todas estas herramientas tienen alguna limitación y están desarrolladas a nivel poblacional, por lo que para su interpretación a nivel individual es conveniente tener en cuenta las condiciones específicas de cada sujeto. Además, los estudios a partir de los que se desarrollan estas herramientas están desarrollados en poblaciones específicas, cuyas características puedan diferir de otras, por lo que las recomendaciones a cerca de qué método usar pueden variar en función de la población sobre la que se quiera calcular el riesgo, considerándose más apropiados aquellos que han sido desarrollados en poblaciones con características similares a los pacientes que queramos evaluar.

Las guías europeas de 2016[16] recomendaban aplicar el algoritmo SCORE (Systemic Coronary Risk Estimation), desarrollado a partir de los datos de distintas cohortes de 11 países europeos, y que estima el riesgo de sufrir un evento cardiovascular fatal en 10 años. Esta recomendación cambió en la última versión de esta guía, publicada en 2021 [15], que recomienda el uso de una adaptación de la herramienta anterior, SCORE2, que estima el riesgo de sufrir en 10 años un evento fatal o no fatal en personas sanas, entre 40 y 69 años con FRCV sin tratar. Sin embargo, las guías americanas[16] recomiendan el

uso de una herramienta diferente, la calculadora de riesgo PCE (pooled cohort equations), también conocida como calculadora ASCVD y la cual fue desarrollada a partir 4 cohortes distintas de población americana.

Otra de estas herramientas bien conocida es la calculadora de riesgo Framingham[16], la cual se desarrolló a partir de la cohorte con el mismo nombre y cuyo uso recomiendan las guías canadienses y NCEP (NAtional Cholesterol Education Program). Originariamente esta calculadora estimaba el riesgo de tener una enfermedad arterial coronaria en 10 años, pero posteriormente se adaptó para calcular el riesgo de evento cardiovascular. Esta calculadora ha sido adaptada y validada para otras poblaciones distintas a la americana, como por ejemplo para la población de Nueva Zelanda [16] o la española[44]. La adaptación española, llamada Framingham-REGICOR, se desarrolló a partir de la cohorte REGICOR y fue validada para la población de este país.

Otras herramientas desarrolladas con este fin son QRISK 2 y QRISK 3, desarrolladas para población inglesa, FINRISK, para la población danesa, y otras desarrolladas con población americana como Reynolds Risk y Globorisk[16].

Técnicas de machine learning para la estimación del riesgo cardiovascular

Como se ha mencionado previamente, los scores desarrollados hasta ahora se han realizado para poblaciones específicas, sin tener en cuenta si los sujetos reciben tratamiento para algún FRCV, y además tienen ciertas limitaciones metodológicas [45–47] debidas a la correlación entre variables, la no linealidad de las mismas y la posibilidad de sobreajuste. Por otro lado, la utilización de grandes cantidades de datos médicos generados en la práctica clínica diaria constituye una nueva y rica fuente de información a explorar en la investigación médica [48,49].

Las técnicas de aprendizaje automático o machine learning son una alternativa novedosa a las calculadoras tradicionales de riesgo y recientemente se han aplicado en algunas cohortes [48,50] para analizar estos enormes conjuntos de datos y superar algunas de las limitaciones mencionadas de las calculadoras tradicionales, incluso mejorando los resultados obtenidos con estas. El aprendizaje automático puede utilizarse para generar

modelos que predigan mejor el riesgo, aumentando así la eficiencia, objetividad y fiabilidad del proceso diagnóstico [45,46,48,51] . En concreto, estas técnicas de aprendizaje supervisado utilizan datos existentes para entrenar modelos mediante el aprendizaje de patrones que posteriormente se aplicarán para predecir otra variable. Al aplicar estas técnicas a la investigación de enfermedades, existen algunos problemas que deben tenerse en cuenta durante el análisis.

Estos problemas están relacionados con la interpretabilidad de los modelos y con la existencia de datos desbalanceados, la calidad y la cantidad de los datos. La escasa interpretabilidad se debe a que funcionan como cajas negras, lo que dificulta la interpretación de sus resultados, aunque se están desarrollado distintas herramientas para mejorar este aspecto. La presencia de datos desbalanceados es bastante común en investigación sanitaria, ya que siempre hay menos personas enfermas que sanas, por lo que normalmente los datos están desequilibrados y la probabilidad de evento es distinta a la de no evento. Por último, la calidad y la cantidad de los datos médicos suelen ser bajas debido a las fuentes de información disponibles y a cuestiones de accesibilidad y ética [52].

Existen distintos tipos de aprendizaje automático supervisado. Cuando la variable a predecir es categórica (por ejemplo, un evento cardiovascular), se utilizan técnicas de aprendizaje automático de clasificación [45,46] como los algoritmos Naïve Bayes (NB) y los métodos ensemble [53]. Los algoritmos NB utilizan la regla de Bayes para estimar la probabilidad de que los nuevos datos pertenezcan a cada una de las categorías de la variable a predecir, y los asignan a la categoría para la que se ha calculado la probabilidad más alta. Los métodos ensemble, que incluyen los métodos bagging y boosting, combinan múltiples árboles de decisión para hacer la clasificación. Uno de los métodos de ensamblaje más utilizados es Random Forest (RF), mediante el cual se aprenden múltiples árboles de decisión en paralelo y la predicción final se basa en la respuesta más frecuente [48,53].

Justificación del estudio

La alta incidencia de la enfermedad cardiovascular y la gravedad de sus consecuencias hace que este grupo de enfermedades sean un problema de salud pública. La mayor parte de guías recomiendan el cálculo del riesgo cardiovascular para establecer las medidas preventivas a tomar. Por lo tanto, la estimación precisa del riesgo es esencial para la prevención y la gestión efectiva de las recomendaciones y tratamientos a tomar.

Por otra parte, la evolución de la enfermedad cardiovascular y sus consecuencias están influidas por la variabilidad en la exposición a los FRCV a lo largo de la vida. El análisis longitudinal de la evolución de estos factores es una herramienta para comprender mejor esta relación y contribuir al desarrollo de estrategias de prevención más efectivas. La utilización de herramientas como el análisis de clusters longitudinal para la caracterización de perfiles de pacientes con enfermedad cardiovascular se presenta como una buena oportunidad. Este método permite la agrupación de individuos de acuerdo con similitudes en variables clave, lo que facilita la identificación de perfiles basados en factores de riesgo cardiovascular y el riesgo cardiovascular. Además, la caracterización de estos perfiles no solo resulta valiosa para la identificación de estrategias preventivas, sino que también constituye una herramienta para abrir nuevas líneas de investigación en el ámbito de la enfermedad cardiovascular.

Para la estimación del riesgo, tradicionalmente se han utilizado distintos scores desarrollados en poblaciones específicas. Sin embargo, las técnicas de aprendizaje automático ofrecen un enfoque prometedor para mejorar la precisión de estas predicciones. Entre otras ventajas, estas técnicas permiten la utilización de una gran cantidad de datos de vida real para predecir la aparición de eventos cardiovasculares, e incluir variables fundamentales en el manejo del riesgo cardiovascular, como la adherencia a tratamientos para la prevención cardiovascular.

Este enfoque innovador busca analizar la capacidad de varios algoritmos para predecir eventos cardiovasculares y explorar cómo la inclusión de variables relacionadas con los FRCV y manejo del riesgo mejora estas predicciones. La evaluación de diferentes algoritmos y su capacidad para incorporar variables relacionadas con los FRCV en la predicción de ECV constituye un aporte novedoso y relevante.

Por otra parte, la literatura muestra una desigualdad de género en los estudios cardiovasculares, con una atención históricamente centrada en hombres. Además, la inclusión de determinantes sociales de la salud ha sido limitada, y los factores que explican las desigualdades entre sexos en los factores de riesgo cardiovascular han sido escasamente explorados. El análisis contrafactual se presenta como un enfoque novedoso para estudiar el impacto de las diferencias de factores de riesgo cardiovascular y nivel socioeconómico entre hombres y mujeres en el desarrollo de enfermedad cardiovascular. Identificar estas fuentes de desigualdad puede ofrecer oportunidades para reducir disparidades entre grupos.

Finalmente, en Aragón se han desarrollado distintas cohortes para el estudio de la enfermedad cardiovascular. La cohorte AWHS está compuesta por 5400 trabajadores de una fábrica, mayoritariamente hombres con características sociodemográficas homogéneas. De esta cohorte se obtienen datos tanto en los reconocimientos médicos anuales como del Sistema Aragonés de Salud. Por otra parte, la cohorte CARhES está compuesta por hombres y mujeres de Aragón con algún factor de riesgo cardiovascular. Esta segunda cohorte cuenta con características más heterogéneas, aumentando la validez externa de los resultados que se obtengan del estudio con esta cohorte.

En síntesis, a través del análisis de la evolución de factores de riesgo cardiovascular, del empleo de enfoques avanzados para la caracterización de perfiles de pacientes, de la predicción del riesgo de enfermedad cardiovascular, y de la comprensión de las disparidades de género y sociales vinculadas a esta enfermedad, se podría obtener información crucial. Esto no solo permitiría el diseño de estrategias preventivas más personalizadas y eficaces, y que también aportaría valiosas perspectivas para la formulación de políticas de salud pública.

Hipótesis

La prevalencia de factores de riesgo cardiovascular es elevada, mostrando variaciones notables en función de las características particulares de la población estudiada. Además, la prevalencia de factores de riesgo cardiovascular es más alta entre aquellas personas que sufren un evento cardiovascular.

Los algoritmos de machine learning son una alternativa adecuada a los métodos tradicionales para estimar el riesgo y predecir la aparición de eventos cardiovasculares. La edad tiene un peso importante en la predicción y el riesgo de sufrir un evento cardiovascular. La inclusión de variables relacionadas con la adherencia a los tratamientos mejora la precisión de las predicciones de riesgo.

Los sujetos incluidos en la cohorte AWHS se agrupan según la evolución de sus factores de riesgo cardiovascular y su riesgo cardiovascular, que aumentan con el tiempo.

Existen diferencias significativas por sexo en la prevalencia de factores de riesgo cardiovascular y en la incidencia de eventos cardiovasculares mayores (MACE) en la cohorte CARhES. Las diferencias en la incidencia de MACE por sexo pueden atribuirse a disparidades entre hombres y mujeres en la prevalencia de hipertensión, hipercolesterolemia, diabetes, y en el nivel socioeconómico.

2. OBJETIVOS

2. OBJETIVOS

Objetivo general

Estimar la frecuencia de factores de FRCV en dos poblaciones diferentes de Aragón, cohortes AWHS y CARhES, describiendo su evolución en función de perfiles de pacientes; estudiar la adherencia a tratamientos; aplicar diferentes técnicas para estimar el riesgo de enfermedad cardiovascular, y analizar las diferencias entre sexos en la incidencia de MACE, estudiando posibles factores que influyen en estas diferencias.

Objetivos específicos

1.- Trabajo con datos de la cohorte AWHS

1.1- Realizar un análisis descriptivo de la frecuencia de factores de riesgo cardiovascular (hipertensión, hipercolesterolemia, diabetes y estado físico), exposición a tratamientos preventivos e incidencia de eventos cardiovasculares en la cohorte AWHS.

1.2- Describir los valores analíticos y variables médicas relacionadas con factores de riesgo cardiovascular y el SCORE, analizando la evolución de los mismos.

1.3- Analizar la capacidad de diferentes métodos de machine learning para predecir la aparición de eventos cardiovasculares y describir la influencia de distintos factores de riesgo cardiovascular y la exposición a tratamientos preventivos, en la incidencia de evento mediante la aplicación de dichos modelos incluyendo las siguientes variables: edad, hipertensión, hipercolesterolemia, diabetes, estado físico y exposición al tratamiento.

1.4- Identificar perfiles de participantes en la cohorte AWHS en función de la evolución de factores de riesgo cardiovascular y del SCORE utilizando la información de tres momentos, aplicando técnicas de cluster longitudinal.

2.- Para el trabajo con datos de la cohorte CARhES

2.1- Describir la prevalencia de factores de riesgo en la población de Aragón, así como la frecuencia de factores de riesgo cardiovascular, adherencia a tratamientos e incidencia de MACE en la cohorte CARhES.

2.2- Analizar la diferencia en la prevalencia de factores de riesgo cardiovascular y nivel socioeconómico y en la incidencia de MACE entre hombres y mujeres.

2.3- Analizar la capacidad de distintos métodos de machine learning para predecir la incidencia de MACE en la cohorte CARhES de manera separada para hombres y mujeres, analizando la influencia de 4 grupos de variables (Edad, FRCV, valores analíticos y mediciones de tensión arterial y adherencia a tratamientos antihipertensivos, antidiabéticos e hipolipemiantes) en dicha predicción.

2.4- Estudiar el impacto que tienen las diferencias en la distribución de hipertensión, hipercolesterolemia, diabetes y nivel socioeconómico entre sexos en las diferencias observadas en la incidencia de MACE.

3. MATERIAL Y MÉTODOS

3. MATERIAL Y MÉTODOS

En este apartado se presentará, en primer lugar, la metodología aplicada en el estudio centrado en la cohorte AWHS y, a continuación, la utilizada en el análisis de la cohorte CARhES para alcanzar los objetivos planteados.

Estudio en la cohorte AWHS

Descripción de la cohorte

Como se ha comentado, algunos de los objetivos propuestos corresponden al estudio realizado en el marco del AWHS, un estudio prospectivo longitudinal de cohortes que incluye trabajadores de una planta de montaje de automóviles situada en Figueruelas (Zaragoza, España). Esta cohorte se diseñó con el objetivo principal de evaluar la evolución de FRCV, tanto tradicionales como emergentes, y su asociación con la prevalencia y progresión de la aterosclerosis subclínica. En el estudio se analizaron anualmente los FRCV y parámetros clínicos mientras que cada tres años se llevaron a cabo pruebas más exhaustivas, realizadas en una muestra de la cohorte, que incluían imágenes para detectar aterosclerosis subclínica. Además, se realizó un seguimiento de la utilización de los servicios sanitarios por parte de los participantes. El reclutamiento de sujetos se llevó a cabo entre febrero de 2009 y diciembre de 2010 y hoy en día su seguimiento continúa.

Dentro de los objetivos específicos del estudio AWHS se incluyeron a) la creación de un biobanco con las muestras biológicas de suero, plasma, sangre, orina y ADN; b) la identificación de nuevos determinantes genéticos, conductuales y ambientales de la progresión de la adiposidad y del desarrollo de anomalías metabólicas y FRCV; y c) la caracterización de la prevalencia y la progresión de la ECV subclínica mediante técnicas de imagen no invasivas y sus determinantes genéticos, conductuales y ambientales.

La cohorte estuvo compuesta por 5048 hombres y 351 mujeres y su media de edad (DE) fue de 49,3 (8,7) años para los hombres y 40,8 (11,6) años para las mujeres. En el momento del reclutamiento, entre los hombres, el 55% tenían sobrepeso; 23,1%

obesidad; el 37,1% eran fumadores en ese momento; 40,3% tenían HTA; el 75% HC y el 7,4% DM. Estos porcentajes fueron menores entre las mujeres: 23,7%; 8,3%; 45,0%; 12,1%; 59,5% y 0,6%, respectivamente. Las pruebas de imagen iniciales se realizaron sobre 587 sujetos, la mayoría hombres, y mostraron que 67,7% de ellos tenía, al menos, una placa aterosclerótica en las arterias carótida o femoral. Puede encontrarse más información sobre el estudio AWHS en Casasnovas et al[43].

Selección de sujetos para la presente tesis

Para el análisis realizado en la tesis que se presenta, solo se consideraron los hombres incluidos en la cohorte, debido al bajo número de mujeres (N=380). Además, se excluyeron a los trabajadores diagnosticados de cualquier enfermedad cardiovascular antes de la inclusión en el AWHS, y aquellos que carecían de una tarjeta sanitaria emitida por el sistema sanitario público de Aragón, entre la fecha de inclusión en la cohorte y el 31 de mayo de 2019. También se excluyeron ocho hombres de los que no se pudieron obtener datos en el registro del sistema de salud de Aragón (SALUD). Estos criterios comunes a todos los análisis realizados dentro del AWHS se muestran en la Figura 6. Además, en esta figura también se indican los criterios de selección específicos para cada objetivo y que se desarrollarán más adelante.

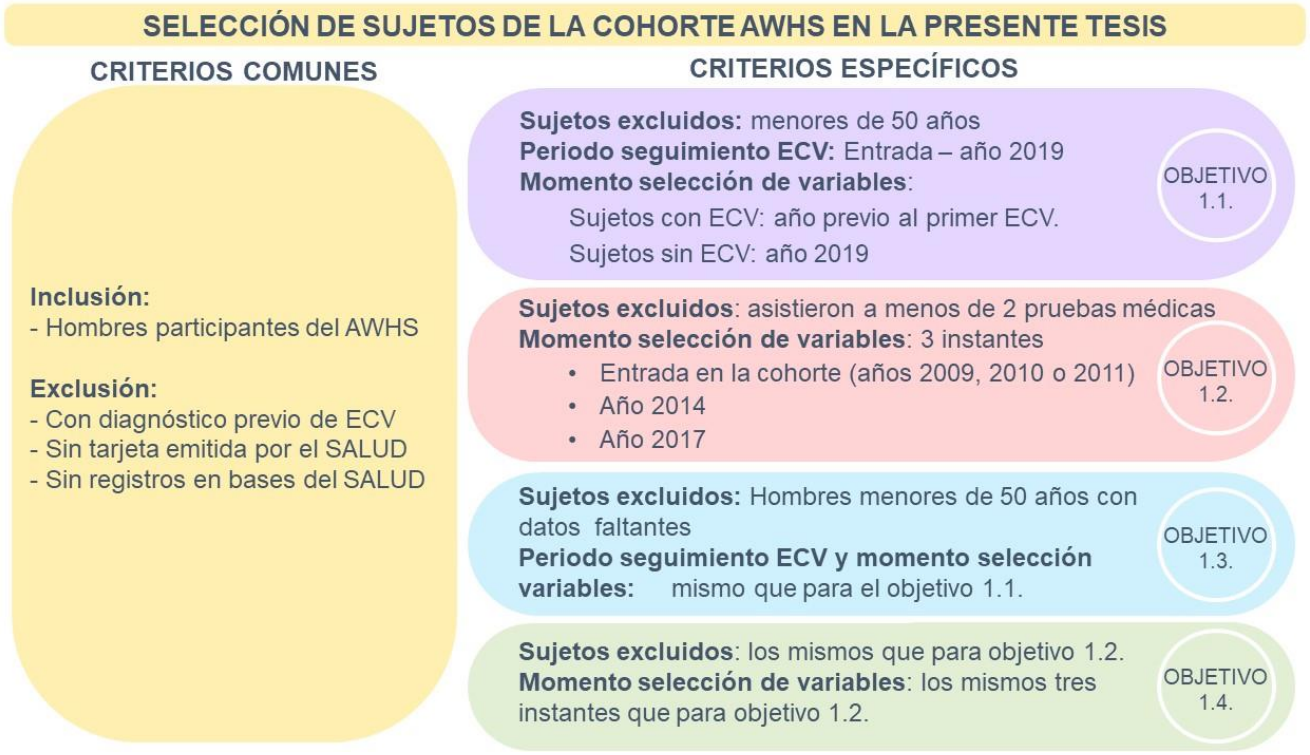


Figura 6: Criterios de selección de sujetos y periodos de estudio de la cohorte AWHS.

Variables del estudio y fuentes de información utilizadas para la presente tesis

Variables obtenidas de las bases de datos

Se utilizaron datos recogidos de tres fuentes distintas de información: la propia cohorte del AWHS; BIGAN, un repositorio de datos sanitarios puesto a disposición de investigadores; y la Dirección General de Salud Pública. A continuación, se especifican todas las bases de datos utilizadas, su procedencia, y la información extraída de cada una de ellas (Tabla 1).

Tabla 1: Bases de datos consultadas e información obtenida de la cohorte AWHS.

Bases de datos	Variables
AWHS	Altura, peso, perímetro de cintura, TAS y TAD, IMC, glucosa sérica en ayunas, colesterol y colesterol HDL, hábito tabáquico.
Dispensación farmacéutica	Código ATC, DDD, número envases dispensados.
CMBD	Código CIE-10 y fecha de ingreso
Urgencias	Código CIE-9 y fecha del ingreso
Registro de Mortalidad de Aragón	Causa de fallecimiento y fecha

TAS: tensión arterial sistólica; TAD: tensión arterial diastólica; IMC: índice de masa corporal; ATC: código anatómico terapéutico químico; DDD: número de dosis diarias definidas; CMBD: Conjunto Mínimo de Datos; CIE-9: Clasificación Internacional de Enfermedades, 9ª revisión; CIE-10: Clasificación Internacional de Enfermedades, 10ª revisión.

Del estudio AWHS se incluyeron datos de las pruebas médicas anuales de los trabajadores (incluidos los análisis de sangre). Estos datos fueron recogidos por los médicos y enfermeras de los servicios médicos de la fábrica, quienes recibieron formación previa. Todos los procedimientos del estudio fueron estandarizados.

BIGAN es un repositorio de datos sanitarios que recoge información del Servicio Aragonés de Salud y que están disponibles para la investigación previa solicitud. De este repositorio se obtuvieron datos de a) la Base de Datos de Dispensación Farmacéutica, que recoge información sobre la fecha de dispensación, el código Anatómico Terapéutico Químico (ATC), el número de dosis diarias definidas (DDD) y el número de envases dispensados por las farmacias y financiados por el Servicio Aragonés de Salud; b) el CMBD (Conjunto Mínimo Básico de Datos), que registra diagnósticos y fechas de hospitalizaciones; y c) la base de datos de Urgencias, que registra diagnósticos y fechas correspondientes a la utilización de servicios de urgencias hospitalarias.

Por último, la información sobre la fecha y la causa de muerte se obtuvo del Registro de Mortalidad de Aragón a través de la Dirección General de Salud Pública.

Todos los datos estuvieron pseudoanonimizados, utilizando un código único que vincula la información del paciente entre las distintas fuentes, pero que impide su identificación personal

Las variables procedentes del AWHS fueron recogidas en la revisión médica anual realizada en la empresa, en las que los participantes en el estudio proporcionaron datos sobre su historia clínica anual y se sometieron a un examen físico, incluyendo antropometría (altura, peso y perímetro de cintura) y se les realizaron analíticas.

Los niveles de glucosa sérica en ayunas, colesterol total y colesterol HDL se registraron a partir de análisis de laboratorio, previa obtención de muestras de sangre y orina de los trabajadores tras ayuno nocturno (>8 horas). Las muestras se procesaron el mismo día de la extracción, se midieron mediante espectrofotometría y se registraron como variables cuantitativas en mg/dL.

La TAS y la TAD se registraron como variables cuantitativas, tras medirse 3 veces consecutivas, con un esfigmomanómetro automático y con el participante sentado tras un descanso de 5 minutos.

La estatura, el peso y el perímetro de cintura se registraron como variables cuantitativas. El índice de masa corporal (IMC) se calculó a partir de la estatura y el peso.

El hábito de tabáquico fue autodeclarado, utilizando las categorías de fumador actual, exfumador o no fumador y se registró como variable cualitativa, teniendo datos hasta el año 2014. Los datos sobre la actividad física y la ingesta de alcohol no se registraron para todos los sujetos incluidos en la cohorte, ni en todos los años de seguimiento. Se valoró incluir la ingesta de alcohol como variable explicativa junto con el hábito tabáquico, y se analizó la correlación de ambas variables para los años e individuos que estuvieron disponibles. Este análisis mostró que ambas variables estaban altamente correlacionadas, por lo que la ingesta de alcohol no fue considerada para incluirla en los análisis.

Variables calculadas

Además de estas variables recogidas directamente de los exámenes médicos y entrevistas anuales, se calcularon otras necesarias para dar respuesta a los distintos objetivos fijados en esta tesis que se explican a continuación.

Para la **identificación de FRCV** se examinaron los resultados de los análisis médicos y sanguíneos y los datos de la Base de Datos de Dispensación Farmacéutica. Se clasificó a los sujetos como afectados por el FRCV si cumplían los criterios de al menos una de esas bases de datos como tales. A partir de los datos médicos y de análisis de sangre, la clasificación de los sujetos con FRCV se realizó aplicando los siguientes puntos de corte, tal como recomiendan las directrices europeas de prevención de la ECV [16]:

- HTA se definió como una TAD ≥90 mmHg y/o una TAS ≥140 mmHg.
- HC como un colesterol total ≥200 mg/dl o un colesterol LDL ≥115 mg/dl.
- DM como una glucosa sérica en ayunas ≥126 mg/d.
- En cuanto al estado físico, se consideró que un sujeto tenía un peso normal si IMC <25, sobrepeso se definió como un IMC ≥25 y <30, y obesidad como un IMC ≥30.

A partir de la información de la Base de Datos de Dispensación Farmacéutica, se consideró que los individuos eran hipertensos en un año dado si se les había dispensado al menos una receta en ese año, correspondiente a los siguientes códigos ATC: C02 (antihipertensivos), C03 (diuréticos), C07 (betabloqueantes), C08 (antagonistas del calcio) y C09 (agentes que actúan sobre el sistema renina-angiotensina). Dado que los diuréticos y los betabloqueantes también se prescriben para otras indicaciones, la dispensación de estos fármacos sólo se consideró un indicador de HTA si el individuo cumplimentaba al menos tres dispensaciones distintas en el mismo año [54]. Se consideró que los participantes tenían HC si se les había dispensado al menos un fármaco correspondiente al código ATC C10 (agentes modificadores de lípidos) y DM si se les había dispensado al menos un fármaco correspondiente al código ATC A10 (fármacos utilizados en la DM).

La variable **exposición al tratamiento** se determinó cuantificando la adherencia a tres tratamientos preventivos distintos (antihipertensivos, antidiabéticos e hipolipemiantes), teniendo en cuenta datos de la base de dispensación farmacéutica.

Para estandarizar la terminología relacionada con la adherencia a las terapias farmacológicas, el proyecto europeo ABC propuso una Taxonomía de la Adherencia que

consta de tres fases: iniciación, implementación y discontinuación. La iniciación del tratamiento corresponde a la toma de la primera dosis de un medicamento prescrito. El proceso continúa con la implementación del régimen de dosificación, definido como el grado en que la dosificación real del paciente se corresponde con el régimen de dosificación prescrito, desde el inicio hasta la toma de la última dosis. La discontinuación marca el final de la terapia, cuando se omite la siguiente dosis a tomar y no se toman más dosis a partir de entonces. La persistencia es el tiempo transcurrido entre el inicio y la última dosis, que precede inmediatamente a la interrupción.

De las diferentes fases de adherencia al tratamiento, el presente análisis se centró en la fase de implementación, definido por el proyecto ABC [35] como el grado en que la dosificación real de un paciente se corresponde con el régimen de dosificación prescrito, desde el inicio del tratamiento hasta el consumo de la última dosis.

La adherencia a los tratamientos de HC, HTA y DM se determinó por separado para cada participante y se calculó a través de la Proporción de Días Cubiertos (PDC), calculada como porcentaje. La PDC es un índice calculado como el número de días cubiertos por los medicamentos dispensados por la farmacia dividido por el número de días que el sujeto debería haber tenido cubiertos [55]. En este estudio, el denominador para el PDC fue de 365 días, excepto en los casos en los que los sujetos iniciaron el tratamiento una vez que el periodo de seguimiento ya había comenzado. En estos casos, el denominador fue el número de días transcurridos desde el inicio del tratamiento hasta el final del periodo de seguimiento. El número de días cubiertos se calculó a partir de la DDD dispensada a cada sujeto. Sin embargo, un estudio previo de nuestro grupo [55] demostró que el uso de un valor subrogado de la dosis diaria de cada fármaco, calculado a partir de la posología y forma de presentación habituales, proporcionaba resultados más precisos. Por lo tanto, en el estudio se utilizaron valores subrogados para las dosis diarias.

Para cada sujeto, la PDC obtenida para los tres FRCV se resumió en una nueva variable: la exposición al tratamiento. Esta variable se clasificó en 3 posibles categorías: totalmente expuestos, individuos a los que se les dispensaron recetas para el tratamiento de todos los FRCV identificados y tuvieron una PDC ≥80% para todos ellos; no

expuestos, individuos a los que no se les dispensaron recetas para ninguno de los FRCV identificados o tuvieron una PDC <80% para todos los tratamientos tomados; parcialmente expuestos, individuos que no tuvieron recetas dispensadas para al menos un FRCV identificado y tuvieron una PDC ≥80% para otros o una PDC <80% para algún tratamiento y ≥80% para otros de los tratamientos tomados.

Para estimar el **riesgo cardiovascular** de cada participante se utilizó la herramienta SCORE[56], que da una puntuación del riesgo individual de sufrir ECV a 10 años y que fue diseñada para aplicarse en población europea con bajo RCV. Para calcular esta variable, se tuvieron en cuenta las variables hábito tabáquico, colesterol total y TAS. Atendiendo al hábito tabáquico se dividió a los trabajadores en 2 grupos: a) fumadores actuales y b) no fumadores y ex fumadores.

Variable resultado

Finalmente, la variable resultado considerada en esta parte de la presente tesis fue la incidencia de un ECV. Para ello se identificó a los individuos que experimentaron una ECV en cualquier momento entre la inclusión en la cohorte y el 31 de diciembre de 2019, y se registró la fecha y la naturaleza de cada ECV. Los eventos se identificaron a partir de los datos del CMBD, la base de datos de Urgencias hospitalarias y el Registro de Mortalidad de Aragón aplicando los siguientes criterios:

(i) Para los hombres con registros sólo en una de las bases de datos, CMBD o Urgencias, se seleccionaron los datos del primer registro disponible en estas.

(ii) Para los hombres con registros tanto en la base de datos de CMBD como en la de Urgencias, se comprobó si los primeros registros de cada base de datos coincidían en tiempo y diagnóstico. En caso afirmativo, se eligieron los datos de la base de datos del CMBD. En caso contrario, y si el registro del CMBD era anterior al de la base de datos de Urgencias, se seleccionaba el primero. Por el contrario, si el registro de la base de datos de Urgencias era anterior al de la base de datos de la CMBD, se comprobó si correspondía con un registro de la base de datos de la CMBD que no era ECV: en caso

afirmativo, se rechazaba el registro de la base de datos de Urgencias; en caso negativo, se seleccionaba el registro de la base de datos de Urgencias.

(iii) Por último, se analizó el Registro de Mortalidad de Aragón para identificar a los sujetos cuyo primer evento fue un ECV mortal.

Los diagnósticos se registraron según la Clasificación Internacional de Enfermedades, 9ª revisión (CIE-9) en la base de datos de Urgencias, y según la Clasificación Internacional de Enfermedades, 10ª revisión (CIE-10) en el CMBD y el Registro de Mortalidad de Aragón. Se consideraron los siguientes códigos CIE-9 y CIE-10: CIE-9 410-415 y CIE-10 I20-I25 (cardiopatías); CIE-9 415-417 y CIE-10 I26-I28 (cardiopatías pulmonares y enfermedades de la circulación pulmonar); CIE-9 427.4, 427.5, 428, 429.2 y CIE-10 I46, I49. 0, I50 (otras cardiopatías); CIE-9 430-438 y CIE-10 G45-G46 y I60-I69 (enfermedades cerebrovasculares); CIE-9 440-445 y CIE-10 I70-I79 (enfermedades de las arterias, arteriolas y capilares).

En la Tabla 2 se muestran los objetivos para los que cada variable fue utilizada así como su procedencia.

Tabla 2: Relación variables y objetivos en la cohorte AWHS.

	OBJETIVO 1.1.-	OBJETIVO 1.2.	OBJETIVO 1.3.	OBJETIVO 1.4.	Procedencia
Edad	X	X	X	X	AWHS
Hipertensión	X		X		Calculadas a partir de CMBD y dispensación de farmacia
Hipercoleste-rolemia	X		X		
Diabetes	X		X		
TAS y TAD		X		X	AWHS
Peso		X		X	AWHS
Perímetro cintura		X		X	AWHS
IMC		X		X	AWHS
Colesterol total, HDL y LDL		X		X	AWHS
Glucosa		X		X	AWHS
Hábito tabáquico		X		X	AWHS
Estado físico	X		X		Calculada a partir de IMC
Exposición tratamiento	X		X		Calculada a partir de dispensación de farmacia
SCORE		X		X	Calculada a partir de hábito tabáquico, colesterol total y TAS
Evento cardiovascular	X		X		Calculada a partir de CMBD, Urgencias y Registro de Mortalidad de Aragón

TAS: tensión arterial sistólica; TAD: tensión arterial diastólica; AWHS: Aragon Workers' Health Study; IMC: índice de masa corporal; CMBD: conjunto minimo de datos

Selección de sujetos y análisis realizados para dar respuesta a los objetivos planteados

En esta sección se mostrará la información por separado para cada uno de los objetivos propuestos. En la Figura 6 se muestran los criterios de selección de sujetos comunes a todos los análisis realizados dentro de la cohorte AWHS y los criterios específicos considerados para cada análisis, así como los periodos de seguimiento y momentos en los que se han tomado las variables.

Metodología utilizada para dar respuesta al OBJETIVO 1.1.- Análisis descriptivo de la frecuencia de factores de riesgo cardiovascular (hipertensión, hipercolesterolemia, diabetes y estado físico), exposición a tratamientos preventivos e incidencia de ECV en la cohorte AWHS.

Dada las características especiales de sujetos con ECV en edades tempranas, y su baja incidencia en estas edades, para la consecución del presente objetivo se seleccionaron los hombres sin ECV previa y mayores de 49 años.

Las variables incluidas fueron la edad, los FRCV, la exposición a tratamientos preventivos y ECV. Los factores de riesgo considerados fueron HTA, HC, DM y estado físico. No se tuvo en cuenta el hábito tabáquico ya que no se disponía de los datos correspondientes para el periodo analizado. Los FRCV se identificaron para el año anterior al primer ECV en los individuos que sufrieron un evento, y para 2019 en los individuos sin ECV durante el periodo de estudio.

Para la descripción de las variables de interés, las variables continuas se expresaron como media y desviación estándar (DE) y las variables categóricas como porcentajes

Metodología utilizada para dar respuesta al OBJETIVO 1.2.- Describir los valores analíticos y variables médicas relacionadas con FRCV y el SCORE, analizando la evolución de los mismos.

Para alcanzar este objetivo, se seleccionaron los datos de los sujetos en 3 momentos diferentes. Los hombres seleccionados para la presente tesis, no todos tuvieron datos de analíticas de los tres momentos temporales seleccionados para el estudio, por lo que el número de trabajadores de los que se obtuvieron datos para cada momento fue

diferente. Se dispuso de datos de 5122 trabajadores para el momento 1; 3891 para el momento 2 y 3545 para el momento 3. Por último, se excluyeron los hombres que no habían asistido al menos a 2 de las 3 pruebas médicas realizadas en el tiempo de seguimiento (N = 975). La población final del estudio estaba formada por 4147 individuos.

El momento primero correspondió al primer registro de datos disponible para cada individuo después de proporcionar el consentimiento informado (es decir, datos recopilados en 2009, 2010 o 2011, dependiendo de cuándo se firmó el formulario de consentimiento y de la disponibilidad de datos); el momento 2 correspondió a la mitad del periodo de estudio (2014); y el momento 3 correspondió al último registro de datos disponible para cada participante (se priorizaron los datos del año 2017, pero si no estaban disponibles, se seleccionaron los datos del año 2016).

Las variables descritas para este objetivo fueron: edad, TAS y TAD, peso, perímetro de cintura, IMC, colesterol total, HDL y LDL, glucosa en sangre, hábito tabáquico y SCORE, que mide el riesgo de sufrir ECV en 10 años.

Dado que en este caso se disponía del hábito tabáquico para los dos primeros momentos, de los tres seleccionados, se realizó una imputación de esta variable para ese momento mediante un procedimiento de proyección [57]. Con este fin, se generó una matriz de probabilidad de transición para el hábito tabáquico basada en los datos disponibles para los momentos 1 y 2, y se utilizó para estimar el hábito tabáquico de cada trabajador para el momento 3, asumiendo estacionariedad en el comportamiento de fumar.

La descripción de las variables se realizó mediante la media y la DE para las variables cuantitativas, y porcentajes para las variables categóricas. Además, se realizó un análisis de la evolución por cuartiles. En este se identificó el porcentaje de individuos que pasaron de un cuartil a otro entre los momentos 1 y 2 y los momentos 2 y 3 para las siguientes variables: índice de masa corporal, glucemia y SCORE de riesgo de enfermedad cardiovascular.

Metodología utilizada para dar respuesta al OBJETIVO 1.3.- Analizar la capacidad de diferentes métodos de machine learning para predecir la aparición de ECV y describir la influencia de distintos FRCV y la exposición a tratamientos preventivos en la incidencia de evento mediante el análisis de dichos modelos, incluyendo las siguientes variables: edad, HTA, HC, DM, estado físico y exposición al tratamiento.

Para dar respuesta a este objetivo se partió de los mismos sujetos que para el objetivo 1.1. Debido al bajo número de sujetos con datos perdidos, para este objetivo se excluyeron los sujetos sin datos en cualquier variable (N=89). En cuanto a las variables consideradas, fueron las mismas que para el objetivo 1.1.

Los métodos aplicados en este objetivo fueron algoritmos supervisados de aprendizaje automático o algoritmos de machine learning para determinar la utilidad de las distintas variables en la predicción de ECV. Estos algoritmos incluyeron RF, XGBoost y NB [45,46,49]. Estos métodos utilizan un conjunto de datos de entrenamiento para crear modelos que den lugar a una salida lo más parecida a la realidad posible. Los modelos son capaces de tomar la información disponible en el conjunto de datos de entrenamiento para ir aprendiendo de los fallos que va cometiendo. Por esto, se necesitan dos conjuntos de datos distintos: el primero, un conjunto de datos para crear y entrenar a los modelos que se denomina conjunto de entrenamiento; el segundo conjunto de datos se utiliza para evaluar la validez y precisión de los modelos desarrollados con datos distintos a los utilizados en la fase de entrenamiento.

Así pues, la población de estudio se dividió aleatoriamente en dos grupos: el 70% de la muestra formó el grupo de entrenamiento y el 30% restante formó el grupo de prueba. También se utilizó la muestra completa para probar los distintos algoritmos, ya que el tamaño de la muestra del grupo de prueba era demasiado pequeño para validar los modelos obtenidos. Para ajustar los hiperparámetros y evitar el sobreajuste, se probó la precisión de predicción de todos los modelos mediante validación cruzada de 5 y 10 iteraciones para estimar el F1-score. Los resultados fueron similares al aplicar 5 y 10 iteraciones, por lo que los resultados mostrados corresponden a la validación cruzada de 5 iteraciones. Debido a que los datos disponibles eran muy desequilibrados, el umbral se movió y seleccionó en función del máximo de F1 score en las curvas P-R, siendo 0,1 en todos los modelos.

Se ajustaron los siguientes parámetros para los modelos RF: número de árboles, número de variables a considerar en cualquier división, profundidad máxima, que es el número máximo de particiones en la rama más larga del árbol, y número mínimo de observaciones en un nodo. En el caso de XGBoost, los parámetros ajustados fueron el número de variables a considerar en cada división, la profundidad máxima y el número mínimo de observaciones en un nodo. Para el método NB, las probabilidades a priori fueron 0,9 para la no ocurrencia de CVE y 0,1 para la ocurrencia de CVE. Para medir la validez de los modelos, se tuvieron en cuenta las siguientes medidas: exactitud, sensibilidad, especificidad, valor predictivo positivo (VPP) y valor predictivo negativo (VPN). A continuación, se aplicaron tres pruebas diferentes para evaluar el rendimiento del modelo: área bajo la curva de precisión-recall (AUC-PR), Log Loss y F1 score [58]. Tras conseguir modelos válidos y precisos, se extrajo la contribución de cada variable a la predicción y se normalizó en una escala de 0-1 para facilitar la comparabilidad. Cada método se aplicó dos veces: primero incluyendo sólo los FRCV como variables, y de nuevo incluyendo tanto los FRCV como la exposición al tratamiento.

Metodología utilizada para dar respuesta al OBJETIVO 1.4.- Identificar perfiles de participantes en la cohorte AWHS en función de la evolución de FRCV y del SCORE utilizando la información de tres momentos, aplicando técnicas de cluster longitudinal.

Para el presente objetivo se planteó la identificación de perfiles de pacientes que reunieran a sujetos con trayectorias de la evolución de los FRCV similares. Para la Identificación de estos perfiles, sobre los sujetos seleccionados para alcanzar el objetivo 1.2., descrito anteriormente, se realizó un análisis de conglomerados o clusters**,** teniendo en cuenta las variables: edad, IMC, perímetro de cintura, colesterol HDL, niveles de glucosa y la puntuación SCORE. Los FRCV elegidos se seleccionaron específicamente para el análisis porque no se utilizaron en el cálculo del SCORE. También se realizó un análisis de correlación de las variables incluidas en el análisis de conglomerados.

Las técnicas de cluster son una forma de aprendizaje no supervisado que tratan de agrupar los elementos a estudio en grupos homogéneos basándose en similitudes entre ellos [59]. Como se ha comentado, esta parte del estudio fue un estudio longitudinal de cohortes en el que cada variable se midió en diferentes momentos y cambió con el tiempo

para cada individuo. El método estándar para agrupar las trayectorias que siguen las variables a lo largo del periodo de seguimiento consiste en hacer grupos para cada variable por separado, pero en estudios donde se trata de agrupar las trayectorias de más de una variable, el análisis de cluster permite analizar la evolución conjunta de las variables de interés.

Para el análisis se utilizó el paquete Kml3D en el software estadístico R [60]. Este algoritmo, basado en el método de k-medias, utiliza una noción generalizada de distancia entre trayectorias individuales, y se utilizó para agrupar a los individuos según la evolución de los FRCV y el SCORE estimado en 3 momentos distintos. Se utilizó el criterio Calinski-Harabasz para determinar el número de grupos en que debían dividirse los participantes. La aplicación del algoritmo Kml3D requiere que los datos de todas las variables incluidas en el análisis estén disponibles para todos los participantes. Dado que algunos trabajadores no acudieron a todas las pruebas médicas no se disponía de todos los datos en los tres momentos estudiados, por lo que fue necesario imputar algunos valores de estas variables antes de aplicar el algoritmo [60,61]. Para ello se utilizó la función de imputación del paquete R longitudinalData[57], concretamente el comando "linearInterpol.bisector", que distingue entre datos perdidos monótonos e intermitentes. En el primer caso, crea la bisectriz de a) la línea que une los dos primeros o los dos últimos datos no ausentes (dependiendo de si los datos ausentes se recogen al principio o al final del periodo de estudio) y b) la línea que une el primer y el último dato no ausente. En el segundo caso, crea una línea que une los valores que rodean inmediatamente al valor que falta.

En la Tabla 3 se muestran las medias de las variables del estudio calculadas tanto con los datos disponibles como con los imputados. En ellos se observa que la imputación se puede considerar correcta, ya que las medias de las variables no variaron de forma importante. La adecuación de la imputación se confirmó mediante análisis descriptivos y comparación de medias mediante una prueba t de Student (todos obtuvieron valores $p<0,05$).

Tabla 3: Comparación entre los datos de las distintas variables reales e imputados de la cohorte AWHS.

VARIABLE	MOMENTO 1		MOMENTO 2		MOMENTO 3	
	Imputado	Real	Imputado	Real	Imputado	Real
IMC (kg/m^2)	27,6 (3,5)	27,6 (3,5)	27,8 (3,7)	27,8 (3,7)	27,9 (3,8)	27,8 (3,8)
Perímetro cintura (cm)	96,8 (9,7)	96,8 (9,6)	97,4 (10,0)	97,3 (10,0)	98,1 (10,8)	97,7 (10,5)
Colesterol HDL (mg/dL)	52,4 (10,9)	52,4 (11,0)	53,8 (11,3)	54,1 (11,3)	51,5 (12,9)	51,0 (12,4)
Glucosa (mg/dL)	97,7 (18,7)	97,7 (18,7)	98,4 (19,5)	96,5 (19,5)	89,9 (21,5)	88,1 (18,6)
SCORE	1,6 (1,4)	1,6 (1,4)	2,1 (1,7)	2,1 (1,7)	2,4 (2,2)	2,1 (1,7)

IMC: índice de masa corporal; SCORE: estimación del riesgo cardiovascular [56].

Estudio en la cohorte CARhES

Descripción de la cohorte

La cohorte CARhES es una cohorte poblacional que utiliza datos de mundo real (Real World Date [RWD]) para estudiar el impacto del uso de servicios sanitarios y la utilización de fármacos en los resultados de salud de pacientes con FRCV e identificar posibles desigualdades en la atención sanitaria. El objetivo final de dicha cohorte es desarrollar estrategias que mejoren la gestión médica de estos pacientes promoviendo una atención eficaz y equitativa. Es una cohorte dinámica abierta cuyo seguimiento se inicia en 2017 y está previsto que dure hasta, al menos, 2026.

Criterios de inclusión en la cohorte

Está formada por todos los individuos con algún FRCV, mayores de 16 años, registrados como usuarios del sistema público de salud en Aragón. Ésta es una región española con alrededor de 1,3 millones de habitantes, los cuales son atendidos en su gran mayoría por el sistema público.

Los participantes tenían al menos uno de los siguientes FRCV: HTA, HC o DM. Los FRCV se identificaron a partir de un diagnóstico médico de HTA, DM o HC y/o la prescripción de, al menos, un fármaco antidiabético o hipolipemiante durante el periodo de estudio. Los datos sociodemográficos, sobre condiciones clínicas, medicación y uso de servicios sanitarios se recogen en BIGAN, un repositorio de datos sanitarios que son recopilados desde la sanidad pública y privada aragonesa y que pone esta información a disposición de los investigadores que la solicitan. Este repositorio está compuesto por un número importante de bases de datos que se explicarán más adelante.

La cohorte CARhES comenzó en 2017 y en aquel momento estaba formada por 446.998 individuos (50,64% mujeres), de los cuales 252.508 tenían hipertensión (56,5%), 332.644 hipercolesterolemia (74,4%) y 96.709 DM (21,6%). La mayoría de los individuos (57,8%) sufrían un solo FRCV, el 31,8% tenía dos y el 10,4% de los pacientes incluidos en la cohorte CARhES tenía los tres FRCV analizados.

Comparación cohorte CARhES y población aragonesa

Dadas las limitaciones de la población incluida en el estudio AWHS, en el que la mayoría de sujetos fueron hombres en edad laboral y encargados de realizar trabajos manuales, perteneciendo a un mismo nivel socioeconómico en su mayoría, se planteó extender el estudio a una población más amplia con características más similares a la de la población de Aragón y que incluyera a un número representativo de mujeres.

Se analizaron las características sociodemográficas de la población de Aragón y de la cohorte CARhES. A continuación, se muestra una tabla comparativa de la distribución por sexo, edad, lugar de residencia y nivel socioeconómico, de la población aragonesa y la población de la cohorte (Tabla 4). Además, se muestra el porcentaje de aragoneses que han entrado en la cohorte para cada variable y categoría.

Tabla 4: Distribución por sexo, edad, nivel socioeconómico y zona de residencia de la cohorte CARhES y de la población de Aragón mayor de 16 años.

	Aragón	**CARhES**	**Porcentaje de aragoneses en CARhES**
Sexo			
Hombres	532959 (48,8%)	220618 (49,4%)	41,4%
Mujeres	558258 (51,2%)	226380 (50,6%)	40,6%
Edad			
16-44	446299 (40,9%)	46272 (10,4%)	10,4%
45-64	370697 (34,0%)	170037 (38,0%)	45,9%
65-79	172943 (15,8%)	140789 (31,5%)	81,4%
>=80	101278 (9,28%)	89900 (20,1%)	88,8%
Nivel socioeconómico			
Activos <18000	411980 (37,8%)	94849 (21,2%)	23,0%
Activos>18000	224084 (20,5%)	70426 (15,8%)	31,4%
Especiales/Otros	83520 (7,65%)	14945 (3,34%)	17,9%
Mutualista	8475 (0,78%)	3368 (0,75%)	39,7%
Pensionistas<18000 y con Farmacia Gratuita	267313 (24,5%)	190346 (42,6%)	71,2%
Pensionistas >18000	95700 (8,77%)	73060 (16,3%)	76,3%
Lugar de residencia			
No Urbana	303750 (27,8%)	132989 (29,8%)	43,8%
Urbana	787467 (72,2%)	314007 (70,3%)	39,9%

Selección de sujetos para la presente tesis

Para los análisis realizados en esta tesis dentro de la cohorte CARhES, se partió de aquellos sujetos con algún FRCV en 2017 y susceptibles de recibir prevención primaria, excluyendo a los sujetos con un diagnóstico de MACE previo al inicio del periodo de estudio (enero 2018). Las bases y fuentes de información que se utilizaron para excluir a los sujetos con evento previo se explicarán en el siguiente apartado. Estos criterios comunes a todos los análisis realizados dentro de la cohorte CARhES se muestran en la

Figura 7. Además, en esta figura también se muestran los criterios de selección específicos para cada objetivo y que se desarrollarán más adelante.

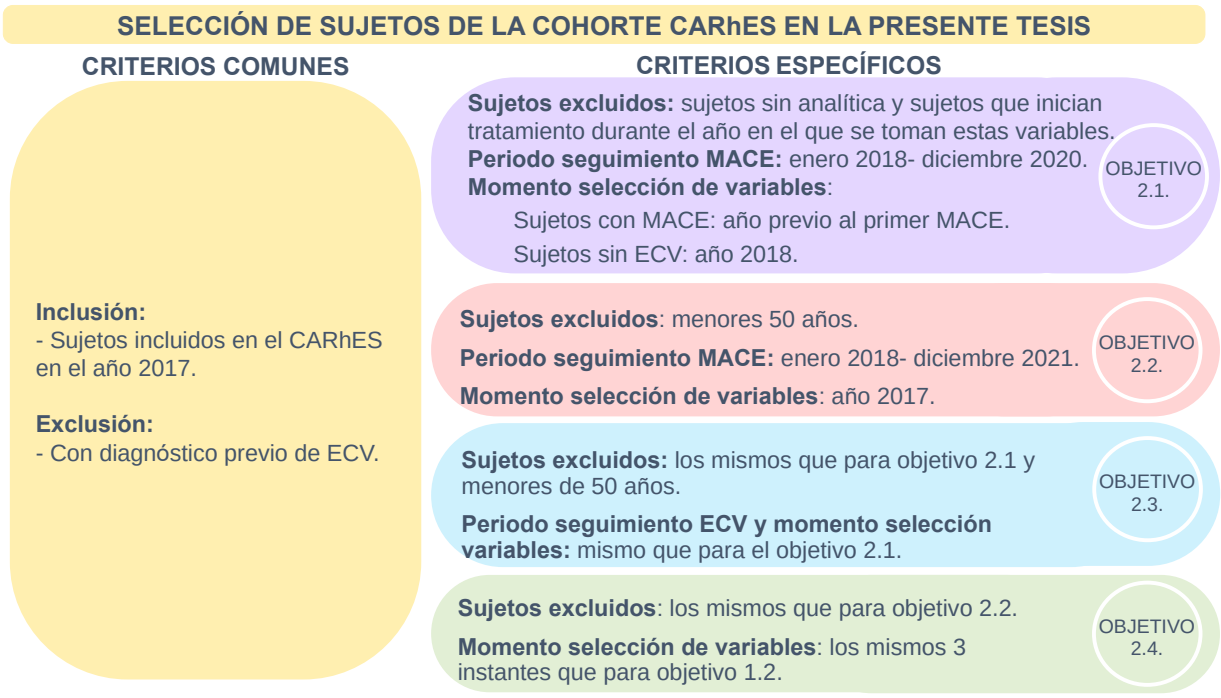


Figura 7: Criterios de selección de sujetos y periodos de estudio de la cohorte CARhES.

Variables de estudio y fuentes de información utilizadas en la presente tesis

Como se ha mencionado anteriormente, toda la información de los sujetos incluidos en la cohorte se obtuvo de BIGAN [62,63] un repositorio donde los datos se recopilan en distintas bases. Las bases de datos a las que se accedió fueron:

- BDU (base de datos de usuarios del sistema sanitario), que proporciona información sobre edad, afiliación al sistema sanitario público de Aragón y datos sociodemográficos de cada sujeto.
- CMBD (conjunto mínimo básico de datos), que recoge información sobre todos los ingresos y altas hospitalarias.
- Base de datos de atención primaria, que recoge información de los pacientes que acuden a un centro de atención primaria.

- GMA (grupos ajustados por morbilidad), que registra información sobre todos los diagnósticos médicos disponibles en la atención primaria y en la base de datos del CMBD y urgencias hospitalarias.
- Base de datos de urgencias hospitalarias, que almacena información sobre diagnósticos y procedimientos de los pacientes procesados a través del sistema de urgencias hospitalarias.
- Base de datos del sistema de prescripción electrónica, que registra todos los tratamientos farmacológicos prescritos a los pacientes.
- Base de datos de dispensación farmacéutica, que recoge información sobre la medicación dispensada en las farmacias a cada paciente.

Todos los datos están pseudoanonimizados mediante un código único que vincula la información del paciente entre las distintas fuentes, pero impide su identificación personal.

Las bases de datos consultadas y las variables extraídas de cada una de ellas se muestran en la Tabla 5.

Tabla 5: Bases de datos consultadas e información obtenida en la cohorte CARhES.

Bases de datos	**Variables**
BDU	Edad, sexo, nivel de ingresos y tipo de actividad económica
GMA	HTA, DM y HC
Atención primaria	Tensión arterial, colesterol total, colesterol HDL y LDL y glucosa en sangre
Dispensación farmacéutica	Código ATC, DDD, número envases dispensados.
Prescripción farmacéutica	Código ATC
CMBD	Código CIE-10 y fecha de ingreso
Urgencias	Código CIE-9 y fecha del ingreso

BDU: bases de datos de usuario; GMA: grupos de morbilidad ajustada; HTA: hipertensión; DM: diabetes mellitus; HC: hipercolesterolemia ATC: código anatómico terapéutico químico; DDD: número de dosis diarias definidas; CIE-9: Clasificación Internacional de Enfermedades, 9ª revisión; CIE-10: Clasificación Internacional de Enfermedades, 10ª revisión.

Antes de comenzar los análisis y para obtener la población de estudio de la tesis, como se ha comentado previamente, fue necesario identificar a aquellos sujetos que habían sufrido un MACE previo, se consultó la base de datos de GMA y/o del CMBD durante

2016 y 2017. En ambas bases de datos se comprobó si tenían diagnóstico de ictus o IAM. En la base de datos de CMBD los diagnósticos se registraron siguiendo la taxonomía CIE-10, y los códigos que se correspondían con estos diagnósticos fueron I21 e I60-I63 (correspondientes a IAM, hemorragia subaracnoidea no traumática, hemorragia intracerebral no traumática, otra hemorragia intracraneal no traumática e ictus isquémico agudo, respectivamente). En GMA se consideraron los registros que constaban como "ictus" y "cardiopatía isquémica".

A continuación, se describen las variables utilizadas para los análisis de la tesis enmarcados en la cohorte CARhES, así como su procedencia y cómo se han obtenido en caso de no estar recogidas directamente de las fuentes de información.

Variables obtenidas de las bases de datos

La edad, el sexo, el nivel de ingresos y el tipo de actividad económica se obtuvieron de BDU.

Como se ha explicado anteriormente, en la cohorte, se consideró que un sujeto tenía un FRCV si había registros del diagnóstico médico de HTA, DM o HC y/o la prescripción de, al menos, un fármaco antidiabético o hipolipemiante durante el periodo de estudio. Para esta identificación de FRCV se consultaron las bases de datos GMA y de prescripción farmacéutica electrónica. GMA se utilizó para identificar a los sujetos con un diagnóstico médico correspondiente a cualquiera de los 3 FRCV de interés. La base de prescripción electrónica se utilizó para identificar los tratamientos farmacológicos que correspondían a los siguientes códigos ATC y que se habían prescrito a los pacientes: A10 para DM y C10 para HC. Finalmente, se consideró que un sujeto sufrió HC o DM si cumplía los criterios de GMA y/o prescripción, y HTA si cumplió los criterios de GMA.

La TA y los parámetros relacionados con la analítica: colesterol total, colesterol HDL y LDL y glucosa en sangre, se obtuvieron de la base de atención primaria, ya que estas variables fueron recogidas durante las consultas o de los análisis realizados en estos centros.

Variables calculadas

El nivel de ingresos y la actividad económica, obtenidas de BDU, se utilizaron para calcular una nueva variable denominada nivel socioeconómico. Las dos variables se combinaron para obtener cinco categorías diferentes de nivel socioeconómico: empleados que ganan más de 18000€ al año; empleados que ganan menos de 18000€ al año; individuos con una pensión contributiva que ganan más de 18000€; individuos con una pensión contributiva que ganan menos de 18000€ y personas con medicamentos gratuitos; y otros, incluyendo principalmente a aquellos con un régimen especial de farmacia y con bajos ingresos.

Para dar respuesta a alguno de los objetivos de este trabajo fue necesario el cálculo de la variable adherencia al tratamiento para fármacos antihipertensivos, hipolipemiantes y antidiabéticos, que se hizo partiendo de algunas de las variables previamente nombradas. Este cálculo para cada grupo de fármacos se realizó mediante el cálculo del PDC, y para ello, de la base de datos de dispensación farmacéutica, se utilizaron las variables: fecha de dispensación, código ATC, las dosis diarias definidas (DDD) y el número de envases dispensados. Esta variable se calculó por separado para cada sujeto mediante el PDC. El PDC, como ya se ha comentado, es un índice calculado como el número de días cubiertos por los medicamentos dispensados por la farmacia dividido por el número de días que el sujeto debería haber tenido cubiertos. En este estudio, el denominador para el PDC fue de 365 días. El número de días cubiertos se calculó a partir de la DDD dispensada a cada sujeto. Al igual que en los estudios dentro de la cohorte AWHS, en el presente estudio se utilizaron valores subrogados para las dosis diarias. Por ejemplo, para las estatinas se utilizó siempre una DDD de 28 en lugar del valor de 37,3 utilizado en otros estudios.

Variable resultado

Dentro de los objetivos planteados en la tesis para esta cohorte, la incidencia de MACE durante el seguimiento fue la variable resultado. Los episodios de MACE se identificaron a través de la base de datos de CMBD y de la base de datos de urgencias hospitalarias. Se consideró que un episodio era un MACE si el primer diagnóstico en la base de datos

de CMBD correspondía a uno de los siguientes códigos CIE-10 I21, I60-I63. En la base de datos de urgencias, se consideraron MACE los episodios con el mismo diagnóstico, correspondientes a los códigos 410 y 430-433 en CIE-9, y que causaron la muerte.

En la tabla 6 se muestran los objetivos para los que cada variable fue utilizada, así como la base de datos de su procedencia.

Tabla 6: Relación variables y objetivos en la cohorte CARhES

	OBJETIVO 2.1.-	**OBJETIVO 2.2.**	**OBJETIVO 2.3.**	**OBJETIVO 2.4.**	**Procedencia**
Edad	X	X	X	X	BDU
Sexo	X	X	X	X	BDU
Nivel socioeconómico		X		X	Calculada a partir de nivel de ingresos y actividad económica de BDU
Hipertensión	X	X	X	X	Calculadas a partir de CMBD y prescripción de farmacia
Hipercolesterol-emia	X	X	X	X	
Diabetes	X	X	X	X	
TAS y TAD	X		X		Atención primaria
Colesterol total, HDL y LDL	X		X		Atención primaria
Glucosa	X		X		Atención primaria
Adherencia antihipertensivos	X		X		PDC calculado a partir de la base de dispensación de farmacia
Adherencia hipolipemiantes	X		X		
Adherencia antidiabéticos	X		X		
Evento cardiovascular mayor	X	X	X	X	Calculada a partir de CMBD y Urgencias

TAS: tensión arterial sistólica; TAD: tensión arterial diastólica; CARhES: CArdiovascular Risk factors for hEalth Service research; CMBD: conjunto mínimo de datos; BDU: base de datos de usuarios; PDC: proporción de días cubiertos.

Selección de sujetos y análisis estadísticos por objetivos

En esta sección se mostrará la información por separado para cada uno de los objetivos propuestos. En la Figura 7 se muestran los criterios de selección de sujetos comunes a todos los análisis realizados dentro de la cohorte CARhES y los criterios específicos

considerados para cada análisis, así como los periodos de seguimiento y momentos en los que se han tomado las variables.

Metodología utilizada para dar respuesta al objetivo 2.1.- Describir la prevalencia de FR en la población de Aragón, así como la frecuencia de FRCV, adherencia a tratamientos e incidencia de MACE en la cohorte CARhES.

Para este objetivo, la incidencia de MACE se analizó durante el periodo enero 2018 y diciembre 2020. El resto de variables analizadas fueron valores analíticos, TA y adherencia a tratamiento de antihipertensivos, hipolipemiantes y antidiabéticos. La selección de estas variables se hizo en función a la incidencia de MACE: para aquellos sujetos que sufrieron un MACE se tomaron los datos del año previo al evento, y para los que no tuvieron MACE, los recogidos en el año 2018. Debido a este criterio, aquellos sujetos con MACE a los que no se les había realizado analítica el año previo al evento y los sujetos sin MACE sin analíticas en 2018, fueron excluidos.

En cuanto a la adherencia a fármacos, obtenida a través del PDC, se calculó para el año 2018 para los sujetos sin MACE, y para el año anterior al evento en los sujetos con MACE. Por esto, los sujetos sin MACE que iniciaron tratamiento a algún fármaco antihipertensivo, antidiabético o hipolipemiante durante 2018 fueron excluidos. Entre los sujetos con MACE, los que iniciaron tratamiento a alguno de los fármacos nombrados anteriormente el año previo al evento, fueron también eliminados.

La descripción de las variables se realizó mediante la media y la DE para las variables cuantitativas, y porcentajes para las variables categóricas.

Metodología utilizada para dar respuesta al objetivo 2.2.- Analizar diferencia en la prevalencia de FRCV y nivel socioeconómico y la incidencia de MACE entre hombres y mujeres.

Para dar respuesta a este objetivo, se incluyeron aquellos sujetos de la cohorte CARhES mayores de 49 años y se analizaron las variables: edad, nivel socioeconómico, HTA, HC y DM, con datos correspondientes al año 2017, y la incidencia de MACE para el periodo enero 2018-diciembre 2021.

Se realizaron dos análisis descriptivos por sexo. El primero para describir las variables para la población total y estratificada por sexo. El segundo análisis muestra las características de hombres y mujeres que sufrieron un MACE. En ambos análisis se calcularon los porcentajes y media y DE para la edad.

Metodología utilizada para dar respuesta al objetivo 2.3.- Analizar la capacidad de distintos métodos de machine learning para predecir la incidencia de MACE en la cohorte CARhES de manera separada para hombres y mujeres, analizando la influencia de 4 grupos de variables (Edad, FRCV, valores analíticos y TA y adherencia a tratamientos antihipertensivos, antidiabéticos e hipolipemiantes) en dicha predicción.

Para este objetivo se utilizaron las variables recopiladas para alcanzar el objetivo 2.1. en los mismos periodos de tiempo. Los sujetos incluidos en el estudio fueron los mismos que para el objetivo 2.1 excluyendo a aquellos menores de 50 años. Para el análisis de machine learning, se utilizaron los métodos Random Forest (RF) y XG Boost para determinar la utilidad de diferentes variables para predecir la incidencia de MACE[49,64,65]. Ambos se aplicaron por separado para hombres y mujeres incluyendo como variables predictivas la edad y 3 grupos de variables combinados de diferentes formas en función del modelo:

- Modelo 1: Edad, análisis de sangre y medición de la TA, factores de riesgo cardiovascular y adherencia a la medicación.
- Modelo 2: Edad, análisis de sangre y medición de la TA, y adherencia a la medicación.
- Modelo 3: Edad, factores de riesgo cardiovascular y adherencia a la medicación.

Siguiendo la literatura, la población del estudio se dividió aleatoriamente en dos grupos: el 80% de la muestra se asignó al grupo de entrenamiento y el 20% restante al grupo de prueba. Para entrenar y ajustar los modelos se aplicó al conjunto de datos de entrenamiento una validación cruzada de 10 iteraciones para evitar el sobreajuste. Para ambos algoritmos, los hiperparámetros se determinaron mediante una búsqueda en rejilla (grid search) en la validación cruzada con 10 iteraciones del conjunto de entrenamiento para identificar los valores que conducían a un rendimiento óptimo.

Cuando la incidencia de eventos es baja, se considera que los datos están desbalanceados[66,67]. Observamos una incidencia de MACE del 1,12%, lo que indica que se puede considerar que los datos estaban muy desbalanceados. Para resolver este problema, se utilizó el método Random Over Sampling Examples (ROSE) [67,68] con reemplazo para sobremuestrear la clase minoritaria y equilibrar los datos en el conjunto de entrenamiento. Para evitar estimaciones erróneas del rendimiento de los modelos, el proceso de remuestreo se aplicó a cada una de las 10 submuestras creadas durante el proceso de validación cruzada, independientemente de las demás submuestras.

El rendimiento de los modelos se evaluó utilizando el conjunto de prueba, y se utilizó el índice de Youden para establecer el umbral óptimo de clasificación. En casos de datos desequilibrados, ciertas medidas como la precisión, el valor predictivo positivo y el valor predictivo negativo pueden verse notablemente alteradas. Por lo tanto, para evaluar el rendimiento de los modelos creados se calcularon cuatro parámetros distintos[69]: (i) AUC, que proporciona información sobre la precisión del modelo; (ii) F1 score, que refleja la capacidad del modelo para captar la sensibilidad y la precisión (es decir, para ser exacto en los casos que capta); (iii) sensibilidad, que indica la proporción de casos clasificados como de alto riesgo de sufrir un acontecimiento; (iv) y especificidad, que refleja la proporción de no casos clasificados como tales. Por último, se extrajo la contribución de cada variable a la predicción y se estandarizó utilizando una escala de 0-1 para facilitar la comparabilidad.

Metodología utilizada para dar respuesta al objetivo 2. 4.- Estudiar el impacto que tienen las diferencias en la distribución de hipertensión, hipercolesterolemia, diabetes y nivel socioeconómico entre sexos en las diferencias observadas en la incidencia de MACE.

Los sujetos incluidos y las variables analizadas para alcanzar este objetivo fueron las mismas que en el objetivo 2.2.: edad, nivel socioeconómico, HTA, HC y DM, fueron las variables explicativas incluidas en el estudio. El sexo se incluyó como variable de agrupación, los MACE como resultado, la edad como factor de confusión y el nivel socioeconómico, la HTA, la HC y la DM como factores de confusión y explicativos según el modelo aplicado.

Se realizaron cuatro modelos diferentes considerando un factor explicativo causal diferente en cada uno, ajustado por el resto de factores explicativos (utilizados como factores de confusión) y la edad.

Para estimar la contribución causal de cada factor explicativo a la diferencia entre hombres y mujeres en la incidencia de MACE, se realizó una descomposición de la razón de riesgos (RR), ajustada por edad, para los hombres en relación con las mujeres (grupo no expuesto) [70,71]. Así pues, la principal medida de resumen del resultado por grupos fue el riesgo de incidencia de MACE calculándose a partir de la comparación del RR de los hombres respecto a las mujeres. Las mujeres fueron tomadas como grupo de referencia ya que la incidencia global de MACE era mayor en los hombres que en las mujeres y los RR se estimaron aplicando regresiones de Poisson ajustadas por edad.

Cuando los FRCV se consideraron factores explicativos, la categoría de referencia fue no tener FRCV. Para el análisis en el que el nivel socioeconómico se consideró factor explicativo, esta variable se recalculó en dos categorías y la categoría de referencia fue la de aquellos que ganaban más de 18000€/año.

La estimación de la contribución de los factores causales se realizó comparando el RR observado con la distribución real de los factores causales con el que se observaría si estableciéramos que los hombres tienen la misma distribución de factores causales que las mujeres. Para ello, se necesita un riesgo de incidencia de MACE contrafactual en los hombres, que se obtuvo aplicando la fórmula G y la integración de Monte-Carlo (Apéndice 1). Así, se crearon dos pseudopoblaciones: una denominada población de curso natural que se crea utilizando los coeficientes obtenidos del análisis de los datos observados; y una pseudopoblación contrafactual que se creó con los coeficientes del análisis utilizando valores simulados de factores explicativos. La diferencia entre las dos pseudopoblaciones corresponde a la contribución causal de la desigualdad.

Todos los análisis se realizaron con la versión 4.2.2 de R utilizando el paquete cfdecomp[72].

Consideraciones éticas

Todos los datos recogidos en ambas bases de datos fueron pseudonimizados. Tanto los estudios realizados dentro de la cohorte AWHS como los realizados dentro de CARhES fueron aprobados por el Comité Ético de Investigación Clínica de Aragón (CEICA) (código de identificación de los proyectos PI17/00042 y PI21/148, respectivamente).

Financiación

La presente tesis ha sido financiada parcialmente por el Gobierno de Aragón con una de las subvenciones destinadas a la contratación de personal investigador predoctoral en formación (IIU/796/2019). También se contó con el apoyo del Grupo de Investigación de Servicios Sanitarios de Aragón (GRISSA) [B09-23R], que forma parte de Instituto de Investigación Sanitaria en Aragón (IIS Aragón) y que es financiado por el Gobierno de Aragón. Los estudios enmarcados dentro de la cohorte AWHS fueron financiados por Proyecto del Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (Ministerio de Ciencia e Innovación) y el Fondo Europeo de Desarrollo Regional (FEDER) (PI17/01704). Los estudios dentro de la cohorte CARhES, fueron financiados por el Proyecto del Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (Ministerio de Ciencia e Innovación), y el Fondo Europeo de Desarrollo Regional (FEDER) (PI22/01193).

4. RESULTADOS

4. RESULTADOS

En este apartado se presentarán los resultados obtenidos con los análisis realizados en los sujetos de las dos cohortes descritas en el apartado anterior. A continuación, se presentan los resultados siguiendo el orden de los objetivos planteados.

Resultados del estudio en la cohorte AWHS

Resultados que dan respuesta al OBJETIVO 1.1.- Análisis descriptivo de factores de riesgo cardiovascular (hipertensión, hipercolesterolemia, diabetes y estado físico), exposición a tratamientos preventivos e incidencia de ECV en la cohorte AWHS.

En el análisis descriptivo se incluyeron 3.746 sujetos (edad media, 61,6 años), todos ellos varones debido al bajo número de mujeres en la cohorte (N=380). La prevalencia de HTA fue del 66,1%, la de HC del 81,0% y la de DM del 17,0%. El porcentaje de participantes que recibieron tratamiento para estas enfermedades fue del 74,3% para HTA, del 52,7% para HC y del 83,5% para DM. Se registró sobrepeso en el 54,4% de los participantes y obesidad en el 30%. El número de FRCV que tuvieron los sujetos fue: 1 FRCV, 46,9%; 2 FRCV, 41,9%; 3 FRCV, 11,2%.

El número de eventos cardiovasculares registrados entre enero de 2010 y diciembre de 2019 fue de 298 (7,9%).

La evaluación de la adherencia por grupo farmacológico por separado indicó una PDC ≥80% en el 63,3% de los participantes que tomaban antidiabéticos, el 78,1% de los que tomaban antihipertensivos y el 64,4% de los que tomaban hipolipemiantes.

La media de edad, la prevalencia de FRCV y el tratamiento en los sujetos con y sin ECV se muestran en la Tabla 7. En comparación con el grupo sin ECV, los grupos con ECV tenían una edad media más elevada (1,4 años más) y una mayor prevalencia de HTA, HC, DM y obesidad. Por el contrario, la prevalencia de sobrepeso fue ligeramente superior en el grupo sin ECV (sin llegar a ser esas diferencias estadísticamente

significativas). Entre los que no presentaban ECV, el 45,60% se clasificaron como totalmente expuestos al tratamiento, y entre los que tuvieron evento el 24,30%.

Tabla 7: Análisis descriptivo de las variables estratificado según la incidencia de evento cardiovascular en la cohorte AWHS.

	No ECV	ECV	p-value
N (%)	3448 (92,05)	298 (7,95)	
EDAD*	61,50 (4,82)	62,90 (4,20)	<0,001
HIPERTENSIÓN	2252 (65,60)	213 (71,70)	0,039
HIPERCOLESTEROLEMIA	2765 (80,30)	264 (89,2)	<0,001
DIABETES	567 (16,50)	65 (22,00)	0,019
ESTADO FÍSICO			0,297
Sobrepeso	1854 (54,50)	157 (53,40)	
Obesidad	1009 (29,70)	98 (33,30)	
EXPOSICIÓN AL TRATAMIENTO			<0,001
Totalmente expuestos	1075 (45,60)	51 (24,30)	
No expuestos	485 (20,60)	74 (35,20)	
Parcialmente expuestos	798 (33,80)	85 (40,50)	

Datos expresados como número (%); *En este caso se refleja media (DE). ECV, evento cardiovascular. Totalmente expuestos: sujetos con recetas dispensadas para el tratamiento de todos los FRCV identificados y PDC ≥80% para todos ellos; no expuestos: participantes sin recetas dispensadas para ninguno de los FRCV identificados o PDC <80% para todos los tratamientos tomados; parcialmente expuestos: participantes sin recetas dispensadas para al menos un FRCV identificado y PDC ≥80% para otros o una PDC <80% para algún tratamiento y ≥80% para otros de los tratamientos tomados.

Estos resultados han sido publicados en el artículo que se presenta en el Anexo I.

Síntesis de los resultados que dan respuesta al Objetivo 1.1

En este análisis descriptivo se encontró que el FRCV más prevalente en la población del AWHS fue la HC, seguido de HTA, y que la mitad de la población sufría de sobrepeso. La incidencia de ECV fue del 7,9% en 10 años de seguimiento. Aquellos sujetos con ECV fueron mayores, con prevalencias más altas de HTA, HC y DM, pero un porcentaje ligeramente menor de sobrepeso. Finalmente, los sujetos que se consideraron totalmente expuestos al tratamiento fueron un 20% menos en los que tuvieron un ECV que en los que no lo presentaron.

Resultados que dan respuesta al OBJETIVO 1.2.- Describir los valores analíticos y variables médicas relacionadas con FRCV y el SCORE, analizando la evolución de los mismos.

Se analizó la correlación entre las variables incluidas en este análisis descriptivo para cada momento. En el momento 1, los índices de correlación más bajos correspondieron al perímetro de cintura y el colesterol HDL (-0,21) y al colesterol HDL y el IMC (-0,21). El índice de correlación más alto (0,87) se obtuvo para el valor de perímetro de cintura y el IMC. Se obtuvieron resultados similares para los otros dos momentos. El índice de correlación fue significativamente distinto de 0 para todos los pares de variables, excepto para el colesterol HDL y SCORE en el momento 1, y para el colesterol HDL y la edad en los momentos 2 y 3, indicando que estas variables en esos momentos mostraron no tener una correlación estadísticamente significativa.

En la Tabla 8 se muestra un análisis descriptivo de los datos de la cohorte AWHS para los tres momentos estudiados. En el primer momento analizado, el peso medio de los hombres era de 81,64 kg y más de la mitad de la población del estudio tenía sobrepeso. Los valores medios del resto de variables se encontraban todos dentro de los rangos recomendados.

La comparación de los valores medios entre los momentos 1 y 2 reveló pocos cambios. El mayor cambio se observó en la media de colesterol total, que fue inferior en el momento 2 que en el 1. El análisis del hábito tabáquico reveló un aumento del porcentaje de exfumadores y una disminución del número de fumadores y no fumadores. La comparación del momento 3 con el momento 2 reveló cambios en los niveles medios de colesterol total y glucosa, ambos disminuyeron. Finalmente, la evolución total entre el inicio y final del seguimiento reveló una reducción de los niveles totales de colesterol y glucosa, y del porcentaje de tabaquismo, mientras que el único porcentaje que aumentó fue el de la obesidad.

Tabla 8: Análisis descriptivo de las variables de estudio

Variables cuantitativas		**MOMENTO 1 Años 2009-2011** Media (DE)	**MOMENTO 2 Año 2014** Media (DE)	**MOMENTO 3 Año 2016-2017** Media (DE)
Edad (años)		48,00 (8,42)	51,49 (8,27)	53,00 (8,25)
Tensión arterial sistólica (mmHg)		126,00 (14,14)	124,00 (14,25)	128,89 (15,00)
Tensión arterial diastólica (mmHg)		83,44 (9,82)	79,80 (9,39)	81,36 (9,68)
Peso (kg)		81,64 (11,47)	82,10 (11,92)	82,66 (12,38)
Perímetro de cintura (cm)		96,81 (9,61)	97,30 (10,00)	97,73 (10,53)
Índice de masa corporal (kg/m^2)		27,61 (3,54)	27,77 (3,67)	27,84 (3,80)
Colesterol HDL (mg/dL)		52,45 (11,00)	54,07 (11,30)	51,00 (12,40)
Colesterol Total (mg/dL)		212,18 (37,62)	205,93 (34,75)	187,96 (32,85)
Glucosa (mg/dL)		97,70 (18,75)	96,51 (19,46)	88,06 (18,60)
SCORE DE RIESGO CV		1,56 (1,40)	2,05 (1,73)	2,09 (1,74)
Variables categóricas		N (%)	N (%)	N (%)
Hábito tabáquico	Fumador	1488 (36,82)	1235 (32,19)	1156 (32,65)
	No fumador	1087 (26,90)	925 (24,11)	853 (24,09)
	Exfumador	1466 (36,28)	1677 (43,71)	1532 (43,26)
Índice de masa corporal	Normopeso	938 (23,05)	846 (22,01)	762 (22,38)
	Sobrepeso	2223 (54,63)	2088 (54,33)	1813 (54,25)
	Obesidad	908 (22,32)	909 (23,65)	830 (24,38)

SCORE calculado a partir de la puntuación del riesgo individual de sufrir ECV a 10 años, diseñada para aplicarse en población europea con bajo RCV [56].

Los resultados del análisis utilizando cuartiles en los tres momentos para algunas variables se muestran en Tabla 9. Para el momento 1, el análisis de cuartiles del IMC mostró que los trabajadores obesos se situaban en el cuartil (Q) 4, los que tenían un peso normal, en el Q1, y los que tenían sobrepeso, en el Q2 y Q3. En cuanto al valor de

perímetro de cintura y glucosa en sangre, los trabajadores con niveles superiores a los valores recomendados se situaron en el Q4.

En el caso del IMC y el perímetro de cintura, más del 80% de los trabajadores que se encontraban en el Q4 en el momento 1 permanecieron en este cuartil en el momento 2. Una proporción similar permaneció en el Q4 entre los momentos 2 y 3. Para los niveles de glucosa en sangre, aproximadamente el 50% de los participantes que estaban en el Q4 en el momento 1 permanecieron en este cuartil en el momento 2, y se observó un efecto similar comparando los momentos 2 y 3. Para los niveles de colesterol HDL, el porcentaje de trabajadores que permanecieron en el Q1, que se corresponde con valores inferiores a los recomendados, fue del 60% en el momento 2 frente al momento 1, y del 80% en el momento 3 frente al momento 2. Para el SCORE, el porcentaje de trabajadores que permanecieron en el Q1 en el momento 2 frente al momento 1 fue del 70%, y del 89% los que permanecieron en el Q4 (puntuación de riesgo, <0,63 y >2,13, respectivamente). Para el momento 3 frente al momento 2, el porcentaje de individuos que permanecieron en el Q1 de la variable SCORE fue del 84%, y del 88% para el Q4.

Estos resultados han sido publicados en el artículo que se presenta en el Anexo II.

Síntesis de los resultados que dan respuesta al Objetivo 1.2

Como síntesis de los resultados destacar que el colesterol total y glucemia se redujeron a lo largo del seguimiento, al igual que el porcentaje de fumadores, mientras que el porcentaje de personas con obesidad aumentó. El análisis por cuartiles mostró que alrededor del 80% de las personas en el cuartil más alto de IMC al principio del estudio y 50% de personas con valores altos de glucemia en sangre no cambiaron su estado durante el seguimiento. En cuanto a la evolución del RCV de los sujetos, más del 80% de los sujetos con un riesgo bajo al principio del estudio se mantuvieron en ese cuartil, mientras que más del 80% de los que tenían un riesgo alto no lo redujeron durante el tiempo del estudio.

Tabla 9:Resultados del análisis por cuartiles en el AWHS. Los valores representan el porcentaje de individuos que pasaron de un cuartil a otro entre los momentos 1 y 2 y los momentos 2 y 3 para las siguientes variables: índice de masa corporal, glucemia y SCORE de riesgo

ÍNDICE DE MASA CORPORAL

		Momento 1						Momento 2					
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p
Momento 2	Q. 1	814 (79,0%)	143 (13,8%)	15 (1,4%)	2 (0,2%)	<0,001	Momento 3 Q. 1	819 (84,1%)	156 (14,8%)	11 (1,07%)	1 (0,1%)	<0,001
	Q. 2	203 (19,7%)	660 (63,9%)	182 (17,3%)	10 (1,0%)		Q.2	140 (14,4%)	679 (64,4%)	155 (15,1%)	9 (0,8%)	
	Q. 3	14 (1,4%)	221 (21,4%)	658 (62,7%)	134 (13,0%)		Q. 3	13 (1,3%)	214 (20,3%)	678 (66,0%)	109 (10,0%)	
	Q. 4	0 (0,0%)	9 (0,9%)	194 (18,5%)	888 (85,9%)		Q. 4	2 (0,2%)	6 (0,6%)	183 (17,8%)	972 (89,1%)	

GLUCOSA

		Momento 1						Momento 2					
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p
Momento 2	Q. 1	713 (64,0%)	356 (35,2%)	180 (16,8%)	49 (5,2%)	<0,001	Momento 3 Q. 1	1103 (85,0%)	790 (74,0%)	458 (45,8%)	107 (13,7%)	<0,001
	Q. 2	260 (23,3%)	351 (34,8%)	359 (33,5%)	98 (10,3%)		Q.2	140 (10,8%)	181 (16,9%)	279 (27,9%)	113 (14,5%)	
	Q. 3	106 (9,5%)	239 (23,7%)	372 (34,7%)	282 (29,7%)		Q. 3	40 (3,1%)	72 (6,7%)	169 (16,9%)	158 (20,2%)	
	Q. 4	35 (3,1%)	64 (6,3%)	161 (15,0%)	522 (54,9%)		Q. 4	15 (1,2%)	25 (2,3%)	93 (9,3%)	404 (51,7%)	

SCORE DE RIESGO CARDIOVASCULAR

		Momento 1						Momento 2					
		Q. 1 N(%)	Q. 2 N(%)	Q. 3 N(%)	Q. 4 N(%)	p		Q. 1 N(%)	Q. 2 N(%)	Q. 3 N(%)	Q. 4 N(%)	p
Momento 2	Q. 1	722 (69,6%)	15 (1,5%)	0 (0,0%)	0 (0,0%)	<0,001	Momento 3 Q. 1	618 (83,9%)	28 (4,0%)	6 (0,5%)	10 (0,6%)	<0,001
	Q. 2	272 (26,2%)	360 (34,8%)	65 (6,3%)	9 (0,9%)		Q.2	115 (15,6%)	353 (50,0%)	85 (7,5%)	18 (1,2%)	
	Q. 3	38 (3,7%)	544 (52,6%)	453 (43,8%)	108 (10,4%)		Q. 3	3 (0,4%)	296 (41,9%)	636 (55,6%)	154 (9,9%)	
	Q. 4	6 (0,6%)	116 (11,2%)	517 (50,0%)	922 (88,7%)		Q. 4	1 (0,1%)	29 (4,1%)	416 (36,4%)	1379 (88,3%)	

Abreviaturas: Q, cuartil; N, numero; p, p-valor (Chi-cuadrado y test U Mann-Whitney).

Resultados que dan respuesta al OBJETIVO 1.3.- Analizar la capacidad de diferentes métodos de machine learning para predecir la aparición de ECV y describir la influencia de distintos FRCV y la exposición a tratamientos preventivos en la incidencia de evento mediante el análisis de dichos modelos, incluyendo las siguientes variables: edad, hipertensión, hipercolesterolemia, diabetes, estado físico y exposición al tratamiento.

A continuación, se presentan los modelos desarrollados con técnicas de machine learning para la predicción de incidencia de ECV. En primer lugar, se muestran los resultados desarrollados utilizando como variables predictivas los siguientes FRCV: edad, HTA, DM, HC y actividad física. En segundo lugar, los modelos que incluyen los FRCV mencionados más la variable exposición al tratamiento.

1. Resultados obtenidos para los modelos que utilizan sólo factores de riesgo cardiovascular como variables predictivas

En este apartado se muestran los resultados para los modelos desarrollados considerando solo los FRCV. Para facilitar la comparación de la capacidad predictiva de cada una de las variables probadas en los algoritmos XGBoost, RF y NB, los valores se normalizaron en una escala de 0-1 (Figura 8), donde 0 y 1 indican la capacidad predictiva mínima y máxima, respectivamente.

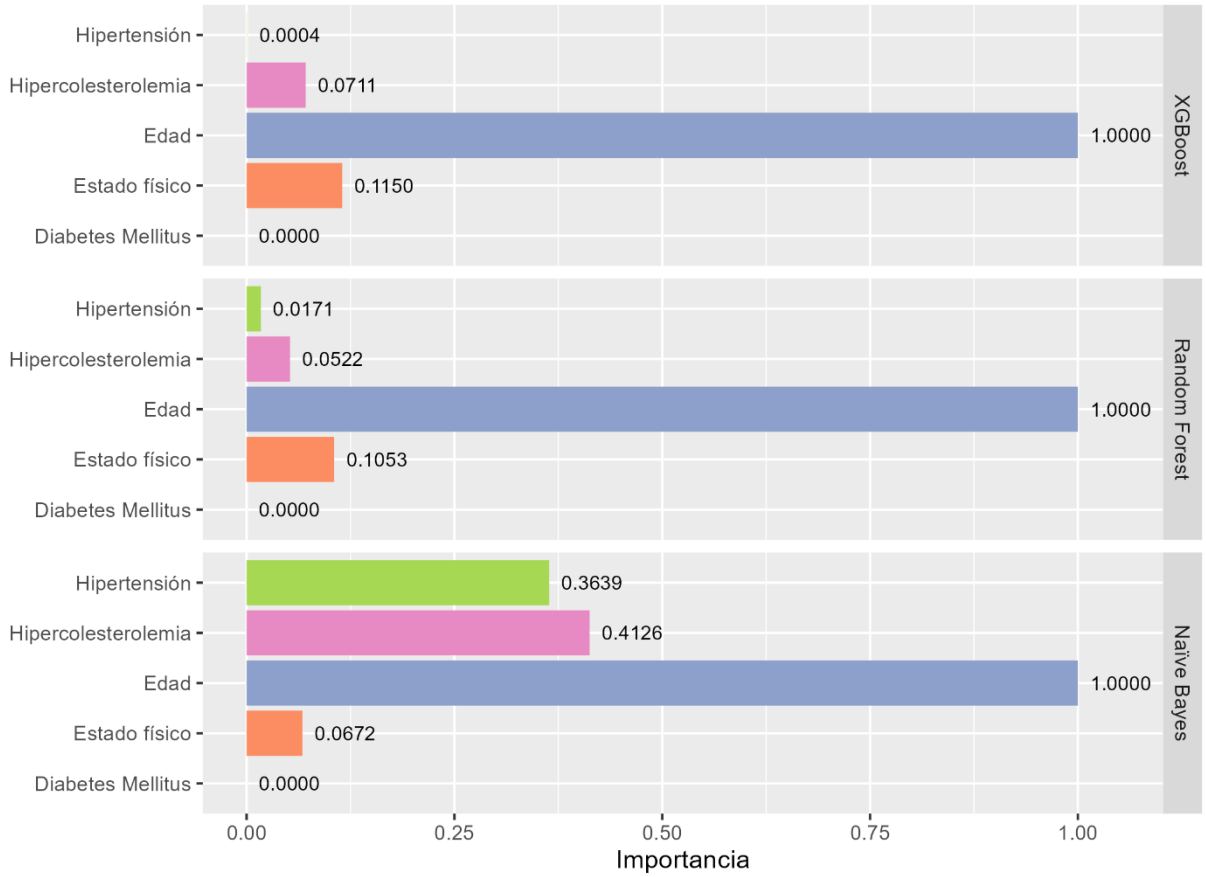


Figura 8: Capacidad predictiva de las variables incluidas en el estudio según los tres métodos empleados: XG Boost, Random Forest y Naïve Bayes en la cohorte AWHS.

En todos los modelos desarrollados, la variable que mejor predijo la ocurrencia de un ECV fue la edad. El siguiente mejor predictor fue el estado físico (tener peso dentro de la normalidad, sobrepeso u obesidad) en el caso de los modelos desarrollados con XGBoost y RF, y la HC (seguida de cerca por la HTA) en el caso de NB.

En la Tabla 10 se comparan las diferentes medidas utilizadas para evaluar la validez y el rendimiento de los modelos. En términos de precisión y sensibilidad, los mejores resultados se obtuvieron para el método RF (73,96% y 75,66%, respectivamente). El único parámetro en el que el método XG Boost superó a RF fue la especificidad (53,00% y 52,05%, respectivamente). Los peores resultados en términos de validez se obtuvieron con el método NB (precisión, 68,29%; sensibilidad, 69,96%). Los tres parámetros utilizados para evaluar el rendimiento de los modelos (AUC-PR, Log Loss y F1-Score) indicaron que el método RF fue el que mejor rendimiento tuvo, mientras que NB fue el que peor (AUC-PR y F1-Score más bajos y Log Loss más alto). Esto quiere decir que el

modelo que obtuvo mayor validez y rendimiento fue el desarrollado con RF, seguido de cerca por el desarrollado con XGBoost, mientras que el desarrollado con NB fue el que peor validez mostró.

Tabla 10: Evaluación de la validez y rendimiento utilizando solo factores de riesgo cardiovascular como variables predictivas en la cohorte AWHS.

Modelo	PRECISIÓN (%)	SENSIBILIDAD (%)	ESPECIFICIDAD (%)	VPP (%)	VPN (%)	AUC-PR	Log-loss	F1-Score
XGBoost	71,29	72,71	53,00	95,20	13,16	0,15	0,24	0,83
Random Forest	73,96	75,66	52,05	95,29	14,30	0,17	0,24	0,84
Naïve bayes	68,29	69,96	47,00	94,42	10,88	0,11	0,26	0,80

VPP, valor predictivo positivo; VPN, valor predictivo negativo; AUC-PR, Área bajo la curva precision-recall.

2. Resultados obtenidos para los modelos que utilizan los factores de riesgo cardiovascular y la exposición al tratamiento como variables predictivas

A continuación, se muestran los resultados para los modelos desarrollados considerando los FRCV más la variable exposición al tratamiento. Al incluirla como predictor (Figura 9), la edad siguió siendo la variable que mejor predecía en los modelos desarrollados con XGBoost y con RF, seguida de la exposición al tratamiento. En el modelo NB, la exposición al tratamiento fue la variable con mayor capacidad predictiva, seguida de cerca por la edad. Mientras que la edad y la exposición al tratamiento tuvieron una puntuación en su capacidad para predecir de 1 y 0,40, respectivamente, en el modelo RF y en el XGBoost, los valores correspondientes en el modelo NB fueron mucho más cercanos entre sí (0,96 y 1, respectivamente). En los modelos con RF y XG Boost las demás variables tuvieron muy poca influencia, aunque el estado físico tuvo una mayor capacidad predictiva en el modelo RF frente al XGBoost. En el modelo NB, tras la exposición al tratamiento y la edad, las variables HC y HTA tuvieron mayor capacidad predictiva que el estado físico y la DM.

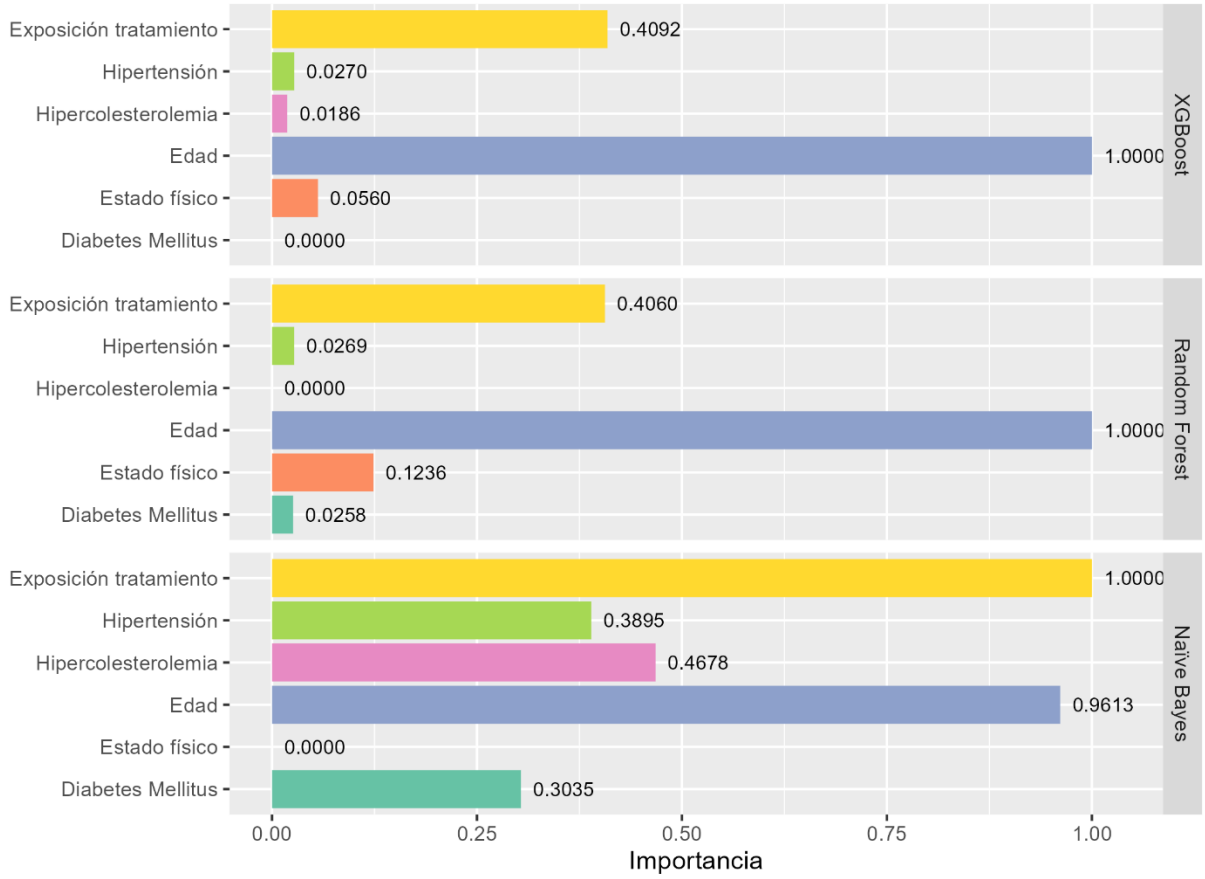


Figura 9: Capacidad predictiva de las variables incluidas en el estudio según los tres métodos empleados: XG Boost, Random Forest y Naïve Bayes, en la cohorte AWHS.

En la Tabla 11 se comparan los distintos parámetros utilizados para evaluar la validez y el rendimiento de los modelos. El modelo desarrollado con RF mostró las puntuaciones más altas en precisión, sensibilidad, especificidad, VPP y VPN, mientras que el modelo con NB mostró las puntuaciones más bajas en estos parámetros. Resultados similares mostraron las pruebas calculadas para evaluar el rendimiento de los modelos: los mejores resultados fueron para el algoritmo RF y los peores para el NB, siendo las puntuaciones de XGBoost similares a las de RF. Esto quiere decir que, al igual que en los modelos desarrollados sin exposición al tratamiento, el modelo que obtuvo mayor validez fue el desarrollado con RF, seguido de cerca por el desarrollado con XGBoost, mientras que el desarrollado con NB fue el que peor validez mostró.

Tabla 11: Evaluación de la validez y rendimiento utilizando solo factores de riesgo cardiovascular y exposición al tratamiento como variables predictivas en la cohorte AWHS.

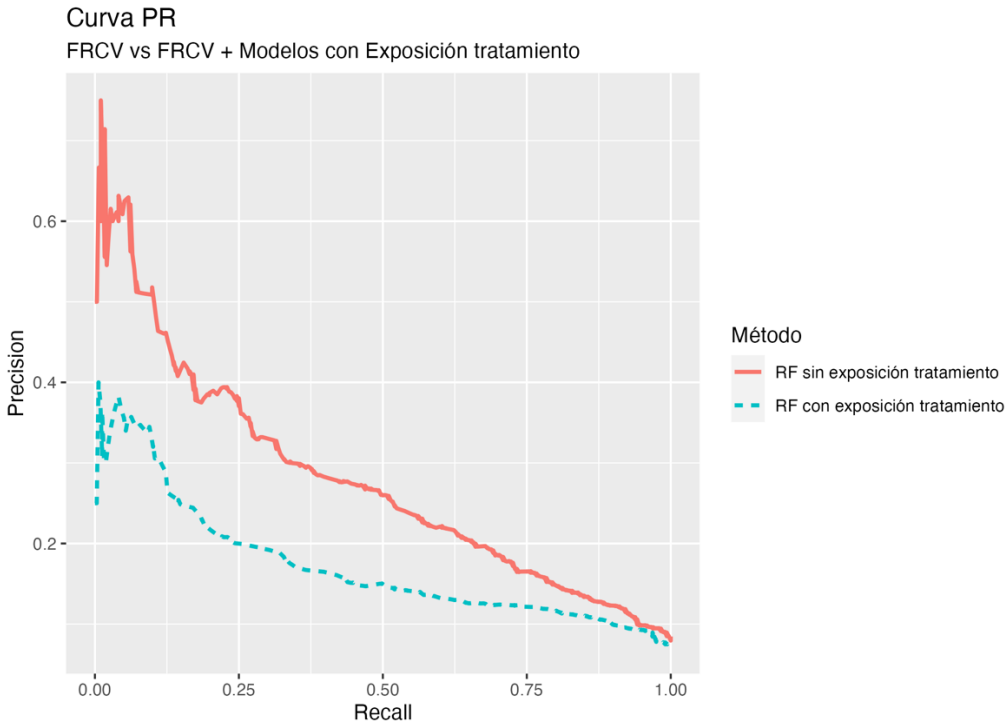
Modelo	PRECISIÓN (%)	SENSIBILIDAD (%)	ESPECIFICIDAD (%)	VPP (%)	VPN (%)	AUC-PR	Log-loss	F1-Score
XGBoost	72,39	73,17	63,36	95,85	16,94	0,24	0,25	0,83
Random forest	73,35	73,68	69,52	96,55	18,57	0,28	0,24	0,84
Naïve bayes	61,78	61,88	60,62	94,79	12,07	0,13	0,28	0,75

VPP, valor predictive positivo; VPN, valor predictive negativo; AUC-PR, Área bajo la curva precision-recall.

3. Comparación de los modelos desarrollados con Random Forest

A continuación, se muestra una comparación de los modelos desarrollados con RF, que fueron los que mejores medidas de validez y rendimiento obtuvieron, al considerarse sólo FRCV o considerando FRCV más exposición al tratamiento.

Los modelos realizados con el algoritmo RF obtuvieron los mejores resultados para la predicción de eventos, independientemente de los grupos de variables incluidos. La Figura 10 muestra la curva PR de este método cuando se consideran como variables predictivas tanto los FRCV por si solos como acompañados por la exposición al tratamiento, mostrando mejores resultados cuando se incluyeron ambos grupos de variables.



RF: random forest; FRCV: factores de riesgo cardiovascular; PR: Precision-Recall.

Figura 10: Curva PR para los modelos desarrollados con Random Forest, considerando factores de riesgo solos o factores de riesgo y exposición al tratamiento como variables predictivas en la cohorte AWHS.

Los parámetros Log-Loss y el F1-score fueron muy similares independientemente de las variables incluidas (0,24 y 0,84 para los FRCV y los FRCV + exposición al tratamiento, respectivamente). Los parámetros utilizados para evaluar la validez de los modelos (precisión, sensibilidad y VPP) fueron muy similares al utilizar ambos grupos de variables (alrededor del 73%, 74% y 96%). Sin embargo, la especificidad y el VPN fueron considerablemente mayores cuando se incluyó la exposición al tratamiento como variable predictiva (69,52% y 18,57%, respectivamente, frente a 52,05% y 14,30%, respectivamente, cuando se excluyó la exposición al tratamiento). Es decir, la capacidad de los modelos con y sin exposición al tratamiento para identificar los casos de MACE fue similar, sin embargo, los modelos con exposición al tratamiento identificaron mejor a las personas que no sufrieron MACE.

Estos resultados han sido publicados en el artículo que se presenta en el Anexo I.

Síntesis de los resultados que dan respuesta al Objetivo 1.3

La evaluación del rendimiento de los modelos mostró que el modelo RF obtuvo los mejores resultados, independientemente de las variables incluidas, seguido de XGBoost. Ambos modelos obtuvieron mejores resultados cuando se incluyeron los FRCV y la exposición al tratamiento como variables predictivas, en comparación con los FRCV solos (Figura 10). Salvo en uno de los modelos, la variable que mayor capacidad predictiva mostró fue la edad, seguida de cerca por la exposición al tratamiento. En el modelo en el que la exposición al tratamiento fue la variable con mayor capacidad predictiva, ésta estuvo seguida muy de cerca por la edad.

Resultados que dan respuesta al OBJETIVO 1.4.- Identificar perfiles de participantes en la cohorte AWHS en función de la evolución de FRCV y del SCORE utilizando la información de tres momentos, aplicando técnicas de cluster longitudinal.

Como se ha presentado en el apartado de metodología, para dar respuesta al objetivo 1.4., se realizó un análisis de cluster para evaluar la evolución conjunta de las siguientes variables: edad, perímetro de cintura, IMC, glucemia, niveles de colesterol HDL y SCORE. Basándonos en distintos índices de calidad para determinar el número de clusters (Figura 11), dividimos la cohorte en 2 y 3 grupos. La Figura 11 muestra cómo varían tres índices distintos de calidad según el número de clusters. Cuando se realizaron análisis para ambos escenarios, los resultados se justificaron mejor cuando la cohorte se dividió en 2 grupos en lugar de en 3. Además, la puntuación de uno de los índices para medir la validez interna de los cluster, Calinski-Harabasz, fue de 1197 cuando la cohorte se dividió en 2 conglomerados, y de 1067 cuando se dividió en 3. Por lo tanto, nos centramos en los resultados obtenidos utilizando 2 conglomerados.

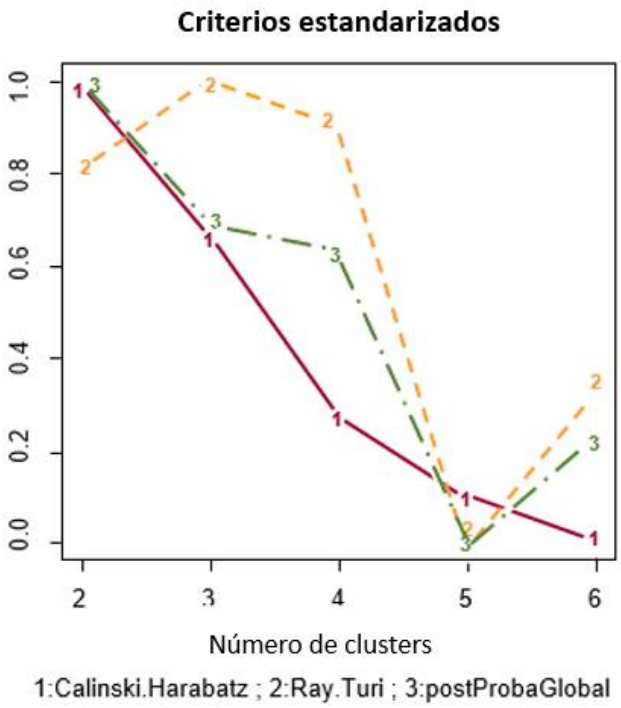
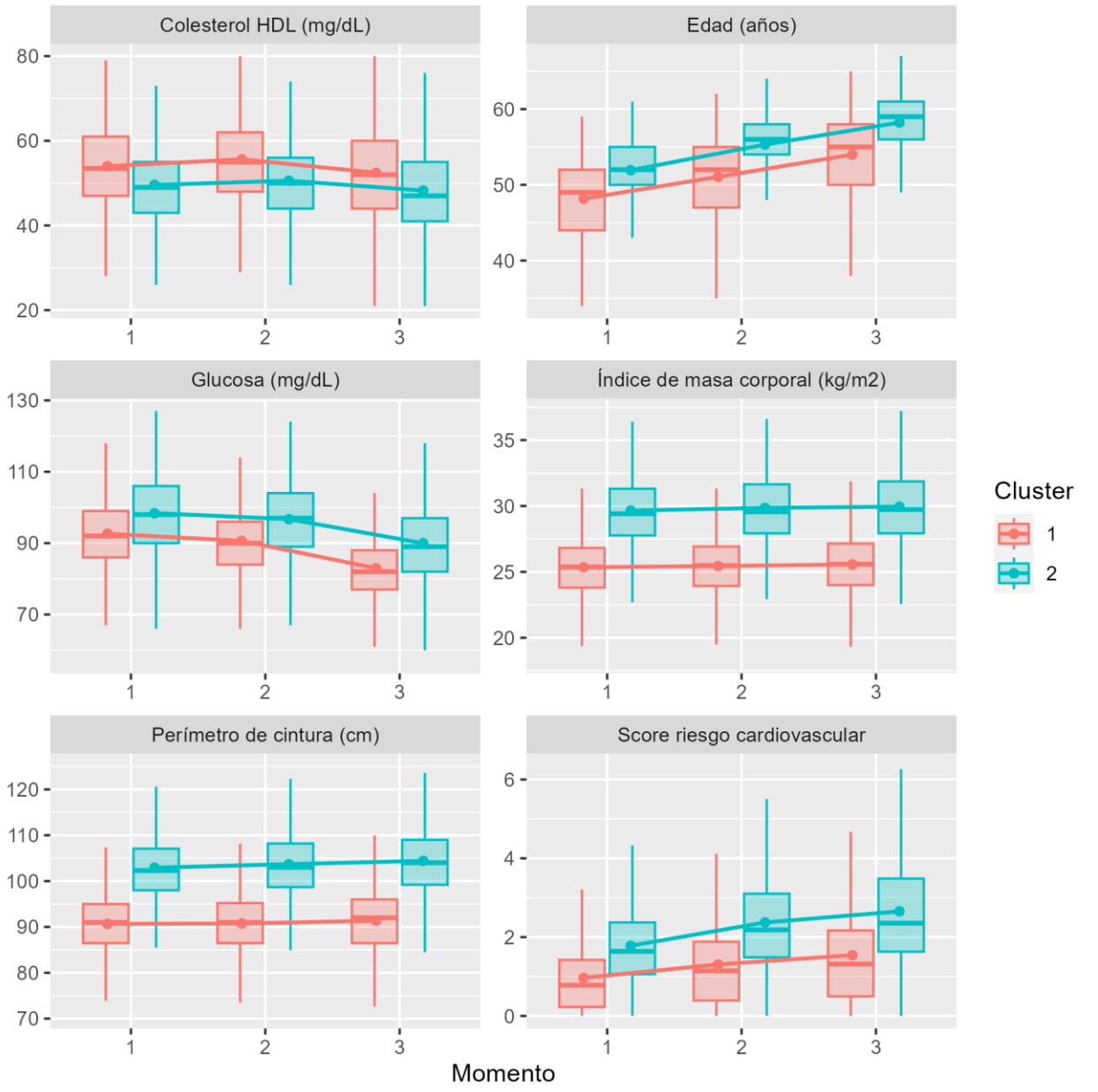


Figura 11: Índices de calidad en función del número de conglomerados. Cada línea representa un índice de calidad diferente y muestra los cambios resultantes en función del número de conglomerados. Cada índice de calidad se ha normalizado a un valor entre 0 y 1.

La Figura 12 muestra los valores medios obtenidos para cada variable, estratificados por clusters. Las diferencias encontradas entre clusters para todos los momentos y variables fueron significativas ($p<0,01$).

El cluster 1 estuvo formado por trabajadores más jóvenes, con valores medios de IMC, perímetro de cintura, glucemia y SCORE más bajos, y valores medios de colesterol HDL más altos. El análisis de la evolución de cada variable a lo largo de los tres momentos reveló patrones similares en ambos conglomerados. Los valores medios de perímetro de cintura e IMC aumentan ligera y progresivamente con el tiempo, aunque este aumento fue algo mayor en el cluster 2. Los niveles de glucosa en sangre disminuyeron de la misma forma con el tiempo en ambos clusters. En los cluster 1 y 2, los niveles de colesterol HDL aumentaron entre los dos primeros momentos, y disminuyeron entre los momentos 2 y 3. Por último, en ambos clusters, el SCORE aumentó durante el periodo de estudio, probablemente por la influencia de la edad sobre el SCORE, aunque este aumento fue mayor en el cluster 2.



Momento 1: datos de 2009,2010 o 2011; Momento 2: datos de 2014; Momento 3: Datos de 2016 o 2017

Figura 12: Gráficos de caja que representan los valores medios (puntos) de cada variable por cluster en la cohorte AWHS. Las líneas representan la evolución de la media para cada cluster a lo largo de los tres momentos analizados.

La evolución de las variables por cuartiles a lo largo de los tres momentos, estratificados para los clusters identificados, se muestra en la Tabla 12. En el caso del IMC, tanto en el cluster 1 como en el 2, más de la mitad de los individuos estuvieron en el mismo cuartil en los momentos 1 y 2, con la excepción del Q4. Los sujetos del cluster 2 que permanecieron en el Q4 entre los momentos 1 y 2 fue el doble que los del cluster 1, es

decir, más sujetos del cluster 1 redujeron su IMC entre los dos primeros momentos que del cluster 2. La comparación de los valores de IMC en los momentos 2 y 3 fue similar. La evolución del perímetro de cintura fue parecida a la del IMC.

Para el colesterol HDL, en el grupo 1, el 54% de los trabajadores que se encontraban en el Q1 en el primer momento permanecieron en este cuartil en el momento 2. El porcentaje de trabajadores que permanecieron en el Q1 en los momentos 2 y 3 fue del 78%. En el cluster 2, estos porcentajes fueron del 65% entre los momentos 1 y 2 y del 82% entre los momentos 2 y 3.

El análisis de los valores de glucosa en sangre mostró que en el cluster 2, la proporción de trabajadores que se encontraban en el Q4 en el momento 1 y permanecían en este cuartil en el momento 2 era el doble que el porcentaje en el cluster 1. Para ambos clusters, entre los momentos 2 y 3, se observó una disminución de la proporción de trabajadores que no cambiaron de cuartil, excepto para el Q1, para el que se observó un aumento significativo.

Por último, para el SCORE, en el cluster 1, el porcentaje de trabajadores que se encontraban en Q1 y Q4 en los momentos 1 y 2 (74% y 82%, respectivamente) era superior al observado para Q2 y Q3 (38% y 48%, respectivamente). La comparación de los momentos 2 y 3 mostró que, para todos los cuartiles, el porcentaje de trabajadores que no cambiaron de cuartil fue superior al observado entre los momentos 1 y 2. En el cluster 2, los resultados obtenidos fueron similares a los observados para el cluster 1, aunque entre los trabajadores en el Q1 en el momento 1, sólo el 41% permaneció en este cuartil en el momento 2, mientras que el 46% pasó al Q2.

Tabla 12: Resultados del análisis de cuartiles por clusters en la cohorte AWHS. Los valores representan el porcentaje de individuos que pasaron de un cuartil a otro entre los momentos 1 y 2 y los momentos 2 y 3 para las siguientes variables: índice de masa corporal, glucemia y SCORE

ÍNDICE DE MASA CORPORAL

		Momento 1						Momento 2				
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*
		Cluster 1				<0.001		**Cluster 1**				<0,001
Momento 2	Q. 1	789 (80,4%)	125 (17,2%)	10 (3,0%)	1 (1,7%)		**Momento 3**	786 (85,0%)	126 (16,3%)	3 (0,8%)	0 (0,0%)	
	Q. 2	179 (18,2%)	481 (66,1%)	106 (31,9%)	5 (8,6%)			127 (13,7%)	515 (66,8%)	76 (22,1%)	4 (6,8%)	
	Q. 3	13 (1,3%)	119 (16,3%)	187 (56,3%)	25 (43,1%)			11 (1,2%)	126 (16,3%)	220 (64,0%)	18 (30,5%)	
	Q. 4	0 (0,0%)	3 (0,4%)	29 (8,7%)	27 (46,6%)			1 (0,1%)	4 (0,5%)	45 (13,1%)	37 (62,7%)	
		Cluster 2				<0,001		**Cluster 2**				<0,001
Momento 2	Q. 1	25 (50,0%)	18 (5,9%)	5 (0,7%)	1 (0,1%)		**Momento 3**	33 (67,3%)	30 (10,6%)	8 (1,2%)	1 (0,1%)	
	Q. 2	24 (48,0%)	179 (58,7%)	76 (10,6%)	5 (0,5%)			13 (26,5%)	164 (57,7%)	79 (11,6%)	5 (0,5%)	
	Q. 3	1 (2,0%)	102 (33,4%)	471 (65,7%)	109 (11,2%)			2 (4,1%)	88 (31,0%)	458 (67,1%)	91 (8,8%)	
	Q. 4	0 (0,0%)	6 (2,0%)	165 (23,0%)	861 (88,2%)			1 (2,0%)	2 (0,7%)	138 (20,2%)	935 (90,6%)	

GLUCOSA

		Momento 1						Momento 2				
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*
		Cluster 1				<0,001		**Cluster 1**				<0,001
Momento 2	Q. 1	487 (67,5%)	249 (39,6%)	96 (19,5%)	27 (10,5%)		**Momento 3**	765 (89,1%)	520 (81,4%)	257 (58,4%)	40 (24,8%)	
	Q. 2	172 (23,9%)	234 (37,2%)	193 (39,2%)	40 (15,6%)			67 (7,8%)	92 (14,4%)	108 (24,5%)	37 (23,0%)	
	Q. 3	52 (7,2%)	133 (21,1%)	162 (32,9%)	93 (36,2%)			19 (2,2%)	23 (3,6%)	56 (12,7%)	35 (21,7%)	
	Q. 4	10 (1,4%)	13 (2,1%)	41 (8,3%)	97 (37,7%)			8 (0,9%)	4 (0,6%)	19 (4,3%)	49 (30,4%)	

Q: cuartil; N: número; p: p-valor (Chi-squared y Mann-Whitney U-test).

Tabla 12 (Continuación): Resultados del análisis de cuartiles por clusters en la cohorte AWHS. Los valores representan el porcentaje de individuos que pasaron de un cuartil a otro entre los momentos 1 y 2 y los momentos 2 y 3 para las siguientes variables: índice de masa corporal, glucemia y SCORE

GLUCOSA

		Momento 1						Momento 2				
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	p		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*
		Cluster 2				<0,001		**Cluster 2**				<0,001
Momento 2	Q. 1	226 (57,5%)	107 (28,1%)	84 (14,5%)	22 (3,2%)		**Momento 3**	338 (77,0%)	270 (62,9%)	201 (36,0%)	67 (10,8%)	
	Q. 2	88 (22,4%)	117 (30,7%)	166 (28,6%)	58 (8,4%)			73 (16,6%)	89 (20,7%)	171 (30,6%)	76 (12,2%)	
	Q. 3	54 (13,7%)	106 (27,8%)	210 (36,2%)	189 (27,2%)			21 (4,8%)	49 (11,4%)	113 (20,2%)	123 (19,8%)	
	Q. 4	25 (6,4%)	51 (13,4%)	120 (20,7%)	425 (61,2%)			7 (1,6%)	21 (4,9%)	74 (13,2%)	355 (57,2%)	

SCORE DE RIESGO CARDIOVASCULAR

		Momento 1						Momento 2				
		Q, 1 N (%)	Q, 2 N (%)	Q, 3 N (%)	Q, 4 N (%)	p		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N(%)	*p*
		Cluster 1				<0,001		**Cluster 1**				<0,001
Momento 2	Q. 1	657 (74,7%)	10 (1,7%)	0 (0,0%)	0 (0,0%)		**Momento 3**	578 (86,7%)	19 (4,2%)	1 (0,2%)	2 (0,5%)	
	Q. 2	200 (22,7%)	215 (38,0%)	31 (7,8%)	4 (1,6%)			86 (12,9%)	241 (53,6%)	52 (9,4%)	3 (0,7%)	
	Q. 3	23 (2,6%)	298 (52,7%)	192 (48,2%)	40 (15,7%)			2 (0,3%)	177 (39,3%)	322 (58,2%)	54 (12,6%)	
	Q. 4	0 (0,0%)	43 (7,6%)	175 (44,0%)	211 (82,7%)			1 (0,2%)	13 (2,9%)	178 (32,2%)	370 (86,2%)	
		Cluster 2				<0,001		**Cluster 2**				<0,001
Momento 2	Q. 1	65 (41,1%)	5 (1,1%)	0 (0,0%)	0 (0,0%)		**Momento 3**	40 (57,1%)	9 (3,5%)	5 (0,9%)	8 (0,7%)	
	Q. 2	72 (45,6%)	145 (30,9%)	34 (5,3%)	5 (0,6%)			29 (41,4%)	112 (43,8%)	33 (5,6%)	15 (1,3%)	
	Q. 3	15 (9,5%)	246 (52,5%)	261 (41,0%)	68 (8,7%)			1 (1,4%)	119 (46,5%)	314 (53,2%)	100 (8,8%)	
	Q. 4	6 (3,8%)	73 (15,6%)	342 (53,7%)	711 (90,7%)			0 (0,0%)	16 (6,3%)	238 (40,3%)	1009 (89,1%)	

Q: cuartil; N: número; p: p-valor (Chi-squared y Mann-Whitney U-test).

Estos resultados han sido publicados en el artículo que se presenta en el Anexo II.

Síntesis de los resultados que dan respuesta al Objetivo 1.4

El cluster 1 estuvo formado por trabajadores más jóvenes, con valores medios de IMC, perímetro de cintura, glucemia y SCORE más bajos, y valores medios de colesterol HDL más altos. En cuanto a la evolución de las variables analizadas, el porcentaje de sujetos con valores más altos de IMC y valores más bajos de los recomendados de HDL que no cambiaron su estado a lo largo del estudio fue sustancialmente mayor en el cluster 2 que en el 1. La evolución de la glucosa en sangre siguió el mismo patrón, aunque entre los momentos 2 y 3 se observó un aumento significativo de personas que permanecieron en el primer cuartil (personas con niveles de glucosa dentro de los recomendados) en ambos grupos. En cuanto al SCORE, la evolución fue similar en ambos grupos, pero el porcentaje de sujetos que cambiaron del Q1 al Q2 fue mayor en el grupo 2 que en el 1.

Resultados cohorte CARhES

Al igual que en el apartado anterior, a continuación, se muestran los resultados de los análisis realizados dentro de la cohorte CARhES siguiendo los objetivos planteados.

Resultados que dan respuesta al objetivo 2.1.- Describir la prevalencia de factores de riesgo en la población de Aragón, así como la frecuencia de FRCV, adherencia a tratamientos e incidencia de MACE en la cohorte CARhES.

Se calcularon las prevalencias de los FRCV en Aragón (Tabla 13). Para el total de la población aragonesa, el FRCV más prevalente fue la hipercolesterolemia, tanto para hombres como para mujeres. En la comparación por sexos, la hipertensión fue el único FRCV más prevalente en mujeres que en hombres, HC y DM fueron más comunes entre ellos, siendo estas diferencias estadísticamente significativas.

Tabla 13: Prevalencias de los FRCV en la población de Aragón.

FRCV		**Total**	**HOMBRES N=539.181**	**MUJERES N=564.170**
HTA	N (%)	252.508 (22,89)	119.812(22,22)	132.696(23,52)
HC	N (%)	332.644 (30,15)	166.512(30,88)	166.132(29,45)
DM	N (%)	96.709 (8,77)	52.867(9,81)	43.842(7,77)

N: número; HTA: hipertensión arterial; HC: hipercolesterolemia; DM: diabetes mellitus.

Para dar respuesta a este objetivo, siguiendo los criterios descritos en el apartado de material y método (sujetos con HTA, HC y/o DM con registros de analíticas y que no iniciaron tratamiento con antihipertensivos, hipolipemiantes o antidiabéticos en el año de estudio), se incluyeron 52.393 sujetos de la cohorte CARhES, el 57,3% eran mujeres, y la edad media era de 70,2 años. Las mujeres eran mayores que los hombres (edad media, 71,6 y 68,3 años, respectivamente) (Tabla 14).

En ambos sexos, dentro de los sujetos incluidos, el FRCV más prevalente fue HTA, seguido de HC. Las proporciones de individuos con 1, 2 y 3 FRCV fueron similares en ambos sexos. Alrededor del 40% de los incluidos tenían un solo FRCV.

Los valores medios de colesterol total, HDL y LDL fueron más altos en las mujeres que en los hombres, y los de glucemia, TAS y TAD fueron superiores en los hombres.

La adherencia al tratamiento fue mayor para los fármacos antihipertensivos y menor para los antidiabéticos, tanto en el conjunto de la población como tras estratificar por sexo. En el caso de los antidiabéticos y los hipolipemiantes, los hombres mostraron una mayor adherencia media, aunque mayor desviación estándar (DE). En el caso de los antihipertensivos, la adherencia media fue mayor en las mujeres.

Entre los sujetos seleccionados 581 (1,1%) experimentaron un MACE: 282 hombres y 299 mujeres entre enero de 2018 y diciembre de 2020. En 12 casos (8 hombres y 4 mujeres), el evento causó la muerte del sujeto. El MACE más frecuente fue el ictus, que representó el 57,5% de todos los eventos, seguido del IAM (26,2%). Estratificando por sexo, el ictus fue más frecuente en mujeres que en hombres (61,5% y 53,2%, respectivamente), mientras que el IAM fue más frecuente en hombres que en mujeres (32,3% y 20,4%, respectivamente).

Tabla 14: Análisis descriptivo de los sujetos incluidos en este análisis de CARhES.

Variables	Unidad	Total N=52,393	HOMBRES N=22,383	MUJERES N=30,010	P value
Edad	media (DE)	70,2 (12,8)	68,3 (12,6)	71,6 (12,8)	<0,001
FACTORES DE RIESGO CARDIOVASCULAR					
DM	N (%)	14.181 (27,1)	7.162 (32,0)	7.019 (23,4)	<0,001
HTA	N (%)	38.253 (73,0)	15.964 (71,3)	22.289 (74,3)	<0,001
HC	N (%)	37.316 (71,2)	15.877 (70,9)	21.439 (71,4)	0,209
Número FRCVs	N (%)				<0,001
1		22.508 (43,0)	9.406 (42,0)	13.102 (43,7)	
2		22.413 (42,8)	9.334 (41,7)	1.079 (43,6)	
3		7.472 (14,3)	3.643 (16,3)	3.829 (12,8)	
VALORES ANALÍTICOS Y TENSIÓN ARTERIAL					
Colesterol total (mg/dL)	media (DE)	195 (36,1)	186 (35,5)	201 (35,1)	0,000
Colesterol HDL (mg/dL)	media (DE)	53,7 (13,4)	48,8 (11,6)	57,3 (13,5)	0,000
Colesterol LDL (mg/dL)	media (DE)	118 (31,5)	114 (31,7)	121 (31,1)	<0,001
Glucosa (mg/dL)	media (DE)	104 (24,8)	107 (26,5)	101 (23,1)	<0,001
TAS (mm Hg)	media (DE)	133 (15,8)	134 (15,4)	133 (16,2)	<0,001
TAD (mm Hg)	media (DE)	76,8 (13,9)	77,8 (16,5)	76,0 (11,4)	<0,001
ADHERENCIA A TRATAMIENTO, PDC					
Antihipertensivos	media (DE)	58,3 (44,0)	57,5 (44,7)	58,9 (43,6)	<0,001
Antidiabéticos	media (DE)	17,3 (33,4)	21,0 (36,1)	14,5 (31,0)	<0,001
Hipolipemiantes	media (DE)	38,5 (42,0)	40,1 (42,5)	37,3 (41,5)	<0,001
CARACTERÍSTICAS DE MACE *					
Frecuencia	N (%)	581 (1,1)	282 (1,3)	299 (1,0)	0,005
Diagnósticos	N (%)				0,011
Infarto agudo de miocardio		152 (26.2)	91 (32,3)	61 (20,4)	
Hemorragia subaracnoidea no traumática		14 (2.4)	4 (1,4)	10 (3,3)	
Hemorragia intracerebral no traumática		57 (9.8)	24 (8,5)	33 (11,0)	
Otra hemorragia intracraneal no traumática		24 (4.1)	13 (4,6)	11 (3,7)	
Accidente cerebrovascular isquémico		334 (57.5)	150 (53,2)	184 (61,5)	

TAS: tensión arterial sistólica; TAD: tensión arterial diastólica; DE: desviación estándar; N: número; DM: diabetes mellitus; HTA: hipertensión arterial; HC: hipercolesterolemia; MACE: evento cardiovascular mayor; PDC: proporción de días cubiertos. * Periodo de seguimiento: enero 2018-diciembre 2020

Características de las personas con y sin MACE

Del total de MACE, el 51% fueron experimentados por mujeres, aunque la incidencia de MACE fue mayor en los hombres. La edad media fue mayor entre los individuos que experimentaron un MACE: 78,9 y 70,1 años en los individuos que experimentaron y no experimentaron un MACE, respectivamente (Tabla 15).

Las frecuencias de DM y HTA fueron mayores entre los individuos que experimentaron un MACE. No hubo diferencias significativas en la proporción de pacientes con HC entre los individuos con o sin MACE. Además, los que experimentaron un MACE presentaron con mayor frecuencia 2 o 3 FRCV, y los que no lo experimentaron presentaron con mayor frecuencia 1 FRCV.

No se observaron diferencias en la adherencia a los fármacos hipolipemiantes entre los individuos con o sin un MACE, mientras que los que experimentaron un MACE presentaban una mayor adherencia a los fármacos antihipertensivos y antidiabéticos.

Tabla 15: Análisis descriptivo según la incidencia de MACE, de los sujetos incluidos de CARhES.

		Sin MACE N=51.812	**Con MACE N=581**	**P value**
Edad	media (DE)	70,1 (12,8)	78,9 (9,92)	<0,001
Sexo	N (%)			0,005
Hombres		22.101 (42,7)	282 (48,5)	
Mujeres		29.711 (57.3)	299 (51,5)	
FACTORES DE RIESGO CARDIOVASCULAR				
HTA	N (%)	37.767 (72,9)	486 (83,6)	<0,001
HC	N (%)	36.906 (71,2)	410 (70,6)	0,761
DM	N (%)	13.952 (26,9)	229 (39,4)	<0,001
Número de FRCV	N (%)			<0,001
1		22.330 (43,1)	178 (30,6)	
2		22.151 (42,8)	262 (45,1)	
3		7.331 (14,1)	141 (24,3)	
VALORES ANALÍTICOS Y TENSIÓN ARTERIAL				
Colesterol total (mg/dL)	media (DE)	195 (36,1)	187 (35,2)	<0,001
Colesterol HDL (mg/dL)	media (DE)	53,7 (13,4)	50.9 (12,9)	<0,001
Colesterol LDL (mg/dL)	media (DE)	118 (31,5)	111 (30,6)	<0,001
Glucosa en sangre (mg/dL)	media (DE)	104 (24,6)	109 (38,0)	<0,001
TAS (mm Hg)	media (DE)	133 (15,8)	137 (16,6)	<0,001
TAD (mm Hg)	media (DE)	76,8 (13,9)	75,0 (10,7)	<0,001
ADHERENCIA AL TRATAMIENTO, PDC				
Antihipertensivos	media (DE)	58.2 (44.1)	66,2 (40,9)	<0,001
Antidiabéticos	media (DE)	17,2 (33,4)	27,0 (39,2)	<0,001
Hipolipemiantes	media (DE)	38,5 (42,0)	38,7 (42,0)	0,908

MACE: Evento cardiovascular mayor; N: Número; DE: desviación estándar; FRCV: Factores de riesgo cardiovascular; TAS: tensión arterial sistólica; TAD: tensión arterial diastólica; N: número; HTA: hipertensión arterial; HC: hipercolesterolemia; DM: diabetes mellitus; PDC: proporción de días cubiertos.

Estos resultados han sido publicados en el artículo que se presenta en el Anexo III.

Síntesis de los resultados que dan respuesta al Objetivo 2.1

Se incluyeron en este análisis 52.393 sujetos, los cuales sufrían HTA, HC y/o DM, con registros de analíticas y sin iniciar tratamiento durante el periodo de seguimiento, su edad media fue de 70,2 años y una mayoría fueron mujeres. El FRCV más prevalente entre los sujetos de CARhES incluidos en este estudio fue la HTA y un 40% de los sujetos sólo tuvieron un FRCV. Sin embargo, al considerar la prevalencia para el total de la población aragonesa el FRCV más común fue la HC. Además, el grupo de fármacos al que los sujetos fueron más adherentes fue a los antihipertensivos. La incidencia de MACE fue mayor en hombres, y los sujetos con MACE fueron mayores, con mayor frecuencia de HTA y DM y más de un FRCV. Sin embargo, quienes sufrieron un MACE, tuvieron adherencias mayores a fármacos antihipertensivos y antidiabéticos que los que no lo sufrieron.

Resultados que dan respuesta al objetivo 2.2.- Analizar diferencia en la prevalencia de FRCV y nivel socioeconómico y la incidencia de MACE entre hombres y mujeres.

La Tabla 16 muestra las características de la población al inicio del estudio por sexo y según la incidencia de MACE. Una vez excluidos los sujetos menores de 50 años que formaban parte de la cohorte CARhES y sin antecedentes de ECV, el número de individuos incluidos en este estudio fue 278.515, de los cuales el 44,7% eran varones. Las mujeres eran de mayor edad que los hombres. El FRCV más prevalente en ambos sexos fue la HC, seguida de la HTA y, por último, la DM. Los dos primeros FRCV tenían una prevalencia similar en hombres y mujeres, pero la DM era más prevalente en los hombres.

En cuanto al nivel socioeconómico, se encontraron diferencias entre sexos especialmente en los grupos extremos. La mitad de las mujeres incluidas en el estudio estaban jubiladas, ganaban menos de 18.000 euros al año o pertenecían al grupo de farmacia gratuita. Este grupo era también el más numeroso entre los hombres, y representaba el 35,9% de su población. Los asalariados que ganaban más de 18.000 euros al año representaban el 17,6% de la población masculina, mientras que entre las

mujeres eran el 7,7%. El tercer grupo con mayores diferencias entre sexos fue el de los jubilados que ganaban más de 18.000 euros anuales (24,3% en los hombres frente al 17% en las mujeres). Por último, la incidencia de MACE durante el periodo de seguimiento (enero 2018-diciembre 2021) fue del 2,5% en los hombres frente al 1,7% en las mujeres.

Tabla 16: Características de la población total y por sexo al inicio del periodo del estudio de los sujetos incluidos de CARhES.

	Total	**Hombres N=124,602**	**Mujeres N=153,912**
Edad	67,6 (10,5)	65,9 (10)	69,1 (10,8)
Nivel socioeconómico			
Empleados que ganan >18000	33.812 (12,1%)	21.969 (17,6%)	11.843 (7,7%)
Empleados que ganan <18000	39.769 (14,3%)	18.838 (15,1%)	20.931 (13,6%)
Jubilados que ganan <18000 y farmacia gratuita	125.692 (45,1%)	44.738 (35,9%)	80.954 (52,6%)
Jubilados ganando >18000	56.398 (20,2%)	30284 (24,3%)	26.114 (17,0%)
Otros	22.843 (8,20%)	8.773 (7,0%)	14.070 (9,1%)
Hipertensión	171.339 (61,5%)	76.544 (61,4%)	94.795 (61,6%)
Hipercolesterolemia	205.700 (73,9%)	90.640 (72,7%)	115.060 (74,8%)
Diabetes Mellitus	57.612 (20,7%)	30.509 (24,5%)	27.103 (17,6%)
MACE en los 5 años de seguimiento	5.732 (2,06%)	3.169 (2,5%)	2.563 (1,7%)

Información mostrada como número (%) para variables categóricas y media en años (desviación estándar) para Edad. Los porcentajes mostrados se calculan por columnas. MACE: Evento Cardiovascular Mayor

La Tabla 17 muestra las características de los hombres y las mujeres que sufrieron un MACE. Las mujeres con MACE eran mayores que los hombres. En ambos sexos, la incidencia de MACE fue mayor entre los que padecían DM y HTA. Sin embargo, en hombres y mujeres la incidencia de MACE fue mayor en los que no tenían HC que en los que sí la tenían. Finalmente, también en ambos sexos el grupo de nivel socioeconómico con mayor incidencia de MACE fue el de jubilados con ingresos inferiores a 18000€/año o con farmacia gratuita.

Tabla 17: Incidencia de MACE en función de los factores explicativos entre los sujetos incluidos de CARhES

	MACE Hombres 3169 (2.5%)	**MACE Mujeres 2563 (1.7%)**
Edad	70,5 (10,8)	77,2 (10,3)
Nivel socioeconómico		
Empleados que ganan >18000	306 (1,39%)	55 (0,46%)
Empleados que ganan <18000	318 (1,69%)	117 (0,56%)
Jubilados que ganan <18000 y farmacia gratuita	1599 (3,57%)	1922 (2,37%)
Jubilados ganando >18000	764 (2,52%)	351 (1,34%)
Otros	182 (2,07%)	118 (0,84%)
Hipertensión		
No	985 (2,05%)	560 (0,95%)
Si	2184 (2,85%)	2003 (2,11%)
Hipercolesterolemia		
No	958 (2,82%)	773 (1,99%)
Si	2211 (2,44%)	1790 (1,56%)
Diabetes Mellitus		
No	2048 (2,18%)	1795 (1,42%)
Si	1121 (3,67%)	768 (2,83%)
La información se muestra como número (%) para las variables categóricas y media en años (desviación estándar) para la edad. Los porcentajes mostrados se calculan por fila.		

Estos resultados han sido publicados en el artículo que se presenta en el Anexo IV.

Síntesis de los resultados que dan respuesta al Objetivo 2.2

Como respuesta a este objetivo se encontró que las mujeres fueron mayores que los hombres, y que la prevalencia de DM fue la que mayores diferencias mostró entre ambos sexos. En cuanto al nivel socioeconómico, las mayores diferencias entre sexos se encontraron en pensionistas con rentas inferiores a 18000€ anuales o farmacia gratuita, con mayor porcentaje entre las mujeres, y en activos que ganan más de 18000€ anuales, con más porcentaje entre los hombres. En cuanto a las diferencias entre hombres y mujeres que tuvieron un MACE, las mujeres con MACE fueron mayores que los hombres y en ambos sexos la mayor incidencia de MACE estuvo entre los pensionistas que ganaban menos de 18000€/año o con farmacia gratuita.

Resultados que dan respuesta al objetivo 2.3.- Analizar la capacidad de distintos métodos de machine learning para predecir la incidencia de MACE en la cohorte CARhES de manera separada para hombres y mujeres, analizando la influencia de 4 grupos de variables (edad, FRCV, valores analíticos y mediciones de TA y adherencia a tratamientos antihipertensivos, antidiabéticos e hipolipemiantes) en dicha predicción.

Este análisis se realizó con los sujetos incluidos en el objetivo 2.1: individuos que tenían registros de analíticas y que no iniciaron tratamiento durante el periodo de seguimiento; además, para este objetivo, se seleccionaron los que fueron mayores de 49 años.

A través del desarrollo de modelos de machine learning se puede analizar el papel que han tenido las distintas variables introducidas para la predicción de MACEs, y así saber qué variables son más influyentes en su incidencia. Así pues, para dar respuesta a este objetivo se realizaron diferentes modelos utilizando dos técnicas distintas de machine learning: Random Forest y XG Boost. Con cada uno de los métodos se realizaron tres modelos distintos para hombres y mujeres, incluyendo distintos grupos de variables:

- Modelo 1: Edad, análisis de sangre y medición de la TA, factores de riesgo cardiovascular y adherencia a la medicación.
- Modelo 2: Edad, análisis de sangre y medición de la TA, y adherencia a la medicación.
- Modelo 3: Edad, factores de riesgo cardiovascular y adherencia a la medicación.

A continuación, se presentan los resultados para los modelos creados con RF y XG Boost por separado.

Modelos construidos con Random Forest

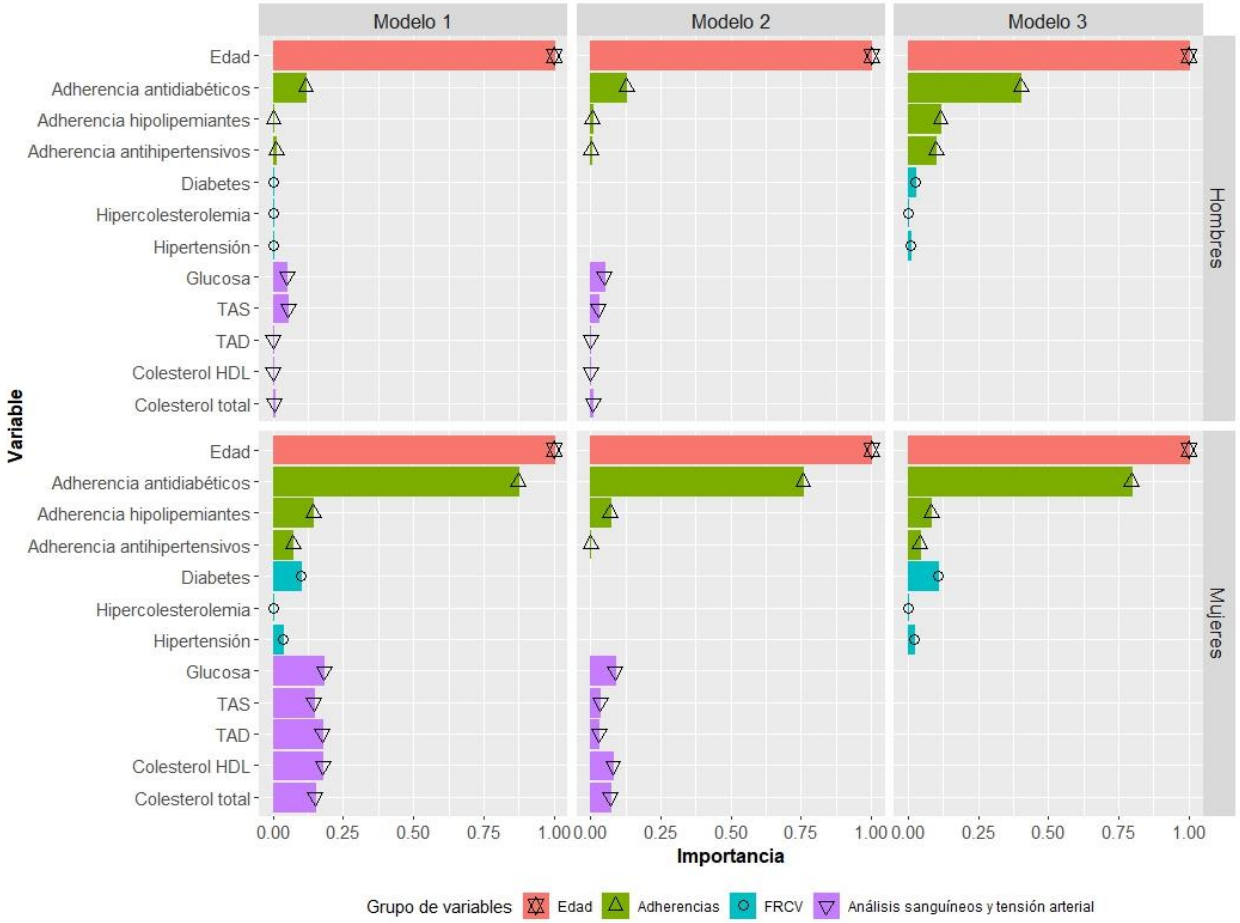
De los modelos construidos considerando a los hombres, el modelo 3 proporcionó el F1 score, la sensibilidad y la especificidad más altas, aunque su exactitud fue la más baja (Tabla 18). En el caso de las mujeres, la puntuación F1 y la sensibilidad más alta se alcanzó con el modelo 3, mientras que todos los modelos alcanzaron una especificidad de 0,75. Las diferencias entre las medidas de validez de los modelos fueron menores entre los modelos generados para la población masculina frente a la femenina.

Tabla 18: Métricas de rendimiento de los modelos creados con Random Forest con la población incluida que formaba CARhES.

	AUC	Índice Youden	F1 SCORE	SENSIBILIDAD	ESPECIFICIDAD
HOMBRES					
MODELO 1	0,70	0,50	0,77	0,62	0,69
MODELO 2	0,70	0,52	0,76	0,61	0,71
MODELO 3	0,69	0,54	0,77	0,62	0,71
MUJERES					
MODELO 1	0,77	0,64	0,71	0,66	0,75
MODELO 2	0,76	0,62	0,81	0,69	0,75
MODELO 3	0,79	0,53	0,84	0,72	0,75

El modelo 1 incluye las variables edad, FRCV, adherencia, y análisis de sangre y mediciones de la tensión arterial. El modelo 2 incluye la edad, la adherencia y las mediciones de la tensión arterial y los análisis de sangre. El modelo 3 incluye la edad, los FRCV y la adherencia al tratamiento. Abreviaciones: AUC: área bajo la curva; FRCV: Factores de riesgo cardiovascular.

En la Figura 13 se muestra la contribución de las variables para los 3 modelos desarrollados aplicando RF para hombres y mujeres por separado. En todos los modelos, tanto para hombres como para mujeres, la edad fue la variable que más contribuyó al riesgo de MACE.



El modelo 1 incluye las variables edad, FRCV, adherencia y análisis de sangre y mediciones de la tensión arterial. El modelo 2 incluye la edad, la adherencia y las mediciones de la tensión arterial y los análisis de sangre. El modelo 3 incluye la edad, los FRCV y la adherencia al tratamiento. Abreviaturas: TAS, tensión arterial sistólica; TAD, tensión arterial diastólica; FRCV, factores de riesgo cardiovascular.

Figura 13: Contribuciones relativas de las variables en los modelos Random Forest para hombres y mujeres de CARhES.

Para los hombres, la edad fue la variable para la que se observó la mayor contribución al riesgo de MACE, seguida de la adherencia a antidiabéticos. La contribución de la adherencia al tratamiento antidiabético en los modelos 1 y 2 fue mucho menor que la contribución de la edad. Además, la contribución de la adherencia al tratamiento antidiabético en el modelo 3 fue mayor que en los modelos anteriores, pero no tan alta como la observada para las mujeres.

En el caso de las mujeres, en términos de contribución relativa al riesgo de MACE, la edad fue seguida de cerca por la adherencia al tratamiento antidiabético. Todas las

demás variables contribuyeron en menor medida. En el modelo 1, las variables de análisis de sangre y medición de la tensión contribuyeron más que el diagnóstico de HTA, DM o HC, y que la adherencia a los fármacos hipolipemiantes o la HTA.

Modelos creados con XG Boost

Los modelos creados para la población masculina utilizando XG Boost alcanzaron niveles de precisión comparables a los de los modelos RF. La puntuación F1 y la sensibilidad fueron superiores a las obtenidas con los modelos RF, mientras que la especificidad fue inferior para los modelos 2 y 3 y superior para el modelo 1.

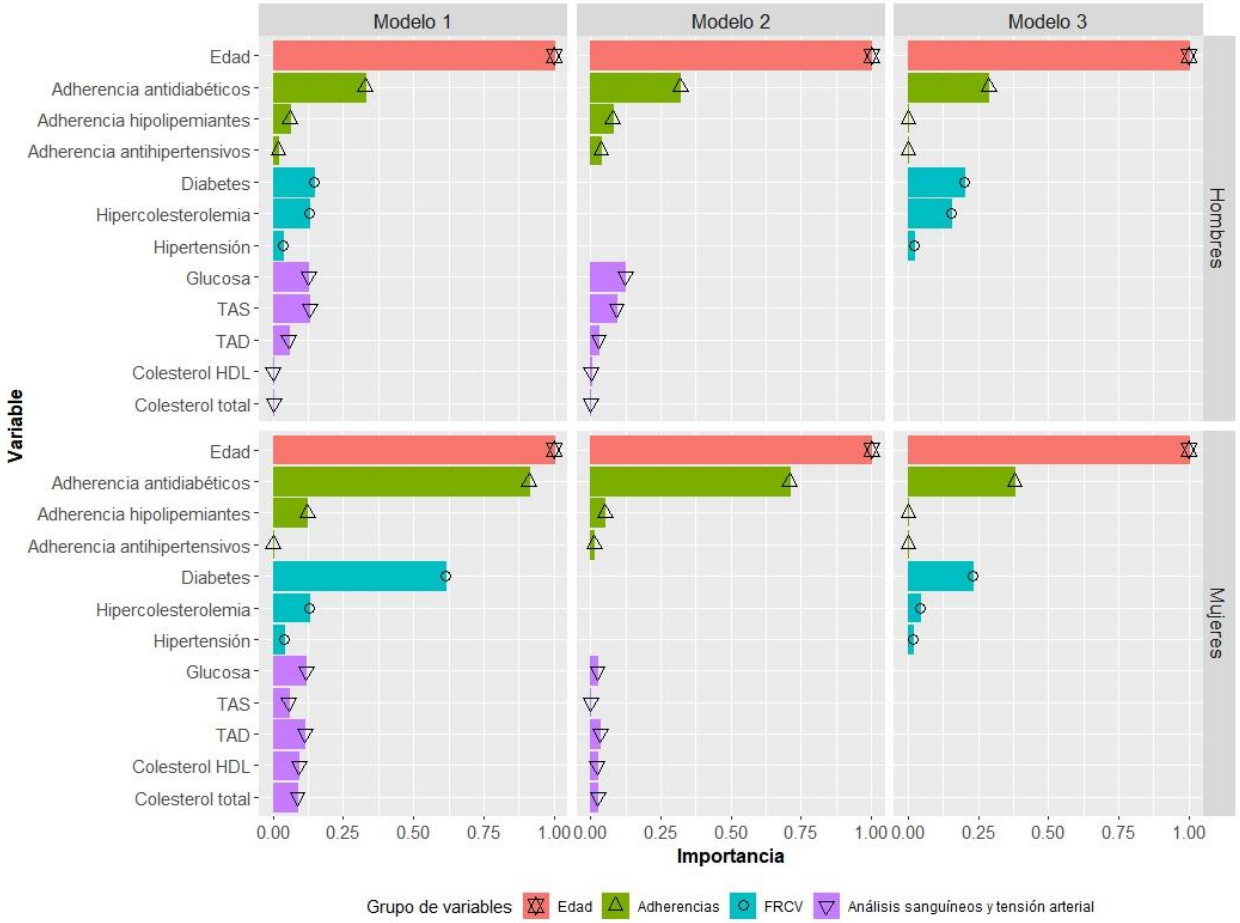
En los modelos creados para la población femenina utilizando XG Boost (Tabla 19), la precisión fue comparable a la de los modelos RF para los modelos 2 y 3, mientras que el AUC fue inferior para el modelo 1 en relación con el modelo RF correspondiente. La puntuación F1 y la sensibilidad fueron las más altas para el modelo 1, mientras que la especificidad fue la más alta para el modelo 2. En comparación con el modelo RF correspondiente, la puntuación F1 y la sensibilidad, pero no la especificidad, fueron mayores en el modelo 1 de XG Boost. Es decir, no se obtuvo un modelo para el que las medidas de validez fueran claramente mejores que en los demás, estas variaron en todos los modelos.

Tabla 19: Métricas rendimiento de los modelos creados con XG Boost con la población incluida de CARhES.

	AUC	Índice Youden	F1 SCORE	SENSIBILIDAD	ESPECIFICIDAD
HOMBRES					
MODELO 1	0,70	0,53	0,78	0,64	0,71
MODELO 2	0,70	0,51	0,79	0,65	0,68
MODELO 3	0,69	0,52	0,79	0,65	0,66
MUJERES					
MODELO 1	0,74	0,58	0,89	0,80	0,56
MODELO 2	0,76	0,54	0,80	0,67	0,81
MODELO 3	0,79	0,50	0,81	0,69	0,78

El modelo 1 incluye las variables edad, FRCV, adherencia, y análisis de sangre y mediciones de la tensión arterial. El modelo 2 incluye la edad, la adherencia y las mediciones de la tensión arterial y los análisis de sangre. El modelo 3 incluye la edad, los FRCV y la adherencia al tratamiento. Abreviaturas: AUC: área bajo la curva; FRCV: Factores de riesgo cardiovascular.

En los modelos construidos con XG Boost, tanto para hombres como para mujeres, las variables que más contribuyeron a predecir un riesgo elevado de ECV fueron la edad seguida de la adherencia al tratamiento antidiabético (Figura 14). Para los hombres, en el modelo 1 de XG Boost, se observaron contribuciones similares para la DM y la HC y para la glucemia y la TAS. En el caso de las mujeres, contrariamente a lo observado en los modelos RF, en el modelo 1 de XG Boost la DM contribuyó mucho más que los análisis de sangre y las mediciones de la TA, con un efecto similar al de la adherencia al tratamiento antidiabético.



El modelo 1 incluye las variables edad, FRCV, adherencia y análisis de sangre y mediciones de la tensión arterial. El modelo 2 incluye la edad, la adherencia y las mediciones de la tensión arterial y los análisis de sangre. El modelo 3 incluye la edad, los FRCV y la adherencia al tratamiento Abreviaturas: TAS, tensión arterial sistólica; TAD, tensión arterial diastólica; FRCV, factores de riesgo cardiovascular.

Figura 14: Contribuciones relativas de las variables en los modelos XG Boost para hombres y mujeres de la cohorte CARhES.

Estos resultados han sido enviados para ser publicados, el manuscrito enviado se presenta en el Anexo III.

Síntesis de los resultados que dan respuesta al Objetivo 2.3

Para dar respuesta al objetivo propuesto, se generaron y evaluaron tres modelos para cada sexo utilizando los algoritmos RF y XG Boost. Ambos algoritmos mostraron medidas de rendimiento similares.

En todos los modelos, para ambos sexos, la edad fue el parámetro que más contribuyó a predecir un riesgo elevado de ECV, seguido de cerca por la adherencia a los

antidiabéticos. La influencia de la adherencia a antidiabéticos en la incidencia de MACE fue mayor en las mujeres que en los hombres.

Resultados que dan respuesta al objetivo 2. 4.- Estudiar el impacto que tienen las diferencias en la distribución de hipertensión, hipercolesterolemia, diabetes y nivel socioeconómico entre sexos en las diferencias observadas en la incidencia de MACE.

Para dar respuesta al último objetivo de esta tesis, considerando los sujetos de CARhES mayores de 49 años, se trató de cuantificar la contribución causal de las diferencias en diabetes, hipertensión, hipercolesterolemia y nivel socioeconómico en la incidencia de MACE por sexo. Esto se hizo a través de distintos análisis contrafactuales, los resultados de estos cuatro análisis se muestran en la Figura 15.

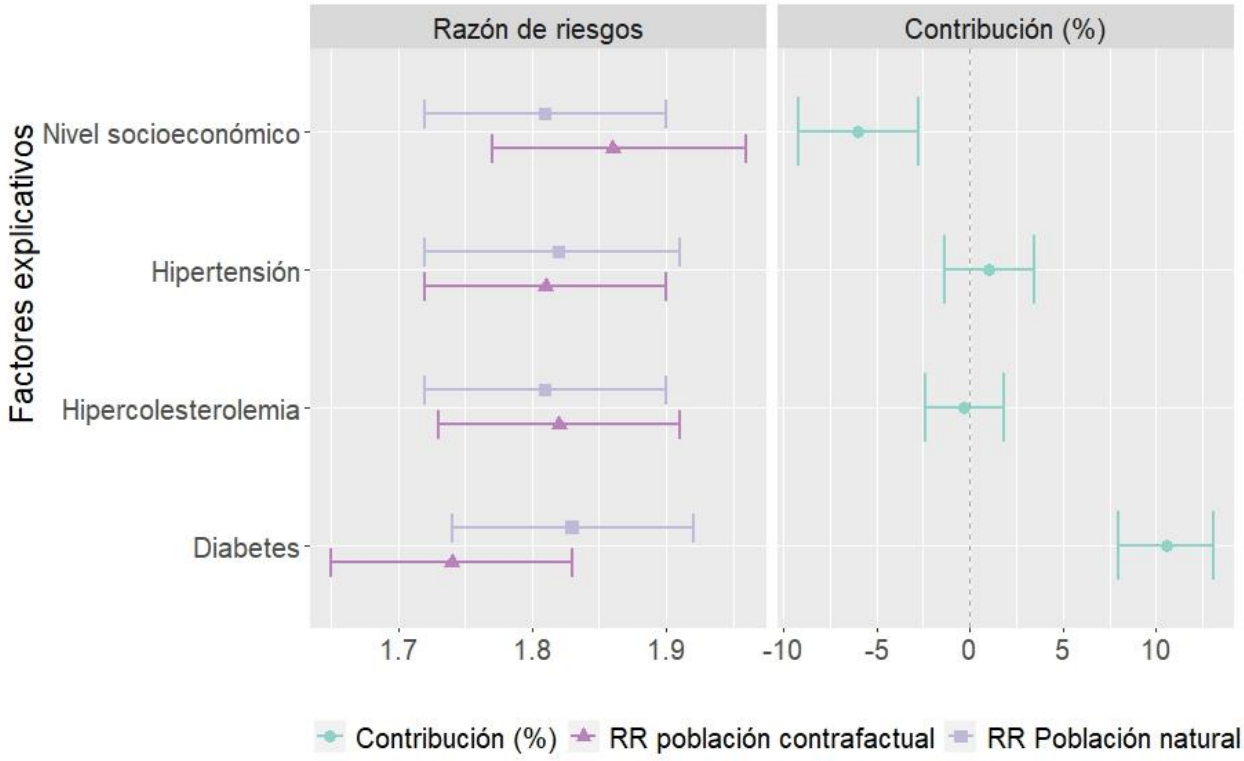


Figura 15: RR para las poblaciones natural y contrafactual y porcentaje de contribución de cada factor explicativo en CARhES (RR: Razón de riesgo)

Cuando la DM fue el factor explicativo, el RR en el análisis del curso natural de sufrir un MACE para los hombres en relación con las mujeres fue de 1,83 (IC 95%: 1,74; 1,92). Es decir, una vez controlado el efecto de los factores de confusión (edad, HTA, HC y nivel socioeconómico) y debido a las diferencias en la prevalencia de DM entre ambos sexos, los hombres tuvieron un 83% más de probabilidades de sufrir un MACE que las mujeres.

El RR contrafactual (tras establecer que los hombres tuvieran la misma distribución de DM que las mujeres) fue de 1,74 [IC 95%: (1,65, 1,83)], lo que corresponde a una contribución causal de la DM a la diferencia entre sexos en la incidencia de MACE del 10,5% (IC 95%: 7,99, 13,08); es decir, si la prevalencia de DM en los hombres fuera la misma que en las mujeres, la incidencia de MACE en los hombres se reduciría del 2,5% al 2,2%.

En el caso de la HTA, se encontraron cocientes de riesgo muy similares en la población del curso natural y en la población contrafactual [1,81; IC 95%: (1,72,1,9) y 1,82; IC 95%: (1,73,1,91)], respectivamente). Por tanto, la contribución porcentual de la HTA a la diferencia en la incidencia de MACE entre sexos fue muy baja [1%, IC 95%: (-1,4, 3,4)] y no estadísticamente significativa. Algo similar ocurrió al analizar la HC, con un porcentaje de contribución cercano a 0.

Por último, al considerar el nivel socioeconómico como factor explicativo, el RR de MACE para los hombres en relación con las mujeres en el análisis la población natural fue menor (RR=1,8) que en el contrafactual (RR=1,9). Esto significaría que si los hombres tuvieran el mismo nivel socioeconómico que las mujeres, el riesgo de sufrir un MACE entre los hombres aumentaría. Por este motivo, la contribución del nivel socioeconómico a la diferencia entre sexos en la incidencia de MACE fue negativa (-6%; IC 95%: (-9.2, -2.77)]. Este efecto compensatorio se observa en este caso porque los MACE son mayores entre los del nivel socioeconómico bajo (Tabla 17) y hay más mujeres que hombres en el nivel socioeconómico bajo (Tabla 16).

Estos resultados han sido publicados en la revista *European Journal of Public Health* y el manuscrito se presenta en el Anexo IV.

Síntesis de los resultados que dan respuesta al Objetivo 2.4

La incidencia de MACE fue mayor en hombres que en mujeres. Se observó que, después de ajustar por factores de confusión, los hombres tenían un 83% más de probabilidades de experimentar MACE que las mujeres debido a la DM. Esta contribuyó en un 10,5% a la diferencia de incidencia de MACE entre sexos. Las contribuciones de la HTA y HC fueron bajas y no significativas. Finalmente, cuando el nivel socioeconómico fue considerado factor explicativo, el RR de MACE para los hombres en relación con las mujeres en el análisis la población natural fue menor que en el contrafactual. Esto indica que el nivel socioeconómico mostró un efecto compensatorio negativo (-6%).

5. DISCUSIÓN

5. DISCUSIÓN

En este apartado se recordarán las principales razones que justifican el trabajo con dos cohortes diferentes, se comentarán los resultados más relevantes y se presentarán las limitaciones y fortalezas del estudio presentado.

Poblaciones de estudio

Para dar respuesta a los objetivos marcados en la presente tesis se ha trabajado con dos cohortes distintas. Los primeros análisis se enmarcaron en una cohorte de trabajadores de una fábrica automovilística donde la población tuvo características homogéneas: el 93,5% de los sujetos que componían la cohorte eran hombres, entre 40 y 60 años en el momento de entrada en la cohorte, que realizaban trabajos manuales y con un nivel socioeconómico medio. En cuanto al nivel de RCV, por su edad y características, la población del AWHS estaba compuesta por sujetos con bajo riesgo.

Dadas las características de esta primera cohorte, y que las mujeres tuvieron que ser excluidas en los estudios por su bajo número (6,5%), los siguientes análisis se realizaron con datos de la cohorte CARhES. Como se ha comentado en otros apartados, esta cohorte recoge la información de todos los aragoneses y aragonesas con algún FRCV. La cantidad de sujetos incluidos fue sustancialmente mayor, con una población más heterogénea y permitió incluir a mujeres en los análisis. En cuanto a la edad, se incluyeron sujetos mayores de 16 años habiendo así representación de todos los grupos de edad, especialmente de personas mayores de 60 años. Así pues, aunque no se incluyó toda la población aragonesa, se cuenta con una parte muy importante de los grupos poblacionales de edad más avanzada. Además, dado que el criterio de entrada fue que tuvieran algún FRCV, consistió en una población con alto RCV.

A continuación, se comentarán los resultados obtenidos para las dos cohortes, comenzando por los resultados de la cohorte AWHS y posteriormente de la cohorte CARhES.

Discusión de los resultados cohorte AWHS

La discusión de los resultados encontrados para el AWHS se presenta en este apartado: en primer lugar, se comentan los resultados de los análisis descriptivos, seguidos de los resultados para los análisis de machine learning y por último los resultados encontrados en el estudio de clusters longitudinal.

Resultados análisis descriptivos cohorte AWHS

En el marco de esta cohorte se realizó un análisis descriptivo de la frecuencia de FRCV, así como de la exposición a fármacos antihipertensivos, hipolipemiantes y antidiabéticos.

Como resultado de estos análisis se encontró que los FRCV más prevalentes en dicha cohorte fueron la obesidad/sobrepeso y la HC, seguidos de HTA. La incidencia de ECV fue del 7,9% en 10 años de seguimiento. En el análisis descriptivo estratificado según incidencia de ECV, aquellos sujetos con ECV tuvieron media de edad más alta, con mayores prevalencias de HTA, HC y DM, y un porcentaje ligeramente menor de sobrepeso pero mayor de obesidad. Finalmente, la exposición al tratamiento fue un 20% menor en los que tuvieron un ECV que en los que no lo presentaron.

La prevalencia de FRCV en este trabajo muestra discrepancias con las prevalencias de los mismos FRCV descritas por otros estudios en España [2,8]. Estas disparidades pueden atribuirse a las características específicas previamente mencionadas de la población del AWHS. No obstante, según dichas fuentes, la obesidad o el sobrepeso se presentaron como el FRCV más prevalente, mientras que la diabetes fue el menos prevalente, resultados que coinciden con los hallazgos del presente estudio. En la comparativa de frecuencia de FRCV se observó una prevalencia mayor entre las personas que habían tenido un ECV. Estos FRCV son determinantes clave en el riesgo de sufrir un ECV [2], por lo que cabía esperar que en aquella población que ha sufrido algún evento su prevalencia sea mayor.

En cuanto a la adherencia a tratamiento, su impacto en la incidencia de ECV está ampliamente descrito [73–76]. Otros estudios [77] indican que un número considerable de ECV se debe a una adherencia deficiente a los tratamientos preventivos

cardiovasculares. Por lo tanto, cabía esperar que nuestra variable exposición al tratamiento fuera más baja en aquellos con evento que en aquellos que no lo hayan sufrido.

Otro de los análisis realizados en el presente estudio tuvo como objetivo analizar la evolución, entre 2009 y 2017, de valores analíticos y variables médicas relacionadas con el riesgo de sufrir un evento cardiovascular.

Las variables cuantitativas que experimentaron mayores cambios a lo largo de todo el periodo de estudio fueron el colesterol total, la glucosa y el SCORE de RCV a 10 años. El SCORE medio aumentó, mientras que los valores medios de colesterol total y glucosa disminuyeron. Estos resultados coinciden con los del análisis realizado por cuartiles. Debe tenerse en cuenta que el aumento observado en el SCORE podría deberse principalmente al aumento de la edad, variable que tiene gran repercusión en el valor de este estimador de RCV. Diferentes autores [56,78–80] han descrito la influencia de la edad en el cálculo del SCORE. Conroy et al. [56] observaron que el SCORE era muy bajo en personas de 30 años, y que aumentaba más rápidamente entre los 50 y los 65 años. En el presente estudio, el SCORE aumentó con el tiempo, pero este aumento fue probablemente inferior al esperado, lo que puede deberse al efecto de la estabilización y mejora de algunos FRCV. Así, los niveles medios de glucosa y colesterol total disminuyeron de 97,7 mg/dl a 88,06 mg/dl, y de 212,18 mg/dl a 187,96 mg/dl, respectivamente, entre el primer y el último momento del estudio realizado (años 2009-2010 y año 2017, respectivamente). La disminución de los niveles de colesterol total y glucosa y la estabilización de la tensión arterial podría deberse al estrecho control al que los sujetos incluidos son sometidos por los servicios médicos de la fábrica, y a que estos FRCV se controlan mediante tratamiento farmacológico.

El IMC no mostró cambios significativos, destacando que más de la mitad de los participantes presentaban sobrepeso durante la duración del estudio. Otro hallazgo destacable fue la disminución a lo largo del tiempo del porcentaje de fumadores, hecho que coincide con otros estudios[2,3,81]. Algunos autores han encontrado una tendencia decreciente en los niveles de colesterol en sangre en los últimos años[2,82]. Estos estudios muestran también un aumento del IMC y de proporción de personas con obesidad.

Es importante considerar que, como se mencionó previamente, todos estos análisis se llevaron a cabo en el contexto de una cohorte de trabajadores con edades similares y un bajo RCV. No obstante, parece que las intervenciones focalizadas en la atención primaria, orientadas hacia personas de alto riesgo según su edad o factores de riesgo, demuestran ser efectivas[83].

Resultados análisis machine learning en la cohorte AWHS

En la estimación del riesgo cardiovascular, las calculadoras tradicionalmente utilizadas para su estimación presentan algunos inconvenientes ya mencionados, como que son desarrolladas en poblaciones específicas y que pueden tener ciertas limitaciones metodológicas por la correlación entre las variables, su no linealidad y la posibilidad de sobreajuste. Por otra parte, la creciente disponibilidad de datos médicos generados en la práctica clínica diaria proporciona una gran cantidad de información que se podría explotar con técnicas de machine learning en los estudios de riesgo[48,49].

Las técnicas de aprendizaje automático se han aplicado ampliamente [48,50] para analizar estos enormes conjuntos de datos y superar algunas de las limitaciones de los sistemas tradicionales. El aprendizaje automático puede utilizarse para generar modelos que predigan mejor el riesgo, aumentando así la eficiencia, objetividad y fiabilidad del proceso diagnóstico [45,48,51,84].

Por ello, el trabajo presentado se propuso desarrollar distintos modelos aplicando técnicas de machine learning para predecir la aparición de ECV y tratar de describir la influencia de distintos FRCV y la exposición a tratamientos preventivos en la incidencia de evento mediante la aplicación de dichos modelos.

Puesto que hay un importante número de variables que pueden influir en el RCV, unas más estudiadas que otras y como se ha presentado en el apartado de metodología, se aplicaron 3 técnicas distintas de machine learning, para comparar su rendimiento, utilizando en cada caso dos combinaciones de variables predictivas: a) FRCV (edad, estado físico, HC, hipertensión y DM); y b) FRCV más exposición al tratamiento. Primero

se discutirán los resultados obtenidos en cuanto al papel de las distintas variables en la predicción de eventos cardiovasculares y después se discutirá el rendimiento de las distintas técnicas de machine learning utilizadas.

En los modelos en los que sólo se incluyeron los FRCV, la edad mostró la mayor capacidad predictiva, mientras que la de otros FRCV fue notablemente inferior, y varió en función del modelo utilizado. La relación entre ECV y edad, HC, hipertensión, DM y estado físico está bien documentada[19,85–87]. Todos los modelos mostraron unánimemente que la edad tiene el mayor poder predictivo, lo que confirma que esta variable es un FRCV clave no modificable[45,84,85,88].

Cuando se añadió la exposición al tratamiento como variable, la edad siguió siendo la variable con mayor poder predictivo, seguida de la exposición al tratamiento. Además, la capacidad predictiva de la exposición al tratamiento se aproximó más a la de la edad que a la de cualquiera de los otros FRCV. Estos resultados indican que la exposición al tratamiento desempeña un papel importante en la determinación de la incidencia de ECV y, por tanto, debe tenerse en cuenta a la hora de gestionar el RCV.

El control farmacológico adecuado de los FRCV es crucial para mitigar el riesgo de sufrir un ECV. A pesar de que las guías clínicas establecen los niveles de TA, colesterol y glucosa en sangre óptimos que deben ser alcanzados con estos tratamientos para disminuir el riesgo [34,89], la adherencia a estos tratamientos es subóptima[73–75]. Es importante destacar que la literatura ha señalado que un control inadecuado de los FRCV se asocia con un considerable número de casos de ECV [77].

Las metodologías convencionales para evaluar el riesgo de ECV generalmente no consideran la adherencia al tratamiento como una variable predictiva, lo que podría ser una limitación significativa de dichos métodos. La exposición a medicamentos utilizados para el control de los FRCV podría ser considerada al determinar el riesgo de experimentar una ECV, dado que esta exposición varía considerablemente entre los individuos y desempeña un papel fundamental en la gestión del riesgo. Esta tesis respalda este enfoque, ya que se observó una mejora sustancial en el rendimiento de los modelos cuando se incluyó la exposición al tratamiento como variable predictiva, la cual se determinó a partir de la adherencia a cada tratamiento por separado.

Como se ha mencionado anteriormente, la capacidad predictiva de los FRCV incluidos varió de un modelo a otro. Además, la combinación de dos o más FRCV se asocia con un aumento del riesgo de mortalidad [87]. Es importante señalar que la influencia de los FRCV individuales varía según las investigaciones [19,86,87], como ocurre en los resultados presentados en esta tesis.

En el contexto de la influencia de los distintos FRCV en la incidencia de ECV, Yandrapalli et al. hallaron que la HC era el más influyente, seguido de cerca por la hipertensión, el tabaquismo y, por último, la DM [86]. En otro estudio [87] se observó que la hipertensión era el FRCV más estrechamente relacionado con la mortalidad por todas las causas, seguido de la DM, la HC y el sobrepeso. Huang et al. [87] concluyeron que los FRCV más importantes eran la obesidad y el tabaquismo en las zonas rurales y urbanas, respectivamente, seguidos de la dislipemia. En este sentido, más allá de la edad, los resultados de diferentes estudios presentan cierta diversidad sobre cuál es el FRCV con mayor poder determinante de la incidencia de ECV.

Los modelos presentados en el presente estudio podrían ser útiles en la práctica clínica para evaluar el riesgo individual de sufrir un ECV, en función de las características del paciente y del cumplimiento terapéutico, por lo que podrían servir para hacer una estimación del RCV. Además, pueden ayudar a orientar la intervención y a identificar las medidas más apropiadas a adoptar (por ejemplo, cambio de conducta enfocado a reforzar la adherencia).

A lo largo de los años, se han invertido muchos esfuerzos en calcular el riesgo de ECV, utilizando diversos métodos, muchos de los cuales presentan limitaciones que pueden superarse con técnicas de aprendizaje automático. Estas técnicas ofrecen diversos enfoques para procesar importantes cantidades de datos con el fin de predecir la incidencia de la ECV, lo que permite a investigadores y clínicos seleccionar el algoritmo que mejor se adapte a sus datos u objetivos.

En estudios anteriores [47,90] en los que se compararon distintos algoritmos se observó que el modelo RF proporcionaba los resultados más precisos, en consonancia con nuestros hallazgos. En una comparación previa de XGBoost y NB [88] se observó que XGBoost obtuvo mejores resultados, como también se observó en el presente estudio.

Entre los resultados obtenidos, destacar que la edad fue la variable con mayor capacidad predictiva de ECV en todos los modelos, excepto en el modelo NB cuando se incluyeron como variables predictivas los FRCV más exposición al tratamiento, en cuyo caso la exposición al tratamiento mostró la mayor capacidad predictiva, seguida de cerca por la edad.

Resultados análisis cluster longitudinal en la cohorte AWHS

Los últimos análisis realizados dentro de la cohorte AWHS estuvieron enfocados a analizar las diferentes trayectorias de los FRCV y el SCORE a lo largo de tres momentos temporales.

Uno de los objetivos de la tesis era identificar perfiles de pacientes en función de la evolución de sus FRCV y de su riesgo de sufrir un ECV. Para identificar esos perfiles, se realizó un análisis de cluster longitudinal. Estas técnicas consisten en dividir la cohorte de estudio en grupos basados en las trayectorias de las distintas variables a lo largo del tiempo. Este algoritmo se aplicó a los valores de FRCV y SCORE en los tres momentos analizados. Los resultados mostraron que los participantes en el estudio podían dividirse en 2 grupos, en función de la evolución de los valores de FRCV y SCORE.

El primer grupo estaba formado por los hombres más jóvenes con valores medios más bajos de glucosa en sangre, IMC y perímetro de cintura, valores SCORE más bajos y valores medios más altos de colesterol HDL. El segundo grupo estaba formado por personas de más edad con valores medios más altos de glucemia, IMC y perímetro de cintura, valores SCORE más altos y valores medios más bajos de colesterol HDL.

El análisis de los cambios en los cuartiles estratificado por los clusters identificados reveló que el porcentaje de individuos que permanecieron en cuartiles con valores de IMC, perímetro de cintura y glucosa superiores a los recomendados, y valores de colesterol HDL inferiores a los recomendados, fue mayor en el cluster 2 que en el cluster 1. Además, en ambos clusters este porcentaje aumentó con el tiempo para todas las variables excepto para la glucosa, para la que se observaron descensos a lo largo de los tres momentos. En el caso del SCORE, se observó una evolución similar en ambos

clusters para los individuos que iniciaron el estudio en el Q2, Q3 o Q4. La mayor diferencia entre clusters se observó para los trabajadores con un SCORE en Q1: la proporción de trabajadores que permanecieron en este cuartil a lo largo de los tres momentos fue mayor en el cluster 1 que en el cluster 2.

Pocos estudios publicados han utilizado una metodología de agrupación similar a la nuestra para analizar la evolución de los FRCV. Varios estudios han analizado las trayectorias de uno [91–94] o varios [95,96] FRCV utilizando una variedad de métodos diferentes, y han intentado identificar correlaciones entre sus hallazgos y otros factores o enfermedades. Uno de estos estudios [92] analizó la evolución de la TAS, para la que se identificaron 4 trayectorias distintas a lo largo del tiempo. Los autores observaron que las trayectorias de la TAS no predecían la ECV ni la mortalidad por cualquier causa mejor que los valores medios de TAS. En otro estudio [94] sobre TAS y TAD en una población anciana se identificaron 3 trayectorias de TA. Las trayectorias de TA también fueron analizadas por Allen et al. [93],que identificaron 5 trayectorias de TA en una población de mediana edad. Rospleszcz et al. [97]analizaron la asociación entre las trayectorias de FRCV y los depósitos de tejido adiposo utilizando una metodología similar a la nuestra e identificaron 3 clusters distintos. El primer cluster agrupaba a los individuos con la edad media más joven y los valores medios de FRCV más bajos, y el tercer cluster agrupaba a los individuos con la edad media más alta y los valores medios de FRCV más elevados. La edad media y los valores medios del FRCV del cluster 2 se situaban entre los de los clusters 1 y 3, excepto para el colesterol total y el colesterol HDL, cuyos valores eran superiores a los de los otros 2 clusters. Por último, Norby et al. [95] utilizaron un modelo de mezcla para identificar por separado las trayectorias de diferentes FRCV. Se identificaron cinco trayectorias distintas para el IMC, la obesidad y la TAS, y 4 para la hipertensión y la diabetes. Así pues, los estudios analizados encuentran distintas conclusiones y agrupaciones para los FRCV analizados.

El análisis de los resultados de los estudios realizados dentro esta tesis en la cohorte AWHS y la comparación con otros estudios muestra que las prevalencias de FRCV en dicha cohorte difieren de las prevalencias encontradas en otros estudios realizados en la población española, pudiéndose deber a las particularidades de la cohorte AWHS, cuyos

sujetos se encontraron en un rango de edad muy limitado, siendo relativamente jóvenes, y compartían características sociodemográficas muy homogéneas.

Tanto en los análisis dentro del AWHS como en los revisados se encontró que la prevalencia de FRCV fue mayor en aquellas personas con ECV que en aquellas que no habían tenido este tipo de eventos. La influencia de la edad en el desarrollo de FRCV y de ECV también ha sido ampliamente documentada y así lo reflejan los análisis de machine learning dentro de la cohorte AWHS. La importancia de la adherencia a los fármacos para controlar FRCV ha quedado plasmada tanto por su influencia a la hora de predecir ECV como por la mejora en el rendimiento de los modelos al introducir esta variable. Finalmente, se encontraron dos perfiles de pacientes distintos según la evolución que tuvieron en el riesgo de sufrir ECV y de sus FRCV. Los estudios observados que tienen como objetivo la identificación de perfiles según la evolución de FRCV, identificaron distinto número de perfiles así como distintas características dentro de estos, lo que subraya la diversidad de enfoques en la investigación de esta temática.

Discusión resultados cohorte CARhES

Se procederá a discutir los resultados obtenidos en los estudios llevados a cabo en la cohorte CARhES, iniciando con los análisis descriptivos, seguidos por los análisis de machine learning, y concluyendo con los resultados derivados del análisis contrafactual.

Resultados análisis descriptivos cohorte CARhES

El estudio realizado con información de la cohorte CARhES comenzó por el análisis descriptivo de los FRCV y la incidencia de MACE, utilizando datos de vida real (RWD) de usuarios del Sistema Sanitario Público de Aragón.

La población del estudio CARhES estaba formada por mayor número de mujeres que de hombres, al igual que la población aragonesa. El FRCV más prevalente en la cohorte fue la hipertensión, mientras que cuando se calcularon las prevalencias para la población aragonesa, el FRCV más prevalente fue la hipercolesterolemia. La incidencia de MACE en la cohorte durante el periodo de estudio fue mayor en hombres que en mujeres, siendo el ictus el MACE más frecuente en ambos sexos, pero con porcentaje más elevado en mujeres. La prevalencia de DM e hipertensión entre los sujetos incluidos en CARhES fue mayor en los que habían sufrido una ECV que en los que no lo tuvieron.

Como ya se ha comentado, el FRCV más prevalente en la población aragonesa fue la hipercolesterolemia, tanto en hombres como en mujeres, mientras que en la cohorte CARhES, en la que todos los incluidos presentan algún FRCV, fue la hipertensión. Según la Encuesta Europea de Salud [2,8], en España, el FRCV más prevalente es la hipertensión, con niveles alrededor del 20%, similares a los encontrados en este estudio para la población aragonesa. Sin embargo, la prevalencia de hipercolesterolemia para el total de la población adulta aragonesa fue casi el doble en este estudio que en los resultados publicados por dicha encuesta. Estas diferencias se pueden deber a que la información de dicha encuesta es recogida de forma autoreferida mientras que los datos de este estudio que provenían de diagnósticos clínicos. En cuanto a la diabetes, los resultados reportados por la Encuesta Europea y la presente tesis para la población de todo Aragón fueron similares. Los porcentajes de HTA, HC y DM en la población de CARhES fueron

diferentes a los estudios consultados[2,8], pero hay que tener en cuenta las particularidades de dicha población, ya que, como se ha mencionado, todos los incluidos presentan algún FRCV.

El sexo/género [98] es considerado un factor no modificable que tiene importancia en el riesgo de desarrollar ECV. Las diferencias entre hombres y mujeres en cuanto a las ECV están relacionadas con distintos factores como que la interacción de los FRCV en el desarrollo de ECV es diferente entre ellos o que las mujeres presentan afecciones específicas, como la preeclampsia, la diabetes gestacional y la menopausia prematura, que se han asociado a un aumento del riesgo de ECV [3,15]. Considerando las diferencias entre sexos en el desarrollo de FRCV y ECV, se llevó a cabo un análisis descriptivo para examinar la prevalencia de FRCV, el nivel socioeconómico y la incidencia de eventos, estratificándolos por sexo.

Se observó que el porcentaje de individuos que sufrieron un MACE entre los pacientes con diabetes e hipertensión fue mayor que entre aquellos que no tuvieron ninguno de esos FRCV. Hay que tener en cuenta que estos sujetos sí que tenían hipercolesterolemia, ya que todos los sujetos incluidos en este estudio tuvieron algún FRCV.

Por otra parte, se observó que en ambos sexos el grupo nivel socioeconómico con mayor incidencia de MACE fue el de jubilados con ingresos inferiores a 18000€/año o con farmacia gratuita.

Un estudio realizado en la misma región de Aragón encontró mayor porcentaje de mujeres con insuficiencia cardiaca que hombres (aunque las mujeres eran mayores) pero los hombres tenían más probabilidad de tener cardiopatía isquémica e IAM [99].

Según la Sociedad Europea de Cardiología [2], tanto en España como en los países de renta alta, la incidencia de cardiopatía isquémica estandarizada por edad tiende a ser mayor en los varones, pero es similar en ambos sexos para el ictus.

En consonancia con nuestros resultados, la literatura [2,21] ha demostrado que los hombres tienen una mayor incidencia de MACE y que las mujeres con MACE tienden a ser mayores que los hombres. Aunque su incidencia es menor, las enfermedades cardiovasculares son también la principal causa de muerte en las mujeres, pero en ellas

están infravaloradas y son menos tratadas tanto a nivel de prevención primaria como de secundaria. Los médicos tienden a pensar que las ECV son predominantemente masculinas, lo que puede llevar a un uso mayor del cribado del riesgo de ECV en los hombres y a un uso insuficiente en las mujeres. Además, se ha demostrado que las mujeres presentan síntomas diferentes cuando sufren un IAM, lo que puede llevar a los médicos a pasar por alto algunos de estos episodios agudos [21]. También hay pruebas de que, tras sufrir un MACE, la consecución de los objetivos relativos a los FRCV y al estilo de vida es menor en las mujeres que en los hombres, lo que podría ser el resultado de un tratamiento diferente para las mujeres en las prescripciones médicas [22].

Resultados análisis de machine learning en la cohorte CARhES

Los análisis de machine learning realizados en la cohorte AWHS se replicaron en el entorno de la cohorte CARhES, permitiendo la integración de nuevas variables en el análisis y el estudio también de las mujeres.

Dado que la incidencia, las interacciones y el control de los FRCV difieren en hombres y mujeres [81,100–102], se realizaron los modelos para cada sexo por separado. En todos los modelos, para ambos sexos, la edad fue el parámetro que más contribuyó a predecir la incidencia de MACE, seguido de cerca por la adherencia a los antidiabéticos. Además, la adherencia a los antidiabéticos contribuyó más a la incidencia de MACE en las mujeres que en los hombres. Por último, tanto para los hombres como para las mujeres, las contribuciones de las demás variables a la predicción del riesgo de ECV diferían según los modelos.

Los resultados confirman la edad como la variable más importante a la hora de predecir el riesgo de ECV. Estudios previos que incluyen la edad como variable predictiva han mostrado consistentemente que este parámetro tiene el mayor poder predictivo, sugiriendo que es un FRCV clave no modificable[18,103–105].

En cuanto a las contribuciones relativas de las demás variables individuales difieren entre hombres y mujeres. Sin embargo, la adherencia a los antidiabéticos fue considerada un determinante clave de la incidencia de MACE en ambos sexos.

Hasta donde sabemos, ningún estudio basado en métodos de machine learning publicado hasta la fecha ha considerado la adherencia a tratamientos cardiovasculares como variable predictiva. Los estudios que consideraron alguna variable relacionada con tratamientos para la hipertensión hipercolesterolemia o diabetes, sólo consideraron si estos tratamientos habían sido prescritos y esta información se recogió a través de cuestionarios auto reportados por los participantes [18,104,105].

Sin embargo, la adherencia a los antihipertensivos y a los fármacos hipolipemiantes mostraron escaso poder predictivo en el presente trabajo. Múltiples estudios han descrito asociaciones entre la adherencia a los antihipertensivos y a los fármacos hipolipemiantes y la incidencia de ECV y el riesgo de mortalidad por todas las causas [106–108]. En cuanto a la influencia sobre el riesgo de diferentes tipos de ECV, la adherencia a los antidiabéticos está menos estudiada que la adherencia a los fármacos hipolipemiantes y antihipertensivos [106]. En su revisión sistemática, Mengying et al. [106] consideraron la adherencia a los antidiabéticos, a los antihipertensivos y a los fármacos hipolipemiantes, y hallaron que una mayor adherencia a estos tres tratamientos estaba asociada a un menor riesgo de ECV.

Los hallazgos anteriores subrayan la importancia de un adecuado control farmacológico de los FRCV modificables para reducir el riesgo de ECV. Las guías clínicas proponen el control de los FRCV para disminuir este riesgo [34,89]. Investigaciones previas [73–75] han demostrado que la adherencia a estos tratamientos es subóptima, mientras que los métodos más utilizados para determinar el riesgo de ECV no incluyen la adherencia a la medicación como variable predictiva. También hay pruebas [77] que sugieren que un número considerable de ECV se debe a una adherencia deficiente a los tratamientos preventivos cardiovasculares. Por lo tanto, medir la adherencia para tomar acciones encaminadas a mejorarla cuando sea necesario podría maximizar la eficacia de las terapias en el ámbito clínico.

De los estudios publicados anteriormente realizados con técnicas de ML, los análisis se realizaron sin estratificar en función del sexo, y en la mayoría [18,103,104] no se identificó el sexo como una variable importante, con la excepción de un estudio [105] en el que el sexo fue el segundo factor que más contribuyó al riesgo global de ECV. En cada estudio se

tuvieron en cuenta diferentes variables, incluidos los resultados de laboratorio, las mediciones de la TA y los factores sociodemográficos, la importancia de cada una de ellas varió según los modelos y las variables consideradas, como en el presente estudio. Sólo en un artículo de los que se han podido revisar identificó el valor de la glucemia como la variable más importante [103]. Dos estudios identificaron la TAS como la segunda variable más importante [18,104].

Los análisis de machine learning incluyeron dos técnicas distintas y como parte del estudio también se comparó su rendimiento. En este caso las métricas para evaluar el rendimiento de los modelos fueron similares para los desarrollados aplicando el algoritmo RF y XG Boost. Algunos estudios previos que compararon diferentes ML [47,90] informaron de que el modelo RF proporcionaba los resultados más precisos. Sin embargo, en su comparación de los modelos RF y XG Boost, Dinh et al. [104] informaron de un AUC ligeramente superior (es decir, mayor precisión) para XG Boost.

Diversos estudios han demostrado que los modelos construidos mediante técnicas de machine learning pueden superar ciertas limitaciones de los métodos tradicionales utilizados para predecir el riesgo CV, además de ofrecer un mayor poder predictivo[18,103,105]. Los resultados de los modelos descritos en este estudio podrían aplicarse en la práctica clínica para evaluar el riesgo individual de ECV en función de las características del paciente y la adherencia a la medicación, desempeñando así un papel importante en los procesos de cribado. Esto es particularmente importante dado que las intervenciones basadas en la atención primaria dirigidas a individuos considerados de alto riesgo, en función de su edad o factores de riesgo, parecen ser eficaces para reducir el riesgo de ECV [83]. Estos modelos pueden ayudar a orientar la intervención y a identificar las medidas más apropiadas a adoptar.

Además de las ventajas descritas, las técnicas de ML ofrecen una variedad de enfoques para procesar grandes cantidades de datos con el fin de predecir la incidencia de la ECV, lo que permite a investigadores y clínicos seleccionar el algoritmo que mejor se adapte a sus datos u objetivos.

Resultados análisis contrafactual en la cohorte CARhES

Finalmente, en esta tesis, se propuso cuantificar la contribución causal de las diferencias en diabetes, hipertensión, hipercolesterolemia y nivel socioeconómico en la incidencia de MACE por sexo.

La diabetes y el nivel socioeconómico fueron los dos factores que influyeron en las diferencias de incidencia de MACE entre hombres y mujeres. Los resultados de este estudio señalan la importancia de reducir los niveles de diabetes, centrando esfuerzos en los hombres dada su mayor prevalencia entre ellos, para disminuir las diferencias en la incidencia de MACE entre los dos sexos.

Se han realizado varios estudios en los que se analiza cómo influye la diabetes en la aparición de ECV en hombres y mujeres [23–26,99], y en ellos se ha observado que las mujeres con diabetes tienen un mayor riesgo de sufrir MACE que los hombres con diabetes. Sin embargo, el presente estudio se centró en el impacto de la diferente prevalencia entre sexos en la incidencia de MACE, por lo que los resultados son distintos y además no se pueden comparar.

La asociación del nivel socioeconómico con la incidencia de ECV se ha constatado en numerosos estudios [24,109,110] en los que un nivel socioeconómico bajo tiende a asociarse con peores resultados en ECV. Además, las personas con un nivel socioeconómico bajo suelen presentar más comorbilidades, pero suelen recibir menos intervenciones.

Los factores que constituyen el nivel socioeconómico de un individuo varían según el lugar y la cultura, pero los cuatro marcadores del nivel socioeconómico que han mostrado una asociación con la ECV son: el nivel de ingresos, el nivel educativo, la situación laboral y los factores ambientales. Todos estos factores están interrelacionados, por lo que deben tenerse en cuenta para estudiar el riesgo de ECV [110].

Se ha descrito que las personas con bajo nivel de ingresos y bajo nivel educativo tienen menos probabilidades de someterse a intervenciones, de recibir la medicación recomendada por las diferentes directrices y de ser derivadas a programas de intervención. La relación entre bajos ingresos y atención deficiente puede explicarse por el menor acceso a la atención en las personas con menor renta. Por otro lado, las

personas con bajo nivel educativo presentan más factores de riesgo conductuales que pueden explicarse por la asociación entre educación y alfabetización en salud. Este resultado identifica a la alfabetización en salud como un área en la que intervenir en las personas con bajo nivel educativo [110,111].

Hay factores ambientales que intervienen en las enfermedades cardiovasculares además del nivel socioeconómico, ya que las personas con un nivel socioeconómico más bajo son más propensas a vivir en barrios desfavorecidos. Vivir en zonas desfavorecidas está relacionado con una mayor incidencia de ECV, y esta relación está impulsada por condiciones físicas y sociales como la presencia de aceras, la disponibilidad y el coste de alimentos saludables, la seguridad, el apoyo social y la falta de cohesión comunitaria[110–112]. Otro factor recientemente relacionado con el desarrollo de enfermedad cardiovascular y el lugar donde viven las personas es la contaminación ambiental. Algunos estudios la sitúan como uno de los principales factores de riesgo para este tipo de enfermedades, incluso por encima de la diabetes [31,33].

Por último, el sexo/género también influye en la relación entre el nivel socioeconómico y la ECV, ya que las mujeres tienden a tener un nivel socioeconómico más bajo y es más probable que se vean afectadas por las disparidades en la distribución de la riqueza, los ingresos y el acceso a los recursos [110–113].

Destacar que, a pesar de que las investigaciones han demostrado que la asociación entre hipertensión grave y la incidencia de IAM es el doble en las mujeres que en los hombres [27] y que la hipercolesterolemia tiene una asociación más fuerte con la incidencia de IAM en los hombres que en las mujeres [20,22,28], ninguno de estos factores contribuyó a explicar la diferencia entre sexos en la incidencia de MACE en el estudio presentado. Debe tenerse en cuenta que la cohorte en la que se realizó el estudio está compuesta por personas con cualquier FRCV, y la prevalencia de hipertensión e hipercolesterolemia es similar en ambos sexos.

Aunque otros estudios han analizado las diferencias entre sexos teniendo en cuenta otros resultados de salud, como la salud autodeclarada o la discapacidad[114,115], aplicando métodos clásicos de descomposición, hasta donde sabemos, éste es el primer estudio

que utiliza un análisis contrafactual para analizar las posibles variables explicativas de la diferencia entre sexos en la incidencia de MACE.

Comparación resultados en las dos cohortes

Como ya se ha explicado, en la tesis presentada se utilizaron dos cohortes, con características distintas, para realizar análisis similares. A continuación, se presenta una comparación de los resultados más destacables en las dos cohortes.

La prevalencia de HTA y DM fue similar en la cohorte AWHS y CARhES, mientras que la de hipercolesterolemia fue mayor en la cohorte AWHS. Cabe destacar que en la cohorte AWHS el FRCV más común fue la HC y en CARhES fue la HTA. Sin embargo, las prevalencias de FRCV en ambas cohortes fueron considerablemente más altas que las prevalencias en la población aragonesa, y que la HC fue el FRCV más prevalente entre los aragoneses. Hay que tener en cuenta que los sujetos de la cohorte AWHS tuvieron una media de edad de 61,6 años con una DE de 4,82, mientras que la media de edad en la cohorte CARhES fue de 67,6 años con una DE de 10,5. Es decir, la media de edad de la cohorte CARhES fue mayor que la de AWHS pero también la dispersión de la edad fue mucho mayor entre los sujetos incluidos en CARhES.

Otro importante factor a tener en cuenta es que todos los sujetos de la cohorte CARhES tuvieron HTA, HC o DM, ya que fue uno de los criterios de entrada, mientras que esto no fue así en la cohorte AWHS. Destacar que los porcentajes de personas con 1, 2 o 3 de estos FRCV, de entre aquellos que tuvieron alguno, fue similar en ambas cohortes. También es importante decir que el nivel socioeconómico y condiciones laborales de los integrantes de la cohorte CARhES fueron mucho más heterogéneas que las de la cohorte AWHS. Además, los sujetos de la cohorte AWHS estaban sometidos al menos a un control médico anual por los servicios médicos de la empresa, mientras que eso no sucedió en la cohorte CARhES.

La incidencia de ECV difiere entre las dos cohortes, y hay que recalcar que los criterios para definir esta variable de resultado fueron distintos. Debido a que en la cohorte AWHS el número de sujetos era menor, se consideró cualquier evento cardiovascular, mientras

que en la cohorte CARhES, al tener muchos más sujetos, se pudieron considerar sólo los eventos cardiovasculares considerados mayores (MACE).

En ambas cohortes se realizaron análisis utilizando técnicas de machine learning incluyendo diferentes grupos de variables para predecir la aparición de ECV o MACE y describir la influencia de distintos FRCV y la adherencia a tratamientos preventivos en la incidencia de evento mediante el análisis de dichos modelos. En los análisis de la cohorte CARhES fue posible incorporar variables relacionadas con analíticas que no fue posible incorporar en la cohorte AWHS. Los resultados obtenidos en los análisis de las dos cohortes demostraron la gran influencia de la edad en el riesgo de tener un ECV. Aunque la adherencia al tratamiento fue medida para cada FRCV por separado en la cohorte CARhES y sintetizada en una sola variable en la cohorte AWHS, los resultados coincidieron en la importancia de considerar la adherencia al tratamiento en la evaluación del RCV.

Limitaciones y fortalezas del estudio

A continuación, se comentan las principales limitaciones y fortalezas del estudio que se presenta, relacionadas especialmente con los datos y poblaciones disponibles, así como con los métodos utilizados.

Limitaciones del estudio

En este apartado se comentarán las limitaciones encontradas en los análisis realizados en las dos cohortes, fundamentalmente relacionadas con el tipo de datos y las poblaciones utilizadas y con los métodos estadísticos aplicados.

Poblaciones de estudio y datos utilizados

Dentro de la cohorte AWHS, las principales limitaciones se encontraron en las características de la población a estudio. En los análisis enmarcados en dicha cohorte, la población de estudio fue exclusivamente masculina, debido al escaso número de mujeres en la cohorte que hizo que fueran eliminadas del estudio. Sin embargo, aunque

no es representativa de la población general, refleja bien a los trabajadores de este tipo de fábricas y en estos rangos de edad, lo cual abarca una parte importante de la población. Estas limitaciones fueron superadas en los análisis realizados con la cohorte CARhES, ya que esta cohorte incorpora tanto a hombres como a mujeres, y están representados todos los grupos de edad.

En cuanto a la cohorte CARhES hay que tener en cuenta que se trata de personas con algún FRCV, por lo que son personas con alto riesgo de sufrir un ECV.

Otras limitaciones están relacionadas con el tipo de datos disponibles y con su recolección. Aunque en los análisis realizados dentro de la cohorte AWHS la mayoría de datos utilizados venían de las bases de la propia cohorte y fueron recogidos por personas entrenadas, hubo algunos datos que dejaron de estar disponibles en algunos años del periodo de estudio, como por ejemplo el hábito tabáquico. Esta variable estuvo recogida para todos los sujetos en los primeros años del estudio pero no en los siguientes, lo que constituye una limitación dado que el tabaquismo es uno de los factores de riesgo modificables más importantes para la ECV. Además, se utilizaron datos de vida real provenientes de registros generados durante la práctica clínica diaria, lo que hizo que algunas variables no estuvieran completamente cumplimentadas, impidiendo su uso adecuado.

Los análisis realizados utilizando datos de la cohorte CARhES también tienen algunas limitaciones. Hay que tener en cuenta que todos los datos de esta cohorte son datos de vida real, recogidos de los sistemas de información utilizados en la práctica clínica diaria y provienen de bases de datos administrativas. Esto hace que algunas variables no estén disponibles o que su calidad sea insuficiente para ser incluidas. Algunos ejemplos son los datos sobre tabaquismo y actividad física, que se registraron en muy pocos sujetos y que, tras un control de calidad, no se consideraron fiables. Así pues, no todas las variables deseables pudieron ser incluidas en el estudio.

Por otra parte, la definición de ECV que se utilizó en los análisis de la cohorte AWHS fue amplia, lo que se traduce en que algunos de los ECV incluidos en el estudio (por ejemplo, las arritmias) pueden no estar claramente relacionados con los FRCV considerados. En la cohorte CARhES la variable resultado fue la incidencia de MACE, que es criterio de

ECV más restrictivo en el que los diagnósticos incluidos fueron sólo eventos cardiovasculares mayores relacionados con los FRCV estudiados.

Finalmente, se puede considerar que los periodos de seguimiento considerados en los análisis dentro de la cohorte CARhES fueron cortos ya que para realizar el análisis descriptivo en función de la incidencia de MACE y los análisis de machine learning (Objetivos 2.1. y 2.3.) fue entre enero de 2019 y diciembre de 2020. Sin embargo, en los análisis realizados por sexos (Objetivos 2.2. y 2.4), el periodo de estudio fue entre enero de 2018 y diciembre 2021. Sin embargo, consideramos que el tamaño de la población de estudio y la cantidad de datos tratados fue suficiente para responder a la pregunta de investigación.

Métodos estadísticos aplicados

Dado que se está trabajando con datos sanitarios y los casos de enfermedad siempre son más bajos que los casos de sujetos sanos, en esta tesis se trabajó con un desequilibrio de clases o datos. Este tipo de datos tienen ciertas limitaciones en la aplicación de métodos de machine learning, incluidos RF, XG Boost y NB, teniendo que aplicar distintas técnicas para abordar estos problemas. En el caso de los análisis de machine learning realizados en la cohorte AWHS (objetivo 1.3.), el problema se abordó cambiando el umbral en el proceso de validación. Al aplicar estas técnicas en la cohorte CARhES (objetivo 2.3), durante el preprocesado se aplicó la técnica de sobremuestreo ROSE (Random Over-Sampling Examples) para submuestrear la clase mayoritaria.

Debido a los datos desbalanceados, en los análisis de la cohorte AWHS, los valores de AUC-PR, parámetro para analizar la capacidad predictiva de la prueba, fueron bajos. Además, el valor VPN fue bajo para todos los modelos, lo que podría significar que se ha omitido un determinado factor en los modelos. No obstante, los análisis revelaron un buen rendimiento para todos los modelos, en particular RF y XGBoost cuando se incluyó la exposición al tratamiento como variable predictiva.

En los análisis realizados en la cohorte AWHS sobre trayectorias de la evolución de FRCV (objetivo 1.4), no todos los participantes acudieron a cada una de las 3 pruebas

médicas de las que se extrajeron los datos del estudio. Además, el algoritmo utilizado no tolera datos faltantes, por lo que se realizó una imputación de datos en algunas variables y se tuvieron que eliminar algunos casos en los casos en que la imputación no fue posible.

Fortalezas del estudio

Finalmente, se considera que el presente estudio presenta algunas fortalezas que se comentan a continuación.

Poblaciones de estudio y datos utilizados

En cuanto a las características de las cohortes utilizadas, aunque la validez externa de la cohorte AWHS puede ser limitada, tiene una alta validez interna debido al control al que todos los sujetos incluidos son sometidos por los servicios médicos de la fábrica y al riguroso proceso realizado para la obtención de datos.

La fortaleza de los datos de la cohorte CARhES radica en la gran cantidad de sujetos incluidos, y a la gran cantidad de datos a los que se tiene acceso. Esta cohorte destaca por incluir datos extraídos de diferentes niveles asistenciales de todos los individuos residentes en Aragón, mayores de 16 años, con cualquier FRCV.

Además, la información es obtenida de múltiples registros, lo que permite la evaluación de las variables de interés en un contexto del mundo real. La cohorte CARhES destaca por incluir datos extraídos de diferentes niveles asistenciales de todos los individuos residentes en Aragón, mayores de 16 años, con cualquier FRCV. Se considera que la calidad de la información utilizada es elevada, ya que el uso de diferentes fuentes ha permitido comprobar la coherencia de los datos utilizados.

Métodos estadísticos aplicados

En cuanto a los métodos utilizados, una parte de esta tesis muestra los resultados de aplicar distintas técnicas de machine learning para predecir la aparición de ECV o MACE y describir la influencia de distintos FRCV y la adherencia a tratamientos preventivos en

la incidencia de evento mediante el análisis de dichos modelos en dos cohortes distintas. Estas técnicas son capaces de integrar todos los datos disponibles y ofrecen varias ventajas sobre las calculadoras de riesgo anteriores, como que son capaces de manejar una gran cantidad de datos, y la comparación de los resultados entre ellas para determinar cuál es más precisa.

En ambos análisis realizados con machine learning se introdujeron variables relacionadas con la adherencia a tratamientos farmacológicos para el control de FRCV tradicionales. En los estudios realizados dentro del AWHS (objetivo 1.3.) la inclusión de la adherencia a diferentes tratamientos integrada en una única variable y en los realizados en la cohorte CARhES considerando por separando la adherencia a antidiabéticos, antihipertensvos e hipolipemiantes, y consideramos que este enfoque es esencial para determinar el verdadero riesgo de experimentar una ECV dada la influencia de la adherencia en el resultado terapéutico. Hasta donde sabemos, pocos estudios de machine learning han examinado el poder predictivo de la adherencia al tratamiento, y los que lo han hecho suelen evaluar la adherencia preguntando a los pacientes si están tomando alguna medicación, sin tener en cuenta si esta medicación está prescrita por un médico o si el paciente recoge realmente su medicación en una farmacia. En el presente estudio se midió la adherencia tomando datos de dispensación de farmacia y calculando el PDC.

Por último, los dos análisis con técnicas de machine learning tuvieron en cuenta múltiples algoritmos y diferentes combinaciones de variables predictivas, lo que nos permitió identificar el modelo que obtuvo mejores resultados en esta población de estudio concreta y evaluar la influencia de diferentes variables en la aparición de la ECV.

En cuanto a los análisis realizados para definir trayectorias en función de la evolución de FRCV (objetivo 1.4). La metodología utilizada se basa en un algoritmo de agrupación de k-means que es capaz considerar simultáneamente diferentes variables y puntos temporales (es decir, trayectorias de múltiples variables). Este estudio presenta agrupaciones en función de la evolución de distintos FRCV al mismo tiempo, a diferencia de estudios anteriores que pretendían crear agrupaciones en función de cada FRCV por

separado. Además, los datos analizados se extrajeron de múltiples fuentes e incluían múltiples variables bien definidas.

Finalmente, para analizar el impacto de distintas variables en la incidencia de MACE entre hombres y mujeres, se utilizó una nueva metodología que permite la descomposición causal de la razón de riesgo ajustada por edad de los hombres respecto a las mujeres en la incidencia de MACE, yendo más allá de los métodos de descomposición anteriores. Además, hasta donde conocemos, es la primera vez que este método se aplica a datos empíricos y utilizando varias variables explicativas.

6. CONCLUSIONES

6. CONCLUSIONES

1. En el estudio descriptivo de la cohorte de trabajadores AWHS se identificó la hipercolesterolemia como el factor de riesgo más prevalente, seguido de la hipertensión y destacando que la mitad de la población presentaba sobrepeso. La incidencia de evento cardiovascular fue del 7,9%, siendo los sujetos que sufrieron un evento mayores y con prevalencias más altas de los factores de riesgo cardiovascular estudiados que los que no lo padecieron.

2. En términos de cambios a lo largo del tiempo, en los trabajadores de la cohorte, se observó una disminución en los valores medios de colesterol total y glucemia, así como en el porcentaje de fumadores. Sin embargo, el porcentaje de personas con obesidad aumentó durante el seguimiento. El análisis por cuartiles reveló patrones estables, con la mayoría de las personas manteniendo su estado inicial en cuanto a obesidad y niveles elevados de glucemia.

3. En la evaluación de los modelos predictivos desarrollados en la población de hombres del AWHS, aplicando técnicas de machine learning, destacó que su capacidad predictiva mejoraba al añadir la variable exposición al tratamiento. La edad y la exposición al tratamiento resultaron ser las variables con mayor capacidad predictiva en la mayoría de los modelos, siendo el algoritmo Random el más efectivo, seguido por XG Boost.

4. En relación con la agregación de la población del AWHS en clusters según la evolución de factores de riesgo cardiovascular, se identificaron diferencias notables entre los grupos en términos de edad, índice de masa corporal, perímetro de cintura, glucemia y SCORE. Los sujetos que se mantuvieron con peores niveles en el IMC y de colesterol HDL fueron mayores en el segundo cluster que en el primero.

5. En el estudio realizado sobre población aragonesa, el FRCV más frecuente fue la hipercolesterolemia. Mientras que, cuando se analizó la prevalencia de FRCV para los sujetos incluidos en la cohorte CARhES, todos ellos presentando algún FRCV, se

encontró que el más prevalente fue la hipertensión. La incidencia de MACE fue más alta en hombres, y aquellas personas que experimentaron MACE fueron más mayores y con mayor número de factores de riesgo. Igualmente, la adherencia a fármacos antihipertensivos y antidiabéticos fue mayor en quienes sufrieron MACE.

6. En cuanto a las diferencias entre sexos en la población aragonesa con algún FRCV, se observaron disparidades en la edad, prevalencia de DM y nivel socioeconómico, con implicaciones específicas para el riesgo de MACE, ya que la incidencia de MACE fue más alta en aquellos sujetos mayores, que sufrían diabetes y con peor nivel socioeconómico. El análisis contrafactual reveló que la diabetes y el nivel socioeconómico son las variables más explicativas de las diferencias en la incidencia de MACE entre hombres y mujeres.

7. La evaluación de modelos predictivos desarrollados en la cohorte CARhES, mostró que la edad y la adherencia a antidiabéticos fueron los factores más influyentes a la hora de sufrir un MACE. Puesto que la edad no es modficable se deben desarrollar estrategias para mejorar el control de los FRCV, especialmente la diabetes.

8. La aplicación de las técnicas de machine learning resulta de utilidad en la predicción de riesgo CV identificando pacientes en los que el seguimiento y control debería ser más exhaustivo para reducir la frecuencia de eventos cardiovasculares.

CONCLUSIONS

1. In the descriptive study of the AWHS worker cohort, hypercholesterolemia was identified as the most prevalent risk factor, followed by hypertension and with half of the population exhibiting overweight. The incidence of cardiovascular events was 7.9%, and individuals who experienced an event had higher prevalences of cardiovascular risk factors than those who did not.

2. Over time, there was a decrease in mean values of total cholesterol and blood glucose, as well as a reduction in the percentage of smokers among cohort workers. However, the percentage of individuals with obesity increased during the follow-up. Quartile analysis revealed stable patterns, with most people maintaining their initial status regarding obesity and elevated blood glucose levels.

3. Evaluation of predictive models developed in the male population of AWHS, using machine learning techniques, highlighted improved predictive capacity with the addition of the treatment exposure variable. Age and treatment exposure proved to be the variables with the highest predictive capacity in most models, with the Random algorithm being the most effective, followed by XG Boost.

4. Concerning the aggregation of the AWHS population into clusters based on the evolution of cardiovascular risk factors, notable differences were identified among groups in terms of age, body mass index, waist circumference, blood glucose, and SCORE. Subjects who stayed with worse levels of IMC and HDL- cholesterol were greater in the second cluster than in the first.

5. In the study conducted on the Aragonese population, hypercholesterolemia was the most frequent cardiovascular risk factor. However, when analyzing the prevalence of cardiovascular risk factors in subjects included in the CARhES cohort, all presenting some risk factor, hypertension was found to be the most prevalent. MACE incidence was higher in men, and people who experienced MACE were older and had a higher number of risk factors. However, adherence to antihypertensive and antidiabetic medications was higher in those who experienced MACE.

6. Regarding gender differences in the Aragonese population with some cardiovascular risk factor, disparities were observed in age, prevalence of diabetes mellitus, and socioeconomic status, with specific implications for MACE risk. MACE incidence was higher in older individuals with diabetes and poorer socioeconomic status. Counterfactual analysis revealed that diabetes and socioeconomic status are the most explanatory variables for differences in MACE incidence between men and women.

7. The evaluation of predictive models developed in the CARhES cohort showed that age and adherence to antidiabetic medications were the most influential factors in experiencing MACE. Since age is not modifiable, strategies should be developed to improve the control of cardiovascular risk factors, especially diabetes.

8. The application of machine learning techniques is useful in predicting cardiovascular risk, identifying patients requiring more thorough monitoring and control to reduce the frequency of cardiovascular events.

7. BIBLIOGRAFÍA

7. BIBLIOGRAFÍA

1. Organización Mundial de la Salud. Enfermedad cardiovascular (ECV) [Internet]. 2017 [citado 3 de Julio de 2023]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. Timmis A, Vardas P, Townsend N, et al. European Society of Cardiology: cardiovascular disease statistics 2021. *Eur Heart J.* 2022; 43(8): 716–799. DOI: 10.1093/EURHEARTJ/EHAB892.

3. Roth GA, Mensah GA, Johnson CO, 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. DOI: 10.1016/J.JACC.2020.11.010.

4. Badimon L. Fisiopatología de la pared arterial y papel del colesterol en el origen y progresión de la placa de ateroma. *Clin Invest Arterioscl.* 2017;29(Supl 1):4-8.

5. Eurostat. Causes of death statistics. Marzo de 2023 [citado Julio de 2023]. Disponible en: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics#Major_causes_of_death_in_the_EU_in_2020

6. Instituto Nacional de Estadística. Tablas de mortalidad por año, sexo, edad y funciones. 2021 [citado Julio de 2023]. Disponible en: https://www.ine.es/jaxiT3/Tabla.htm?t=27153&L=0

7. Instituto Aragonés de estadística. Estadísticas de Defunciones según la causa de muerte. Instituto Aragonés de Estadística; Junio 2023.

8. Ministerio de Sanidad. *Estrategia en Salud Cardiovascular del Sistema Nacional de Salud (ESCAV)*. 2022.

9. Instituto Nacional de Estadística. Estancias causadas según el sexo, el diagnóstico principal, la provincia, Comunidad y Ciudad autónoma de hospitalización. 2021

[citado Julio 2023]. Disponible en: https://www.ine.es/jaxi/Tabla.htm?tpx=58456&L=0

10. Instituto Nacional de Estadística. Altas hospitalarias según el sexo, el diagnóstico principal, la provincia, Comunidad y Ciudad autónoma de residencia. INE; 2021 [citado Julio 2023]. Disponible en: https://www.ine.es/jaxi/Tabla.htm?tpx=58451&L=0

11. Departamento de Sanidad. Gobierno de Aragón. Informe prevalencia cardiopatía isquémica. Aragón 2022 [Internet]. Gobierno de Aragón; 2022. Disponible en: https://www.aragon.es/documents/20127/1650151/2022+MORBILIDAD+CIsqu%C3%A9mica.pdf/554dbe7f-e25d-3140-cc70-95389404e9e9?t=1683288980164

12. Departamento de Sanidad. Gobierno de Aragón. Informe prevalencia accidentes cerebrovasculares. Aragón 2022 [Internet]. Gobierno de Aragón; 2022. Disponible en: https://www.aragon.es/documents/20127/1650151/2022+MORBILIDAD+AVC.pdf/344436cb-b598-4223-0c7a-f44a487c476b?t=1683288979035

13. Departamento de Sanidad. Gobierno de Aragón. Mortalidad en Aragón 2021 [Internet]. Servicio de Vigilancia en Salud Pública; 2023 [citado Julio 2023]. Disponible en: https://www.aragon.es/-/registro-de-mortalidad

14. Instituto Aragonés de Estadística. Estadistica morbilidad hospitalaria 2021 [Internet]. INE; 2023. Disponible en: https://www.ine.es/prensa/emh_2021.pdf

15. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2021; 42(34): 3227–3337. DOI: 10.1093/EURHEARTJ/EHAB484.

16. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2016; 37: 2315–2381. DOI: 10.1093/eurheartj/ehw106.

17. Ministerio de Sanidad y Consumo. Estrategia en cardiopatía isquémica del sistema nacional de salud. 2006.

18. Sajeev S, Champion S, Beleigoli A, et al. Predicting Australian Adults at High Risk of Cardiovascular Disease Mortality Using Standard Risk Factors and Machine Learning. *Int J Environ Res Public Health.* 2021; 18(6): 1–14. DOI: 10.3390/IJERPH18063187.

19. Huang H-J, Lee C-W, Li T-H, et al. Different Patterns in Ranking of Risk Factors for the Onset Age of Acute Myocardial Infarction between Urban and Rural Areas in Eastern Taiwan. *Internat J Environ Res and Public Health.* 2021; 18: 5558. DOI: 10.3390/ijerph18115558.

20. Vallabhajosyula S, Verghese D, Desai VK, et al. Sex differences in acute cardiovascular care: a review and needs assessment. *Cardiovasc Res.* 2022; 118(3): 667–685. DOI: 10.1093/CVR/CVAB063.

21. Woodward M. Cardiovascular Disease and the Female Disadvantage. *Internat J Environ Res and Public Health.* 2019; 16(7): 1165. DOI: 10.3390/IJERPH16071165.

22. Humphries KH, Izadnegahdar M, Sedlak T, et al. Sex differences in cardiovascular disease – Impact on care and outcomes. *Front Neuroendocrinol.* 2017; 46: 46–70. DOI: 10.1016/J.YFRNE.2017.04.001.

23. Huebschmann AG, Huxley RR, Kohrt WM, et al. Sex differences in the burden of type 2 diabetes and cardiovascular risk across the life course. *Diabetologia. 2019 62:10* 2019; 62(10): 1761–1772. DOI: 10.1007/S00125-019-4939-5.

24. Pourfarzi F, Moghadam TZ, Zandian H. Decomposition of Socioeconomic Inequality in Cardiovascular Disease Prevalence in the Adult Population: A Cohort-based Cross-sectional Study in Northwest Iran. *J Prev Med Public Health.* 2022; 55(3): 297–306. DOI: 10.3961/JPMPH.22.051.

25. Angoulvant D, Ducluzeau PH, Renoult-Pierre P, et al. Impact of gender on relative rates of cardiovascular events in patients with diabetes. *Diabetes Metab.* 2021; 47(5): 101226. DOI: 10.1016/J.DIABET.2021.101226.

26. Al-Salameh A, El bouzegaoui N, Saraval-Gross M. Diabetes and cardiovascular risk according to sex: An overview of epidemiological data from the early

Framingham reports to the cardiovascular outcomes trials. *Ann Endocrinol (Paris)*. 2023; 84(1): 57–68. DOI: 10.1016/J.ANDO.2022.09.023.

27. Madsen TE, Howard G, Kleindorfer DO, et al. Sex Differences in Hypertension and Stroke Risk in the REGARDS Study: A Longitudinal Cohort Study. *Hypertension.* 2019; 74(4): 749–755. DOI: 10.1161/HYPERTENSIONAHA.119.12729.

28. Lu Y, Li SX, Liu Y, et al. Sex-Specific Risk Factors Associated With First Acute Myocardial Infarction in Young Adults. *JAMA Netw Open.* 2022; 5(5): e229953–e229953. DOI: 10.1001/JAMANETWORKOPEN.2022.9953.

29. Powell-Wiley TM, Baumer Y, Baah FO, et al. Social Determinants of Cardiovascular Disease. *Circ Res.* 2022; 130(5): 782–799. DOI: 10.1161/CIRCRESAHA.121.319811.

30. Graham H. Social Determinants and Their Unequal Distribution: Clarifying Policy Understandings. *Milbank Q.* 2004; 82(1): 101. DOI: 10.1111/J.0887-378X.2004.00303.X.

31. Pilar Mazón Ramos. Air pollution: a new risk factor for cardiovascular disease. *e-Journal of Cardiology Practice*; 22(20). Disponible en: https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-22/air-pollution-a-new-risk-factor-for-cardiovascular-disease

32. Ballester Díez F, Zafra MS, Alonso Fustel M, et al. The EMECAM project: the Spanish Multicenter Study on the Relationship between Air Pollution and Mortality. The background, participants, objectives and methodology. *Rev Esp Salud Publica.* 1999; 73: 165–175.

33. Brauer M, Casadei B, Harrington RA, et al. Taking a stand against air pollution – the impact on cardiovascular disease: A Joint Opinion from the World Heart Federation, American College of Cardiology, American Heart Association, and the European Society of Cardiology. *Eur Heart J.* 2021; 42(15): 1460–1463. DOI: 10.1093/EURHEARTJ/EHAA1025.

34. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation.* 2019; 140(11): e596–e646. DOI: 10.1161/CIR.0000000000000678/FORMAT/EPUB.

35. Vrijens B, De Geest S, Hughes DA, et al. A new taxonomy for describing and defining adherence to medications. *Br J of Clin Pharmacol*. 2012; 73(5): 691–705. DOI: 10.1111/J.1365-2125.2012.04167.X.

36. Malo S, Kardas P, Menditto E. Some reflections concerning the assessment of patient adherence and persistence to medication. *Curr Med Res Opin.* 2019; 35(1): 3–4. DOI: 10.1080/03007995.2018.1528216.

37. Boston Medical Center. Estudio de Framingham [Internet]. BMC; [citado Julio de 2023]. Disponible en: https://www.bmc.org/es/stroke-and-cerebrovascular-center/research/framingham-study

38. Lind L. Population-based cardiovascular cohort studies in Uppsala. *Ups J Med Sci* 2019; 124(1): 16. DOI: 10.1080/03009734.2018.1515282.

39. REGICOR. Misión y vision [Internet]. REGICOR; 2019 [citado Julio de 2023]. Disponible en: https://regicor.cat/es/presentacion/mision-y-vision/

40. THE HEART HEALTHY HOODS PROJECT: A multifaceted approach to cardiovascular diseases in European Cities [Internet]. HHH Project; 2016 [citado Julio de 2023]. Disponible en: https://www.hhhproject.es/

41. Predimed Plus. Estudio de la dieta mediterránea [Internet]. CIBER; 2016 [citado Julio 2023]. Disponible en: https://www.predimedplus.com/

42. Predimed. Prevención con dieta mediterránea [Internet]. [citado en Julio 2023] Disponible en: http://www.predimed.es/

43. Casasnovas JA, Alcaide V, Civeira F, et al. Aragon workers' health study - design and cohort description. *BMC Cardiovasc Disord.* 2012;12(45). DOI: 10.1186/1471-2261-12-45.

44. Gil-Guillén VF, Orozco-Beltrán D, Maiques-Galán A, et al. Agreement between REGICOR and SCORE scales in identifying high cardiovascular risk in the Spanish population. *Rev Esp Cardiol.* 2007; 60(10): 1042–1050. DOI: 10.1157/13111236.

45. Alaa AM, Bolton T, Angelantonio E di, et al. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One.* 2019;14(5). DOI: 10.1371/journal.pone.0213653.

46. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circ Res.* 2017; 121(9): 1092–1101. DOI: 10.1161/CIRCRESAHA.117.311312.

47. Jamthikar AD, Gupta D, Mantella LE, et al. Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: a 500 participants study. *Int J Cardiovasc Imaging.* 2021; 37(4): 1171–1187. DOI: 10.1007/S10554-020-02099-7/TABLES/5.

48. Vrbaški D, Vrbaški M, Kupusinac A, et al. Methods for algorithmic diagnosis of metabolic syndrome. *Artif Intell Med.* 2019; 101: 101708. DOI: 10.1016/J.ARTMED.2019.101708.

49. Amaratunga D, Cabrera J, Sargsyan D, et al. Uses and opportunities for machine learning in hypertension research. *Int J Cardiol Hypertens.* 2020; 5: 100027. DOI: 10.1016/J.IJCHY.2020.100027.

50. Ghosh P, Azam S, Karim A, et al. A comparative study of different machine learning tools in detecting diabetes. *Procedia Comp Sci.* 2021; 192: 467–477. DOI: 10.1016/j.procs.2021.08.048.

51. Krittanawong C, Virk HUH, Bangalore S, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep.* 2020; 10(1): 16057. DOI: 10.1038/s41598-020-72685-1.

52. Yu Z, Wang K, Wan Z, et al. Popular deep learning algorithms for disease prediction: a review. *Cluster Comput*. DOI: 10.1007/s10586-022-03707-y.

53. Rhys HI. *Machine Learning with R, the tidyverse, and mlr*. 2020.

54. Chaure-Pardos A, Malo S, Rabanaque MJ, et al. Factors Associated with the Prescribing of High-Intensity Statins. *JClin Med.* 2020; 9(12): 3850. DOI: 10.3390/jcm9123850.

55. Malo S, Aguilar-Palacio I, Feja C, et al. Different approaches to the assessment of adherence and persistence with cardiovascular-disease preventive medications. *Curr Med Res and Opin*. 2017; 33(7): 1329–1336. DOI: 10.1080/03007995.2017.1321534.

56. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J.* 2003; 24(11): 987–1003. DOI: 10.1016/S0195-668X(03)00114-3.

57. Genolini C, Falissard B, Fang D, et al. Package 'longitudinalData' [Internet]. CRAN; 2016 [actualizado Febrero 2023; citado Octubre 2020] Disponible en: https://cran.r-project.org/web/packages/longitudinalData/longitudinalData.pdf

58. Dinga R, Penninx BWJH, Veltman DJ, et al. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv*. 2019; 743138.

59. Everitt B, Landau L, Leese M. *Cluster analysis*. 4th ed. London: Hodder Edwar Arnold, 2001.

60. Genolini C, Alacoque X, Sentenac M, et al. Kml and kml3d: R packages to cluster longitudinal data. *J Stat Softw.* 2015; 65(4): 1–34. DOI: 10.18637/jss.v065.i04.

61. Genolini C, Falissard B, Genolini C, et al. KmL: k-means for longitudinal data. *Comput Stat* 2010; 25: 317–328. DOI: 10.1007/s00180-009-0178-4.

62. ORDEN SAN/1355/2018, de 1 de agosto, por la que se crea la plataforma de información BIGAN como elemento del Sistema de Información de Salud de Aragón. Boletín Oficial de Aragón, número 162, (22 de agosto de 2018)

63. Instituto Aragonés de Ciencias de la Salud. Actividad de Tratamiento BIGAN [Internet]. IACS; 2017 [citado Febrero de 2022]. Disponible en: https://www.iacs.es/actividad-tratamiento-bigan/

64. Maniruzzaman M, Rahman J, Ahammed B, et al. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf Sci Syst.* 2020; 8: 7. DOI: 10.1007/s13755-019-0095-z.

65. Ali MM, Paul BK, Ahmed K, et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Comput Biol Med.* 2021; 136: 104672. DOI: 10.1016/J.COMPBIOMED.2021.104672.

66. Haibo He, Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. 1st Editio. Wiley-IEEE Press, 2013.

67. Cordón I, García S, Fernández A, et al. Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowl Based Syst.* 2018; 161: 329–341. DOI: 10.1016/j.knosys.2018.07.035.

68. tidymodels. Subsampling for class imbalances. [citado Febrero de 2022]. Disponible en: https://www.tidymodels.org/learn/models/sub-sampling/

69. Bekkar M, Djemaa HK, Alitouche TA. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *J Inf Eng and Appl.* 2013; 3(10): 27–38.

70. Sudharsanan N, Bijlsma MJ. Educational note: causal decomposition of population health differences using Monte Carlo integration and the g-formula. *Int J Epidemiol*. 2022; 50(6): 2098–2107. DOI: 10.1093/IJE/DYAB090.

71. Sudharsanan N, Bijlsma MJ, De BM, et al. A Generalized Counterfactual Approach to Decomposing Differences Between Populations. 2019 [citado Marzo 2023]. Disponible en: www.demogr.mpg.de

72. Maarten M, Bijlsma J. Package 'cfdecomp'. 2022 [citado Marzo 2023]. Disponible en: https://cran.r-project.org/web/packages/cfdecomp/cfdecomp.pdf

73. Zhao B, He X, Wu J, et al. Adherence to statins and its impact on clinical outcomes: a retrospective population-based study in China. *BMC Cardiovasc Disord.* 2020; 20: 282. DOI: 10.1186/S12872-020-01566-2.

74. Lee H, Yano Y, Cho SMJ, et al. Adherence to Antihypertensive Medication and Incident Cardiovascular Events in Young Adults With Hypertension. *Hypertension.* 2021; 77(4): 1341–1349. DOI: 10.1161/HYPERTENSIONAHA.120.16784.

75. Aguilar-Palacio I, Rabanaque MJ, Maldonado L, et al. New Male Users of Lipid-Lowering Drugs for Primary Prevention of Cardiovascular Disease: The Impact of Treatment Persistence on Morbimortality. A Longitudinal Study. *Internat J Environ Res and Public Health.* 2020; 17(20): 7653. DOI: 10.3390/ijerph17207653.

76. Yang Q, Chang A, Ritchey MD, et al. Antihypertensive Medication Adherence and Risk of Cardiovascular Disease Among Older Adults: A Population-Based Cohort Study. *J Am Heart Assoc.* 2017;6(6). DOI:10.1161/JAHA.117.006056

77. Chowdhury R, Khan H, Heydon E, et al. Adherence to cardiovascular therapy: a meta-analysis of prevalence and clinical consequences. *Eur Heart J.* 2013; 34 (38): 2940–2948. DOI: 10.1093/EURHEARTJ/EHT295.

78. Dhingra R, Vasan RS. Age as a Cardiovascular Risk Factor. *Med Clin North Am.* 2012; 96(1): 87. DOI: 10.1016/J.MCNA.2011.11.003.

79. Wahabi H, Esmaeil S, Zeidan R, et al. Effects of Age, Metabolic and Socioeconomic Factors on Cardiovascular Risk among Saudi Women: A Subgroup Analysis from the Heart Health Promotion Study. *Medicina.* 2023; 59(3): 623.DOI: 10.3390/MEDICINA59030623.

80. Redon J. Global Cardiovascular Risk Assessment: Strengths and Limitations. *High Blood Press and Cardiovasc Prev.* 2016; 23: 87–90. DOI: 10.1007/s40292-016-0139-2.

81. Álvarez-Fernández C, Romero-Saldaña M, Álvarez-López C, et al. Gender differences and health inequality: Evolution of cardiovascular risk in workers. *Arch Environ Occup Health.* 2021; 76: 406–413. DOI: 10.1080/19338244.2021.1891017.

82. He J, Zhu Z, Bundy JD, et al. Trends in Cardiovascular Risk Factors in US Adults by Race and Ethnicity and Socioeconomic Status, 1999-2018. *JAMA.* 2021; 326: 1286–1298. DOI: 10.1001/JAMA.2021.15187.

83. Eriksen CU, Rotar O, Toft U, et al. *What is the effectiveness of systematic population-level screening programmes for reducing the burden of cardiovascular diseases?* Copenhagen: WHO Regional Office for Europe; 2021.

84. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circ Res.* 2017; 121: 1092–1101. DOI: 10.1161/CIRCRESAHA.117.311312.

85. Oppenheimer GM. Framingham Heart Study: the first 20 years. *Prog Cardiovasc Dis.* 2010; 53: 55–61. DOI: 10.1016/J.PCAD.2010.03.003.

86. Yandrapalli S, Nabors C, Goyal A, et al. Modifiable Risk Factors in Young Adults With First Myocardial Infarction. *J Am Coll Cardiol.* 2019; 73: 573–584. DOI: 10.1016/J.JACC.2018.10.084.

87. Bazalar-Palacios J, Jaime Miranda J, Carrillo-Larco RM, et al. Aggregation and combination of cardiovascular risk factors and their association with 10-year all-cause mortality: the PERU MIGRANT Study. *BMC Cardiovasc Disord.* 2021; 21: 582. DOI: 10.1186/S12872-021-02405-8/TABLES/3.

88. Wang K, Tian J, Zheng C, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput Biol Med.* 2021; 137: 104813. DOI: 10.1016/J.COMPBIOMED.2021.104813.

89. Visseren FLJ, MacH F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J.* 2021; 42: 3227–3337. DOI: 10.1093/eurheartj/ehab484.

90. Khan B, Naseem R, Shah MA, Wakil K, Khan A, Uddin MI, Mahmoud M. Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine

Learning Techniques. *J Healthc Eng*. 2021; 2021:8899263. DOI: 10.1155/2021/8899263.

91. Lin H, Cui M, Spatz ES, et al. Heterogeneity in Trajectories of Systolic Blood Pressure among Young Adults in Qingdao Port Cardiovascular Health Study. *Glob Heart.* 2020; 15: 9–11. DOI: 10.5334/gh.764.

92. Tielemans SMAJ, Geleijnse JM, Laughlin GA, et al. Blood pressure trajectories in relation to cardiovascular mortality: The Rancho Bernardo Study. *J Hum Hypertens.* 2017; 31: 515–519. DOI: 10.1038/jhh.2017.20.

93. Allen NB, Siddique J, Wilkins JT, et al. Blood pressure trajectories in early adulthood and subclinical atherosclerosis in middle age. *JAMA.* 2014; 311: 490–497. DOI: 10.1001/jama.2013.285122.

94. Smitson CC, Scherzer R, Shlipak MG, et al. Association of blood pressure trajectory with mortality, incident cardiovascular disease, and heart failure in the cardiovascular health study. *Am J Hypertens*. 2017; 30: 587–593. DOI: 10.1093/ajh/hpx028.

95. Norby FL, Soliman EZ, Chen LY, et al. Trajectories of cardiovascular risk factors and incidence of atrial fibrillation over a 25-year follow-up. *Circulation*. 2016; 134: 599–610. DOI: 10.1161/CIRCULATIONAHA.115.020090.

96. Pebesma J, Martinez-Millana A, Sacchi L, et al. Clustering Cardiovascular Risk Trajectories of Patients with Type 2 Diabetes Using Process Mining. *Annu Int IEEE Eng Med Biol Soc.* 2019: 341–344. DOI: 10.1109/EMBC.2019.8856507.

97. Rospleszcz S, Lorbeer R, Storz C, et al. Association of longitudinal risk profile trajectory clusters with adipose tissue depots measured by magnetic resonance imaging. *Sci Rep.* 2019; 9: 1–12.DOI: 10.1038/s41598-019-53546-y.

98. Springer KW, Mager Stellman J, Jordan-Young RM. Beyond a catalogue of differences: A theoretical frame and good practice guidelines for researching sex/gender in human health. *Soc Sci Med.* 2012; 74: 1817–1824. DOI: 10.1016/j.socscimed.2011.05.033.

99. Gutiérrez AG, Poblador-Plou B, Prados-Torres A, et al. Sex Differences in Comorbidity, Therapy, and Health Services' Use of Heart Failure in Spain: Evidence from Real-World Data. *Int J Environ Res Public Health.* 2020; 17: 2136. DOI: 10.3390/ijerph17062136.

100. Liu W, Tang Q, Jin J, et al. Sex differences in cardiovascular risk factors for myocardial infarction. *Herz.* 2021; 46 (Suppl 1): 115–122. DOI: 0.1007/s00059-020-04911-5.

101. Santilli F, D'Ardes D, Guagnano MT, et al. Metabolic Syndrome: Sex-Related Cardiovascular Risk and Therapeutic Approach. *Curr Med Chem.* 2017;24(24):2602-2627. DOI: 10.2174/0929867324666170710121145.

102. Ministerio de Sanidad. Análisis con perspectiva de género de los registros sobre la enfermedad cardiovascular contenidos en la Base de Datos Clínicos de Atención Primaria. Series 4. [Internet]. 2022.

103. Huang W, Ying TW, Chin WLC, et al. Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. *Sci Rep.* 2022;12(1):1033. DOI: 10.1038/S41598-021-04649-Y.

104. Dinh A, Miertschin S, Young A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak.* 2019; 19: 1–15. DOI: 10.1186/s12911-019-0918-5.

105. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4): e0174944. DOI: 10.1371/journal.pone.0174944.

106. Liu M, Zheng G, Cao X, et al. Better medications adherence lowers cardiovascular events, stroke, and all-cause mortality risk: A dose-response meta-analysis. *J Cardiovasc Dev Dis.* 2021; 8: 146. DOI: 10.3390/jcdd8110146.

107. Chen C, Li X, Su Y, et al. Adherence with cardiovascular medications and the outcomes in patients with coronary arterial disease: "Real-world" evidence. *Clin Cardiol*. 2022; 45(12):1220-1228. DOI: 10.1002/CLC.23898.

108. Donneyong MM, Fischer MA, Langston MA, et al. Examining the Drivers of Racial/Ethnic Disparities in Non-Adherence to Antihypertensive Medications and Mortality Due to Heart Disease and Stroke: A County-Level Analysis. *Int J Environ Res Public Health. 2021*; 18(23):12702. DOI: 10.3390/IJERPH182312702.

109. Mosquera PA, San Sebastian M, Ivarsson A, et al. Decomposition of gendered income-related inequalities in multiple biological cardiovascular risk factors in a middle-aged population. *Int J Equity Health. 2018*; 17(1): 102. DOI: 10.1186/s12939-018-0804-2.

110. Schultz WM, Kelli HM, Lisko JC, et al. Socioeconomic status and cardiovascular outcomes: Challenges and interventions. *Circulation.* 2018; 137: 2166–2178. DOI: 10.1161/CIRCULATIONAHA.117.029652.

111. Clark AM, DesMeules M, Luo W, et al. Socioeconomic status and cardiovascular disease: risks and implications for care. *Nat Rev Cardiol*. 2009; 6: 712–722. DOI: 10.1038/nrcardio.2009.163.

112. Zhang YB o., Chen C, Pan XF, et al. Associations of healthy lifestyle and socioeconomic status with mortality and incident cardiovascular disease: two prospective cohort studies. *BMJ. 2021*; 373: n604. DOI: 10.1136/BMJ.N604.

113. Aarnio E, Martikainen J, Winn AN, et al. Socioeconomic Inequalities in Statin Adherence under Universal Coverage: Does Sex Matter? *Circ Cardiovasc Qual Outcomes.* 2016; 9: 704–713. DOI: 10.1161/CIRCOUTCOMES.116.002728.

114. Hosseinpoor AR, Williams JS, Jann B, et al. Social determinants of sex differences in disability among older adults: a multi-country decomposition analysis using the World Health Survey. *Int J Equity Health. 2012*; 11: 52. DOI: 10.1186/1475-9276-11-52.

115. Boerma T, Hosseinpoor AR, Verdes E, et al. A global assessment of the gender gap in self-reported health with survey data from 59 countries. *BMC Public Health. 2016*; 16: 675. DOI: 10.1186/S12889-016-3352-Y.

8. ANEXOS

ANEXO I

PLOS ONE

RESEARCH ARTICLE

Influence of cardiovascular risk factors and treatment exposure on cardiovascular event incidence: Assessment using machine learning algorithms

Sara Castel-Feced[1,2,3]*, **Sara Malo**[1,2,3], **Isabel Aguilar-Palacio**[1,2,3], **Cristina Feja-Solana**[2,3,4], **José Antonio Casasnovas**[5,6], **Lina Maldonado**[2,3,7‡], **María José Rabanaque-Hernández**[1,2,3‡]

1 Microbiology, Pediatrics, Radiology, and Public Health, University of Zaragoza, Zaragoza, Spain, **2** Fundación Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza, Spain, **3** GRISSA Research Group, Zaragoza, Spain, **4** Directorate of Public Health, Government of Aragon, Zaragoza, Spain, **5** Hospital Universitario Miguel Servet, Instituto de Investigación Sanitaria Aragón (IIS Aragón), CIBERCV, Zaragoza, Spain, **6** Department of Medicine, Psychiatry and Dermatology, University of Zaragoza, Zaragoza, Spain, **7** Department of Applied Economic, University of Zaragoza, Zaragoza, Spain

‡ LM and MJRH also contributed equally to this work and served as senior co-authors.
* scastelf@unizar.es

OPEN ACCESS

Citation: Castel-Feced S, Malo S, Aguilar-Palacio I, Feja-Solana C, Casasnovas JA, Maldonado L, et al. (2023) Influence of cardiovascular risk factors and treatment exposure on cardiovascular event incidence: Assessment using machine learning algorithms. PLoS ONE 18(11): e0293759. https://doi.org/10.1371/journal.pone.0293759

Editor: Chi-Shin Wu, NHRI: National Health Research Institutes, TAIWAN

Received: March 8, 2023

Accepted: October 19, 2023

Published: November 16, 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.0293759

Data Availability Statement: Data were provided by the Aragon Health Sciences Institute (IACS), Spain, so authors do not have permission to share

Abstract

Assessment of the influence of cardiovascular risk factors (CVRF) on cardiovascular event (CVE) using machine learning algorithms offers some advantages over preexisting scoring systems, and better enables personalized medicine approaches to cardiovascular prevention. Using data from four different sources, we evaluated the outcomes of three machine learning algorithms for CVE prediction using different combinations of predictive variables and analysed the influence of different CVRF-related variables on CVE prediction when included in these algorithms. A cohort study based on a male cohort of workers applying populational data was conducted. The population of the study consisted of 3746 males. For descriptive analyses, mean and standard deviation were used for quantitative variables, and percentages for categorical ones. Machine learning algorithms used were XGBoost, Random Forest and Naïve Bayes (NB). They were applied to two groups of variables: i) age, physical status, Hypercholesterolemia (HC), Hypertension, and Diabetes Mellitus (DM) and ii) these variables plus treatment exposure, based on the adherence to the treatment for DM, hypertension and HC. All methods point out to the age as the most influential variable in the incidence of a CVE. When considering treatment exposure, it was more influential than any other CVRF, which changed its influence depending on the model and algorithm applied. According to the performance of the algorithms, the most accurate was Random Forest when treatment exposure was considered (F1 score 0.84), followed by XGBoost. Adherence to treatment showed to be an important variable in the risk of having a CVE. These algorithms could be applied to create models for every population, and they can be used in primary care to manage interventions personalized for every subject.

the data. The permission obtained implies the exclusive use of the data by researchers who authored the present study. Thus, this information cannot be published or shared with other institutions. Data access requests should be addressed to the IACS through https://www.iacs.es/. Source code is openly available at https://github.com/saracf623/machinelearning.

Funding: This study was supported by Proyecto del Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (Ministerio de Ciencia e Innovación) and the European Fund for Regional Development (FEDER) (PI17/01704) and by the Grupo de Investigacion en Servicios Sanitarios de Aragón (GRISSA) [B09-23R] of the IIS Aragon, funded by the regional Government of Aragon, Spain. It was also partly supported to SCF by Gobierno de Aragón with a grant for postgraduate research contracts (IIU/796/2019). 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.

Introduction

Cardiovascular diseases (CVD) are the leading cause of morbidity and mortality worldwide and are responsible for 32% of all global deaths [1]. CVD prevention guidelines emphasize the importance of primary CVD prevention, applying lifestyle changes and medicating according to the individual's overall CV risk, which reflects the contributions of multiple CV risk factors [2]. Several scoring systems are used to predict CVD risk, with the ultimate goal of establishing preventive interventions, both pharmacological and non-pharmacological. These scoring systems, which include the Framingham Risk Score and the Systematic Coronary Risk Evaluation (SCORE), have been developed for specific populations, without considering whether subjects are being treated for any cardiovascular risk factor (CVRF), and also suffer from certain methodological limitations [3–5] due to correlation between variables, non-linearity of variables, and the possibility of over-fitting. On the other hand, the knowledge acquired from an ever-growing body of medical data generated in daily clinical practice is providing researchers with valuable insights into medical conditions [6, 7].

Machine learning techniques have been widely applied [6, 8] to probe these huge datasets and overcome some of the aforementioned limitations of scoring systems. Machine learning can be used to generate models that better predict risk, thereby increasing the efficiency, objectivity, and reliability of the diagnostic process [3, 4, 6, 9]. Specifically, these supervised learning techniques use existing data to train models by learning patterns that will be later applied to predict another variable. When applying these techniques to disease research, there are some problems which have to be considered during the analysis. These problems are related to the interpretability of the models and to data imbalance, quality and quantity. Poor interpretability is due to the fact that they work like black boxes, making it difficult to interpret their results. This problem has been overcome by the development of different methods, implemented in different R libraries, that try to set the importance of each feature in the prediction [10–12]. Imbalance data is because in health research always there are fewer people who are sick than those who are healthy, so usually data are imbalance. Finally, quality and quantity of medical data usually are low because of the sources of information available and because of accessibility and ethical issues [13].

There are different kinds of supervised machine learning. When the variable to be predicted is categorical (e.g. a cardiovascular event), classification machine learning techniques [3, 4] such as Naïve Bayes (NB) algorithms and ensemble methods are used [12, 14, 15]. Ensemble methods include bagging and boosting methods such are Random Forest (RF), XG Boost [12, 14, 15].

Machine learning techniques represent a promising approach for CVD risk prediction [3, 4, 12, 14–16] and the advantages they offer over existing scoring systems. In the present study, we analyzed the ability of three machine learning algorithms, NB, RF and XG Boost, to predict the appearance of cardiovascular events (CVE) and analysed the influence of different CVRF-related variables on CVE.

Methodology

Study design and participants

This longitudinal cohort study was conducted within the framework of the Aragon Workers' Health Study (AWHS), a prospective, longitudinal cohort study of workers at an automobile assembly plant located in Figueruelas (Zaragoza, Spain). Recruitment began in February 2009 and ended in December 2010. Since then, data have been collected from the participant's annual medical tests and their utilization of health services has been monitored. Further information on the AWHS can be found in Casasnovas et al. [17].

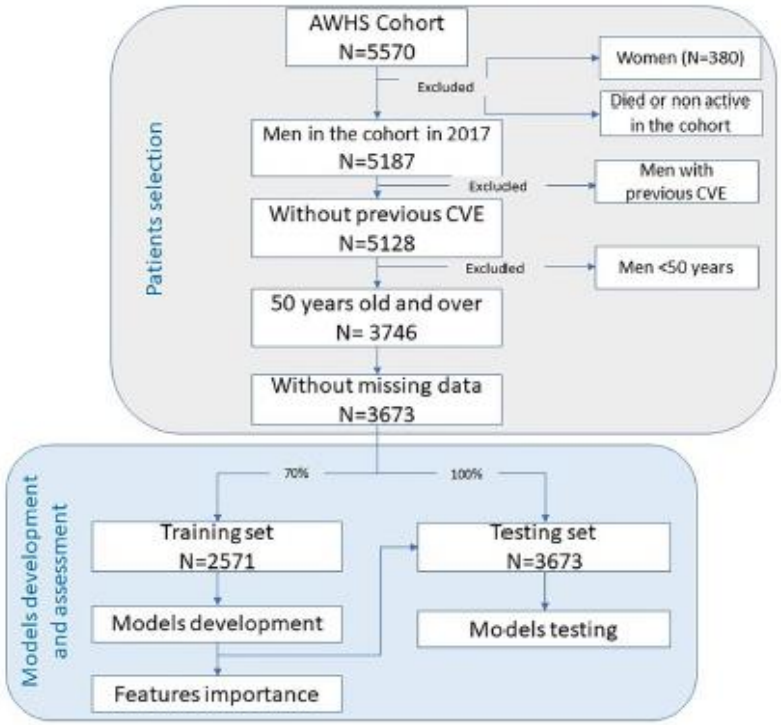


Fig 1. Flowchart depicting the study population and model development.

https://doi.org/10.1371/journal.pone.0293759.g001

All subjects included in our analysis were male (N = 3,746), aged ≥50 years, with no previous medical history of CVE. Females were excluded owing to the low number in the cohort (N = 380) and individuals aged <50 years were excluded owing to the low incidence of CVE in this age group. After selecting the study population, we identified individuals who experienced a CVE at any moment between inclusion in the cohort and December 31, 2019, and recorded the date and nature of each CVE. The selection of patients is explained in Fig 1.

Data source and variables

Data from several sources were used: the AWHS cohort; BIGAN; and the General Direction of Public Health. BIGAN [18] is a health data hub that gathers data from the Aragon Public Health Service. These data are available for research upon request. From the AWHS study we included data from workers' annual medical tests (including blood tests). From BIGAN we obtained data from (i) the Pharmaceutical Dispensation Database, which collects information on the dispensing date, Anatomical Therapeutic Chemical (ATC) code, number of defined daily doses (DDD), and the number of packages dispensed by pharmacies and funded by the Aragon Health Service; (ii) the MBDS (Minimum Basic Data Set), which registers diagnoses and dates of hospitalizations; and (iii) the Emergency database which records diagnoses and dates pertaining to emergency service utilization. Finally, information on the date and cause of death was obtained from the Aragon Mortality Registry via the General Direction of Public Health.

Table 1. Variables sources and missing data.

	AWHS	BIGAN			General Direction of Public Health	
Variables	Annual medical test	Pharmaceutical dispensation database	Minimum Basic Data Set	Emergency Database	Aragon Mortality Registry	NA
Hypertension	X	X				26 (0.5%)
Hypercholesterolemia	X	X				14 (0.3%)
Diabetes	X	X				27 (0.6%)
Physical status	X					57 (1.2%)
Age	X					0
Treatment exposure		X				0
CVE			X	X	X	0

https://doi.org/10.1371/journal.pone.0293759.t001

The source of the variables and number of missing data is summarised in Table 1.

Explanatory variables. *CVRF definition*. The following CVRFs were considered: hypertension, hypercholesterolemia (HC), diabetes mellitus (DM), and physical status, as calculated based on body mass index (BMI). Smoking status was not considered as corresponding data were not available for the period analysed. CVRFs were identified for the year preceding the first CVE in individuals with a CVE, and for 2019 for individuals with no CVE during the study period.

Medical and blood test findings and data from the Pharmaceutical Dispensation Database were examined to identify CVRFs. Subjects were classified as suffering from the CVRF if they were registered in at least one of those databases as such. CVRFs were identified from medical and blood test data applying the following cut-off points, as recommended by European CVD prevention guidelines [2]: overweight was defined as a BMI $\geq$25 and <30, and obesity as a BMI $\geq$30; hypertension as diastolic blood pressure $\geq$90 mmHg and/or systolic blood pressure $\geq$140 mmHg; HC as total cholesterol $\geq$200 mg/dl or LDL-cholesterol $\geq$115 mg/dl; and DM as fasting serum glucose $\geq$126 mg/d.

Based on data from the Pharmaceutical Dispensation Database, individuals were considered to have hypertension if they had filled at least one prescription corresponding to the following ATC codes: C02 (antihypertensives), C03 (diuretics), C07 (beta-blocking agents), C08 (calcium channel blockers), and C09 (agents acting on the renin–angiotensin system). Since diuretics and beta-blocking agents are also prescribed for other indications, dispensation of these drugs was only considered an indicator of hypertension if the individual filled at least three distinct dispensations within the same year [19]. Participants were considered to have HC if they filled at least one dispensation corresponding to ATC code C10 (lipid modifying agents) and DM if they filled least one dispensation corresponding to ATC code A10 (drugs used in diabetes).

Treatment exposure. Treatment exposure was determined by quantifying adherence to the treatment.

To standardise terminology related to adherence to pharmacological therapies, the European Ascertaining Barriers for Compliance (ABC) project proposed a Taxonomy of Adherence [20] consisting of three components: initiation, implementation, and discontinuation. Treatment *initiation* corresponds to intake of the first dose of a prescribed medication. The process continues with *implementation* of the dosing regimen, defined as the extent to which a patient's actual dosing corresponds to the prescribed dosing regimen, from treatment initiation until consumption of the last dose. *Discontinuation* indicates the end of therapy, when the next dose to be taken is omitted and no more doses are taken thereafter.

In the present study, our analysis focused on the implementation phase. Adherence to HC, hypertension, and DM treatments was determined separately for each participant and represented as the Proportion of Days Covered (PDC), calculated as a percentage. PDC is an index calculated as the number of days covered by the medicines dispensed by the pharmacy divided by the number of days that the subject should have had covered. In this study, the denominator for PDC was 365 days, except in cases in which subjects started treatment once the follow-up period had already started. In these cases, the denominator was the number of days from initiation of treatment to the end of the follow-up period. The number of days covered are calculated based on the DDD dispensed to each subject. However, a previous study of our group [21] showed that use of a surrogate value for the daily dose of each drug, calculated based on the usual dosage and form of presentation, provided more accurate results. Therefore, in the present study surrogate values for daily doses were used.

For each subject, the PDC obtained for the three CVRFs was summarized in a new variable: treatment exposure. This variable was classified into 3 possible categories: *fully exposed*, participants who filled prescriptions for the treatment of all identified CVRFs and had a PDC $\geq$80%; *non-exposed*, participants who filled prescriptions for none of the identified CVRFs or had a PDC <80% for all treatments taken; *partially exposed*, participants who did not fill prescriptions for at least one identified CVRF and had a PDC $\geq$80% for others or a PDC <80% for some treatment and $\geq$80% for others.

Evaluated outcome. The primary outcome in the current study was the incidence of a CVE during the study period. This CVE was identified based on data from the MBDS, Emergency database, and the Aragon Mortality Registry as follows:

i. For subjects with records in either the MBDS or the Emergency database, data from the former was selected.

ii. For subjects with records in both the MBDS and Emergency databases, checks were performed to determine whether the first record in each database matched in terms of time and diagnosis. If so, data from the MBDS database were chosen. If not, and the MBDS entry predated that in the Emergency database, the former was selected. Conversely, if the record in the Emergency database predated that in the MBDS database, a check was performed to determine whether record corresponded to a non-CVE record in the MBDS database: if so, the record in the Emergency database was rejected; if not, the record in the Emergency database was selected.

iii. Finally, the Aragon Mortality Registry was analysed to identify subjects with a fatal first CVE.

Diagnoses were recorded according to the International Classification of Diseases, 9th revision (ICD-9) in the Emergency database, and according to the International Classification of Diseases, 10th revision (ICD-10) in the MBDS and Aragon Mortality Registry. The following ICD-9 and ICD-10 codes were considered: ICD-9 410–415 and ICD-10 I20-I25 (heart disease); ICD-9 415–417 and ICD-10 I26-I28 (pulmonary heart disease and diseases of pulmonary circulation); ICD-9 427.4, 427.5, 428, 429.2 and ICD-10 I46, I49.0, I50 (other heart diseases); ICD-9 430–438 and ICD-10 G45-G46 and I60-I69 (cerebrovascular diseases); ICD-9 440–445 and ICD-10 I70-I79 (diseases of arteries, arterioles, and capillaries).

Statistical analyses

For the initial description of the variables included in the study, continuous variables were expressed as the mean and standard deviation and categorical variables as percentages.

Machine learning models development. Supervised machine learning algorithms were used to determine the utility of different variables to predict CVE. The process is depicted in Fig 1. These algorithms included RF, XGBoost, and NB. Due to the low number of subjects with missing data, subjects with missing data in any variable were excluded (N = 89). 70% of the data were randomly split to form the training group. The entire sample was used to test the different algorithms, as the sample size of the testing group was small to validate the models obtained. To fit the hyperparameters and avoid over-fitting, the prediction accuracy of all models was tested using 5 and 10 fold stratified cross validation (cv) to estimate F1-score. Results were similar when applying 5 and 10 fold cv, so results shown in the artide corresponds to the 5-fold cv.

The following parameters were adjusted for RF models: number of trees, number of features to consider at any given split, maximum depth, which is the maximum number of partitions in the longest branch of the tree, and minimum number of observations in one node. For XGBoost, the parameters adjusted were the number of features to consider at any given split, maximum depth, and the minimum number of observations in one node. These parameters are shown in Table 2. For the NB method, the *a priori* probabilities were 0.9 for non-occurrence of CVE and 0.1 for occurrence of CVE.

These parameters were fitted for each algorithm before its implementation to avoid over-fitting.

Each method was applied twice: first including just CVRFs as variables, and again including both CVRFs and treatment exposure.

Machine learning models assessment. Because of the highly imbalanced data that we had, to validate models threshold were moved and selected based on max of f1 score in P-R curves, being 0.1 in all models. To measure the validity of the models, the following measures were taken into account: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Next, three different tests were applied to evaluate the performance of the model: area under the precision-recall curve (AUC-PR), Log Loss, and F1 score. These scores were selected because different studies recommend them for imbalanced data [22–24].

Variables importance. After achieving valid and accurate models, the contribution of each variable to the prediction was extracted using caret R package. Methods applied for each algorithm were different and they give scores in different ranges so, to facilitate comparability, scores obtained for each method were normalized to a scale of 0–1. For RF models, the method applied to compute the contribution of each variable consisted of recoding for each tree the prediction accuracy on the out-of-bag portion. Then each predictor variable was permuted and the same was done. Finally, for all trees, the difference between both accuracies was averaged and normalized by the standard score [10, 11].

Table 2. Random Forest and XGBoost parameters adjusted.

	CVRF	CVRF AND TREATMENT EXPOSURE
RANDOM FOREST	Num. trees: 5000 Number of predictors: 5 Max. depth: 3 Min. node. size: 10	Num. trees: 5000 Number of predictors: 6 Max. depth: 3 Min. node. size: 10
XGBOOST	Number of predictors: 2 Max. depth: 6 Min. node. size: 38	Number of predictors: 1 Max. depth: 9 Min. node. size: 3

CVRF, cardiovascular risk factors; Num, number; Max, maximum; Min, minimum.

https://doi.org/10.1371/journal.pone.0293759.t002

For XG Boost models, the reduction in the loss function attributable to each variable in sum of squared error in predicting the gradient on each iteration was calculated. Finally, the improvement score for each predictor was averaged across all the trees in the ensemble [11, 25].

Finally, for models developed applying NB a ROC curve analysis was conducted on each predictor. Different cutoffs were applied to the predictor data to predict the class. Then, sensitivity and specificity were computed for each cutoff and the area under ROC curve was calculated using the trapezoidal rule. This area was used as the measure of variable importance [11].

All statistical analyses were performed in the R statistical computing environment (version 4.0.5 Foundation for Statistical Computing, Vienna, Austria).

Ethical issues

All participants in the AWHS provided prior written informed consent and all collected data were anonymized according with the Spanish Organic Law 3/2018 and the Declaration of Helsinki. The present study was approved by the Clinical Research Ethics Committee of Aragon (project identification code PI17/00042).

Results

In total, this study included 3,746 participants (mean age, 61.6 years), all of whom were male. The prevalence of hypertension was 66.1%, HC 81.0%, and DM 17.0. The percentage of participants receiving treatment for these conditions was 74.3% for hypertension, 52.7% for HC, and 83.5% for DM. Overweight was recorded in 54.4% of participants and obesity in 30%. The number of CVRFs present was as follows: 1 CVRF, 46.9%; 2 CVRFs, 41.9%; 3 CVRFs, 11.2%. The number of CVEs recorded between January 2010 and December 2019 was 298 (7.9%). Assessment of adherence by pharmacological group indicated a PDC $\geq$80% in 63.3% of participants taking antidiabetics, 78.1% of those taking antihypertensives, and 64.4% of those taking medication for HC.

Mean age, CVRF prevalence, and treatment in subjects with and without a CVE are shown in Table 3. Compared with the non-CVE group, the CVE groups had a higher mean age (1.4 years higher) and a greater prevalence of hypertension, HC, diabetes and obesity. Conversely, the prevalence of overweight was slightly higher in the non-CVE group. Among those with no

Table 3. Descriptive variables stratified according to incidence of cardiovascular events.

	NO CVE	CVE	P-VALUE
N	3448 (92.05)	298 (7.95)	
AGE*	61.50 (4.82)	62.90 (4.20)	<0.001
HYPERTENSION	2252 (65.60)	213 (71.70)	0.039
HC	2765 (80.30)	264 (89.2)	<0.001
DIABETES	567 (16.50)	65 (22.00)	0.019
PHYSICAL STATUS			0.297
OVERWEIGHT	1854 (54.50)	157 (53.40)	
OBESE	1009 (29.70)	98 (33.30)	
TREATMENT EXPOSURE			<0.001
FULLY EXPOSED	1075 (45.60)	51 (24.30)	
NON-EXPOSED	485 (20.60)	74 (35.20)	
PARTIALLY EXPOSED	798 (33.80)	85 (40.50)	

CVE, cardiovascular event; HC, hypercholesterolemia. Data are expressed as the number (%); *In this case, as mean (SD).

https://doi.org/10.1371/journal.pone.0293759.t003

CVRFs, treatment exposure was classified as fully exposed in 45.60% of individuals in the group without CVE and in 24.30% of those in the group with CVE.

Results obtained for each of the models, using different groups of variables, are presented below.

Results obtained for models using cardiovascular risk factors as predictive variables

To facilitate comparison of the predictive capacity of each of the variables tested in XGBoost, RF and NB algorithms, values were normalized on a scale of 0–1 (Fig 2), where 0 and 1 indicate minimum and maximum predictive capacity, respectively.

In all models, the variable that best predicted CVE was age. The next best predictor was physical status in case of the XGBoost and RF models, and HC (followed closely by hypertension) in the case of the NB model.

Table 4 compares the different measures used to evaluate model validity and performance. In terms of F1-score, the best results were obtained for the RF method (0.84). The only parameter for which the XGBoost method outperformed the RF method was specificity (53.00% and 52.05%, respectively). AUC-PR, Log Loss, and F1-Score indicated that the RF method

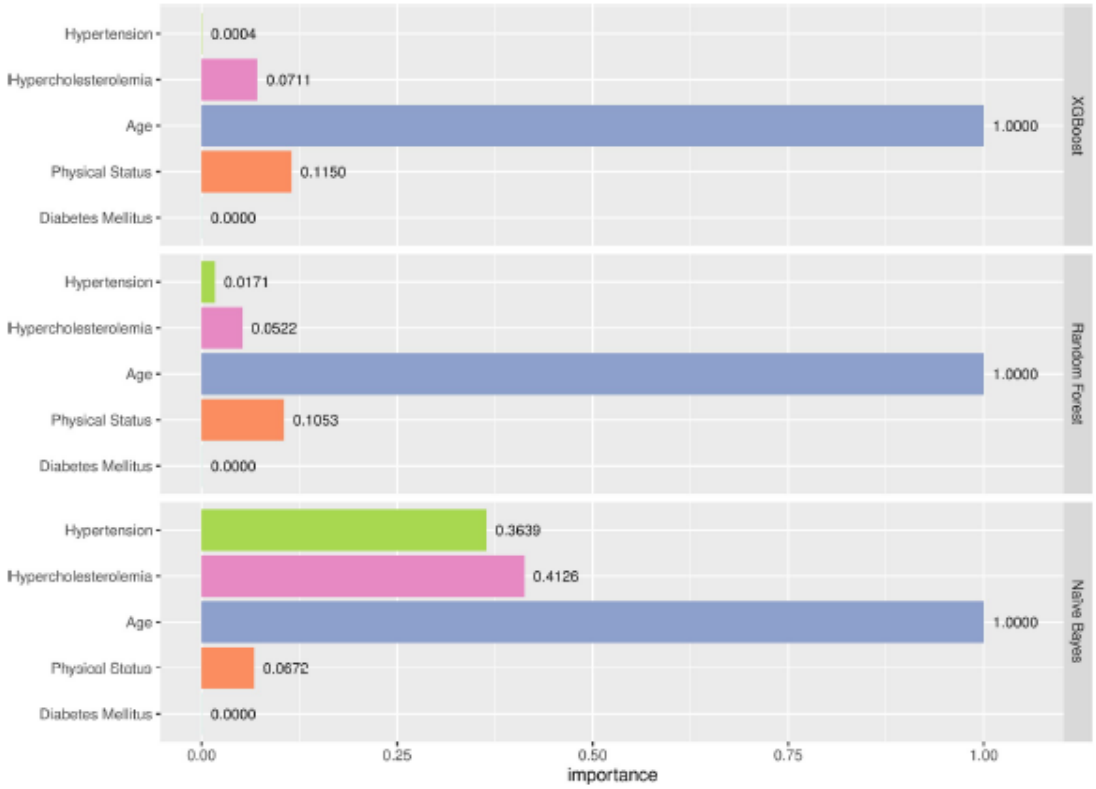


Fig 2. Predictive capacity of the variables included in the study according to the three algorithms applied: XGBoost, Random Forest and Naïve Bayes.

https://doi.org/10.1371/journal.pone.0293759.g002

Table 4. Evaluation of model validity and performance using only cardiovascular risk factors as predictive variables.

MODEL	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	PPV (%)	NPV (%)	AUC-PR	LOG-LOSS	F1-SCORE
XGBOOST	71.29	72.71	53.00	95.20	13.16	0.15	0.24	0.83
RANDOM FOREST	73.96	75.66	52.05	95.29	14.30	0.17	0.24	0.84
NAÏVE BAYES	68.29	69.96	47.00	94.42	10.88	0.11	0.26	0.80

PPV, positive predictable value; NPV, negative predictable value; AUC-PR, Area under the precision -recall curve.

https://doi.org/10.1371/journal.pone.0293759.t004

performed best, while the NB method performed the worst (lowest AUC-PR and F1-Score and highest Log Loss).

Results obtained for models using cardiovascular risk factors and treatment exposure as predictive variables

When treatment exposure was also included as a predictive variable (Fig 3), age remained the variable that best predicted in both the XGBoost and RF models, followed by treatment

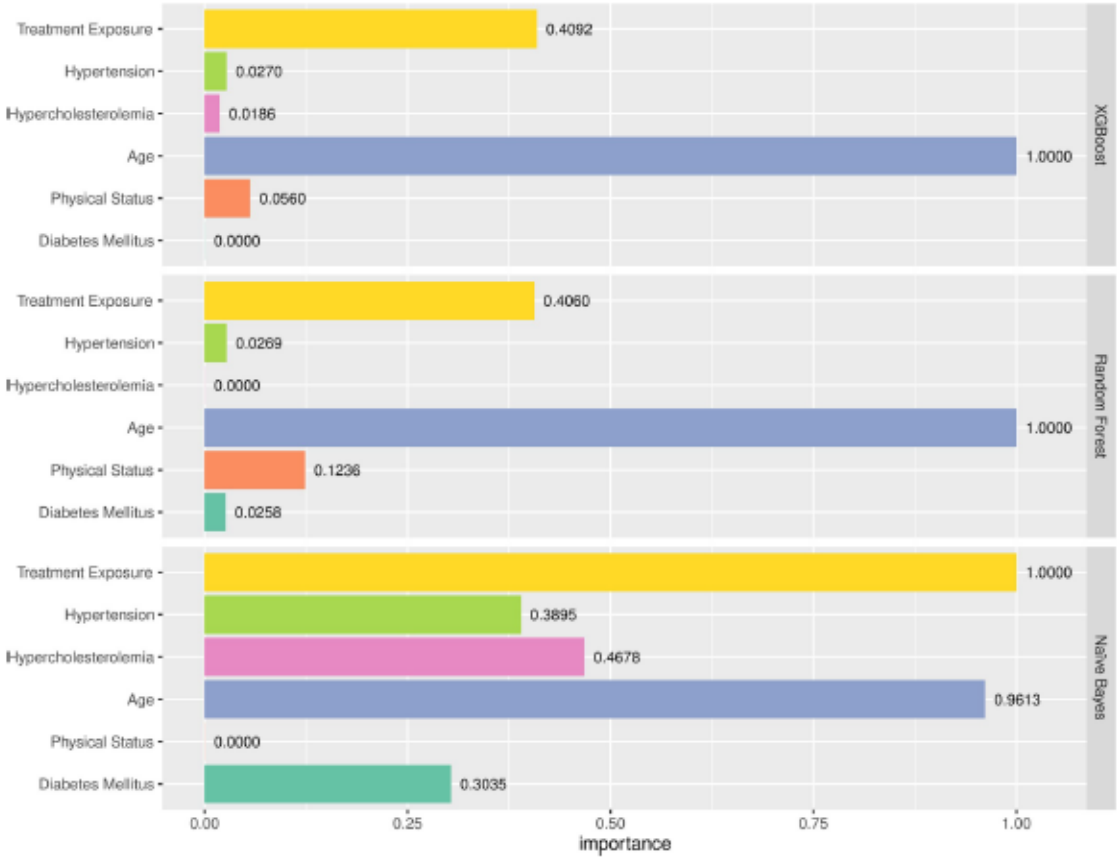


Fig 3. Predictive capacity of the variables included in the study in each of the three algorithms applied: XGBoost, Random Forest, and Naïve Bayes.

https://doi.org/10.1371/journal.pone.0293759.g003

Table 5. Evaluation of model validity and performance using both cardiovascular risk factors and treatment exposure as predictive variables.

MODEL	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	PPV (%)	NPV (%)	AUC-PR	LOG-LOSS	F1-SCORE
XGBOOST	72.39	73.17	63.36	95.85	16.94	0.24	0.25	0.83
RANDOM FOREST	73.35	73.68	69.52	96.55	18.57	0.28	0.24	0.84
NAÏVE BAYES	61.78	61.88	60.62	94.79	12.07	0.13	0.28	0.75

PPV, positive predictable value; NPV, negative predictable value; AUC-PR, Area under the precision-recall curve.

https://doi.org/10.1371/journal.pone.0293759.t005

exposure. In both these models all other variables had very little influence, although physical status had a higher predictive capacity in the RF versus the XGBoost model. In the NB model treatment exposure was the variable with the highest predictive capacity, followed closely by age.

Table 5 compares the different parameters used to evaluate model validity and performance. The RF model showed the highest scores for accuracy, sensitivity, specificity, PPV, and NPV, while the NB model showed the lowest scores for these parameters. Similar results were shown by the tests calculated to evaluate the effectiveness of the models: the best results were for the RF algorithm and the worst for the NB, being the XGBoost scores similar to the RF ones.

Comparison of the models

The RF model performed best, regardless of the groups of variables included. Fig 4 shows the PR curve for this method when considering both CVRFs and treatment exposure as predictive variables. The model performed best when both groups of variables were included.

Log Loss and F1 score were very similar regardless of the variables included (0.24 and 0.84 for CVRFs and CVRFs + treatment exposure, respectively). Parameters used to evaluate model validity (accuracy, sensitivity, and PPV) were very similar in both models (about 73%, 74%, and 96%). However, specificity and NPV were considerably higher when treatment exposure was included as a predictive variable (69.52% and 18.57%, respectively, versus 52.05% and 14.30%, respectively, when treatment exposure was excluded).

Finally, age and treatment exposure had the highest predictive capacity. Age was the variable that best predicted CVE in all models, except in the NB model when both CVRFs and treatment exposure were included as predictive variables: in that scenario, treatment exposure was the most important predictor of CVE, followed closely by age. When treatment exposure was excluded from the RF and XGBoost models, age was the variable with the greatest predictive capacity. While the predictive capacity of age and treatment exposure were 1 and 0.40, respectively, in both the RF and XGBoost models, the corresponding values in the NB model were much closer to one another (0.96 and 1, respectively). In the NB model, after age, the variables HC and hypertension had greater predictive capacity than physical status and DM.

Evaluation of model performance showed that the RF model performed best, regardless of the variables included, followed by XGBoost. Both models performed better when CVRFs and treatment exposure were included as predictive variables, compared with CVRFs alone (Fig 4).

Discussion

In the present study we compared the CVE prediction performance of three machine learning algorithms, in each case using two distinct combinations of predictive variables: (i) CVRFs (age, physical status, HC, hypertension and DM); and (ii) CVRFs plus treatment exposure.

Age was the variable with the greatest capacity to predict CVE in all models, except for the NB model when CVRFs + treatment exposure were included as predictive variables, in which

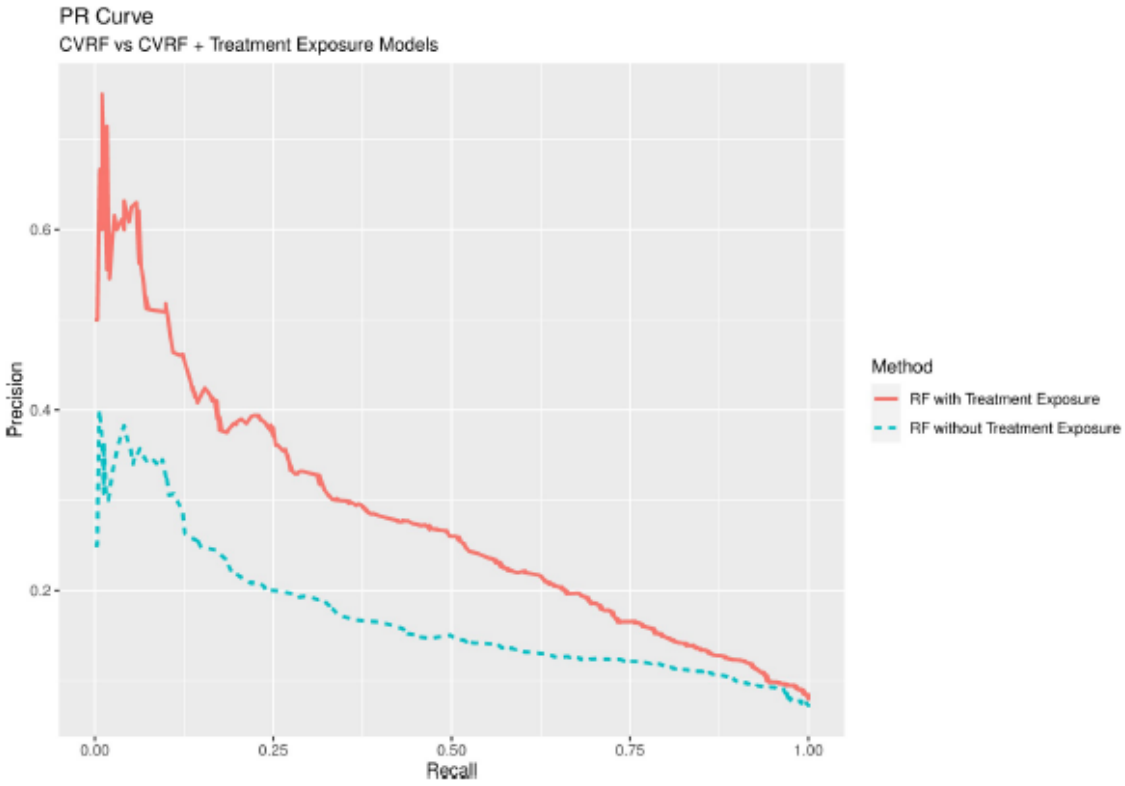


Fig 4. PR curves obtained for RF model, considering cardiovascular risk factors alone or cardiovascular risk factors and treatment exposure as predictive variables.

https://doi.org/10.1371/journal.pone.0293759.g004

case treatment exposure showed the greatest predictive capacity, followed closely by age. When only CVRFs were included, age showed the greatest predictive capacity, while that of other CVRFs was markedly lower, and varied depending on the model used. When treatment exposure was added as an additional predictive variable, its predictive capacity was closer to that of age than any of the other CVRFs. These findings indicate that treatment exposure plays an important role in determining the incidence of CVE and should therefore be taken into account when managing cardiovascular risk.

As said before, age was the variable that best predicted CVE occurrence, while the predictive capacity of other CVRFs differed across models. The relationship between CVD and age, HC, hypertension, DM, and physical status is well documented [26–29]. Studies that included age as a predictive variable have unanimously shown that this parameter has the greatest predictive power, suggesting that age is a key CVRF [3, 4, 12, 26]. Furthermore, the combination of two or more CVRFs increases the risk of mortality [29], although the influence of individual CVRFs varies between studies [17–19]. In their study of the prevalence of CVRFs in individuals who experienced a CVE, Yandrapalli et al. found that HC was the most prevalent, followed closely by hypertension, smoking, and finally DM [27]. Another study [29] found that hypertension was the CVRF most closely associated with all-cause mortality, followed by DM, HC, and overweight. Huang et al. [28] concluded that the most important CVRFs were obesity and

smoking in rural and urban areas, respectively, followed by dyslipidemia. Thus, apart from age, it remains unclear which other CVRF is the greatest determinant of the incidence of CVE.

Proper pharmacological control of modifiable CVRFs is essential to reduce the risk of a CVE. Clinical guidelines propose the objectives to control CVRFs in order to decrease this risk [30, 31]. Previous research [32–34] has shown that adherence to these treatments is suboptimal, and the methods most commonly used to determine the risk of CVE do not include treatment adherence as a predictive variable. There is also evidence [35] that a considerable number of CVEs are due to poor adherence to cardiovascular preventive treatments. Therefore, measuring adherence could maximize the potential of effective cardiac therapies in clinical settings. Exposure to drugs for control of CVRFs could be an important factor to consider when determining the risk of experiencing a CVE, as treatment exposure varies considerably between individuals and its plays a key role in risk management. This view is borne out in our study, in which model performance improved considerably when treatment exposure (determined based on the adherence to each individual treatment) was included as a predictive variable.

Evidence indicates that population-level screening for CVD risk and CVRFs, including screenings performed in a work setting, is not effective in decreasing CVD morbidity and mortality [36]. However, primary-care-based interventions targeting individuals at high risk, based on their age or risk factors, seem to be effective [36]. The models presented in the present study could be applied in clinical practice to assess the individual risk of CVE based on patient characteristics and treatment adherence, and could therefore fulfil this screening role. Furthermore, they can help orient the intervention and identify the most appropriate measures to take (e.g. behavioral change versus reinforcement of adherence).

Over the years, much effort has been invested in calculating CVE risk, using a variety of methods, many of which have limitations that can be overcome with machine learning technics. These techniques offer a variety of approaches to process huge amounts of data to predict the incidence of CVE, thus allowing researchers and clinicians to select the algorithm that better suits their data or objectives.

Previous studies [5, 37] comparing different algorithms found that the RF model provided the most accurate results, in line with our findings. A previous comparison of XGBoost and NB [12] found that XGBoost performed better, as also observed in the present study. Finally, when comparing performance of RF and XGBoost algorithms with SCORE and Framingham risk score, the first ones had better results [5].

This study has some limitations, data were highly imbalanced and some methods were applied to deal with that. Because of that, AUC-PR values were low. In addition, the NPV value was low for all models, this could mean that a certain factor has been omitted from the models. Furthermore, some of the CVEs included in the study (e.g. arrhythmias) may be unrelated to the CVRFs considered. Nonetheless, our study revealed good performance for all the models, in particular RF and XGBoost when treatment exposure was included as a predictive variable. Other limitations, related to the kind of data available, include the absence of smoking data for the entire study period, since smoking is one of the most important modifiable risk factors for CVE, as well as the absence of women in the cohort, as sex also influences CVE risk.

This study has several strengths. First is the use of three different machine learning techniques, which integrate all available data and offer several advantages over earlier algorithms, as explained above, and compare the results between them to determine which is more accurate. In the context of machine learning studies, to our knowledge this is the first study to consider these specific groups of variables in the prediction of CVD. Another key strength is the inclusion of adherence to different treatments integrated into a single variable: this is also the first study to use this approach to predict CVE, and we consider this approach essential to

determine the true risk of experiencing a CVE, given the influence of adherence on therapeutic outcome. Finally, our analysis considered multiple algorithms and different combinations of predictive variables, allowing us to identify the model that performed best in this particular study population and to evaluate the influence of different variables on CVE occurrence.

Further studies that consider additional CVRFs, including smoking habits, and to include more heterogeneous populations to better reflect the situation in terms of presence of CVRF and CV pathology will be needed to evaluate the contribution of CVRF to CVE risk. Moreover, it will be essential to include women in future studies given the differences in the incidence and relevance of CVRFs between sexes.

Conclusions

We found that the age was the most influential variable to predict the occurrence of a CVE followed by the treatment exposure. The rest of variables considered changed its importance depending on the algorithm and the model implemented. The use of machine learning techniques can be of great help to assess the risk of suffering a CVE including a huge amount of data and can be applied for personalized medicine to prevent CVE. The usefulness of machine learning techniques has been proven and the algorithm that better results gave in our case was RF, that improved its results adding the treatment exposure as variable. This study brings to light the importance of considering treatment exposure, estimated based on the adherence to therapy, when trying to assess the risk of suffering from a CVE.

Author Contributions

Conceptualization: Sara Castel-Feced, Lina Maldonado, María José Rabanaque-Hernández.

Data curation: Sara Castel-Feced, Sara Malo, Isabel Aguilar-Palacio, Cristina Feja-Solana, José Antonio Casasnovas.

Formal analysis: Sara Castel-Feced, Lina Maldonado, María José Rabanaque-Hernández.

Funding acquisition: Sara Malo, Isabel Aguilar-Palacio.

Methodology: Sara Castel-Feced, Lina Maldonado, María José Rabanaque-Hernández.

Supervision: María José Rabanaque-Hernández.

Writing – original draft: Sara Castel-Feced, María José Rabanaque-Hernández.

Writing – review & editing: Sara Castel-Feced, Sara Malo, Isabel Aguilar-Palacio, Cristina Feja-Solana, José Antonio Casasnovas, Lina Maldonado, María José Rabanaque-Hernández.

References

1. WHO. Cardiovascular diseases [Internet]. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
2. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J. 2016; 37:2315–81. https://doi.org/10.1093/eurheartj/ehw106 PMID: 27222591
3. Alaa AM, Bolton T, Angelantonio E Di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 2019 May 1; 14(5). https://doi.org/10.1371/journal.pone.0213653 PMID: 31091238
4. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Gregory Hundley W, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 2017 Oct 13; 121(9):1092–101. https://doi.org/10.1161/CIRCRESAHA.117.311312 PMID: 28794054

5. Jamthikar AD, Gupta D, Mantella LE, Saba L, Laird JR, Johri AM, et al. Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: a 500 participants study. Int J Cardiovasc Imaging. 2021 Apr 1; 37(4):1171–87. https://doi.org/10.1007/s10554-020-02099-7 PMID: 33184741

6. Vrbaški D, Vrbaški M, Kupusinac A, Ivanović D, Stokić E, Ivetić D, et al. Methods for algorithmic diagnosis of metabolic syndrome. Artif Intell Med. 2019 Nov 1; 101:101708. https://doi.org/10.1016/j.artmed.2019.101708 PMID: 31813488

7. Amaratunga D, Cabrera J, Sargsyan D, Kostis JB, Zinonos S, Kostis WJ. Uses and opportunities for machine learning in hypertension research. Int J Cardiol Hypertens. 2020 Jun 1; 5:100027. https://doi.org/10.1016/j.ijchy.2020.100027 PMID: 33447756

8. Ghosh P, Azam S, Karim A, Hassan M, Roy K, Jonkman M. A comparative study of different machine learning tools in detecting diabetes. Procedia Comp Sci. 2021; 192:467–77. https://doi.org/10.1016/j.procs.2021.08.048

9. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020 Sep 29; 10(1):16057. https://doi.org/10.1038/s41598-020-72685-1 PMID: 32994452

10. Greenwell BM, Boehmke BC. Variable Importance Plots-An Introduction to the vip Package.

11. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008; 28(5). Available from:http://www.jstatsoft.org/

12. Wang K, Tian J, Zheng C, Yang H, Ren J, Liu Y, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. Comput Biol Med. 2021 Oct 1; 137:104813. https://doi.org/10.1016/j.compbiomed.2021.104813 PMID: 34481185

13. Yu Z, Wang K, Wan Z, Xie S, Lv Z. Popular deep learning algorithms for disease prediction: a review. Cluster Comput. https://doi.org/10.1007/s10586-022-03707-y PMID: 36120180

14. Tsarapatsani K, Sakellarios AI, Pezoulas VC, Tsakanikas VD, Kleber ME, Marz W, et al. Machine Learning Models for Cardiovascular Disease Events Prediction. Annu Int Conf IEEE Eng Med Biol Soc. 2022; 1066–9. https://doi.org/10.1109/EMBC48229.2022.9871121 PMID: 36085658

15. Garavand A, Salehnasab C, Behmanesh A, Aslani N, Zadeh AH, Ghaderzadeh M. Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms. J Healthc Eng. 2022; 2022. https://doi.org/10.1155/2022/5359540 PMID: 36304749

16. Aziz F, Malek S, Ibrahim KS, Shariff RER, Wan Ahmad WA, Ali RM, et al. Short- and long-term mortality prediction after an acute ST-elevation myocardial infarction (STEMI) in Asians: A machine learning approach. PLoS One. 2021 Aug 1; 16(8):e0254894. https://doi.org/10.1371/journal.pone.0254894 PMID: 34339432

17. Casasnovas JA, Alcaide V, Civeira F, Guallar E, Ibañez B, Borreguero JJ, et al. Aragon workers' health study—design and cohort description. BMC Cardiovasc Disord. 2012; 12(45). https://doi.org/10.1186/1471-2261-12-45 PMID: 22712826

18. IACS. Actividad de Tratamiento BIGAN [Internet]. Available from:https://www.iacs.es/actividad-tratamiento-bigan/

19. Chaure-Pardos A, Malo S, Rabanaque MJ, Arribas F, Moreno-Franco B, Aguilar-Palacio I. Factors Associated with the Prescribing of High-Intensity Statins. JClin Med. 2020 Nov 27; 9(12):3850. https://doi.org/10.3390/jcm9123850 PMID: 33260835

20. Vrijens B, de Geest S, Hughes DA, Przemyslaw K, Demonceau J, Ruppar T, et al. A new taxonomy for describing and defining adherence to medications. Br J of Clin Pharmacol. 2012 May 1; 73(5):691–705. https://doi.org/10.1111/j.1365-2125.2012.04167.x PMID: 22486599

21. Malo S, Aguilar-Palacio I, Feja C, Lallana MJ, Rabanaque MJ, Armesto J, et al. Different approaches to the assessment of adherence and persistence with cardiovascular-disease preventive medications. Curr Med Res and Opin. 2017 Jul 3; 33(7):1329–36. https://doi.org/10.1080/03007995.2017.1321534 PMID: 28422521

22. Gaudreault JG, Branco P, Gama J. An Analysis of Performance Metrics for Imbalanced Classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021; 12986 LNAI:67–77. https://doi.org/10.1007/978-3-030-88942-5_6/COVER

23. Miao J, Zhu W. Precision–recall curve (PRC) classification trees. Evol Intell. 2022 Sep 1; 15(3):1545–69. https://doi.org/10.1007/S12065-021-00565-2/TABLES/32

24. Bekkar M, Djemaa HK, Alitouche TA. Evaluation Measures for Models Assessment over Imbalanced Data Sets. Journal of Information Engineering and Applications. 2013; 3(10):27–38. Available from: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633

25. Yuan Jiaming. Package 'xgboost'. 2023; https://doi.org/10.1145/2939672.2939785

26. Oppenheimer GM. Framingham Heart Study: the first 20 years. Prog Cardiovasc Dis. 2010 Jul; 53 (1):55–61. https://doi.org/10.1016/j.pcad.2010.03.003 PMID: 20620427

27. Yandrapalli S, Nabors C, Goyal A, Aronow WS, Frishman WH. Modifiable Risk Factors in Young Adults With First Myocardial Infarction. J Am Coll Cardiol. 2019 Feb 12; 73(5):573–84. https://doi.org/10.1016/j.jacc.2018.10.084 PMID: 30732711

28. Huang HJ, Lee CW, Li TH, Hsieh TC. Different Patterns in Ranking of Risk Factors for the Onset Age of Acute Myocardial Infarction between Urban and Rural Areas in Eastern Taiwan. Internat J Environ Res and Public Health. 2021; 18:5558. https://doi.org/10.3390/ijerph18115558 PMID: 34067428

29. Bazalar-Palacios J, Jaime Miranda J, Carrillo-Larco RM, Gilman RH, Smeeth L, Bernabe-Ortiz A. Aggregation and combination of cardiovascular risk factors and their association with 10-year all-cause mortality: the PERU MIGRANT Study. BMC Cardiovasc Disord. 2021 Dec 1; 21:582. https://doi.org/10.1186/s12872-021-02405-8 PMID: 34876013

30. Visseren FLJ, MacH F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J. 2021 Sep 7; 42(34):3227–337. https://doi.org/10.1093/eurheartj/ehab484 PMID: 34458905

31. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2019 Sep 10; 140(11):e596–646. https://doi.org/10.1161/CIR.0000000000000678/FORMAT/EPUB

32. Zhao B, He X, Wu J, Yan S. Adherence to statins and its impact on clinical outcomes: a retrospective population-based study in China. BMC Cardiovasc Disord. 2020 Jun 10; 20:282. https://doi.org/10.1186/s12872-020-01566-2 PMID: 32522146

33. Lee H, Yano Y, Cho SMJ, Heo JE, Kim DW, Park S, et al. Adherence to Antihypertensive Medication and Incident Cardiovascular Events in Young Adults With Hypertension. Hypertension. 2021; 77 (4):1341–9. https://doi.org/10.1161/HYPERTENSIONAHA.120.16784 PMID: 33641364

34. Aguilar-Palacio I, Rabanaque MJ, Maldonado L, Chaure A, Abad-Díez JM, León-Latre M, et al. New Male Users of Lipid-Lowering Drugs for Primary Prevention of Cardiovascular Disease: The Impact of Treatment Persistence on Morbimortality. A Longitudinal Study. Internat J Environ Res and Public Health. 2020 Oct 2; 17(20):7653. https://doi.org/10.3390/ijerph17207653 PMID: 33092211

35. Chowdhury R, Khan H, Heydon E, Shroufi A, Fahimi S, Moore C, et al. Adherence to cardiovascular therapy: a meta-analysis of prevalence and clinical consequences. Eur Heart J. 2013 Oct 7; 34 (38):2940–8. https://doi.org/10.1093/eurheartj/eht295 PMID: 23907142

36. Eriksen CU, Rotar O, Toft U, Jørgensen T. What is the effectiveness of systematic population-level screening programmes for reducing the burden of cardiovascular diseases? 2021.

37. Nadakinamani RG, Reyana A, Kautish S, Vibith AS, Gupta Y, Abdelwahab SF, et al. Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. Comput Intell Neurosci. 2022 Jan 11; 2022. https://doi.org/10.1155/2022/2973324 PMID: 35069715

ANEXO II

International Journal of
Environmental Research and Public Health

MDPI

Article

Evolution of Cardiovascular Risk Factors in a Worker Cohort: A Cluster Analysis

Sara Castel-Feced [1,2,3,*], Lina Maldonado [4], Isabel Aguilar-Palacio [1,2,3], Sara Malo [1,2,3], Belén Moreno-Franco [1,2], Eusebio Mur-Vispe [5], José-Tomás Alcalá-Nalvaiz [6,7,†] and María José Rabanaque-Hernández [1,2,3,†]

1 Department of Preventive Medicine and Public Health, University of Zaragoza, 50009 Zaragoza, Spain; iaguilar@unizar.es (I.A.-P.); smalo@unizar.es (S.M.); mbmoreno@unizar.es (B.M.-F.); rabanake@unizar.es (M.J.R.-H.)
2 Fundación Instituto de Investigación Sanitaria de Aragón (IIS Aragón), 50009 Zaragoza, Spain
3 GRISSA Research Group, 50009 Zaragoza, Spain
4 Department of Economic Structure, Economic History and Public Economics, University of Zaragoza, 50005 Zaragoza, Spain; lmguaje@unizar.es
5 Prevention Department, Stellantis Spain, 50639 Figueruelas, Spain; eusebio.mur@stellantis.com
6 Department of Statistical Methods, University of Zaragoza, 50005 Zaragoza, Spain; jtalcala@unizar.es
7 Institute of Mathematics and Applications (IUMA), 50009 Zaragoza, Spain
* Correspondence: scastelf@unizar.es
† These authors contributed equally to this work and served as senior co-authors.

check for updates

Citation: Castel-Feced, S.; Maldonado, L.; Aguilar-Palacio, I.; Malo, S.; Moreno-Franco, B.; Mur-Vispe, E.; Alcalá-Nalvaiz, J.-T.; Rabanaque-Hernández, M.J. Evolution of Cardiovascular Risk Factors in a Worker Cohort: A Cluster Analysis. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5610. https://doi.org/10.3390/ijerph18115610

Academic Editor: Paul B. Tchounwou

Received: 25 March 2021
Accepted: 22 May 2021
Published: 24 May 2021

Abstract: The identification of the cardiovascular risk factor (CVRF) profile of individual patients is key to the prevention of cardiovascular disease (CVD), and the development of personalized preventive approaches. Using data from annual medical examinations in a cohort of workers, the aim of the study was to characterize the evolution of CVRFs and the CVD risk score (SCORE) over three time points between 2009 and 2017. For descriptive analyses, mean, standard deviation, and quartile values were used for quantitative variables, and percentages for categorical ones. Cluster analysis was performed using the Kml3D package in R software. This algorithm, which creates distinct groups based on similarities in the evolution of variables of interest measured at different time points, divided the cohort into 2 clusters. Cluster 1 comprised younger workers with lower mean body mass index, waist circumference, blood glucose values, and SCORE, and higher mean HDL cholesterol values. Cluster 2 had the opposite characteristics. In conclusion, it was found that, over time, subjects in cluster 1 showed a higher improvement in CVRF control and a lower increase in their SCORE, compared with cluster 2. The identification of subjects included in these profiles could facilitate the development of better personalized medical approaches to CVD preventive measures.

Keywords: longitudinal study; cluster analysis; real-world data

1. Introduction

Cardiovascular diseases (CVD) are the leading cause of death worldwide [1], and result in disability and considerable economic costs on healthcare systems [2]. The global burden of disease (GBD) study in 2019 [2] showed that, during the last three decades, the prevalence of total CVD has nearly doubled, the number of CVD deaths has increased by more than 6 million and that the years of living with disability has doubled during this period. The most common CVDs are coronary heart disease and stroke [1,2].

Hypertension, dyslipidemia, diabetes mellitus, and smoking are well known as cardiovascular risk factors (CVRF) [3]. Obesity, another CVRF that affects children, adolescents, and adults, has doubled in prevalence over the last 3 decades [4], and numerous studies have associated abdominal obesity with insulin resistance and an increased risk of CVD [5].

A recent study performed in countries with different economic levels [6] has shown that approximately 70% of CVD events and deaths could be attributed to modifiable risk

factors. Moreover, tobacco was the most associated with CVD among the behavioral risk factors, and hypertension among the metabolic risk factors, followed by diabetes.

CVD morbidity and mortality can be significantly reduced by preventive strategies that target both high-risk individuals and the general population [7]. Since the 1970s, a variety of population-level programs have been implemented in developed countries to promote lifestyle changes that can reduce the incidence of CVRFs [7]. A systematic review of the effectiveness of preventive programs reported different patterns of CVRF incidence, depending on the study consulted, although most of the studies included in the review indicated a positive trend [8]. Other studies of a population-level program implemented in Europe reported that after implementation of the program there was an increase in the frequency of some CVRFs, including body mass index (BMI) and glucose levels, and in physical activity, while the prevalence of hypercholesterolemia decreased [7].

Different methods are used to characterize CVD patient profiles, including so-called unsupervised learning cluster analysis, whereby individuals are grouped according to similarities in variables of interest [9]. Identification of patient profiles based on CVRFs and the risk of developing CVD can help to identify new prevention strategies, as well as identifying new lines of research. However, a study of the association between CVRFs and CVD occurrence is complicated by the fact that exposure to CVRFs might not remain constant throughout life. In this sense, the analysis of their evolution can also help in the development and improvement of prevention strategies for control of the CVRFs associated with the profiles identified. For this analysis, longitudinal data are used, where each variable is measured more than once. By clustering the trajectories of the variables into groups with similar characteristics, it is possible to convert multiple continuous variables into a single categorical variable [10], and this facilitates the classification of subjects and the subsequent analysis.

The Aragon Worker's Health Study (AWHS) is a longitudinal cohort study designed to evaluate the evolution of CVRFs and their association with the incidence of CVD in a middle-aged population of factory workers in Spain [11]. In the present study, we hypothesized that the individuals included in the cohort will gather in different groups according to their CVRF, and that both their CVRFs and the CVD risk score will increase with time. Therefore, we sought to characterize the profiles of participants in the AWHS cohort and the evolution of CVRFs and cardiovascular risk over three time points between 2009 and 2017.

2. Materials and Methods

2.1. Study Design and Participants

This study was conducted within the framework of the AWHS, a prospective, longitudinal cohort study of workers at an automobile assembly plant located in Figueruelas (Zaragoza, Spain). Recruitment began in February 2009 and ended in December 2010. For the purpose of our analyses, we selected data collected at annual medical and blood tests that AWHS participants underwent at 3 different time points.

The different time points for which data were extracted are described below. Time point 1 corresponded to the first data record available for each individual after providing informed consent (i.e. data collected in 2009, 2010 or 2011, depending on when the consent form was signed and on the data availability); time point 2 corresponded to the middle of the study (2014); and time point 3 corresponded to the last data record available for each participant (it prioritized data from the year 2017, but if this was not available, data were selected from the year 2016).

Figure 1 depicts how the study population was selected. Our analysis was limited to men, owing to the limited number of women ($N = 380$) in the cohort. Workers diagnosed with any CVD before inclusion in the AWHS, and those who lacked a medical card issued by the Aragón public health system between the date of inclusion and 31 May 2019, were excluded from our analysis. Eight individuals for whom data could not be located in the registry of the Aragon health system (SALUD) were also excluded.

From those selected, medical test results were not available for every year, and therefore the number of workers for which data were acquired for each time point differed. Data were available for 5122 workers for time point 1, 3891 for time point 2, and 3545 for time point 3. Finally, we excluded individuals who had not attended at least 2 of the 3 medical tests from which data were acquired for the present study (N = 975). The final study population consisted of 4147 individuals.

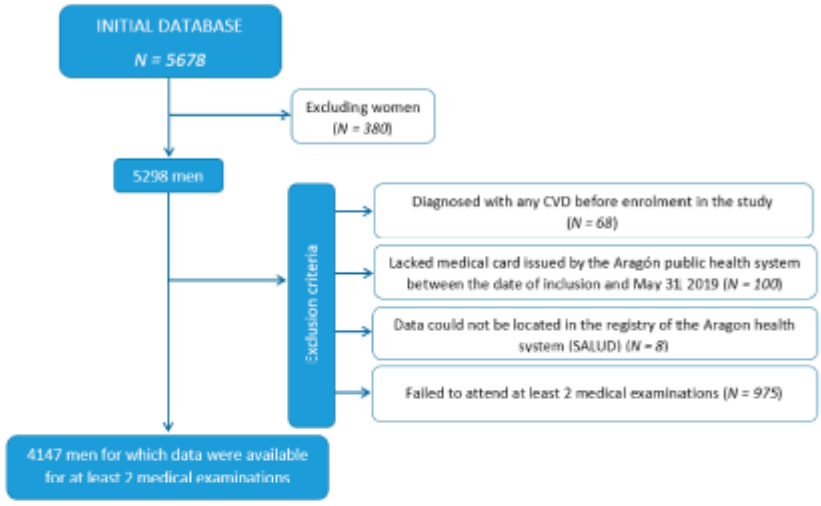


Figure 1. Flowchart depicting the study population.

2.2. *Variables and Data Collection*

Data on BMI, waist circumference (WC), high-density lipoprotein (HDL) cholesterol, and glucose levels, collected at the annual medical examination undergone by all workers at the factory, were obtained from the AWHS database. These data were collected by the physicians and nurses of the factory's medical services, all of whom received prior training. All study procedures were standardized. We also explored the incidence of major CVD events from the CMBD and hospital emergencies databases from the start date of the cohort study until 2017.

At each annual exam, study participants provided a clinical history and underwent a physical exam, including anthropometry (height, weight, and WC). For laboratory analyses, workers provided a sample of blood and urine after overnight (>8 hours) fasting. These samples were processed on the day of the extraction. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were recorded as quantitative variables after being measured 3 consecutive times using an automatic sphygmomanometer, with the participant sitting after a 5-minute rest. Height, weight, and WC were objectively recorded as quantitative variables. BMI was calculated from height and weight. Levels of fasting serum glucose, cholesterol, and HDL cholesterol were measured by spectrophotometry and recorded as quantitative variables in mg/dL.

Smoking status was self-reported (current smoker, ex-smoker, or non-smoker) and was recorded as a qualitative variable. Because smoking status data were not available for time point 3, a projection procedure was applied. To this end, a transition probability matrix for smoking status was generated based on the data available for time points 1 and 2, and used to estimate the smoking status of each worker for time point 3, assuming stationarity in smoking behavior.

Data on physical activity and alcohol intake could not be taken into account in the study, since it was neither recorded for all the subjects included in the cohort nor for all

time points. Furthermore, the correlation between alcohol intake and smoking status was analyzed, and they were highly correlated.

The 10-year CVD risk score (SCORE) was estimated based on the systematic coronary risk evaluation for a European population with low cardiovascular risk [12]. To calculate this variable, smoking status was taken into account, and workers divided into 2 groups: (i) current smokers and (ii) non-smokers and ex-smokers.

Descriptive analyses were performed, taking into account quantitative variables including age, SBP, DBP, weight, WC, BMI, blood levels of HDL cholesterol, total cholesterol, and blood glucose, and SCORE, as well as some categorical variables (smoking status and group BMI).

Cluster analysis was performed, taking into account 4 CVRFs (BMI, WC, HDL cholesterol, and glucose levels), age, and the 10-year SCORE. These CVRFs were specifically selected for analysis because they were not used in the calculation of SCORE.

2.3. *Analysis*

The description of the variables was carried out using the mean and standard deviation (SD) for quantitative variables, and percentages for categorical variables. A correlation analysis of the variables included in the cluster analysis was also performed. The evolution over the study period of the quantitative variables included in the cluster analysis was analyzed using means and quartiles.

Clustering techniques are a form of unsupervised learning that gather elements in homogeneous groups based on similarities between them. Our study was a longitudinal cohort study in which each variable was measured at different time points and changed over time for each individual. The standard method of clustering variable trajectories is to cluster each variable separately. In studies involving more than one variable, cluster analysis enables analysis of the joint evolution of the variables of interest [13].

We applied the Kml3D package in R statistical software [10,14]. This k-means algorithm uses a generalized notion of distance between individual trajectories, and was used to group individuals according to the evolution of CVRF and estimated SCORE over 3 distinct time points. The Calinski–Harabasz criterion was used to determine the number of groups into which participants should be divided. Implementation of the kml3d algorithm requires that data for all variables included in the analysis be available for all participants. Because some workers did not attend all medical tests, some of these variables had to be imputed before applying the algorithm [10,15]. This was achieved using the imputation function in the R longitudinalData package [16], specifically using the "linearInterpol.bisector" command, which differentiates between monotone and intermittent missing data. In the first case, it creates the bisector of (i) the line joining the two first or the two last non-missing data (depending on whether the missing data gathers at the start or at the end of the study period) and (ii) the line joining the first and last non-missing data. In the second case, it creates a line joining the values immediately surrounding the missing value. Table 1 shows the means of the study variables calculated with both available and imputed data. They showed that the imputation worked well, since the mean variables did not change excessively. Successful imputation was confirmed by descriptive analyses and comparison of means using a Student's *t*-test.

All analyses were performed using RStudio and R version 4.0.2, R Foundation for Statistical Computing, Vienna, Austria (22 June 2020).

Table 1. Comparison between real and imputed data.

Variables	Time Point 1		Time Point 2		Time Point 3	
Data	Imputed	Real	Imputed	Real	Imputed	Real
BMI (kg/m^2)	27.6 (3.5)	27.6 (3.5)	27.8 (3.7)	27.8 (3.7)	27.9 (3.8)	27.8 (3.8)
Wc- Cholesterol (cm)	96.8 (9.7)	96.8 (9.6)	97.4 (10.0)	97.3 (10.0)	98.1 (10.8)	97.7 (10.5)
HDL (mg/dL)	52.4 (10.9)	52.4 (11.0)	53.8 (11.3)	54.1 (11.3)	51.5 (12.9)	51.0 (12.4)
Glucose (mg/dL)	97.7 (18.7)	97.7 (18.7)	98.4 (19.5)	96.5 (19.5)	89.9 (21.5)	88.1 (18.6)
SCORE	1.6 (1.4)	1.6 (1.4)	2.1 (1.7)	2.1 (1.7)	2.4 (2.2)	2.1 (1.7)

Abbreviations: WC, waist circumference; BMI, body mass index; SCORE, cardiovascular disease risk score.

2.4. Ethical Issues

Individuals who participated in the AWHS provided prior written informed consent, and all collected data were anonymized. The present study was approved by the Clinical Research Ethics Committee of Aragon; it has not been previously conducted, and current results are not overlapped with other previously published or ongoing reports.

3. Results

A descriptive analysis of the cohort data for the three time points studied is shown in Table 2.

Table 2. Descriptive analysis of the study variables.

Quantitative Variables		Time Point 1 Mean (SD)	Time Point 2 Mean (SD)	Time Point 3 Mean (SD)
Age (years)		48.00 (8.42)	51.49 (8.27)	53.00 (8.25)
Systolic blood pressure (mmHg)		126.00 (14.14)	124.00 (14.25)	128.89 (15.00)
Diastolic blood pressure (mmHg)		83.44 (9.82)	79.80 (9.39)	81.36 (9.68)
Weight (kg)		81.64 (11.47)	82.10 (11.92)	82.66 (12.38)
Waist circumference (cm)		96.81 (9.61)	97.30 (10.00)	97.73 (10.53)
Body mass index (kg/m^2)		27.61 (3.54)	27.77 (3.67)	27.84 (3.80)
HDL cholesterol (mg/dL)		52.45 (11.00)	54.07 (11.30)	51.00 (12.40)
Total cholesterol (mg/dL)		212.18 (37.62)	205.93 (34.75)	187.96 (32.85)
Glucose (mg/dL)		97.70 (18.75)	96.51 (19.46)	88.06 (18.60)
SCORE		1.56 (1.40)	2.05 (1.73)	2.09 (1.74)
Categorical Variables		**Time Point 1 *N* (%)**	**Time Point 2 *N* (%)**	**Time Point 3 *N* (%)**
Smoking status	Smoker	1488 (36.82)	1235 (32.19)	1156 (32.65)
	Non-smoker	1087 (26.90)	925 (24.11)	853 (24.09)
	Ex-smoker	1466 (36.28)	1677 (43.71)	1532 (43.26)
Body mass index groups	Normal weight	938 (23.05)	846 (22.01)	762 (22.38)
	Overweight	2223 (54.63)	2088 (54.33)	1813 (54.25)
	Obesity	908 (22.32)	909 (23.65)	830 (24.38)

Abbreviations: SD, standard deviation; SCORE, cardiovascular disease risk score; *N*, number. Smoking status data for the time point 3 were estimated by imputation.

The correlation between variables was analyzed for each time point. At time point 1, the lowest correlation indices were for WC and HDL cholesterol (−0.21) and for HDL cholesterol and BMI (−0.21). The highest correlation index (0.87) was obtained for WC and BMI. Similar results were obtained for the other 2 time points. The correlation index was significantly different from 0 for all variable pairs, except for HDL cholesterol and SCORE at time point 1, and HDL cholesterol and age at time points 2 and 3.

At the first time point analyzed, the mean weight was 81.64 kg, over half the study population was overweight, and 3 individuals included in the normal-weight group were underweight. Mean values were all within the recommended range.

Comparison of data collected at time point 2 revealed little change in mean values relative to time point 1. The greatest change observed was in the mean total cholesterol level, which was lower at time point 2 versus time point 1. The mean BMI values were

practically the same at both time points. Analysis of smoking status revealed an increase in the percentage of ex-smokers and a decrease in the number of smokers and non-smokers.

Comparison of time point 3 versus time point 2 revealed changes in the mean total cholesterol and glucose levels, both of which were decreased.

Analysis of the overall evolution of the variables of interest over the three time points revealed a reduction in total cholesterol and glucose levels, and in smoking, and an increase in obesity.

The results of quartile analysis over the three time points for some variables are shown in Appendix A. For time point 1, quartile analysis of BMI showed that obese workers fell in quartile (Q) 4, those with normal weight, in Q1, and those who were overweight, in Q2 and Q3. For WC and blood glucose, workers with levels above the recommended values fell in Q4, and for HDL cholesterol, those with lower than recommended levels fell in Q1.

For BMI and WC, over 80% of workers who were in Q4 at time point 1 remained in this quartile at time point 2. A similar proportion remained in Q4 between time points 2 and 3. For blood glucose levels, approximately 50% of participants who were in Q4 at time point 1 remained in this quartile at time point 2, and a similar effect was observed comparing time points 2 and 3. For HDL cholesterol levels, the percentage of workers who remained Q1 was 60% at time point 2 versus time point 1, and 80% at time point 3 versus time point 2. For SCORE, the percentage of workers who remained in the same quartile at time point 2 versus time point 1 was 70% for Q1, and 89% for Q4 (risk score, <0.63 and >2.13, respectively). For time point 3 versus time point 2, the percentage of individuals remaining in Q1 was 84%, and 88% for Q4.

A cluster analysis was performed to evaluate the joint evolution of the following variables: age, waist circumference, BMI, blood glucose, HDL cholesterol levels, and 10-year SCORE. Based on the quality index to determine the number of clusters (Figure 2), we divided the cohort into 2 and 3 clusters. When analyses were performed for both scenarios, results were better justified when the cohort was divided into 2 rather than 3 groups. Moreover, the Calinski–Harabasz score was 1197 when the cohort was divided into 2 clusters, and 1067 when divided into 3. Therefore, we focused on the results obtained using 2 clusters.

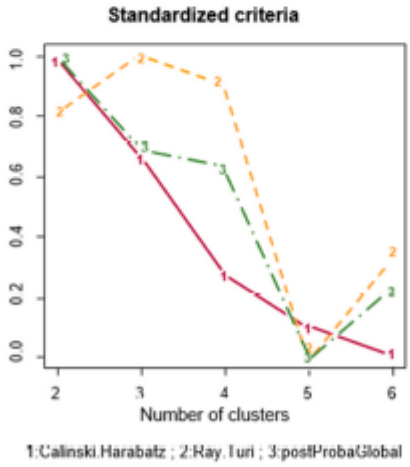


Figure 2. Quality index according to the number of clusters. Each line represents a different quality index, and depicts the resulting changes according to the number of clusters. Each quality index has been normalized to a value between 0 and 1.

Figure 3 shows the mean values obtained for each variable, stratified by clusters. Table 3 depicts the descriptive analysis of the cohort, stratified by cluster and time point.

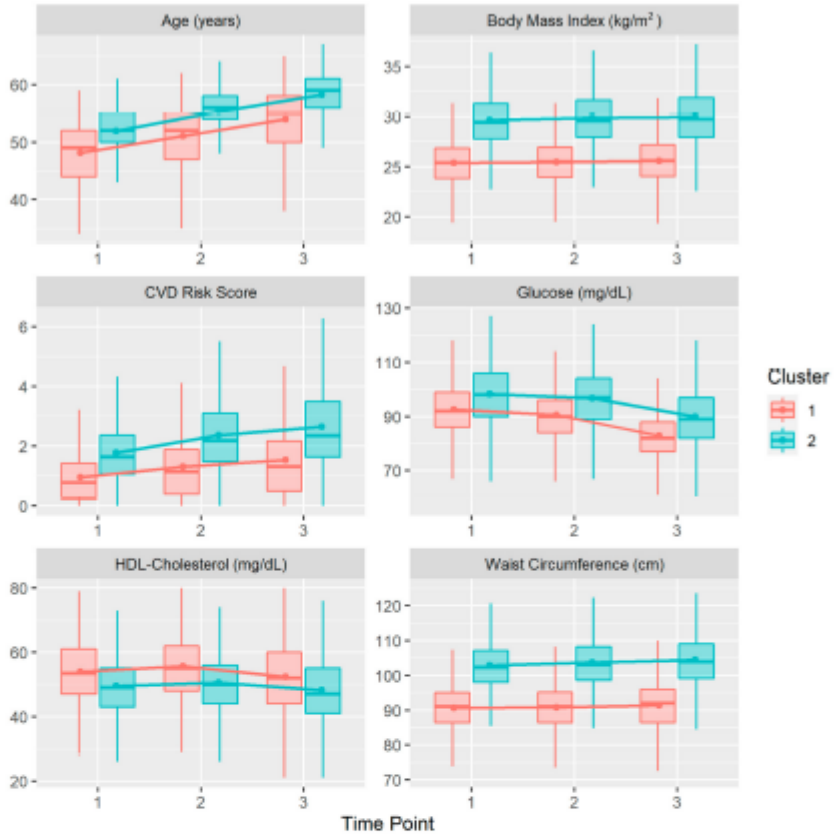


Figure 3. Box plots depicting mean values (dots) for each variable per cluster. Lines represent the evolution of the mean for each cluster over the three time points analyzed. Abbreviation: CVD, cardiovascular disease.

Cluster 1 consisted of younger workers with a lower mean BMI, WC, blood glucose values, and SCORE, and higher mean HDL cholesterol values. Analysis of the evolution of each variable over the three time points revealed similar patterns in both clusters. The mean WC and BMI values increase slightly and progressively over time, although this increase was slightly greater in cluster 2. Blood glucose levels decreased slightly and progressively over time in both clusters. In both clusters, HDL cholesterol levels increased between time points 1 and 2 and decreased between time points 2 and 3. Finally, in both clusters, SCORE increased over the study period, although this effect was greater in cluster 2.

The evolution per cluster of each quartile over the three time points is shown in Appendix A. For BMI, in both clusters 1 and 2, more than half of the individuals in a given BMI quartile at time point 1 remained in the same quartile at time point 2, with the exception of Q4. The percentage of individuals who remained in Q4 at time point 2 was twice as high in cluster 2 versus cluster 1. Moreover, in cluster 1, the percentage of individuals in Q4 who remained in this quartile at time point 2 was similar to the percentage that switched to Q3.

Comparison of BMI values at time points 2 and 3 revealed similar findings to the comparison of time points 1 and 2 in both clusters. However, in cluster 1, the percentage of subjects who remained in Q4 between time points 2 and 3 was greater than the percentage of subjects who remained in Q4 between time points 1 and 2. The evolution of WC was very similar to that of BMI.

Table 3. Descriptive analysis of the variables included in the cluster analysis, stratified by cluster and time point.

Variables	Time Point	Cluster 1 N = 2099	Cluster 2 N = 2048	*p*
		Mean (SD)	Mean (SD)	
Age	Time point 1	44.2 (9.58)	51.7 (4.59)	<0.001
	Time point 2	47.6 (9.58)	55.2 (4.63)	<0.001
	Time point 3	50.6 (9.55)	58.1 (4.58)	<0.001
WC	Time point 1	90.5 (6.80)	103 (7.72)	<0.001
	Time point 2	90.6 (6.61)	104 (8.06)	<0.001
	Time point 3	91.2 (7.31)	105 (8.97)	<0.001
BMI	Time point 1	25.3 (2.30)	30.0 (3.04)	<0.001
	Time point 2	25.4 (2.25)	30.2 (3.22)	<0.001
	Time point 3	25.6 (2.41)	30.4 (3.49)	<0.001
Glucose	Time point 1	92.9 (11.7)	103.0 (22.9)	<0.001
	Time point 2	91.0 (11.5)	102.0 (24.1)	<0.001
	Time point 3	83.4 (11.7)	96.5 (26.7)	<0.001
HDL	Time point 1	55.0 (11.3)	49.9 (9.94)	<0.001
	Time point 2	56.8 (11.8)	50.9 (9.96)	<0.001
	Time point 3	54.1 (13.4)	48.9 (11.8)	<0.001
SCORE	Time point 1	1.02 (1.03)	2.12 (1.55)	<0.001
	Time point 2	1.38 (1.27)	2.74 (1.85)	<0.001
	Time point 3	1.62 (1.49)	3.27 (2.49)	<0.001

Abbreviations: WC, waist circumference; BMI, body mass index; SCORE, cardiovascular disease risk score; *p*, *p*-value (unpaired 2-sample Student's t-test).

For HDL cholesterol, in cluster 1, 54% of workers who were in Q1 at time point 1 remained in this quartile at time point 2. The percentage of workers in Q1 at both time points 2 and 3 was 78%. In cluster 2, 65% of workers in Q1 at time point 1 remained in this quartile at time point 2. In this same cluster, the percentage of workers in Q1 at both time points 2 and 3 was 82%.

Analysis of blood glucose values showed that in cluster 2, the proportion of workers who were at Q4 at time point 1 and remained in this quartile at time point 2 was double that observed for cluster 1. For both clusters, between time points 2 and 3, we observed a decrease in the proportion of workers that did not switch quartiles, except for Q1, for which a significant increase was observed.

Finally, for SCORE, in cluster 1 the percentage of workers who were in Q1 and Q4 at both time points 1 and 2 (74% and 82%, respectively) was higher than that observed for Q2 and Q3 (38% and 48%, respectively). Comparison of time points 2 and 3 showed that for all quartiles, the percentage of workers who did not switch quartile was higher than that observed between time points 1 and 2. In cluster 2, the results obtained were similar to those observed for cluster 1, although among workers in Q1 at time point 1, only 41% remained in this quartile at time point 2, whereas 46% moved to Q2.

In an initial study about CVD events, it was found that of the 45 individuals who suffered a major CVD event, 13 belonged to cluster 1, and 32 to cluster 2.

4. Discussion

In this study, we analyzed the different trajectories of CVRFs and SCORE in a cohort of factory workers across three time points. The quantitative variables that underwent the greatest changes over the entire study period were total cholesterol, glucose, and SCORE. The mean SCORE increased, while the mean total cholesterol and glucose values decreased. These results are in line with those of our quartile analysis. It should be borne in mind, first, that the observed increase in SCORE could be mainly due to the increasing age. Second, the assessments of the individuals included in the cohort are periodically followed up

intensively by the factory medical services. Thus, the decrease in the total cholesterol and glucose levels, and the stabilization of blood pressure could be due to the close control of these patients, as these CVRFs are managed through pharmacological treatment. The BMI did not show significant changes, although more than half of all participants were overweight for the duration of the study. Another noteworthy finding was the decrease over time in the percentage of smokers.

Different authors [12,17] have reported the influence of age on calculating SCORE. Conroy et al. [12] found that SCORE was very low in people aged 30, and that it increased most rapidly between 50 and 65 years. In the present study, SCORE increased over time from 1.56 to 2.09. This increase was probably lower than expected, which may be due to the effect of the stabilization and improvement of some CVRFs. Thus, glucose and total cholesterol mean levels decreased from 97.7 mg/dl to 88.06 mg/dl, and from 212.18 mg/dl to 187.96 mg/dl, respectively, between the first and last time points, and the BMI mean remained stable over the three time points (27.61 at the first, versus 27.84 at the third).

To create clusters, we used an algorithm that divided the study cohort into groups based on the trajectories of different variables over time. This algorithm was applied to the mean CVRF values and SCORE over the three time points analyzed. The results showed that study participants could be divided into 2 clusters, based on the evolution of CVRF values and SCORE.

The first cluster consisted of younger individuals with lower mean blood glucose, BMI, and WC values, higher SCORE values, and higher mean HDL cholesterol values. The second cluster consisted of older individuals with higher mean blood glucose, BMI, and WC values, higher SCORE, and lower mean HDL cholesterol values.

The analysis of changes in quartiles measured by clusters revealed that the percentage of individuals who remained in quartiles with higher than recommended BMI, WC, and glucose values, and lower than recommended HDL cholesterol values, was higher in cluster 2 than in cluster 1. Furthermore, in both clusters this percentage increased over time for all variables except glucose, for which decreases over the three time points were observed. For SCORE, a similar evolution was observed in both clusters for individuals who began the study in Q2, Q3 or Q4. The greatest difference between clusters was observed for workers with a SCORE in Q1: the proportion of workers who remained in this quartile over the three time points was greater in cluster 1 than in cluster 2.

Few published studies have used clustering methodology similar to ours to analyze the evolution of CVRFs. Several studies have analyzed the trajectories of one [18–21] or several [22,23] CVRFs using a variety of different methods, and have attempted to identify correlations between their findings and other factors or diseases. One such study [19] analyzed the evolution of SBP, for which 4 distinct trajectories were identified over time. The authors found that SBP trajectories predicted CVD and all-cause mortality no better than did mean SBP values. Another study [21] of SBP and DBP in an elderly population identified 3 blood pressure (BP) trajectories. BP trajectories were also analyzed by Allen et al. [20], who identified 5 BP trajectories in a middle-aged population. Rospleszcz et al. [24] analyzed the association between CVRF trajectories and adipose tissue deposits using a methodology similar to ours and identified 3 distinct clusters. The first cluster grouped individuals with the youngest mean age and lowest mean CVRF values, and the third cluster grouped those with the highest mean age and highest mean CVRF values. Mean age and mean CVRF values in cluster 2 fell between those of clusters 1 and 3, except for total and HDL cholesterol, for which values were higher than the other 2 clusters. Finally, Norby et al. [22] used a mixture model to separately identify the trajectories of different CVRFs. Five distinct trajectories were identified for BMI, obesity, and SBP, and 4 for hypertension and diabetes.

Although the analysis of the incidence of CVD events was preliminary, more events were detected in cluster 2 than in cluster 1.

Our study has some limitations. First, the study population was exclusively male, owing to the low number of women in the AWHS cohort. Although it is not representative

of the general population, it represents workers well in this type of factory and in these age ranges, which represents an important part of the population. Second, despite the large study population, not all participants attended each of the 3 medical tests for which the study data were extracted. Moreover, the algorithm used does not tolerate missing data, and therefore we were obliged to impute data in some cases and eliminate data in cases in which imputation was not possible. Third, the selected study period was relatively short for a trajectory study. Fourth, due to the young age of the workers, the number of events detected in the preliminary study of the incidence of CVD events was low. Several strengths of our study should also be noted. The methodology used is based on a k-means clustering algorithm that can simultaneously consider different variables and time points (i.e. trajectories of multiple variables). This study presents clusters according to the evolution of different CVRFs at the same time, in contrast to previous studies aimed at creating clusters according to each CVRF separately. Furthermore, the data analyzed were extracted from multiple sources, and included multiple well-refined variables. Nevertheless, further studies conducted over longer time periods will be required to evaluate differences between clusters in the incidence of CVD events.

Finally, regarding the worker cohort, these results could help the development of personalized medicine by the factory medical services, to improve the cardiovascular health of the workers and to implement preventive measures.

5. Conclusions

Using clustering analysis, we found that our cohort could be divided into 2 groups. The profile of cluster 1 was a lower age, BMI, WC, blood glucose level, and SCORE, and higher HDL cholesterol levels. Cluster 2 consisted of individuals with higher BMI, WC, blood glucose levels, and SCORE, and lower HDL cholesterol levels. Finally, although no significant changes in CVRFs were detected during the study period, the worsening of CVRFs was greater in cluster 2 than in cluster 1. Individuals in both groups increased their SCORE, but this increase was greater in the ones in cluster 2, as their worsening of CVRFs was higher. Thus, the identification of subjects included in these profiles could facilitate the development of better, personalized medicine approaches to CVD treatment and preventive measures, especially in those profiles showing the worst CVRF control.

Author Contributions: Study design, M.J.R.-H. and S.C.-F.; formal analysis, J.-T.A.-N., L.M., M.J.R.-H. and S.C.-F.; data curation, B.M.-F., E.M.-V., I.A.-P. and S.M.; writing—original draft preparation, M.J.R.-H. and S.C.-F.; writing—review and editing, all authors.; funding acquisition, I.A.-P. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Proyecto del Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (Ministerio de Ciencia e Innovación) and the European Fund for Regional Development (FEDER) (PI17/01704).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Clinical Research Ethics Committee of Aragon (protocol code PI07/09 on 27 April 2016).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not public due to ethical reasons.

Acknowledgments: The authors thank the study participants, doctors and health-care professionals at the Opel factory, Zaragoza, as well as the AWHS technical staff for their participation and work.

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

Appendix A

Table A1. Results of the quartile analysis. Values represent the percentage of individuals that moved from one quartile to another between time points 1 and 2 and time points 2 and 3 for the following variables: body mass index, blood glucose, and cardiovascular disease risk score.

BODY MASS INDEX

		Time Point 1						Time Point 2				
		Q. 1	Q. 2	Q. 3	Q. 4	*p*		Q. 1	Q. 2	Q. 3	Q. 4	*p*
		N (%)	N (%)	N (%)	N (%)			N (%)	N (%)	N (%)	N (%)	
Time point 2	Q. 1	814 (79.0%)	143 (13.8%)	15 (1.4%)	2 (0.2%)	<0.001	Time point 3	819 (84.1%)	156 (14.8%)	11 (1.07%)	1 (0.1%)	<0.001
	Q. 2	203 (19.7%)	660 (63.9%)	182 (17.3%)	10 (1.0%)			140 (14.4%)	679 (64.4%)	155 (15.1%)	9 (0.8%)	
	Q. 3	14 (1.4%)	221 (21.4%)	658 (62.7%)	134 (13.0%)			13 (1.3%)	214 (20.3%)	678 (66.0%)	109 (10.0%)	
	Q. 4	0 (0.0%)	9 (0.9%)	194 (18.5%)	888 (85.9%)			2 (0.2%)	6 (0.6%)	183 (17.8%)	972 (89.1%)	

BLOOD GLUCOSE

		Time point 1						Time point 2				
		Q. 1	Q. 2	Q. 3	Q. 4	*p*		Q. 1	Q. 2	Q. 3	Q. 4	*p*
		N (%)	N (%)	N (%)	N (%)			N (%)	N (%)	N (%)	N (%)	
Time point 2	Q. 1	713 (64.0%)	356 (35.2%)	180 (16.8%)	49 (5.2%)	<0.001	Time point 3	1103 (85.0%)	790 (74.0%)	458 (45.8%)	107 (13.7%)	<0.001
	Q. 2	260 (23.3%)	351 (34.8%)	359 (33.5%)	98 (10.3%)			140 (10.8%)	181 (16.9%)	279 (27.9%)	113 (14.5%)	
	Q. 3	106 (9.5%)	239 (23.7%)	372 (34.7%)	282 (29.7%)			40 (3.1%)	72 (6.7%)	169 (16.9%)	158 (20.2%)	
	Q. 4	35 (3.1%)	64 (6.3%)	161 (15.0%)	522 (54.9%)			15 (1.2%)	25 (2.3%)	93 (9.3%)	404 (51.7%)	

CARDIOVASCULAR DISEASE RISK SCORE

		Time point 1						Time point 2				
		Q. 1	Q. 2	Q. 3	Q. 4	*p*		Q. 1	Q. 2	Q. 3	Q. 4	*p*
		N(%)	N(%)	N(%)	N(%)			N(%)	N(%)	N(%)	N(%)	
Time point 2	Q. 1	722 (69.6%)	15 (1.5%)	0 (0.0%)	0 (0.0%)	<0.001	Time point 3	618 (83.9%)	28 (4.0%)	6 (0.5%)	10 (0.6%)	<0.001
	Q. 2	272 (26.2%)	360 (34.8%)	65 (6.3%)	9 (0.9%)			115 (15.6%)	353 (50.0%)	85 (7.5%)	18 (1.2%)	
	Q. 3	38 (3.7%)	544 (52.6%)	453 (43.8%)	108 (10.4%)			3 (0.4%)	296 (41.9%)	636 (55.6%)	154 (9.9%)	
	Q. 4	6 (0.6%)	116 (11.2%)	517 (50.0%)	922 (88.7%)			1 (0.1%)	29 (4.1%)	416 (36.4%)	1379 (88.3%)	

Abbreviations: Q, quartile; N, number; *p*, *p*-value (Chi-squared and Mann-Whitney U-test).

Table A2. Results of the quartile analysis by clusters. Values represent the percentage of individuals that moved from one quartile to another between time points 1 and 2 and time points 2 and 3 for the following variables: body mass index, blood glucose, and cardiovascular disease risk score.

BODY MASS INDEX

		Time Point 1						Time Point 2				
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*
Time point 2	**Cluster 1**					<0.001	Time point 3	**Cluster 1**				<0.001
	Q. 1	789 (80.4%)	125 (17.2%)	10 (3.0%)	1 (1.7%)			Q. 1	786 (85.0%)	126 (16.3%)	3 (0.8%)	0 (0.0%)
	Q. 2	179 (18.2%)	481 (66.1%)	106 (31.9%)	5 (8.6%)			Q. 2	127 (13.7%)	515 (66.8%)	76 (22.1%)	4 (6.8%)
	Q. 3	13 (1.3%)	119 (16.3%)	187 (56.3%)	25 (43.1%)			Q. 3	11 (1.2%)	126 (16.3%)	220 (64.0%)	18 (30.5%)
	Q. 4	0 (0.0%)	3 (0.4%)	29 (8.7%)	27 (46.6%)			Q. 4	1 (0.1%)	4 (0.5%)	45 (13.1%)	37 (62.7%)
Time point 2	**Cluster 2**					<0.001	Time point 3	**Cluster 2**				<0.001
	Q. 1	25 (50.0%)	18 (5.9%)	5 (0.7%)	1 (0.1%)			Q. 1	33 (67.3%)	30 (10.6%)	8 (1.2%)	1 (0.1%)
	Q. 2	24 (48.0%)	179 (58.7%)	76 (10.6%)	5 (0.5%)			Q. 2	13 (26.5%)	164 (57.7%)	79 (11.6%)	5 (0.5%)
	Q. 3	1 (2.0%)	102 (33.4%)	471 (65.7%)	109 (11.2%)			Q. 3	2 (4.1%)	88 (31.0%)	458 (67.1%)	91 (8.8%)
	Q. 4	0 (0.0%)	6 (2.0%)	165 (23.0%)	861 (88.2%)			Q. 4	1 (2.0%)	2 (0.7%)	138 (20.2%)	935 (90.6%)

BLOOD GLUCOSE

		Time point 1						Time point 2				
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*
Time point 2	**Cluster 1**					<0.001	Time point 3	**Cluster 1**				<0.001
	Q. 1	487 (67.5%)	249 (39.6%)	96 (19.5%)	27 (10.5%)			Q. 1	765 (89.1%)	520 (81.4%)	257 (58.4%)	40 (24.8%)
	Q. 2	172 (23.9%)	234 (37.2%)	193 (39.2%)	40 (15.6%)			Q. 2	67 (7.8%)	92 (14.4%)	108 (24.5%)	37 (23.0%)
	Q. 3	52 (7.2%)	133 (21.1%)	162 (32.9%)	93 (36.2%)			Q. 3	19 (2.2%)	23 (3.6%)	56 (12.7%)	35 (21.7%)
	Q. 4	10 (1.4%)	13 (2.1%)	41 (8.3%)	97 (37.7%)			Q. 4	8 (0.9%)	4 (0.6%)	19 (4.3%)	49 (30.4%)
Time point 2	**Cluster 2**					<0.001	Time point 3	**Cluster 2**				<0.001
	Q. 1	226 (57.5%)	107 (28.1%)	84 (14.5%)	22 (3.2%)			Q. 1	338 (77.0%)	270 (62.9%)	201 (36.0%)	67 (10.8%)
	Q. 2	88 (22.4%)	117 (30.7%)	166 (28.6%)	58 (8.4%)			Q. 2	73 (16.6%)	89 (20.7%)	171 (30.6%)	76 (12.2%)
	Q. 3	54 (13.7%)	106 (27.8%)	210 (36.2%)	189 (27.2%)			Q. 3	21 (4.8%)	49 (11.4%)	113 (20.2%)	123 (19.8%)
	Q. 4	25 (6.4%)	51 (13.4%)	120 (20.7%)	425 (61.2%)			Q. 4	7 (1.6%)	21 (4.9%)	74 (13.2%)	355 (57.2%)

Table A2. *Cont.*

CARDIOVASCULAR DISEASE RISK SCORE												
		Time point 1						Time point 2				
		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*		Q. 1 N (%)	Q. 2 N (%)	Q. 3 N (%)	Q. 4 N (%)	*p*
Time point 2	**Cluster 1**					<0.001	**Cluster 1**					<0.001
	Q. 1	657 (74.7%)	10 (1.7%)	0 (0.0%)	0 (0.0%)		Time point 3 Q. 1	578 (86.7%)	19 (4.2%)	1 (0.2%)	2 (0.5%)	
	Q. 2	200 (22.7%)	215 (38.0%)	31 (7.8%)	4 (1.6%)		Q. 2	86 (12.9%)	241 (53.6%)	52 (9.4%)	3 (0.7%)	
	Q. 3	23 (2.6%)	298 (52.7%)	192 (48.2%)	40 (15.7%)		Q. 3	2 (0.3%)	177 (39.3%)	322 (58.2%)	54 (12.6%)	
	Q. 4	0 (0.0%)	43 (7.6%)	175 (44.0%)	211 (82.7%)		Q. 4	1 (0.2%)	13 (2.9%)	178 (32.2%)	370 (86.2%)	
Time point 2	**Cluster 2**					<0.001	Time point 3 **Cluster 2**					<0.001
	Q. 1	65 (41.1%)	5 (1.1%)	0 (0.0%)	0 (0.0%)		Q. 1	40 (57.1%)	9 (3.5%)	5 (0.9%)	8 (0.7%)	
	Q. 2	72 (45.6%)	145 (30.9%)	34 (5.3%)	5 (0.6%)		Q. 2	29 (41.4%)	112 (43.8%)	33 (5.6%)	15 (1.3%)	
	Q. 3	15 (9.5%)	246 (52.5%)	261 (41.0%)	68 (8.7%)		Q. 3	1 (1.4%)	119 (46.5%)	314 (53.2%)	100 (8.8%)	
	Q. 4	6 (3.8%)	73 (15.6%)	342 (53.7%)	711 (90.7%)		Q. 4	0 (0.0%)	16 (6.3%)	238 (40.3%)	1009 (89.1%)	

Abbreviations: Q, quartile; N, number; *p*, *p*-value (Chi-squared and Mann-Whitney U-test).

References

1. WHO. Enfermedades Cardiovasculares (cvds). Available online: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases- (accessed on 10 December 2020).
2. Roth, G.A.; Mensah, G.A.; Johnson, C.O.; Addolorato, G.; Ammirati, E.; Baddour, L.M.; Barengo, N.C.; Benjamin, E.J.; Benziger, C.P.; Bonny, A.; et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. In *JACC*; 2020; Volume 76, pp. 2982–3021. [CrossRef] [PubMed]
3. Townsend, N.; Wilson, L.; Bhatnagar, P.; Wickramasinghe, K.; Rayner, M.; Nichols, M. Cardiovascular Disease in Europe: Epidemiological Update 2016. *Eur. Heart J.* **2016**, *37*, 3232–3245. [CrossRef] [PubMed]
4. Balakumar, P.; Maung-U, K.; Jagadeesh, G. Prevalence and Prevention of Cardiovascular Disease and Diabetes Mellitus. *Pharmacol. Res.* **2016**, *113*, 600–609. [CrossRef] [PubMed]
5. Shirasawa, T.; Ochiai, H.; Yoshimoto, T.; Nagahama, S.; Kobayashi, M.; Ohtsu, I.; Sunaga, Y.; Kokaze, A. Associations between Normal Weight Central Obesity and Cardiovascular Disease Risk Factors in Japanese Middle-Aged Adults: A Cross-Sectional Study. *J. Health Popul. Nutr.* **2019**, *38*, 1–7. [CrossRef] [PubMed]
6. Yusuf, S.; Joseph, P.; Rangarajan, S.; Islam, S.; Mente, A.; Hystad, P.; Brauer, M.; Kutty, V.R.; Gupta, R.; Wielgosz, A.; et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): A prospective cohort study. *Lancet* **2020**, *395*, 795–808. [CrossRef]
7. San Sebastián, M.; Mosquera, P.A.; Gustafsson, P.E. Do Cardiovascular Disease Prevention Programs in Northern Sweden Impact on Population Health? An Interrupted Time Series Analysis. *BMC Public Health.* **2019**, *19*, 1–10. [CrossRef] [PubMed]
8. Pennant, M.; Davenport, C.; Bayliss, S.; Greenheld, W.; Marshall, T.; Hyde, C. Community Programs for the Prevention of Cardiovascular Disease: A Systematic Review. *Am. J. Epidemiol.* **2010**, *172*, 501–516. [CrossRef] [PubMed]
9. Peña, D. *Análisis de Datos Multivariantes*; McGraw-Hill: Madrid, Spain, 2002.
10. Genolini, C.; Alacoque, X.; Sentenac, M.; Arnaud, C. Kml and Kml3d: R Packages to Cluster Longitudinal Data. *J. Stat. Softw.* **2015**, *65*, 1–34. [CrossRef]
11. Casasnovas, J.A.; Alcaide, V.; Civeira, F.; Guallar, E.; Ibañez, B.; Borreguero, J.J.; Laclaustra, M.; León, M.; Peñalvo, J.L.; Ordovás, J.M.; et al. Aragon Workers' Health Study-Design and Cohort Description. *BMC Cardiovasc. Disord.* **2012**, *12*, 1–11. [CrossRef] [PubMed]
12. Conroy, R.M.; Pyörälä, K.; Fitzgerald, A.P.; Sans, S.; Menotti, A.; De Backer, G.; De Bacquer, D.; Ducimetière, P.; Jousilahti, P.; Keil, U.; et al. Estimation of Ten-Year Risk of Fatal Cardiovascular Disease in Europe: The SCORE Project. *Eur. Heart J.* **2003**, *24*, 987–1003. [CrossRef]
13. Everitt, B.; Landau, L.; Leese, M. *Cluster Analysis*, 4th ed.; Hodder Edwar Arnold: London, UK, 2001.
14. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
15. Genolini, C.; Falissard, B.; Genolini, C.; Falissard, B.; Falissard, B. KmL: K-Means for Longitudinal Data. *Comput. Stat.* **2010**, *25*, 317–328. [CrossRef]
16. Genolini, C.; Falissard, B.; Fang, D.; Tierney, L. Package LongitudinalData: Longitudinal Data; R Package Version 2.4.1; 2016. Available online: https://CRAN.R-project.org/package=longitudinalData (accessed on 10 December 2020).
17. Redon, J. Global Cardiovascular Risk Assessment: Strengths and Limitations. *High Blood Press. Cardiovasc. Prev.* **2016**, *23*, 87–90. [CrossRef] [PubMed]
18. Lin, H.; Cui, M.; Spatz, E.S.; Wang, Y.; Lu, J.; Li, J.; Li, S.; Huang, C.; Liu, X.; Jiang, L.; et al. Heterogeneity in Trajectories of Systolic Blood Pressure among Young Adults in Qingdao Port Cardiovascular Health Study. *Glob. Heart.* **2020**, *15*, 9–11. [CrossRef] [PubMed]
19. Tielemans, S.M.A.J.; Geleijnse, J.M.; Laughlin, G.A.; Boshuizen, H.C.; Barrett-Connor, E.; Kromhout, D. Blood Pressure Trajectories in Relation to Cardiovascular Mortality: The Rancho Bernardo Study. *J. Hum. Hypertens.* **2017**, *31*, 515–519. [CrossRef] [PubMed]
20. Allen, N.B.; Siddique, J.; Wilkins, J.T.; Shay, C.; Lewis, C.E.; Goff, D.C.; Jacobs, D.R.; Liu, K.; Lloyd-Jones, D. Blood Pressure Trajectories in Early Adulthood and Subclinical Atherosclerosis in Middle Age. *JAMA J. Am. Med. Assoc.* **2014**, *311*, 490–497. [CrossRef] [PubMed]
21. Smitson, C.C.; Scherzer, R.; Shlipak, M.G.; Psaty, B.M.; Newman, A.B.; Sarnak, M.J.; Odden, M.C.; Peralta, C.A. Association of Blood Pressure Trajectory with Mortality, Incident Cardiovascular Disease, and Heart Failure in the Cardiovascular Health Study. *Am. J. Hypertens.* **2017**, *30*, 587–593. [CrossRef] [PubMed]
22. Norby, F.L.; Soliman, E.Z.; Chen, L.Y.; Bengtson, L.G.S.; Loehr, L.R.; Agarwal, S.K.; Alonso, A. Trajectories of Cardiovascular Risk Factors and Incidence of Atrial Fibrillation over a 25-Year Follow-Up. *Circulation* **2016**, *134*, 599–610. [CrossRef] [PubMed]
23. Pebesma, J.; Martinez-Millana, A.; Sacchi, L.; Fernandez-Llatas, C.; De Cata, P.; Chiovato, L.; Bellazzi, R.; Traver, V. Clustering Cardiovascular Risk Trajectories of Patients with Type 2 Diabetes Using Process Mining. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS.* **2019**, 341–344. [CrossRef]
24. Rospleszcz, S.; Lorbeer, R.; Storz, C.; Schlett, C.L.; Meisinger, C.; Thorand, B.; Rathmann, W.; Bamberg, F.; Lieb, W.; Peters, A. Association of Longitudinal Risk Profile Trajectory Clusters with Adipose Tissue Depots Measured by Magnetic Resonance Imaging. *Sci. Rep.* **2019**, *9*, 1–12. [CrossRef] [PubMed]

ANEXO III

PREDICTION OF CARDIOVASCULAR RISK USING MACHINE-LEARNING METHODS AND SEX-SPECIFIC DIFFERENCES

Authors: Sara Castel-Feced [a,b,c,*] , Isabel Aguilar-Palacio [a,b,c], Sara Malo [a,b,c], Juan González-García[d], Lina Maldonado [c,e,†], María José Rabanaque-Hernández [a,b,c,†]

[a] Department of Preventive Medicine and Public Health, University of Zaragoza, Zaragoza 50009, Spain; iaguilar@unizar.es (I.A.-P.), smalo@unizar.es (S.M), rabanake@unizar.es (M. J.R.-H.)

[b] Fundación Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza 50009, Spain.

[c] GRISSA Research Group, Zaragoza 50009, Spain.

[d] Instituto Aragonés Ciencias de la Salud

[e] Department of Economic Structure, Economic History and Public Economics, University of Zaragoza, Zaragoza 50005, Spain; lmguaje@unizar.es (L.M.)

*Correspondence: scastelf@unizar.es

[†]These authors contributed equally to this work and served as senior co-authors.

ABSTRACT

Machine learning (ML) algorithms offer many advantages over traditional scoring systems to assess the influence of cardiovascular risk factors (CVRFs) on the risk of cardiovascular event (CVE), and are better suited to personalized medicine approaches for cardiovascular prevention. These algorithms can also be trained using a growing body of real world data (RWD). Applying the XG Boost and Random Forest ML methods to RWD, we evaluated the outcomes of these two algorithms for CVE risk prediction using different combinations of predictive variables and analysed the influence of distinct CVR-related variables on CVE prediction, stratifying the study population by sex. For each algorithm, we generated 3 models using distinct combinations of variables: (1) age, blood test and blood pressure measurements, CVRFs, and medication adherence; (2) age, blood test and blood pressure measurements, and medication adherence; (3) age, CVRFs, and medication adherence. In all models age was the greatest relative contributor to the risk of CVE, followed by adherence to antidiabetics. Treatment adherence was also identified as a major contributor to the risk of CVE. These algorithms could be used to create models for specific populations and applied in primary care settings to manage interventions in a personalized medicine context.

INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death and disability. It is estimated that CVD accounts for approximately 17.9 million deaths per year, and one third of these deaths occur prematurely in people aged less than 70 years[1]. Some cardiovascular risk factors (CVRF) can be controlled, and CVD prevention guidelines highlight the importance of early diagnosis and intervention in high-risk individuals to prevent CVD mortality and morbidity[2]. Lifestyle changes are among the most important primary prevention interventions. If these prove insufficient, pharmacological preventive treatment selected according to the individual's overall CV risk is indicated.

Different risk estimation tools are widely applied and recommended by CVD prevention guidelines to identify at-risk individuals who should be targeted for primary prevention, both pharmacological and behavioural. These tools, which include the Framingham Risk Score and the Systematic COronary Risk Evaluation (SCORE), estimate the individual's global CV risk based on the individual contributions of multiple CVRFs, but are not without their limitations[3–6]. First, these tools have been developed for specific populations, and therefore have limited generalizability to predict risk in other populations and countries. Second, methodological limitations of these approaches include (i) the fact that they are based on simple regression fitting approaches and cannot assume a nonlinear relationship between predictors and outcomes; (ii) correlation between variables, and (iii) the risk of overfitting. Moreover, although pharmacological treatment and its adherence are related to CV risk[7–9], the aforementioned tools do not consider whether subjects are being treated for any CVRF or whether they correctly adhere to their treatment. While this issue can be addressed thanks to the growing availability of medical data generated in daily clinical practice[10], incorporation of these data in this context remains challenging, and requires an initial debugging process. Finally, the incidence of CVRFs, and how they interact and are controlled, differs between men and women[2,11–14]and therefore CVRFs should be analysed separately for each sex.

To improve the accuracy of traditional systems to overcome some of the aforementioned limitations, machine learning (ML) techniques have been applied and tested in several cohorts to identify individuals with high CV risk [6,15–18]. ML techniques use existing medical data obtained from daily clinical practice to train models to learn patterns that are

later applied to the prediction of other variables. The techniques used include ensemble methods, which enable a kind of supervised ML, and include bagging and boosting methods that combine multiple decision trees to reach a decision[19]. One of the most commonly used bagging methods is Random Forest (RF), whereby multiple decision trees are learned in parallel and the final prediction is based on the most frequent answer[15,19]. Boosting models also train multiple individual models, in this case sequentially, and each model seeks to correct the mistakes of the previous ensemble model. Two well-known boosting methods are AdaBoost and XG Boost.

ML techniques have shown great promise in calculating CVD risk in different cohorts, improving upon the results obtained using traditional scoring methods. In this study, we compared the prediction of CV risk using ML methods applied in men and women together versus separately, and analysed the influence of different traditional CVRFs together with medication adherence when included in these algorithms.

Las técnicas de ML han demostrado ser muy prometedoras para calcular el riesgo de ECV en diferentes cohortes, mejorando los resultados obtenidos con los métodos de puntuación tradicionales. En este estudio, comparamos la predicción del riesgo CV mediante métodos de ML aplicados en hombres y mujeres juntos frente a por separado, y analizamos la influencia de diferentes FRCV tradicionales junto con la adherencia a la medicación cuando se incluyen en estos algoritmos.

METHODS

STUDY COHORT AND DATA SOURCE

This longitudinal cohort study was conducted using the CARhES cohort. This dynamic open cohort has been followed since 2017, and includes all individuals aged 16 and above registered as users of the public health system in Aragón, a Spanish region with about 1.3 million inhabitants that are overwhelmingly attended to by the public health system. Participants had at least one of the following CVRFs: hypertension, hypercholesterolaemia, or diabetes mellitus (DM). CVRFs were identified based on a medical diagnosis of hypertension, DM, or hypercholesterolaemia and/ or a prescription of at least one antihypertensive, antidiabetic, or lipid-lowering drug during the study

period. The CARhES cohort was established in 2017 and consisted of 446,998 individuals (50.64% female), of whom 252,508 had hypertension (56.5%), 332,644 had hypercholesterolaemia (74.4%), and 96,709 had DM (21.6%).

All information necessary to identify patients who met the inclusion criteria was obtained from BIGAN[20], a health data hub that gathers data from the Aragon public health service and makes this information available for research purposes upon request. Data from this cohort were stored in several databases: the BDU (health system users database), which provides information on age and affiliation to the Aragón public health system; the minimum basic dataset database, which gathers data on hospital discharge; the primary care database, which records information from patients who attend a primary health care centre; GMA (morbidity adjusted groups), which records information on all medical diagnoses available in primary healthcare and in the minimum basic dataset database; the emergency database, which stores diagnostic and procedural information on patients processed via the hospital emergency system; the electronic prescribing system database, which records all pharmacological treatments prescribed to patients; and the pharmacy claims database, which gathers information about medication dispensed in pharmacies to each patient. All data in these databases are pseudonymized using a unique code that links patient information across the different data sources but prevents personal identification.

The GMA database was queried to identify subjects with a medical diagnosis corresponding to any of the 3 CVRFs of interest. Pharmacological treatments that corresponded to the following ATC codes and were prescribed to patients were extracted from the electronic prescribing system database: A10 (diabetes); C02, C03, C07, C08, and C09 (hypertension); and C10 (hypercholesterolaemia).

INCLUSION AND EXCLUSION CRITERIA

The process of selecting patients from the CARhES cohort to participate in the present study is depicted in Figure 1. First, as we focused on subjects with primary prevention, we excluded those with a diagnosis of major cardiovascular event (MACE) in the minimum basic dataset database in 2016 or 2017 and/or in the GMA during 2017 (as GMA data were not available in 2016).

Also excluded were patients who died during the follow-up period and for whom MACE was not recorded as the cause of the death. Next, we identified subjects in primary prevention who had experienced MACE between 2018 and 2020. Of those who had, subjects who began treatment corresponding to any of the three pharmacological groups of interest during the year preceding the event were excluded. Of those who did not experience MACE during the follow-up period, we excluded those who began treatment during 2018. Finally, among those who experienced MACE we included those for whom blood test and blood pressure data were available for the year preceding the event and, among those who did not experience CVE, we included those for whom blood test and blood pressure data were available for the year 2018.

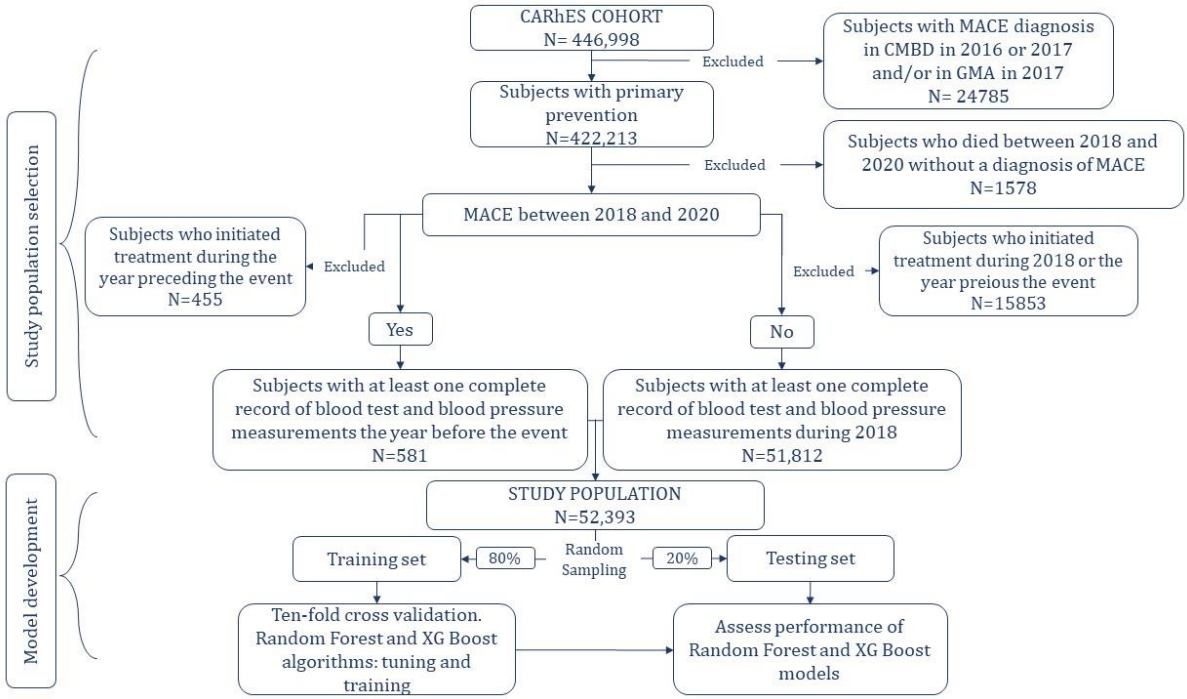


Figure 1: Study population selection and model development

STUDY VARIABLES

Variables included in the present study were age, blood test- and blood pressure-related parameters, CVRFs, and adherence to medication taken for CVRFs. The variables included and the corresponding data sources are summarized in Table 1.

Adherence to antihypertensive, antidiabetic, or lipid-lowering drugs was calculated separately for each subject using the Proportion of Days Covered (PDC). Adherence was calculated as a percentage using data from the year 2018 for those who did not experience an event, and using data from the year preceding the event in all other cases. PDC is an index calculated as the number of days covered by the medicines dispensed by the pharmacy divided by the number of days that the subject should have had covered. In this study, the denominator for PDC was 365 days. The number of days covered were calculated based on the Defined Daily Dose (DDD) dispensed to each subject. However, a previous study by our group[9] showed that use of a surrogate value for the daily dose of each drug, calculated based on the usual dosage and form of presentation, provided more accurate results. Therefore, in the present study surrogate values for daily doses were used. For example, for statins we always used a DDD of 28 rather than the value of 37.3 used in other studies.

Table 1: Variables included in the study

GROUP	VARIABLE	DATABASE	VALUES/ UNITS
Age	Age	BDU	years
CVRFs	Hypertension	GMA	Yes/no
	Hypercholesterolaemia	GMA + Electronic prescribing system	Yes/no
	Diabetes mellitus	GMA+ Prescription	Yes/no
Blood test and blood pressure measurement	Total cholesterol	Primary care database	mg/dL
	HDL-cholesterol	Primary care database	mg/dL
	LDL-cholesterol	Primary care database	mg/dL
	Blood glucose	Primary care database	mg/dL
	Systolic blood pressure	Primary care database	mm Hg
	Diastolic blood pressure	Primary care database	mm Hg
Medication adherence	Adherence to antihypertensives	Pharmacy claims	%
	Adherence to lipid-lowering drugs	Pharmacy claims	%
	Adherence to antidiabetics	Pharmacy claims	%

The primary outcome in this study was MACE incidence during follow up. Episodes were identified in the minimum basic dataset database and the emergency database. An episode was considered a MACE if the first diagnosis in the minimum basic dataset database corresponded to one of the following ICD-10 codes: I21, I60-I63, corresponding to myocardial infarction, nontraumatic subarachnoid haemorrhage, nontraumatic intracerebral haemorrhage, other nontraumatic intracranial haemorrhage, and acute ischemic stroke, respectively. In the emergency database, episodes considered MACE

were those with the same diagnosis, corresponding to ICD-9 codes 410 and 430-433, and that caused death.

ANALYSIS

Random Forest (RF) and XG Boost were used to determine the utility of different variables to predict the risk of MACE. Both were applied separately for men and women to age plus 3 different groups of variables:

- Model 1: Age, blood test and blood pressure measurement, cardiovascular risk factors, and medication adherence.
- Model 2: Age, blood test and blood pressure measurement, and medication adherence.
- Model 3: Age, cardiovascular risk factors, and medication adherence.

As shown in Figure 1, and as usually done when using these techniques, the study population was randomly split into two groups: 80% of the sample was assigned to the training group and the remaining 20% to the testing group. To train and tune the models 10-fold cross validation was applied to the training dataset to avoid overfitting. For both algorithms, hyperparameters were determined using a grid search in the 10-fold cross validation of the training set to identify values that led to optimal performance.

When event incidence is low, data are considered to be imbalanced. We observed a MACE incidence of 1.12%, indicating that the data were highly imbalanced. To resolve this problem, the Random Over Sampling Examples (ROSE) method with replacement was used to oversample the minority class and balance the data in the training set. To avoid poor estimates of model performance, the resampling process was applied to each of the 10 subsamples created during the cross-validation process irrespective of the other subsamples.

The performance of the models was assessed using the test set, and Youden's Index used to establish the optimal threshold for classification. In cases of imbalanced data, certain measures such as accuracy, positive predictive value, and negative predictive value can be markedly altered. Therefore, to assess the performance of the models created we calculated four distinct parameters: (i) AUC, which provides information about

the accuracy of the model; (ii) F1 score, which reflects the ability of the model to capture sensitivity and precision (i.e. to be accurate in the cases that it does capture); (iii) sensitivity, which indicates the proportion of cases classified as at high risk of an event; (iv) and specificity, which reflects the proportion of non-cases classified as such. Finally, the contribution of each variable to the prediction was extracted and standardized using a scale of 0–1 for ease of comparability.

ETHICAL ISSUES

All collected data were anonymized. The present study was approved by the Clinical Research Ethics Committee of Aragon (CEICA) in 2021 (project identification code PI21/148).

FUNDING

This study was supported by Proyecto del Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (ISCIII), "PI22/01193", co-funded by the European Union. It was also partly supported by the Gobierno de Aragón with a grant for postgraduate research contracts (IIU/796/2019).

RESULTS

Descriptive analysis of total population and by sex

Of the 52,393 individuals included in the present study, 57.3% were women, and the mean age was 70.2 years. Female participants were older than their male counterparts (mean age, 71.6 and 68.3 years, respectively) (Table 1).

For both sexes, the most prevalent CVRF was hypertension, followed by hypercholesterolaemia. The proportions of individuals with 1, 2, and 3 CVRFs were similar in both sexes. Around 40% of participants had just one CVRF.

Mean values of total, HDL, and LDL cholesterol were higher in women than men, and mean blood glucose, systolic blood pressure (SBP), and diastolic blood (DBP) pressure were higher in men than women.

Adherence to treatment was highest for antihypertensives and lowest for antidiabetics, both in the total population and after stratifying by sex. For antidiabetics and lipid-lowering drugs, men showed higher mean adherence, but also greater dispersion. For antihypertensives, mean adherence was higher in women but dispersion higher in men.

A MACE was experienced by 581 (1.1%) participants: 282 men and 299 women. In 12 cases (8 men, 4 women) the event resulted in death. The most common MACE was stroke, representing 57.5% of all events, followed by myocardial infarction (MI) (26.2%). Stratifying by sex, stroke was more frequent in women than men (61.5% and 53.2%, respectively), while MI was more frequent in men than women (32.3% and 20.4%, respectively).

Table 1: Descriptive statistics for study population

Variables	Units	Total N=52,393	MEN N=22,383	WOMEN N=30,010	P value
Age	mean (SD)	70.2 (12.8)	68.3 (12.6)	71.6 (12.8)	<0.001
CARDIOVASCULAR RISK FACTORS					
DM	N (%)	14,181 (27.1)	7162 (32.0)	7019 (23.4)	<0.001
Hypertension	N (%)	38,253 (73.0)	15,964 (71.3)	22,289 (74.3)	<0.001
Hypercholesterolaemia	N (%)	37,316 (71.2)	15,877 (70.9)	21,439 (71.4)	0.209
Number of CVRFs	N (%)				<0.001
1		22,508 (43.0)	9406 (42.0)	13,102 (43.7)	
2		22,413 (42.8)	9334 (41.7)	13,079 (43.6)	
3		7472 (14.3)	3643 (16.3)	3829 (12.8)	
BLOOD TEST AND BLOOD PRESSURE MEASUREMENTS					
Total cholesterol levels (mg/dL)	mean (SD)	195 (36.1)	186 (35.5)	201 (35.1)	0.000
HDL cholesterol levels (mg/dL)	mean (SD)	53.7 (13.4)	48.8 (11.6)	57.3 (13.5)	0.000
LDL cholesterol levels (mg/dL)	mean (SD)	118 (31.5)	114 (31.7)	121 (31.1)	<0.001
Blood glucose levels (mg/dL)	mean (SD)	104 (24.8)	107 (26.5)	101 (23.1)	<0.001
Systolic blood pressure (mm Hg)	mean (SD)	133 (15.8)	134 (15.4)	133 (16.2)	<0.001
Diastolic blood pressure (mm Hg)	mean (SD)	76.8 (13.9)	77.8 (16.5)	76.0 (11.4)	<0.001
MEDICATION ADHERENCE, PDC					
Antihypertensives	mean (SD)	58.3 (44.0)	57.5 (44.7)	58.9 (43.6)	<0.001
Antidiabetics	mean (SD)	17.3 (33.4)	21.0 (36.1)	14.5 (31.0)	<0.001
Lipid-lowering drugs	mean (SD)	38.5 (42.0)	40.1 (42.5)	37.3 (41.5)	<0.001
MACE CHARACTERISTICS					
Frequency	N (%)	581 (1.1)	282 (1.3)	299 (1.0)	0.005
Diagnosis	N (%)				0.011
Myocardial infarction		152 (26.2)	91 (32.3)	61 (20.4)	
Nontraumatic subarachnoid haemorrhage		14 (2.4)	4 (1.4)	10 (3.3)	

Nontraumatic intracerebral haemorrhage	57 (9.8)	24 (8.5)	33 (11.0)
Other nontraumatic intracranial haemorrhage	24 (4.1)	13 (4.6)	11 (3.7)
Acute Ischemic stroke	334 (57.5)	150 (53.2)	184 (61.5)

SD, standard deviation; N, number; DM, diabetes mellitus; HDL, high density lipoprotein; LDL, low density lipoprotein; MACE, major cardiovascular event; PDC, proportion of days covered.

Characteristics of individuals with MACE

Of the total number of MACEs, 51% were experienced by women, although the incidence of MACE was higher in men. Mean age was higher among individuals who experienced a MACE: 78.9 and 70.1 years in individuals who did and did not experience MACE, respectively (Table 2).

The frequencies of DM and hypertension were higher among individuals who experienced a MACE. There were no significant differences in the proportion of patients with hypercholesterolaemia between individuals with or without MACE. Moreover, those who experienced MACE more frequently presented 2 or 3 CVRFs, and those who did not more frequently presented 1 CVRF.

We observed no difference in adherence to lipid-lowering drugs between individuals with or without a MACE. Those who did experience a MACE were more adherent to antihypertensive and antidiabetic drugs.

Table 2: Descriptive statistics for MACE

	Units	No MACE N=51,812	MACE N=581	P value
Age	mean (SD)	70.1 (12.8)	78.9 (9.92)	<0.001
Sex	N (%)			0.005
MEN		22,101 (42.7)	282 (48.5)	
WOMEN		29,711 (57.3)	299 (51.5)	
CARDIOVASCULAR RISK FACTORS				
Diabetes	N (%)	13,952 (26.9)	229 (39.4)	<0.001
Hypertension	N (%)	37,767 (72.9)	486 (83.6)	<0.001
Hypercholesterolaemia	N (%)	36,906 (71.2)	410 (70.6)	0.761
Number of CVRF	N (%)			<0.001
1		22,330 (43.1)	178 (30.6)	
2		22,151 (42.8)	262 (45.1)	
3		7331 (14.1)	141 (24.3)	
BLOOD TEST AND BLOOD PRESSURE TESTS				
Total cholesterol levels (mg/dL)	mean (SD)	195 (36.1)	187 (35.2)	<0.001
HDL cholesterol levels (mg/dL)	mean (SD)	53.7 (13.4)	50.9 (12.9)	<0.001
LDL cholesterol levels (mg/dL)	mean (SD)	118 (31.5)	111 (30.6)	<0.001
Blood glucose levels (mg/dL)	mean (SD)	104 (24.6)	109 (38.0)	<0.001
Systolic blood pressure (mm Hg)	mean (SD)	133 (15.8)	137 (16.6)	<0.001
Diastolic blood pressure (mg Hg)	mean (SD)	76.8 (13.9)	75.0 (10.7)	<0.001
MEDICATION ADHERENCE, PDC				
Antihypertensives	mean (SD)	58.2 (44.1)	66.2 (40.9)	<0.001
Antidiabetics	mean (SD)	17.2 (33.4)	27.0 (39.2)	<0.001
Lipid-lowering drugs	mean (SD)	38.5 (42.0)	38.7 (42.0)	0.908

MACE, major cardiovascular event; SD, standard deviation; N, number; DM, diabetes mellitus; HDL, high density lipoprotein; LDL, low density lipoprotein; PDC, proportion of days covered.

CARDIOVASCULAR RISK PREDICTION

MODELS BUILT WITH RANDOM FOREST

After stratification by sex using the RF method, accuracy (measured through AUC) was higher for women than men (Table 3). For women, the highest level of accuracy was obtained for model 3, when using CVRF variables together with adherence. However, for men, accuracy was slightly higher for model 1 and 2 than for model 3. Differences in accuracy between models were higher for women than for men.

Of the models built for men, model 3 provided the highest F1 score, sensitivity, and specificity, although its accuracy was the lowest. For women, the highest F1 score and sensitivity were achieved with model 3, while all models achieved a specificity of 0.75. As also observed for AUC, differences between F1 score, sensitivity, and specificity were smaller across models generated for the male versus female population.

Table 3: Performance metrics for Random Forest models

	AUC	Youden's index	F1 SCORE	SENSITIVITY	SPECIFICITY
			MEN		
MODEL 1	0.70	0.50	0.77	0.62	0.69
MODEL 2	0.70	0.52	0.76	0.61	0.71
MODEL 3	0.69	0.54	0.77	0.62	0.71
			WOMEN		
MODEL 1	0.77	0.64	0.71	0.66	0.75
MODEL 2	0.76	0.62	0.81	0.69	0.75
MODEL 3	0.79	0.53	0.84	0.72	0.75

Model 1 includes the variables age, CVRFs, adherence, and blood test and blood pressure measurements. Model 2 includes age, adherence, and blood test and blood pressure measurements. Model 3 includes age, CVRFs, and treatment adherence. Abbreviations: AUC, area under the curve; CVRF, cardiovascular risk factor.

Relative contributions of variables

In all RF models, for both men and women, age was the variable that contributed most to the risk of MACE (Figure 2).

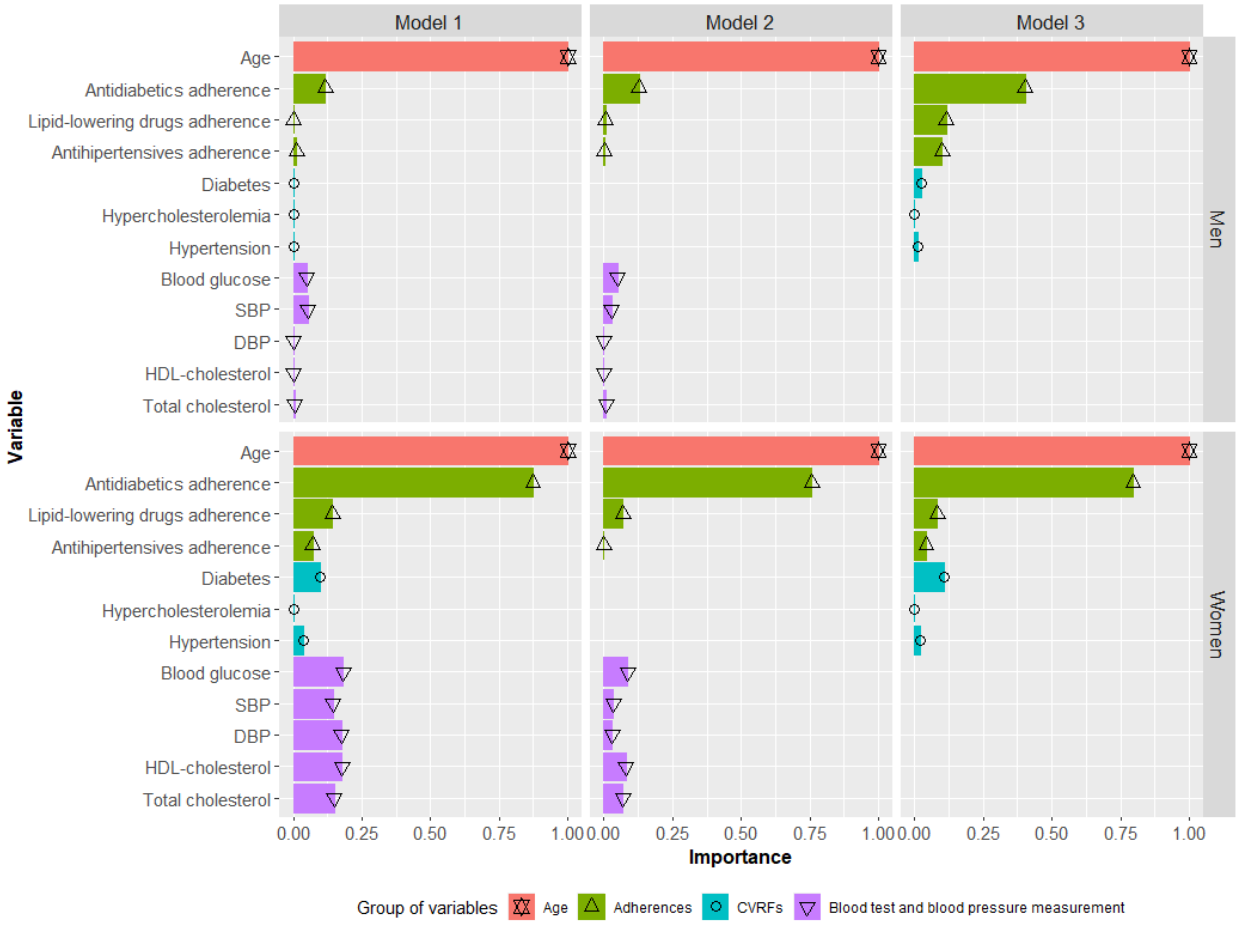


Figure 2: Relative contributions of variables in Random Forest models for men and women.
Model 1 includes the variables age, CVRFs, adherence, and blood test and blood pressure measurements. Model 2 includes age, adherence, and blood test and blood pressure measurements. Model 3 includes age, CVRFs, and treatment adherence.
Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL-cholesterol, high density lipoprotein cholesterol; CVRF, cardiovascular risk factor.

For men, age was the variable for which the greatest contribution to risk of MACE was observed, followed by antidiabetic adherence. The contribution of antidiabetic treatment adherence in models 1 and 2 was much smaller than the contribution of age. Furthermore, the contribution of antidiabetic treatment adherence in model 3 was higher than in models 1 and 2, but not as high as observed for women.

For women, in terms of relative contributions to the risk of MACE, age was closely followed by antidiabetic treatment adherence. All other variables contributed less. For model 1, blood test and pressure measurements variables were greater contributors than a diagnosis of hypertension, DM, or hypercholesterolaemia, and than adherence to lipid-lowering drugs or hypertension.

MODELS BUILT WITH XG BOOST

Models created for the male population using XG Boost achieved levels of accuracy comparable to those of RF models. F1 score and sensitivity were higher than those obtained with RF models, while specificity was lower for models 2 and 3 and higher for model 1.

In models created for the female population using XG Boost (Table 4), accuracy was comparable to that of RF models for models 2 and 3, while AUC was lower for model 1 relative to the corresponding RF model. F1 score and sensitivity were highest for model 1, while specificity was highest for model 2. Compared with the corresponding RF model, F1 score and sensitivity, but not specificity, were higher in XG Boost model 1.

Table4: Performance metrics for XG Boost models

	AUC	Youden's index	F1 SCORE	SENSITIVITY	SPECIFICITY
			MEN		
MODEL 1	0.70	0.53	0.78	0.64	0.71
MODEL 2	0.70	0.51	0.79	0.65	0.68
MODEL 3	0.69	0.52	0.79	0.65	0.66
			WOMEN		
MODEL 1	0.74	0.58	0.89	0.80	0.56
MODEL 2	0.76	0.54	0.80	0.67	0.81
MODEL 3	0.79	0.50	0.81	0.69	0.78

Model 1 includes the variables age, CVRFs, treatment adherence, and blood test and blood pressure measurements. Model 2 includes age, treatment adherence, and blood test and blood pressure measurements. Model 3 includes age, CVRFs, and treatment adherence. Abbreviations: AUC, area under the curve; CVRF, cardiovascular risk factor.

Relative contributions of variables

In models built using XG Boost, for both men and women, the variables that contributed most to a predicted high risk of CVE were age followed by antidiabetic treatment adherence (Figure 3). For men, in XG Boost model 1, similar contributions were observed for DM and hypercholesterolaemia and for blood glucose and SBP. For women, contrary to that which was observed for RF models, in XG Boost model 1 DM was a much more important contributor than blood tests and blood pressure measurements, with an effect similar to that of antidiabetic adherence.

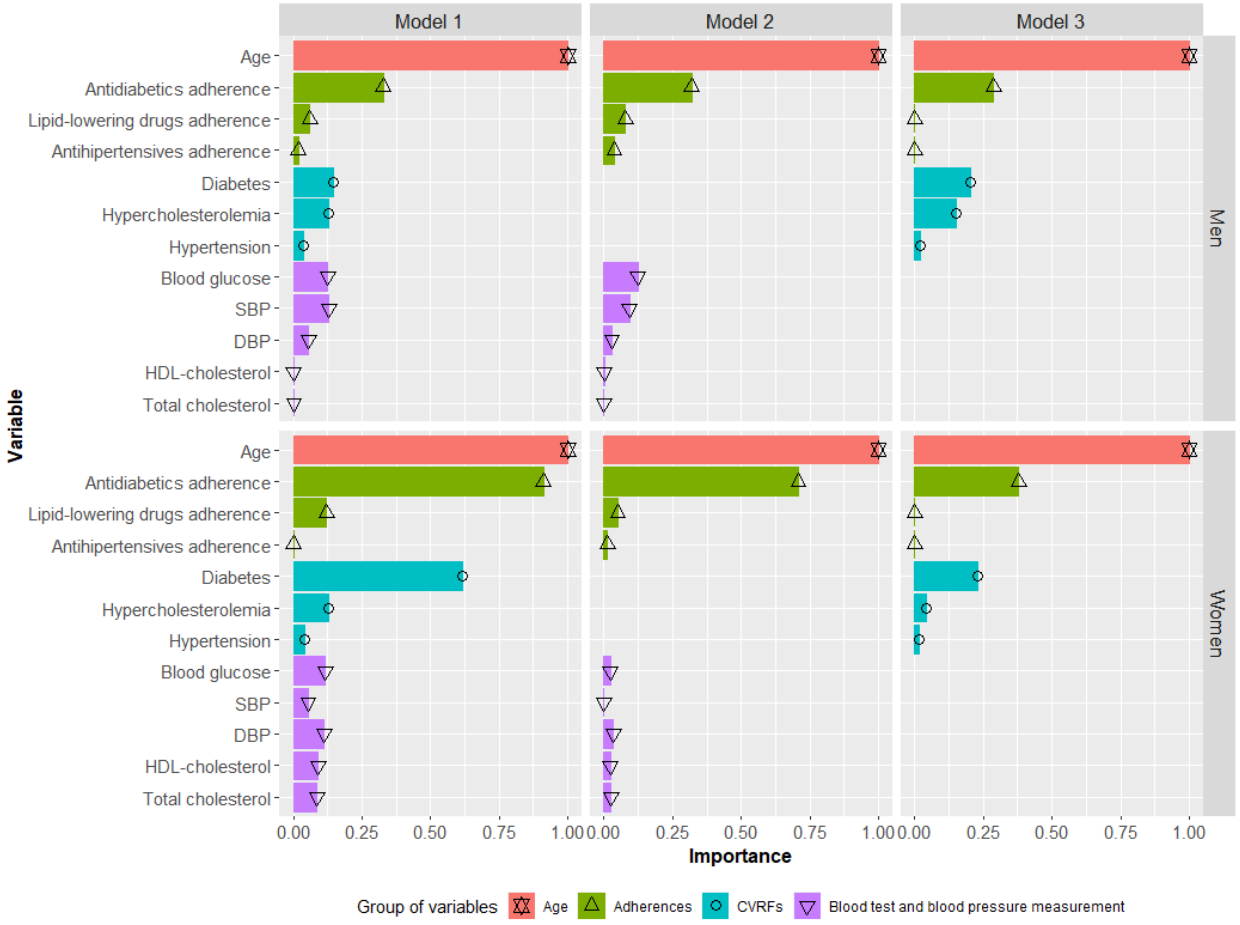


Figure 3: Relative contributions of variables in XG Boost models, for both men and women.
Model 1 includes the variables age, CVRFs, treatment adherence, and blood test and blood pressure measurements. Model 2 includes age, treatment adherence, and blood test and blood pressure measurements. Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL-cholesterol, high density lipoprotein cholesterol; CVRF, cardiovascular risk factor.

DISCUSSION

In the present study we performed a descriptive analysis of CVRFs and MACE incidence using real-world data (RWD) from users of the Spanish public health system. Using different combinations of predictive variables, we evaluated the abilities of two ML algorithms to predict CVE risk, and analysed the relative contributions of distinct CVR-related variables, stratifying the study population by sex.

The study population consisted of more women than men. The most prevalent CVRF was hypertension, and incidence of MACE was higher in women. The most frequent MACE in both sexes was stroke. The prevalence of DM and hypertension were higher among those who had experienced a CVE than those who had not.

Because the incidence, interactions, and control of CVRFs differ in men versus women[11–14], we generated and evaluated three models for each sex using RF and XG Boost algorithms. In all models, for both sexes, age was the parameter that contributed most to a predicted high risk of CVE, followed closely by adherence to antidiabetics. Adherence to antidiabetics was a greater contributor to CVE in women than in men.

For both men and women, the contributions of individual variables to a predicted risk of CVE differed across models. In most models, age and adherence to antidiabetics were the main contributors, with the exception of XG Boost models for men, in which antidiabetic adherence was closely followed by DM in models 1 and 2, and by blood glucose in model 3.

The present findings indicate that age is the most important variable when predicting CVE risk. Furthermore, the relative contributions of individual variables differ between men and women. Finally, adherence to treatment, especially to antidiabetics, is an important determinant of the risk of CVE, a factor that should be taken into account when managing cardiovascular risk[21,22].

Previous studies that include age as a predictive variable have consistently shown that this parameter has the greatest predictive power, suggesting that age is a key CVRF[6,17,18,23]. To the best of our knowledge, no studies based on ML methods published to date have considered cardiovascular treatment as a predictive variable,

while those that consider adherence either measured this variable using questionnaires or did not consider adherence to antidiabetic treatments [6,18,23].

Our study identified adherence to antidiabetics as a key determinant of CVE in both sexes. Conversely, adherence to antihypertensives and lipid-lowering drugs showed little predictive power. Multiple studies have reported associations between adherence to antihypertensives and lipid-lowering drugs and the incidence of CVE and all-cause mortality risk[22,24,25]. In terms of influence on the risk of different types of CVE, adherence to antidiabetics is less well studied than adherence to lipid-lowering drugs and antihypertensives[24]. In their systematic review, Mengying et al.[24] considered antidiabetics, antihypertensives and lipid-lowering drugs adherence, and found that all three were associated with a higher risk of CVE.

The aforementioned findings underscore the importance of proper pharmacological control of modifiable CVRFs to reduce the risk of CVE. Clinical guidelines propose controlling CVRFs in order to decrease this risk[2,26]. Previous research[27–29] has shown that adherence to these treatments is suboptimal, and the methods most commonly used to determine the risk of CVE do not include medication adherence as a predictive variable. There is also evidence[21] suggesting that a considerable number of CVEs are due to poor adherence to cardiovascular preventive treatments. Therefore, measuring adherence could maximize the efficacy of cardiac therapies in clinical settings.

Of the previously published studies similar to ours, analyses were performed without stratifying according to sex, and in most[17,18,23] sex was not identified as an important variable, with the exception of one study[6] in which sex was the second most important contributor to overall CVE risk. Although each study considered different variables, including lab results, blood pressure measurements, and socio-demographic factors, the importance of each varied across models and studies, as in the present study. Only in one study was blood glucose identified as the most important variable[17]. Two studies identified SBP as the second most important variable[18,23].

The present study compared two different ML methods. Some previous studies comparing different ML[5,30] reported that the RF model provided the most accurate

results. However, in their comparison of RF and XG Boost models, Dinh et[23] al. reported a slightly higher AUC (i.e. greater accuracy) for XG Boost.

Studies have shown that models built using ML techniques can overcome certain limitations of traditional methods used to predict CV risk, as well as offering greater predictive power[6,17,18]. The models described in this study could be applied in clinical practice to assess the individual risk of CVE based on patient characteristics and medication adherence, thereby playing an important role in screening processes. This is particularly important given that primary-care-based interventions targeting individuals considered to be at high risk, based on their age or risk factors, appear to be effective in reducing the risk of CVE[31]. Furthermore, these models can help orient the intervention and identify the most appropriate measures to take.

In addition to the advantages described above, ML techniques offer a variety of approaches to process large amounts of data to predict CVE incidence, thus allowing researchers and clinicians to select the algorithm that best suits their data or objectives.

LIMITATIONS AND STRENGTHS

Some limitations of the present study should be noted. First, the incidence of MACE during the follow-up period was low, resulting in class-imbalanced data. This issue was addressed by applying the Random Over Sampling Examples (ROSE) method to subsample the majority class. Second, the follow-up period was short, owing to the availability of data for the period 2017 to 2019 only. However, we feel that the size of the study population was sufficient to answer the research question. Finally, because this study was conducted using data extracted from administrative databases, some data were unavailable or were of insufficient quality to be included. Examples include smoking and physical activity data, which were recorded in very few subjects, and after quality control were deemed not to be reliable.

A key strength of the study is the fact that it was conducted with RWD, obtained from multiple data registries, enabling evaluation of the variables of interest in a real-world context. Out study is remarkable in that it includes data extracted from different levels of care from all individuals residing in Aragon, aged 16 and older, with any CVRF. Moreover,

we used two different ML techniques, which integrate all available data and offer several advantages over earlier algorithms, as explained above, and compared the results obtained with each to determine the most accurate method. To our knowledge, few ML studies have examined the predictive power of treatment adherence, and those that have typically assess adherence by asking patients whether they are taking any medication, without considering whether this medication is prescribed by a doctor or whether the patient actually collects their medication from a pharmacy. Finally, our analysis considered two algorithms and different combinations of predictive variables, allowing us to identify the model that performed best in this particular study population and to evaluate the influence of different variables on CVE occurrence.

CONCLUSIONS

In the present study we found that, in both men and women, age was the variable that most contributed to the risk of CVE. Comparison of distinct ML methods revealed comparable accuracy for RF and XG Boost algorithms. In all models, age was the main contributor to a predicted risk of CVE, although this effect was greater in men than women. The next greatest contributor was adherence to antidiabetics, the effect of which was greater in women than in men. Our findings suggest that ML techniques offer a valuable means of analysing large amounts of data to help accurately assess the risk of CVE, and could ultimately be applied in CVE prevention programs in a personalized medicine context.

REFERENCES

1. WHO. Cardiovascular diseases [Internet]. [cited 2022 Jan 7]. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

2. Visseren FLJ, MacH F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J. 2021 Sep 7;42(34):3227–337.

3. Sajeev S, Champion S, Beleigoli A, Chew D, Reed RL, Magliano DJ, et al. Predicting australian adults at high risk of cardiovascular disease mortality using standard risk factors and machine learning. Int J Environ Res Public Health. 2021 Mar 2;18(6):1–14.

4. Alaa AM, Bolton T, Angelantonio E di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One [Internet]. 2019 May 1 [cited 2021 May 14];14(5). Available from: https://pubmed.ncbi.nlm.nih.gov/31091238/

5. Jamthikar AD, Gupta D, Mantella LE, Saba L, Laird JR, Johri AM, et al. Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: a 500 participants study. Int J Cardiovasc Imaging [Internet]. 2021 Apr 1 [cited 2022 Jan 27];37(4):1171–87. Available from: https://link.springer.com/article/10.1007/s10554-020-02099-7

6. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? 2017; Available from: https://doi.org/10.1371/journal.pone.0174944

7. Malo S, Rabanaque MJ, Maldonado L, Moreno-Franco B, Chaure-Pardos A, Lallana MJ, et al. Identifying clusters of adherence to cardiovascular risk reduction behaviors and persistence with medication in new lipid-lowering drug users. Impact on healthcare utilization. Nutrients [Internet]. 2021 Mar 1 [cited 2021 May 20];13(3):1–15. Available from: https://pubmed.ncbi.nlm.nih.gov/33668726/

8. Khunti K, Danese MD, Kutikova L, Catterick D, Sorio-Vilela F, Gleeson M, et al. Association of a Combined Measure of Adherence and Treatment Intensity With Cardiovascular Outcomes in Patients With Atherosclerosis or Other Cardiovascular Risk Factors Treated With Statins and/or Ezetimibe. JAMA Netw Open [Internet]. 2018 Dec 7 [cited 2022 Nov 25];1(8):e185554–e185554. Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2717559

9. Malo S, Aguilar-Palacio I, Feja C, Lallana MJ, Rabanaque MJ, Armesto J, et al. Different approaches to the assessment of adherence and persistence with cardiovascular-disease preventive medications. Curr Med Res and Opin [Internet]. 2017 Jul 3 [cited 2021 May 20];33(7):1329–36. Available from: https://pubmed.ncbi.nlm.nih.gov/28422521/

10. Amaratunga D, Cabrera J, Sargsyan D, Kostis JB, Zinonos S, Kostis WJ. Uses and opportunities for machine learning in hypertension research. Int J Cardiol Hypertens. 2020 Jun 1;5:100027.

11. Liu W, Tang Q, Jin J, Zhu T, Dai Y, Shi Y. Sex differences in cardiovascular risk factors for myocardial infarction. Herz [Internet]. 2021 Apr 1 [cited 2022 Dec 12];46(Suppl 1):115–22. Available from: https://pubmed.ncbi.nlm.nih.gov/32377778/

12. Santilli F, D'Ardes D, Guagnano MT, Davi G. Metabolic Syndrome: Sex-Related Cardiovascular Risk and Therapeutic Approach. Curr Med Chem. 2017 Sep 11;24(24).

13. Álvarez-Fernández C, Romero-Saldaña M, Álvarez-López C, Molina-Luque R, Molina-Recio G, Vaquero-Abellán M. Gender differences and health inequality: Evolution of cardiovascular risk in workers. Arch Environ Occup Health [Internet]. 2021 [cited 2022 Dec 12];76(7):406–13. Available from: https://pubmed.ncbi.nlm.nih.gov/33625316/

14. Análisis con perspectiva de género de los registros sobre la enfermedad cardiovascular contenidos en la Base de Datos Clínicos de Atención Primaria.

15. Vrbaški D, Vrbaški M, Kupusinac A, Ivanović D, Stokić E, Ivetić D, et al. Methods for algorithmic diagnosis of metabolic syndrome. Artif Intell Med. 2019 Nov 1;101:101708.

16. Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, et al. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. Cardiovasc Res [Internet]. 2020 Dec 1 [cited 2021 Sep 23];116(14):2216–25. Available from: https://academic.oup.com/cardiovascres/article/116/14/2216/5680420

17. Huang W, Ying TW, Chin WLC, Baskaran L, Marcus OEH, Yeo KK, et al. Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. Sci Rep [Internet]. 2022 Dec 1 [cited 2022 Nov 25];12(1). Available from: https://pubmed.ncbi.nlm.nih.gov/35058500/

18. Sajeev S, Champion S, Beleigoli A, Chew D, Reed RL, Magliano DJ, et al. Predicting Australian Adults at High Risk of Cardiovascular Disease Mortality Using Standard Risk Factors and Machine Learning. Int J Environ Res Public Health [Internet]. 2021 Mar 2 [cited 2022 Nov 25];18(6):1–14. Available from: https://pubmed.ncbi.nlm.nih.gov/33808743/

19. Rhys HI. Machine Learning with R, the tidyverse, and mlr. 2020.

20. IACS. Actividad de Tratamiento BIGAN [Internet]. [cited 2022 Feb 17]. Available from: https://www.iacs.es/actividad-tratamiento-bigan/

21. Chowdhury R, Khan H, Heydon E, Shroufi A, Fahimi S, Moore C, et al. Adherence to cardiovascular therapy: a meta-analysis of prevalence and clinical consequences. Eur Heart J [Internet]. 2013 Oct 7 [cited 2022 Feb 16];34(38):2940–8. Available from: https://pubmed.ncbi.nlm.nih.gov/23907142/

22. Chen C, Li X, Su Y, You Z, Wan R, Hong K. Adherence with cardiovascular medications and the outcomes in patients with coronary arterial disease: “Real-world” evidence. Clin Cardiol. 2022;

23. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak [Internet]. 2019 Nov 6 [cited 2022 Nov 28];19(1):1–15. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5

24. Liu M, Zheng G, Cao X, Chang X, Zhang N, Liang G, et al. Better medications adherence lowers cardiovascular events, stroke, and all-cause mortality risk: A dose-response meta-analysis. J Cardiovasc Dev Dis [Internet]. 2021 Nov 1 [cited 2022 Nov 30];8(11):146. Available from: https://www.mdpi.com/2308-3425/8/11/146/htm

25. Donneyong MM, Fischer MA, Langston MA, Joseph JJ, Juarez PD, Zhang P, et al. Examining the Drivers of Racial/Ethnic Disparities in Non-Adherence to Antihypertensive Medications and Mortality Due to Heart Disease and Stroke: A County-Level Analysis. Int J Environ Res Public Health [Internet]. 2021 Dec 1 [cited 2022 Nov 30];18(23). Available from: https://pubmed.ncbi.nlm.nih.gov/34886429/

26. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation [Internet]. 2019 Sep 10 [cited 2022 Feb 8];140(11):e596–646. Available from: http://ahajournals.org

27. Zhao B, He X, Wu J, Yan S. Adherence to statins and its impact on clinical outcomes: a retrospective population-based study in China. BMC Cardiovasc Disord [Internet]. 2020 Jun 10 [cited 2022 Feb 8];20(1):282. Available from: https://pubmed.ncbi.nlm.nih.gov/32522146/

28. Lee H, Yano Y, Cho SMJ, Heo JE, Kim DW, Park S, et al. Adherence to Antihypertensive Medication and Incident Cardiovascular Events in Young Adults With Hypertension. Hypertension [Internet]. 2021 [cited 2022 Feb 8];77(4):1341–9. Available from: https://pubmed.ncbi.nlm.nih.gov/33641364/

29. Aguilar-Palacio I, Rabanaque MJ, Maldonado L, Chaure A, Abad-Díez JM, León-Latre M, et al. New Male Users of Lipid-Lowering Drugs for Primary Prevention of Cardiovascular Disease: The Impact of Treatment Persistence on Morbimortality. A Longitudinal Study. Internat J Environ Res and Public Health [Internet]. 2020 Oct 2 [cited 2021 May 20];17(20):7653. Available from: www.mdpi.com/journal/ijerph

30. Nadakinamani RG, Reyana A, Kautish S, Vibith AS, Gupta Y, Abdelwahab SF, et al. Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine

Learning Techniques. Comput Intell Neurosci [Internet]. 2022 Jan 11 [cited 2022 Feb 2];2022. Available from: /pmc/articles/PMC8767405/

31. Eriksen CU, Rotar O, Toft U, Jørgensen T. What is the effectiveness of systematic population-level screening programmes for reducing the burden of cardiovascular diseases? 2021.

ANEXO IV

European Journal of Public Health, 1–6

https://doi.org/10.1093/eurpub/ckad227

Exploring sex variations in the incidence of cardiovascular events: a counterfactual decomposition analysis

Sara Castel-Feced[1,2,3], Sara Malo[1,2,3,4], Isabel Aguilar-Palacio[1,2,3,4], Lina Maldonado[2,3,5], María José Rabanaque[1,2,3,4,*], Miguel San Sebastián[6,*]

1 Department of Microbiology, Pediatrics, Radiology, and Public Health, University of Zaragoza, Zaragoza, Spain
2 Institute for Health Research Aragón (IIS Aragón), Zaragoza, Spain
3 GRISSA Research Group, Zaragoza, Spain
4 Network for Research on Chronicity, Primary Care, Health Promotion (RICAPPS), ISCIII, Madrid, Spain
5 Department of Economic Structure, Economic History and Public Economics, University of Zaragoza, Zaragoza, Spain
6 Department of Epidemiology and Global Health, Umeå University, Umeå, Sweden

Correspondence: Sara Castel-Feced, Department of Microbiology, Pediatrics, Radiology, and Public Health, Faculty of Medicine ,University of Zaragoza, Domingo Miral street no number, Zaragoza 50009, Spain, Tel: +34 976 761 691, e-mail: scastelf@unizar.es

*These authors contributed equally to this work.

Background: Some cardiovascular risk factors (CVRFs) that occur differently in men and women can be addressed to reduce the risk of suffering a major adverse cardiovascular event (MACE). Furthermore, the development of MACE is highly influenced by social determinants of health. Counterfactual decomposition analysis is a new methodology that has the potential to be used to disentangle the role of different factors in health inequalities. This study aimed to assess sex differences in the incidence of MACE and to estimate how much of the difference could be attributed to the prevalence of diabetes, hypertension, hypercholesterolaemia and socioeconomic status (SES). **Methods:** Descriptive and counterfactual analyses were conducted in a population of 278 515 people with CVRFs. The contribution of the causal factors was estimated by comparing the observed risk ratio with the causal factor distribution that would have been observed if men had been set to have the same factor distribution as women. The study period was between 2018 and 2021. **Results:** The most prevalent CVRF was hypercholesterolaemia, which was similar in both sexes, while diabetes was more prevalent in men. The incidence of MACE was higher in men than in women. The main causal mediating factors that contributed to the sex differences were diabetes and SES, the latter with an offsetting effect. **Conclusions:** This result suggests that to reduce the MACE gap between sexes, diabetes prevention programmes targeting men and more gender-equal salary policies should be implemented.

Introduction

Cardiovascular diseases (CVDs) are one of the leading causes of death and disability worldwide.[1] CVDs are influenced by several cardiovascular risk factors (CVRFs), some of which are modifiable. Most of those CVRFs are related to behavioural lifestyles that can lead to high blood pressure, high levels of glucose and lipids in the blood, overweight and obesity, all key leading factors of CVDs.[1]

The development of CVRFs and CVDs is highly influenced by social determinants of health (SDoH)[2,3] and the relationship between SDoH such as individual-level socioeconomic factors (e.g. education, income and occupation) and CVD is well established.[2–4] The effect of these SDoH on CVD persists throughout the life course, as having low socioeconomic status (SES) during childhood is related to a higher risk of CVD in adulthood.[2,5,6]

In this regard, people with low SES are more likely to present modifiable and behavioural CVRFs and therefore it is a crucial determinant in which to intervene[7,8]—for example, by incorporating SDoH screening and interventions into chronic disease clinical care.[2]

Apart from the SDoH mentioned above, sex/gender[9,10] also plays a role in the risk of developing CVD. In terms of biological sex, differences between men and women could include the fact that the prevalence of CVRF is different between sexes, that the interaction of CVRFs on the development of CVD is different between them or that women have specific conditions, such as pre-eclampsia, gestational diabetes and premature menopause, which have been associated with an increase in the risk of CVD.[11,12] Gender is also an SDoH that has not been traditionally considered in this field despite disparities having been reported in cardiovascular care, especially in acute cardiovascular care.[13–15] Furthermore, there are other gender-based factors interrelated with the differences in CVD in women compared with men, such as lower SES, lower levels of physical activity and higher stress due to family responsibilities.[14,15] Although biological sex is static, gender is socially constructed, making it possible to intervene and change its effect on CVD.[10]

For some time now, public health policy has broadened its scope to address inequalities in the distribution of health and to reduce health differences between population groups. With this purpose, the concept of SDoH has been expanded to SDoH inequalities requiring specific methods to capture them.[16]

To disentangle the relationship between the SDoH and health inequalities, decomposition methods, such as the Blinder–Oaxaca method, are broadly applied.[17,18] These methods try to quantify the degree of social inequality in health and the contribution of different factors to that inequality.[17] These methods include some limitations—for example, they are based on the decomposition of mean differences between groups and cannot be applied to summary measures of population such as risk ratios (RRs) or disability-adjusted life-years or they do not attempt to estimate causal effect

on the group variable so they have ambiguous causal interpretations.[19,20] New perspectives on decomposition analysis have been developed in epidemiology, situating them in the causal inference and counterfactual theory.[19]

These new approaches use parametric models and Monte-Carlo estimation to expand existing decomposition methods used to solve some of the previous limitations[20] so that they can be applied to decompose any contrast of any summary population measure. However, these new methods also come with some challenges: they need high computational power and they are not based on aggregate data, so they need large-scale individual-level data and require parametric modelling assumptions.

Studies focused on CVDs have been conducted in men, with a lack of studies among women.[21] Furthermore, there is extensive literature on the SDoH related to CVD and on the differences in CVD by sex,[2–4,11,13,15,22–24] but to the best of our knowledge, there is a lack of studies disentangling the contributing factors to the sex differences in CVD inequalities. In this regard, counterfactual analysis is a new approach that can be applied to study the impact of the differences in the distribution of CVRFs and SES between men and women with regard to the development of CVD. By identifying these sources of inequality, it is possible to act on them and thus reduce health differences between social groups.

This study aimed to assess sex differences in the incidence of major adverse cardiovascular adverse events (MACEs) and to estimate how much the observed disparity could be attributed to the differences in the prevalence of diabetes, hypertension, hypercholesterolaemia and SES in the region of Aragón, Spain.

Methods

The present study was conducted within the CARhES cohort, a Spanish dynamic cohort comprising people with hypertension, hypercholesterolaemia or diabetes in the region of Aragón, Spain. Aragón is one of the 17 autonomous regions of the country and has a population of about 1.3 million inhabitants that is overwhelmingly attended to by the public health system (98% of the population). The follow-up of this cohort started in 2017 and includes information from all levels of care (hospitalizations, primary care and pharmacy) for the entire population aged 16 or above that has at least one of the three CVRFs mentioned and that is registered in the public health system of the region.

Data sources

All data for this study were obtained from BIGAN, a health data hub that gathers data from all levels of care of the Aragón public health service through the linkage of several databases. For the present study, the following were used: the Users database, which provides information on age and affiliation to the health system; the Hospital Discharge Records (CMBD), which gathers data on hospital discharge; the Adjusted Morbidity Groups, which records information on all medical diagnoses available in primary health care and in the CMBD; the Emergency Care database, which stores information on patients who attended hospital emergencies; and the Electronic Prescribing System database, which records pharmacological treatments prescribed to patients. In these databases, all information is pseudonymized with a unique code that allows patient information to be linked across the different data sources but prevents personal identification.

Study population, inclusion and exclusion criteria

The selection of the study population for the present study is depicted in figure 1. All subjects who were part of the cohort in 2017 and were aged 50 years or older were included. This decision was based on the lower incidence of a MACE at earlier ages and the potential differences in factors influencing the occurrence of the event.[25] From them, those with a previous MACE and those who died during the study period, from January 2018 to December 2021, from a cause other than a MACE were excluded.

To identify subjects who had suffered a previous MACE, the Morbidity and/or the CMBD database during 2016 and 2017 were consulted. In both databases, a check was made as to whether they had had a diagnosis of stroke or heart attack.

Study variables

Diabetes, hypertension, hypercholesterolaemia, age and SES were the mediating variables included in the study. They were selected based on their well-known relationship with MACE.[11] Sex was included as the exposure variable and MACE as the outcome.

Sex, age and SES were obtained from the Health System Users database in 2017. SES was calculated from two variables: income band and economic activity. These two variables were combined to obtain five different categories of SES: employees earning >18 000€/year, employees earning <18 000€/year, individuals

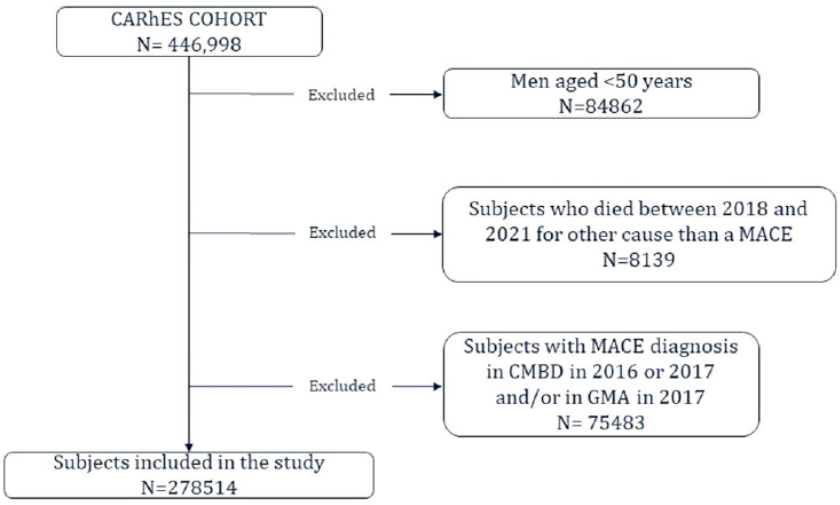


Figure 1 Flowchart depicting the study population.

with a contributory pension earning >18 000€/year, individuals with a contributory pension earning <18 000€/year and people with free medicines and others, including mainly those with a special pharmacy regime and with low income. SES was dichotomized into people earning >18 000€/year and the rest when considered as a causal mediating factor, whereas the five categories were used when included as a potential confounder.

The identification of the CVRFs at baseline was performed according to the medical diagnosis recorded in the Morbidity database [hypertension, diabetes mellitus (DM) and hypercholesterolaemia] and/or the pharmacological prescriptions recorded in the Electronic Prescribing System database (DM and hypercholesterolaemia). Antidiabetic and lipid-lowering drugs were identified through the ATC codes: A10 for antidiabetics and C10 for lipid-lowering drugs.

Events considered to be a MACE were identified from the main diagnosis of hospital admissions in the CMBD database and from episodes in the emergency database that caused death. Diagnosis considered MACE in those databases included myocardial infarction (International Classification of Diseases, 10th revision (ICD-10) code I21; International Classification of Diseases, 9th revision (ICD-9) code 410), subarachnoid, intracerebral and other non-traumatic haemorrhage (ICD-10 code I60-I62; ICD-9 code 430-432) and acute ischaemic stroke (ICD-10 code I63; ICD-9 code 433).

Study analysis

Two descriptive analyses by sex were performed; in one, the variables included in the study for the total population and stratified by sex were described and in the other, the incidence of MACE in men and women in each mediating factor was calculated.

Then, to estimate the causal contribution of each mediating factor to the sex difference in MACE incidence, a decomposition of the age-adjusted RR for men relative to women (non-exposed group) was performed.[19]

Four different models were performed considering one different causal mediating factor in each (diabetes, hypertension, hypercholesterolaemia and SES), adjusted for the rest remaining mediating factors (used as confounders) and age (i.e. in the model that diabetes was considered the mediating factor, hypertension, hypercholesterolemia, SES and age were considered confounders)

The main summary measure of occurrence was the risk of incidence of MACE and the association was the RR for men relative to women, the latter being estimated by applying Poisson regressions.

All these analyses were conducted considering women as the reference category, as the total incidence of MACE was higher in men.

The estimation of the contribution of the causal factors was done by comparing the observed RR with the counterfactual RR. To do this, a counterfactual risk of MACE incidence in men is needed, which was obtained by applying the g-formula and Monte-Carlo integration (Supplementary appendix S1). Thus, two pseudo-populations were created: a so-called 'natural course' population by using the coefficients obtained from analyzing the observed data; and a counterfactual pseudo-population created with the coefficients from the analysis using the simulated mediating factor values. The difference between the two populations corresponds to the causal contribution to the inequality.

All analyses were performed with R version 4.2.2 using *cfdecomp* package.[26]

All collected data were pseudonymized. The present study was approved by the Clinical Research Ethics Committee of Aragon (project identification code PI21/148).

Results

Table 1 shows the characteristics of the total population at baseline and stratified by sex. In total, 278 515 individuals were included in the study, 44.7% were men and on average women were older. The most prevalent CVRF in both sexes was hypercholesterolaemia, followed by hypertension and diabetes. The first two CVRFs had similar prevalence in men and women, but diabetes was more prevalent in men.

In terms of SES, sex differences were identified, particularly among active individuals earning >18 000€/year and retired individuals earning <18 000€/year or receiving free pharmacy benefits. The largest group was the retired, earning <18 000€/year or with free medicines in both men and women, but in women, this group represented 52.6% of their population while in men it was 35.9%. Additionally, among retired individuals earning <18 000€/year or receiving free pharmacy benefits, women surpassed men in representation. Finally, the incidence of MACE during the follow-up period was 2.5% in men vs 1.7% in women.

Women who suffered a MACE were older than men (77.2 years on average and 70.5, respectively). Table 2 shows the incidence of MACE in men and women in each mediating factor. In both sexes, the incidence of MACE was higher among those with diabetes and hypertension. However, the incidence of MACE was higher in those without hypercholesterolaemia than in those with it. Finally, also in both sexes, the SES group with the highest incidence of MACE was the retired earning <18 000€/year or with free pharmacy.

Counterfactual analysis

The results of the four counterfactual analyses are shown in figure 2 and in Supplementary appendix S2.

When diabetes was the mediating factor, the RR in the natural course analysis of having a MACE for men relative to women was 1.83 [95% confidence interval (CI): 1.74–1.92], that is, men had 83% more likelihood of having a MACE than women after adjusting for age, hypertension, hypercholesterolaemia and SES. The counterfactual RR (after setting men to have the same diabetes distribution as women) was 1.74 (95% CI: 1.65–1.83), corresponding to a causal

Table 1 Baseline characteristics in the total population and by sex.

	Total	Men *N* = 124 602	Women *N* = 153 912	*P*-value
Age	67.6 (10.5)	65.9 (10.0)	69.1 (10.8)	0.000
Diabetes	57 612 (20.7%)	30 509 (24.5%)	27 103 (17.6%)	0.000
Hypercholesterolemia	205 700 (73.9%)	90 640 (72.7%)	115 060 (74.8%)	<0.001
Hypertension	171 339 (61.5%)	76 544 (61.4%)	94 795 (61.6%)	0.392
Socioeconomic status				0.000
Employees earning >18 000	33 812 (12.1%)	21 969 (17.6%)	11 843 (7.7%)	
Employees earning <18 000	39 769 (14.3%)	18 838 (15.1%)	20 931 (13.6%)	
Retired earning >18 000	56 398 (20.2%)	30 284 (24.3%)	26 114 (17.0%)	
Retired earning <18 000 and free pharmacy	125 692 (45.1%)	44 738 (35.9%)	80 954 (52.6%)	
Others	22 843 (8.20%)	8773 (7.0%)	14 070 (9.1%)	
MACE in 4-year follow-up period	5732 (2.06%)	3169 (2.5%)	2563 (1.7%)	<0.001

Notes: Information showed as number (%) for categorical variables and mean in years (standard deviation) for age. Percentages shown are calculated by columns.

contribution of diabetes to the relationship between sex and the incidence of MACE of 10.5% (95% CI: 7.99–13.08); that is to say, if the prevalence of diabetes in men was the same as in women, the incidence of MACE in men would be reduced from 2.5% to 2.2% in the 4 years of follow-up.

In the case of hypertension, very similar RRs were found in the natural course and counterfactual populations (RR = 1.81; 95% CI: 1.72–1.9 and RR = 1.82; 95% CI: 1.73–1.91, respectively). Therefore, the percentage contribution of hypertension to sex differences was very low (1%, 95% CI: −1.4 to 3.4). Something similar happened in the case of hypercholesterolaemia, with a percentage of contribution near 0.

Finally, when considering the SES as a mediating factor, the RR of MACE in the natural course analysis was smaller (RR = 1.8) than in the counterfactual one (RR = 1.9). This would mean that if we equalized the SES to women's level, the risk of suffering a MACE among men would increase, with the contribution being negative (−6%). This offsetting effect was observed because MACE was higher among those of low SES (table 2) and more women than men were in the low SES (table 1).

Table 2 Incidence of MACE per explanatory factors by sex.

	MACE men 3169 (2.5%)	MACE women 2563 (1.7%)
Age	77.2 (10.3)	70.5 (10.8)
Diabetes		
No	2048 (2.18%)	1795 (1.42%)
Yes	1121 (3.67%)	768 (2.83%)
Hypertension		
No	985 (2.05%)	560 (0.95%)
Yes	2184 (2.85%)	2003 (2.11%)
Hypercholesterolemia		
No	958 (2.82%)	773 (1.99%)
Yes	2211 (2.44%)	1790 (1.56%)
Socioeconomic status		
Employees earning >18 000	306 (1.39%)	55 (0.46%)
Employees earning <18 000	318 (1.69%)	117 (0.56%)
Retired earning >18 000	764 (2.52%)	351 (1.34%)
Retired earning <18 000 and free pharmacy	1599 (3.57%)	1922 (2.37%)
Others	182 (2.07%)	118 (0.84%)

Notes: Information showed as number (%) for categorical. Percentages shown are calculated by row.

Discussion

This study focused on determining the sex disparity in MACE and disentangling how much of those differences could be attributed to diabetes, hypertension, hypercholesterolaemia and SES. The counterfactual analysis showed that the contributing factors explaining the sex differences in the incidence of MACE were diabetes (10%) and SES (−6%).

The literature shows[14,27] that men have a higher incidence of MACE, so the finding of a sex difference in MACE favouring women was expected. Furthermore, women with MACE tend to be older than men. In this regard, a study conducted in the same region found more women with heart failure than men (though women were older) but men were more likely to have ischaemic heart disease and acute myocardial infarction.[24] Regarding how MACE was identified in the present study, all cases of MACE who were diagnosed by a doctor were registered in CMBD database, irrespective of their severity. Nonetheless, some minor events that were not diagnosed by a doctor could be lost.

While the literature regarding the relationship of diabetes, hypertension, hypercholesterolaemia and SES with MACE is extensive,[11,12,14,28–31] to the best of our knowledge, no previous studies capturing the social determinants of MACE inequalities have been conducted applying the counterfactual decomposition analysis.

In this study, diabetes and SES were the two factors that had an impact on the differences between sexes with regard to the incidence of MACE. Similar to findings in other studies,[23,32] diabetes had a higher prevalence in men. In this case, diabetes emerged as the CVRF with the most pronounced difference between sexes. Hence, it is a plausible explanation for diabetes being the predominant factor in accounting for the differences in the incidence of MACE between men and women. The high impact of diabetes on the risk of suffering a MACE has been reported elsewhere,[11,12,33] thus, by decreasing the incidence of diabetes in men, the incidence of MACE would be reduced and the sex inequalities diminished.

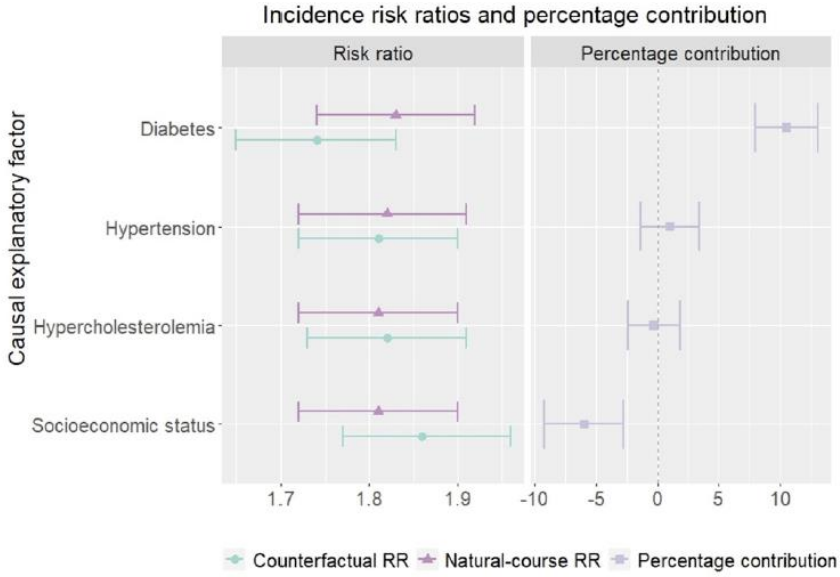


Figure 2 Risk ratio and confidence intervals for natural course and counterfactual populations and percentage of contribution for each explanatory factor.

Similarly, the association of SES with CVD incidence has been found in numerous studies[18,34] where lower SES tends to be related to worse results in CVD, in part because of the association between low SES and unhealthier lifestyles.[6] A systematic review[35] found that women with low SES had a higher risk of suffering a CVD than men with low SES. In the present study, an offsetting effect was identified, indicating that if we equalized the SES in men to that of women, the risk of experiencing a MACE among men would increase. This effect can be explained by the lower SES and incidence of MACE among women. High SES among men is probably acting as a marker of a series of factors in the path towards CVDs.[6,8] For example, although men had better SES, commonly, they have had worst lifestyles and higher levels of metabolic risk factors. This fact leads to lower rates of MACE in women than in men and to women being older when they suffer MACE than men. The claim that men had worse lifestyles regardless of their SES is further supported by the fact that women with any CVRFs are older than men with any CVRFs. This may be attributed to the fact that men tend to have poorer lifestyle habits, leading to the earlier development of CVRFs. In light of the results, while gender-based policies to improve the SES among women should be implemented, they should also be accompanied by CVD prevention gender-specific interventions.

Interestingly, despite previous studies focused on CVRFs have shown that severe hypertension and incidence of stroke are twice as high in women compared with men[22] and that hypercholesterolaemia has a stronger association with the incidence of infarction in men than in women,[13,15] none of these factors contributed to explaining the sex difference in MACE in this study. This apparent discrepancy illustrates the fact that often determinants of health might not be relevant as determinants of health inequalities.

There are studies looking at contributory factors to the sex differences in other health outcomes, such as self-reported health or disability[36,37] applying classical decomposition methods. Furthermore, counterfactual decomposition analysis has been applied to other health outcomes.[38–40] Nonetheless, to the best of our knowledge, this is the first study using a counterfactual approach to analyze the difference between sexes in the incidence of CVD.

Limitations and strengths

This study has several strengths. First, it was conducted with real-world data, data which come from daily clinical practice and that include a great number of registers. It is remarkable that this information came from different levels of care and comprised all subjects from Aragón older than 16 years with any CVRF. Moreover, we applied a new methodology that allows the causal decomposition of the age-adjusted RR for men relative to women in the incidence of MACE, going further than previous decomposition methods. Finally, to the best of our knowledge, this is the first time that this method has been applied to empirical data and using several mediating variables in the counterfactual decomposition analysis.

This study also has some limitations to be considered. First, the follow-up period was short as we only had data from 2017 to 2021, and studies with longer follow-up periods with longer exposures to CVRFs and more MACEs are necessary. However, the size of the population of this study allowed us to conduct the present study as we had information on the whole Aragonese population with any CVRF. Moreover, since the study used data from administrative databases and because some information was not available or was partly registered (i.e. smoking and physical activity), not all desirable variables were possible to be included in the study. Given the methodology applied, it was not possible to include all the different mediating variables together, so separated models considering only one mediator at a time were run. While it was considered analyzing each diagnosis separately, the low number of cases of each diagnosis made us reconsider that option. Finally, all the people included in our cohort had at least one CVRF, so the population of this study had a high risk of suffering a CVD. However, it is important to take into account that our cohort included more than 70% of the Aragonese population aged above 50 years, increasing the external validity of the results.

Conclusion

We found differences between men and women in the incidence of MACE and in the prevalence of diabetes, while the prevalence of hypercholesterolaemia and hypertension were similar in both sexes.

The CVRFs that most influenced the difference between sexes in the risk of MACE were diabetes and SES. These findings suggest that to reduce the MACE gap between men and women in this Spanish population, diabetes prevention programmes targeting men, and more gender-equal policies should be implemented.

Supplementary data

Supplementary data are available at *EURPUB* online.

Funding

This work was supported by the Proyecto del Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (Ministerio de Ciencia e Innovación), and the European Fund for Regional Development (FEDER) (PI22/01193). It was also partly funded by the Government of Aragón with a grant for postgraduate research contracts (IIU/796/2019).

Conflicts of interest: There is no conflict of interest in this work.

Data availability

The data underlying this article cannot be shared publicly due to their sensitive nature. The data will be shared on reasonable request to the corresponding author.

Key points

- Incidence of major adverse cardiovascular events (MACEs) and prevalence of cardiovascular risk factors (CVRFs) are different in men and women.
- By analyzing factors that contribute to sex differences in the occurrence of MACE, we can act on them and thus reduce health differences between groups.
- The incidence of MACE was 0.8 percentage points higher in men than in women.
- The most influential factors in sex differences in the incidence of MACE were diabetes and socioeconomic status.
- Diabetes contributed 10% to the sex differences in the incidence of MACE while socioeconomic status had an offsetting effect.

References

1 WHO. *Cardiovascular Diseases*. Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (9 March 2023, date last accessed).

2 Powell-Wiley TM, Baumer Y, Baah FO, et al. Social determinants of cardiovascular disease. *Circ Res* 2022;130:782–99.

3 Mannoh I, Hussien M, Commodore-Mensah Y, Michos ED. Impact of social determinants of health on cardiovascular disease prevention. *Curr Opin Cardiol* 2021;36:572–9.

4 Jilani MH, Javed Z, Yahya T, et al. Social determinants of health and cardiovascular disease: current state and future directions towards healthcare equity. *Curr Atheroscler Rep* 2021;23:55–11.

5 Kinge JM, Modalsli JH, Øverland S, et al. Association of household income with life expectancy and cause-specific mortality in Norway, 2005-2015. *JAMA* 2019; 321:1916–25.

6 Zhang YO, Chen C, Pan XF, et al. Associations of healthy lifestyle and socioeconomic status with mortality and incident cardiovascular disease: two prospective cohort studies. *BMJ* 2021;373:n604.

7 Stringhini S, Carmeli C, Jokela M, et al.; LIFEPATH Consortium. Socioeconomic status and the 25×25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *Lancet* 2017; 389:1229–37.

8 Clark AM, DesMeules M, Luo W, et al. Socioeconomic status and cardiovascular disease: risks and implications for care. *Nat Rev Cardiol* 2009;6:712–22.

9 Springer KW, Mager Stellman J, Jordan-Young RM. Beyond a catalogue of differences: a theoretical frame and good practice guidelines for researching sex/gender in human health. *Soc Sci Med* 2012;74:1817–24.

10 O'Neil A, Scovelle AJ, Milner AJ, Kavanagh A. Gender/sex as a social determinant of cardiovascular risk. *Circulation* 2018;137:854–64.

11 Visseren FLJ, Mach F, Smulders YM, et al.; ESC Scientific Document Group. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J* 2021;42:3227–337.

12 Roth GA, Mensah GA, Johnson CO, et al.; GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol* 2020;76:2982–3021.

13 Vallabhajosyula S, Verghese D, Desai VK, et al. Sex differences in acute cardiovascular care: a review and needs assessment. *Cardiovasc Res* 2022;118:667–85.

14 Woodward M. Cardiovascular disease and the female disadvantage. *Int J Environ Res Public Health* 2019;16:1165.

15 Humphries KH, Izadnegahdar M, Sedlak T, et al. Sex differences in cardiovascular disease—Impact on care and outcomes. *Front Neuroendocrinol* 2017;46:46–70.

16 Graham H. Social determinants and their unequal distribution: clarifying policy understandings. *Milbank Q* 2004;82:101–24.

17 Amroussia N, Gustafsson E, Mosquera PA. Explaining mental health inequalities in Northern Sweden: a decomposition analysis. *Glob Health Action* 2017;10:1305814.

18 Mosquera PA, San Sebastian M, Ivarsson A, Gustafsson PE. Decomposition of gendered income-related inequalities in multiple biological cardiovascular risk factors in a middle-aged population. *Int J Equity Health* 2018;17:102.

19 Sudharsanan N, Bijlsma MJ. Educational note: causal decomposition of population health differences using Monte Carlo integration and the g-formula. *Int J Epidemiol* 2022;50:2098–107.

20 Sudharsanan N, Bijlsma MJ, De BM, Barclay K. *A Generalized Counterfactual Approach to Decomposing Differences Between Populations*. 2019. Available at: www.demogr.mpg.de (8 March 2023, date last accessed).

21 Norris CM, Yip CYY, Nerenberg KA, et al. State of the science in women's cardiovascular disease: a Canadian perspective on the influence of sex and gender. *J Am Heart Assoc* 2020;9(4): e015634.

22 Madsen TE, Howard G, Kleindorfer DO, et al. Sex differences in hypertension and stroke risk in the REGARDS study: a longitudinal cohort study. *Hypertension* 2019; 74:749–55.

23 Huebschmann AG, Huxley RR, Kohrt WM, et al. Sex differences in the burden of type 2 diabetes and cardiovascular risk across the life course. *Diabetologia* 2019; 62:1761–72.

24 Gutiérrez AG, Poblador-Plou B, Prados-Torres A, et al. Sex differences in comorbidity, therapy, and health services' use of heart failure in Spain: evidence from real-world data. *Int J Environ Res Public Health* 2020;17:2136.

25 Andersson C, Vasan RS. Epidemiology of cardiovascular disease in young individuals. *Nat Rev CardiolCardiol*15:230–40.

26 Bijlsma MJ, Sudharsanan N. *cfdecomp: Counterfactual Decomposition: MC Integration of the G-Formula*. (R package, version 0.4.0). Available at: https://cran.r-project.org/package=cfdecomp (20 February 2023, date last accessed).

27 Timmis A, Vardas P, Townsend N, et al.; Atlas Writing Group, European Society of Cardiology. European Society of Cardiology: cardiovascular disease statistics 2021. *Eur Heart J* 2022;43:716–99.

28 Al-Salameh A, El bouzegaoui N, Saraval-Gross M. Diabetes and cardiovascular risk according to sex: an overview of epidemiological data from the early Framingham reports to the cardiovascular outcomes trials. *Ann Endocrinol (Paris)* 2023;84:57–68.

29 Watts GF, Catapano AL, Masana L, et al. Hypercholesterolemia and cardiovascular disease: focus on high cardiovascular risk patients. *Atheroscler Suppl* 2020; 42:e30–e34.

30 Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet* 2020; 395:795–808.

31 Roth GA, Mensah GA, Johnson CO, et al.; GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol* 2020;76:2982–3021.

32 Tramunt B, Smati S, Grandgeorge N, et al. Sex differences in metabolic regulation and diabetes susceptibility. *Diabetologia* 2019;63:453–61.

33 Arnett DK, Blumenthal RS, Albert MA, et al. ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* 2019;140:e596–646.

34 Schultz WM, Kelli HM, Lisko JC, et al. Socioeconomic status and cardiovascular outcomes: challenges and interventions. *Circulation* 2018;137:2166–78.

35 Backholer K, Peters SAE, Bots SH, et al. Sex differences in the relationship between socioeconomic status and cardiovascular disease: a systematic review and meta-analysis. *J Epidemiol Community Health* 2017;71:550–7.

36 Hosseinpoor AR, Williams JS, Jann B, et al. Social determinants of sex differences in disability among older adults: a multi-country decomposition analysis using the World Health Survey. *Int J Equity Health* 2012;11:52.

37 Boerma T, Hosseinpoor AR, Verdes E, Chatterji S. A global assessment of the gender gap in self-reported health with survey data from 59 countries. *BMC Public Health* 2016;16:675.

38 Wetzel S, Sarker M, Hasan M, et al. Rapidly rising diabetes and increasing body weight: a counterfactual analysis in repeated cross-sectional nationally representative data from Bangladesh. *Epidemiology* 2023;34:732–40.

39 Burgos Ochoa L, Bijlsma MJ, Steegers EAP, et al. Does neighborhood crime mediate the relationship between neighborhood socioeconomic status and birth outcomes? An application of the mediational G-formula. *Am J Epidemiol* 2023;192:939–48.

40 Pitkänen J, Bijlsma MJ, Remes H, et al. The effect of low childhood income on self-harm in young adulthood: mediation by adolescent mental health, behavioural factors and school performance. *SSM Popul Health* 2021;13:100756.

Supplementary material

APPENDIX I: MONTECARLO SIMULATIONS AND G-FORMULA

Parametric g-formula is a general method of standardization used for causal inference. Specifically, it is applied to estimate the average causal effect in the entire population of interest [1]. Therefore, this methodology has been used to estimate various components of the total effect including mediated interaction effects, using Monte Carlo simulations [2].

In this methodology, the objective is to calculate for each subject their nested potential outcomes in a counterfactual scenario where the explanatory factor was the value of the best option. For example, if we were measuring the effects of an intervention, the explanatory factor would have the value supposed after the intervention. To compute points estimate, regressions of the potential outcome on the exposure intervention variable are used and to obtained confidential intervals, bootstrap. Finally, effects are estimated by contrasting different potential outcomes [2].

The steps followed for this analysis are summarised in Appendix I Figure 1 and they are as follows:

Step 1: The relationships between variables observed in the data, represented in Appendix I Figure 2, were estimated applying regression models:

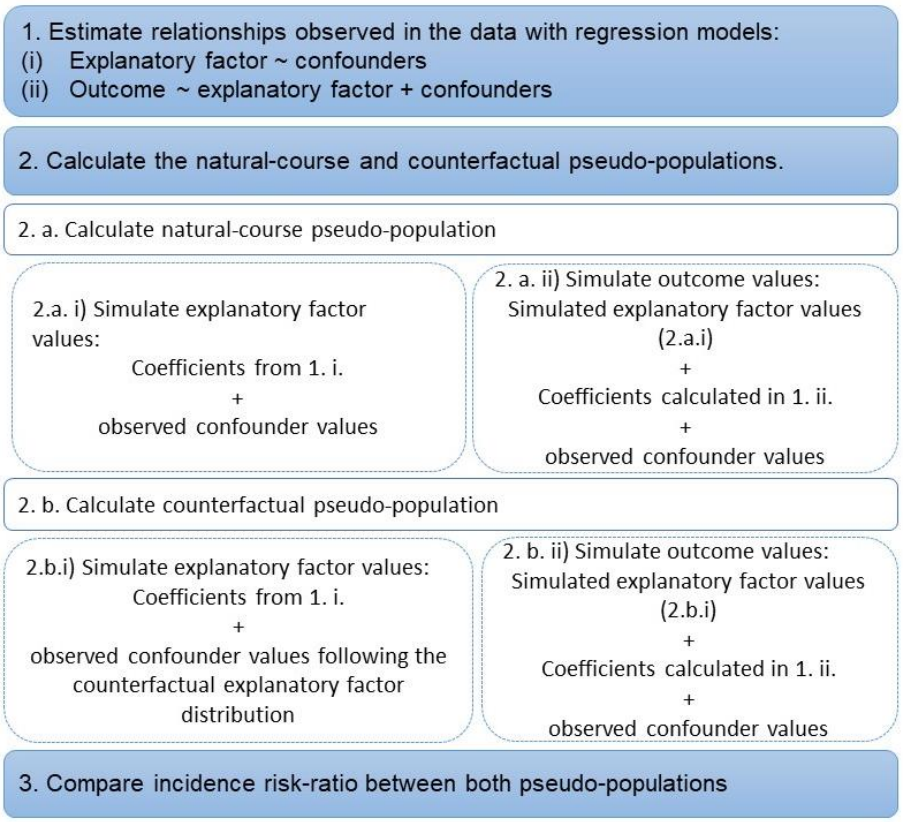
(i) First, the relationship between the causal explanatory factor and confounders.

(ii) Second, the relationship between the outcome, explanatory factor and confounders.

Step 2: The natural and counterfactual pseudo-populations were calculated. The aim of this step was to estimate the incidence risk of MACE for both sexes and then to calculate the risk ratio between men and women in both situations.

a. For the natural course pseudo-population, it was needed to simulate values for the corresponding causal explanatory factor and for the outcome, respectively. To do the first, the coefficients from step 1 (i) and the confounder values were used. To do the second (the outcome values), the simulated causal explanatory factor values, the coefficients from 1 (ii) and the observed confounder values were used.
b. For the counterfactual pseudo-population, simulated values for the causal explanatory factor and for the outcome were also calculated. The process is similar to the previous one (step 2.a) and the coefficients used were the same, but in this case a counterfactual causal explanatory distribution (i.e. for those who have the causal explanatory factor, it will be supposed that they don't have it) was applied.

Step 3: Finally, the incidence risk-ratio in the natural and counterfactual course pseudo-populations was compared to obtain the estimates of the contribution.



Appendix II Figure 1: Steps followed in the analysis

REFERENCES

1. Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

2. Wang A, Arah OA. G-computation demonstration in causal mediation analysis. Eur J Epidemiol. 2015 Oct 1;30(10):1119–27.

APPENDIX II: COUNTERFACTUAL ANALYSIS RESULTS BY EXPLANATORY FACTOR

Appendix II Table 1: Counterfactual analyses for each explanatory factor

Incidence risk ratio	Natural- course RR (95%CI)	Counterfactual RR (95%CI)	Percentage contribution (95%CI)
Diabetes analysis	1.83 (1.74, 1.92)	1.74 (1.65, 1.83)	10.53 (7.99, 13.08)
Hypercholesterolaemia analysis	1.81 (1.72, 1.9)	1.82 (1.73, 1.91)	-0.29 (-2.39, 1.81)
Hypertension analysis	1.82 (1.72, 1.91)	1.81 (1.72, 1.9)	1.02 (-1.38, 3.43)
Socioeconomic status analysis	1.81 (1.72, 1.9)	1.86 (1.77, 1.96)	-5.99 (-9.2, -2.77)

RR: Aged-adjusted Risk Ratio; 95%CI: Confidence Intervals

Percentage contribution is calculated as follows:

$$\text{Percentage contribution} = 1 - \frac{\text{Counterfactual RR} - 1}{\text{Natural} - \text{course RR} - 1}$$