Study Protocol

P4-C3-004

P4-C2-008

P4-C2-009

P4-C2-010

DARWIN EU® - Capturing obesity, obesity-related variables, and changes in weight over time across the DARWIN EU® network

18/12/2025

Version 5.0

Authors: Nicholas Hunt, Julieta Politi, Melissa Leung, Katia Verhamme

Public

This document represents the views of the DARWIN EU® Coordination Centre only and cannot be interpreted as reflecting those of the European Medicines Agency or the European Medicines Regulatory Network.

CONTENTS

Study title	DARWIN EU® - Capturing obesity, obesity-related variables, and changes in weight over time across the DARWIN EU® network
Protocol version	V5.0
Date	18/12/2025
EUPAS number	EUPAS1000000820
Active substance	None
Medicinal product	None
Study objectives	1. To estimate the overall prevalence of individuals with records of obesity-related conditions, measurements, lifestyle factors, and procedures within the DARWIN EU® network. 2. To characterise reporting of anthropometric measurements and lifestyle factors within the DARWIN EU® network within an obese and general population a. To describe the frequency of recording and median number (IQR) of BMI, weight, height, cholesterol, waist circumference, diet, and physical activity records per individual. b. To summarise the first recorded measurement values of BMI, weight, and height measurements by Q05, Q25, Q50 (median), Q75, Q95, and range per individual, in the study period. 3. To estimate the timing between sequential BMI and weight measurement recordings per individual within an obese and general population: a. The mean and median rate of measurement recordings b. The mean time elapsed between each of the first five measurement recordings c. The mean cumulative function of measurement recordings 4. To compare the characteristics of individuals with obesity based on disease codes versus those with obesity defined by BMI cutoff value in terms of demography, comorbidity, use of concomitant medications, procedures, and lifestyle factors.
Countries of study	P4-C3-004: Belgium, Croatia, France, Germany, and The Netherlands, P4-C2-008: Estonia, Italy, Spain, and the United Kingdom P4-C2-009: Denmark, Greece, Portugal, and Spain P4-C2-010: Finland, France, Hungary, Norway, and Sweden
Authors	Nicholas Hunt, n.hunt@darwin-eu.org Julieta Politi, j.politi@darwin-eu.org Melissa Leung, m.leung@darwin-eu.org Katia Verhamme, k.verhamme@darwin-eu.org

LIST OF ABBREVIATIONS

Acronyms/term	Description
AEMPS	Agencia Española de Medicamentos y Productos Sanitarios
APHM	Assistance Publique Hôpitaux de Marseille
ATC	Anatomical Therapeutic Chemical
BIFAP	Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público
BMI	Body mass index
CDM	Common Data Model
CDW Bordeaux	Clinical Data Warehouse of Bordeaux University Hospital
CC	Coordinating centre
CI	Confidence interval
CPRD GOLD	Clinical Practice Research Datalink GOLD
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DK-DHR	Danish Data Health Registries
DKMA	Danish Medicines Agency
DOI	Declaration of Interests
DQ	Data quality
DQD	Data Quality Dashboard
DRE	Digital Research Environment
EBB	Estonian Biobank
EHR	Electronic Health Records
EMA	European Medicines Agency
EMDB-ULSEDV	Egas Moniz Health Alliance database - Entre o Douro e Vouga
EMDB-ULSGE	Egas Moniz Health Alliance database - Gaia E Espinho
EMDB-ULSRA	Egas Moniz Health Alliance database - Baixo Vouga (Região de Aveiro)
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU	European Union
EUPAS	EU Post-Authorisation Studies Register
FinOMOP-ACI Varha	FinOMOP - Auria Clinical Informatics
FinOMOP-THL	FinOMOP - Finnish Care Register for Health Care
FinOMOP-TaUH Pirha	FinOMOP - Tampere University Hospital patient cohort
FISABIO	The Foundation for the Promotion of Health and Biomedical Research of Valencia Region
Erasmus MC	Erasmus Medical Center
GDPR	General Data Protection Regulation
H12O	Hospital Universitario 12 de Octubre
HI-SPEED	Health Impact - Swedish Population Evidence Enabling Data-linkage
ICD	International Classification of Diseases
IDIAP	Institut Universitari d'Investigació en Atenció Primària

IdISBa	Fundación Instituto de Investigación Sanitaria Islas Baleares
IMASIS	Institut Municipal Assistència Sanitària Information System
InGef RDB	Institut für angewandte Gesundheitsforschung Research Database
IP	Inpatient
IPCI	Integrated Primary Care Information
IQR	Interquartile range
IQVIA DA Germany	IQVIA Disease Analyzer Germany
IQVIA LPD Belgium	IQVIA Longitudinal Patient Database Belgium
IRB	Institutional Review Board
NLHR	Norwegian Linked Health Registry data
NAJS	Nacionalni Javnozdravstveni Informacijski Sustav
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OP	Outpatient
PGH	Papageorgiou General Hospital
POLIMI	Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico
PRISIB	Platafoma de Recerca en Informació Sanitària de les Illes Balears
PSMar	Consorci Mar Parc de Salut Barcelona
RWD	Real-world data
RxNorm	Medical prescription normalized
SNOMED	Systematized Nomenclature of Medicine
SUCD	Semmelweis University Clinical Data
ULSM-RT	Unidade Local de Saúde de Matosinhos - Realtime Database
VID	Valencia Health System Integrated Dataset
WHO	World Health Organisation

1. TITLE

DARWIN EU® - Capturing obesity, obesity-related variables, and changes in weight over time across the DARWIN EU® network

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator/epidemiologists	Nicholas Hunt Julieta Politi Melissa Leung Katia Verhamme	Erasmus MC Erasmus MC Erasmus MC Erasmus MC
Data Scientists	Ioanna Nika Maarten van Kessel Ross Williams	Erasmus MC Erasmus MC Erasmus MC
Study Manager	Natasha Yefimenko Nosova	Erasmus MC
Data Partner*	**Names**	**Organisation**
P4-C3-004		
IQVIA LPD Belgium	Dina Vojinovic Hugo Vernooij Gargi Jadhav Isabella Kaczmarczyk	IQVIA
NAJS	Anamaria Jurčević Antea Jezidžić Jakov Vuković Ivan Pristaš	Croatian Institute of Public Health
CDW Bordeaux	Romain Griffer Guillaume Verdy Vianney Jouhet	Bordeaux University Hospital
InGef RDB	Raeleesha Norris Josephine Jacob Annika Vivirito Alexander Harms	Institut für angewandte Gesundheitsforschung Berlin GmbH
IQVIA DA Germany	Dina Vojinovic Gargi Jadhav Isabella Kaczmarczyk	IQVIA
IPCI	Katia Verhamme Mees Mosseveld Guido van Leeuwen	Erasmus MC
P4-C2-008		
EBB	Marek Oja	University of Tartu

	Raivo Kolde	
POLIMI	Gianluigi Galli	Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico
H12O	Paula Rubio Mayo Javier de la Cruz Juan Luis Cruz Bermúdez Noelia García Barrio	Hospital Universitario 12 de Octubre (H12O)
PRISIB	Pau Pericas Pulido	Fundación Instituto de Investigación Sanitaria Islas Baleares (IdISBa)
SIDIAP	Elena Roef Herranz Laura Granés González Agustina Giuliodori Picco Anna Palomar Talita Duarte Salles	Institut Universitari d'Investigació en Atenció Primària (IDIAP Jordi Gol)
CPRD GOLD	Antonella Delmestri Marta Pineda Moncusí	University of Oxford
UKBB	Antonella Delmestri Junqing Xie Mandickel Kamtengeni	University of Oxford
P4-C2-009		
DK-DHR	Elvira Bräuner Susanne Bruun	Danish Medicines Agency (DKMA)
PGH	Achilleas Chytas Alexandros Rekkas Anastasia Farmaki Pantelis Natsiavas	Papageorgiou General Hospital (PGH)
EMDB-ULSEDV	Ana Pinto Luís Ruano Luís Andrade	Clinical Academic Center Egas Moniz Health Alliance
EMDB-ULSGE	Ana Pinto Firmino Machado Natália Araújo	Clinical Academic Center Egas Moniz Health Alliance
EMDB-ULSRA	Ana Pinto Mesquita Bastos Joana Guimarães	Clinical Academic Center Egas Moniz Health Alliance
ULSM-RT	Fernando Montenegro Sá Nuno Silva	Unidade Local de Saúde de Matosinhos (ULSM)
BIFAP	Cristina Justo-Astorgano Elisa Martin-Merino Hermenegildo Martínez-Alcalá García	Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)

	Miguel-Angel Macia-Martinez Ana Llorente-Garcia	
IMASIS	Juan Manuel Ramírez-Anguita Angela Leis Miguel-Angel Mayer	Consorci Mar Parc de Salut Barcelona (PSMar)
VID	Celia Robles Cabaniñas Fran Llopis Cardona Gabriel Sanfélix Gimeno	The Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO)
P4-C2-010		
FinOMOP-ACI Varha	Tiina Wahlfors Pia Tajanen-Doumbouya	Hospital District of Southwest Finland
FinOMOP-THL	Tiina Wahlfors Toni Lehtonen Gustav Klingstedt Petteri Hovi	Finnish Institute for Health and Welfare
FinOMOP-HUS	Tiina Wahlfors Eric Fey Kimmo Porkka	Helsinki University Hospital, Hospital District of Helsinki and Uusimaa
FinOMOP-TaUH Pirha	Tiina Wahlfors Hakkarainen Leena Sampo Kukkurainen Kati Kristiansson	Tampere University Hospital
APHM	Vanessa Pauly Laurent Boyer Dorian Grousset	Assistance Publique Hôpitaux de Marseille
SUCD	Ágota Mészáros Bagyura Zsolt István Kiss Loretta Zsuzsa Héja Tibor	Semmelweis University
NLHR	Hedvig Marie Egeland Nordeng Nhung Trinh Saeed Hayati	University of Oslo
HI-SPEED	Fredrik Nyberg Huiqi Li	Swedish Medical Products Agency - Gothenburg University

*Data partners do not have an investigator role. Data partners execute code at their data source, review and approve their results.

3. ABSTRACT

Title

DARWIN EU® - Capturing obesity, obesity-related variables, and changes in weight over time across the DARWIN EU® network

Rationale and background

Few studies on obesity and weight management using electronic healthcare databases have used anthropometric measurements, largely due to methodological challenges. This study aims to assess the availability and completeness of data on obesity, weight, and obesity-related variables across the network of DARWIN EU® data partners. These findings will inform the viability of conducting further Real-World Data (RWD) studies on obesity and related comorbidities.

Research question and objectives

Research questions

1. What proportion of individuals within the DARWIN EU® network have records of clinical conditions related to obesity, as well as obesity related measurements, observations, and procedures?
2. What are the number of records (repeated measures) per individual and values of anthropometric measurements and lifestyle factors captured over time in the DARWIN EU® network?
3. How do the demographic and obesity-related conditions, medications, procedures, and lifestyle factors of individuals with a high body mass index (BMI) compare to those of individuals with a recorded condition of obesity within the DARWIN EU® network?

Objectives

1. To estimate the overall prevalence of individuals with records of obesity-related conditions, measurements, lifestyle factors, and procedures within the DARWIN EU® network.
2. To characterise reporting of anthropometric measurements and lifestyle factors within the DARWIN EU® network within an obese and general population
    a. To describe the frequency of recording and median number (IQR) of BMI, weight, height, cholesterol, waist circumference, diet, and physical activity records per individual.
    b. To summarise the first recorded measurement values of BMI, weight, and height measurements by Q05, Q25, Q50 (median), Q75, Q95, and range per individual, in the study period.
3. To estimate the timing between sequential BMI and weight measurement recordings per individual within an obese and general population:
    a. The mean and median rate of measurement recordings
    b. The mean time elapsed between each of the first five measurement recordings
    c. The mean cumulative function of measurement recordings
4. To compare the characteristics of individuals with obesity based on disease codes versus those with obesity defined by BMI cutoff value in terms of demography, comorbidity, use of concomitant medications, procedures, and lifestyle factors.

Methods

Study design

We will perform a retrospective cohort study with three components: (i) to describe the proportion of individuals with obesity-related variables (objective 1); (ii) to perform a population-level characterisation to describe the characteristics and frequency of records of anthropometric, lab, and lifestyle measurements

(objective 2 and 3); and (iii) a patient-level characterisation to describe the characteristics of individuals with obesity (objective 4).

Population

The study population will include all individuals present in the database during the study period 01/01/2010 (or start of available data) to 30/06/2025 (or to the end of available data) and with at least 365 days of database history prior to index date (except for individuals in hospital data sources).

Variables

Outcome:

For prevalence estimation, outcomes will include condition records of obesity and measurements of BMI, weight, height, cholesterol, and waist circumference.

Relevant covariates:

The number of occurrences of BMI, weight, height, cholesterol, and waist circumference measurements will be calculated, and the values of the measurements of BMI, weight, and height will be described. Individuals defined as obese will be characterised by the following: condition occurrences of diabetes mellitus type 2, hypertension, ischemic heart disease, chronic kidney disease, hypothyroidism, hypertriglyceridemia, metabolic syndrome X, Cushing's syndrome, knee arthrosis, obstructive sleep apnoea, mental health disorders (including depression and anxiety), dyslipidaemia, metabolic dysfunction-associated steatotic liver disease, non-alcoholic fatty liver disease, stenosis of liver, and cancer; drug records of glucagon-like peptide-1 (GLP-1) or glucose-dependent insulinotropic polypeptide (GIP) receptor agonists, Orlistat, Metformin, and Naltrexone-bupropion; procedure record of bariatric surgery; and observations of diet, physical activity, and smoking status.

Data sources

P4-C3-004:

1. Belgium: IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium)
2. Croatia: Croatian National Public Health Information System (NAJS)
3. France: Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)
4. Germany: InGef Research Database (InGef RDB)
5. Germany: IQVIA Disease Analyzer Germany (IQVIA DA Germany)
6. Netherlands: Integrated Primary Care Information (IPCI)

P4-C2-008:

1. Estonia: Estonian Biobank (EBB)
2. Italy: Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI)
3. Spain: Hospital Universitario 12 de Octubre (H12O)
4. Spain: Platafoma de Recerca en Informació Sanitària de les Illes Balears (PRISIB)
5. Spain: The Information System for Research on Primary Care (SIDIAP)
6. United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)
7. United Kingdom: UK BioBank (UKBB)

P4-C2-009:

1. Denmark: Danish Data Health Registries (DK-DHR)

2. Greece: Papageorgiou General Hospital (PGH)
3. Portugal: Egas Moniz Health Alliance database - Entre o Douro e Vouga (EMDB-ULSEDV)
4. Portugal: Egas Moniz Health Alliance database - Gaia E Espinho (EMDB-ULSGE)
5. Portugal: Egas Moniz Health Alliance database - Baixo Vouga (Região de Aveiro) (EMDB-ULSRA)
6. Portugal: Unidade Local de Saúde de Matosinhos Realtime Database (ULSM-RT)
7. Spain: Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público (BIFAP)
8. Spain: Institut Municipal Assistència Sanitària Information System (IMASIS)
9. Spain: Valencia Health System Integrated Dataset (VID)

P4-C2-010:

1. Finland: Auria Clinical Informatics (FinOMOP-ACI Varha)
2. Finland: Finnish Care Register for Health Care (FinOMOP-THL)
3. Finland: Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)
4. France: Assistance Publique Hôpitaux de Marseille (APHM)
5. Hungary: Semmelweis University Clinical Data (SUCD)
6. Norway: Norwegian Linked Health Registry data (NLHR)
7. Sweden: Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)

Study size

No sample size has been calculated, as this is an exploratory study which will not test a specific hypothesis; instead, the study specifically aims to estimate the frequency of recording and describe the variables under study.

Statistical analysis

The frequency and period prevalence of obesity and obesity-related measurements will be estimated in all individuals in the data sources, overall and stratified by sex and age categories. Characteristics will be described by means of median age, sex, and the covariates of interest, which will be reported as counts and proportions. The statistical analyses will be performed based on OMOP common data model mapped data using the *IncidencePrevalence* and *CohortCharacterisation* R packages. A minimum cell counts of 5 will be used when reporting results.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study milestones and deliverables	Planned dates*
Final Study Protocol	November 2025
Creation of Analytical code	November 2025
Execution of Analytical Code on the data	December 2025
Interim Study Report (P4-C3-004 results)	January 2026
Draft Study Report (P4-C3-004, P4-C2-008, P4-C2-009, P4-C2-010 results)	February 2026
Final Study Report	March 2026

*Planned dates are dependent on obtaining approvals from the internal review boards of the data sources.

6. RATIONALE AND BACKGROUND

Obesity is a multi-factorial disease that has reached epidemic proportions in many countries.[1-3] It is associated with an increased risk for a variety of comorbidities, higher lifetime health care expenditures, and a greater risk for mortality.[3-5]

Research on obesity and related comorbidities (cardiovascular diseases, diabetes, etc.) requires significant resources for the long-term follow-up and detailed and accurate clinical characterisation of patients. Body mass index (BMI), calculated from height and weight, is often seen as the most used measure of obesity, assessed during clinical encounters, and when theoretically entered into electronic healthcare databases. Additionally, BMI, as well as other anthropometric or metabolic measures, is commonly used to assess the effectiveness of weight loss therapies. Many of these variables, including laboratory test results, presence of comorbid conditions, and information on medication use, are however recorded in electronic healthcare databases with different levels of granularity and accuracy.

Few studies on obesity and weight management have used data derived from electronic healthcare databases due to some methodological challenges. For instance, when BMI measures are not available directly, these types of real-world data (RWD) sources may systematically omit patients' heights and weights necessary to calculate BMI or to evaluate changes in patients' weight over time, especially in the long term. Therefore, this study aims to determine the extent to which data on obesity, weight, and obesity-related variables are captured across the network of DARWIN EU® data partners, to inform the viability of conducting further RWD studies to answer research questions related to obesity, its management, and comorbidities.

7. RESEARCH QUESTION AND OBJECTIVES

Research questions

1. What proportion of individuals within the DARWIN EU® network have records of clinical conditions related to obesity, as well as obesity related measurements, observations, and procedures?
2. What are the number of records (repeated measures) per individual and values of anthropometric measurements and lifestyle factors captured over time in the DARWIN EU® network?
3. How do the demographic and obesity-related conditions, medications, procedures, and lifestyle factors of individuals with a high body mass index (BMI) compare to those of individuals with a recorded condition of obesity within the DARWIN EU® network?

Research objectives

The specific objectives of this study are:

1. To estimate the overall prevalence of individuals with records of obesity-related conditions, measurements, lifestyle factors, and procedures within the DARWIN EU® network.
2. To characterise reporting of anthropometric measurements and lifestyle factors within the DARWIN EU® network within an obese and general population
    a. To describe the frequency of recording and median number (IQR) of BMI, weight, height, cholesterol, waist circumference, diet, and physical activity records per individual.
    b. To summarise the first recorded measurement values of BMI, weight, and height measurements by Q05, Q25, Q50 (median), Q75, Q95, and range per individual, in the study period.
3. To estimate the timing between sequential BMI and weight measurement recordings per individual within an obese and general population:
    a. The mean and median rate of measurement recordings
    b. The mean time elapsed between each of the first five measurement recordings
    c. The mean cumulative function of measurement recordings
4. To compare the characteristics of individuals with obesity based on disease codes versus those with obesity defined by BMI cutoff value in terms of demography, comorbidity, use of concomitant medications, procedures, and lifestyle factors.

Results will be stratified by sex and into two age groups: paediatric (aged 2–18 years) and adult (aged 19-plus).

8. RESEARCH METHODS

8.1. Study design

A retrospective cohort study will be conducted using routinely collected health data from thirty data sources from sixteen countries including fourteen EU countries. The study design for objective 1 is illustrated in **Figure 1**. Objectives 2 and 3 are illustrated in both **Figure 2** and **Figure 3**, and objective 4 in **Figure 4**.

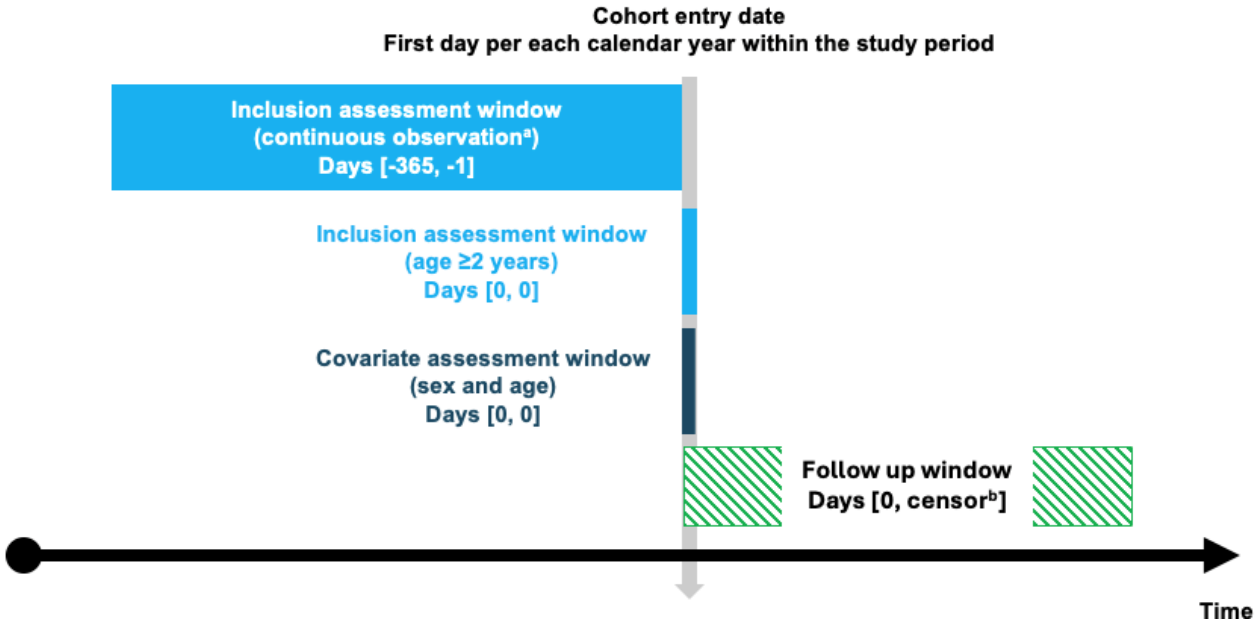


Figure 1. Graphical depiction of the study design for objective 1.

a. Does not apply to individuals in hospital data sources
b. Death, disenrollment, end of data source availability, end of each calendar year (i.e., 31st December), or end of the study period

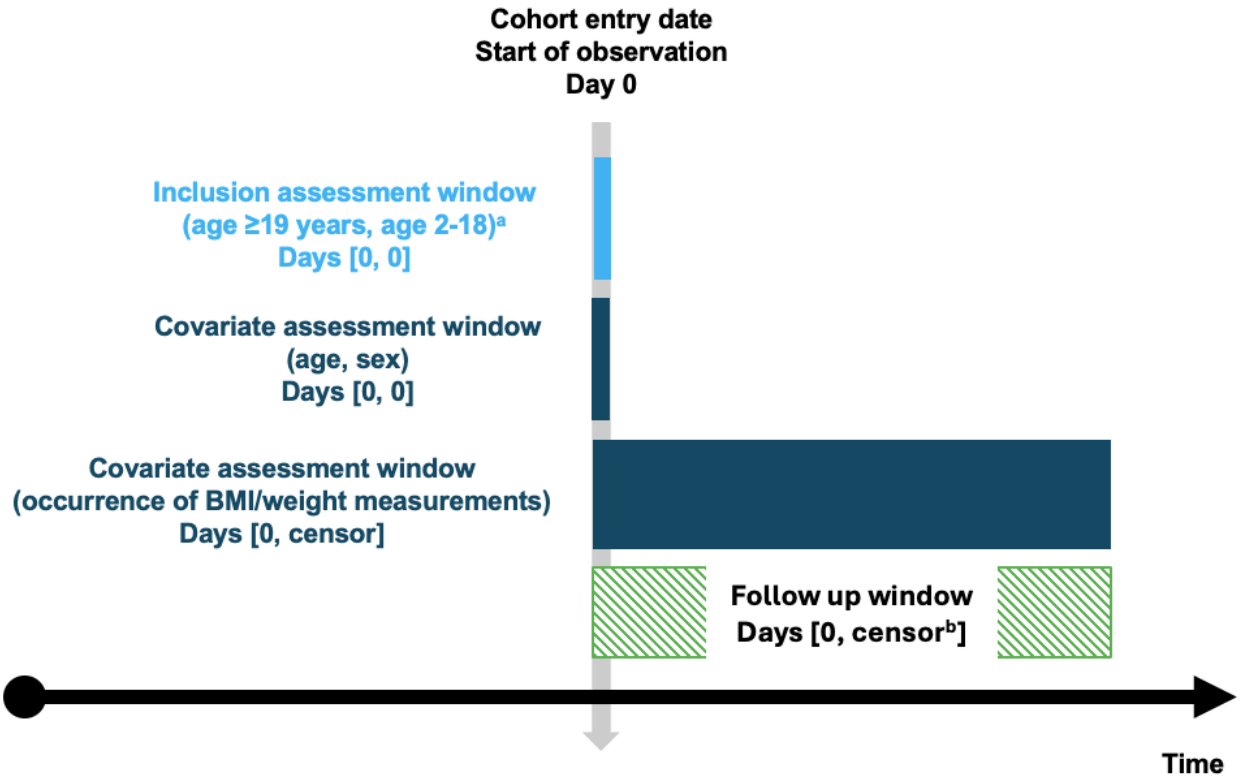


Figure 2. Graphical depiction of the study design for objectives 2 and 3 (general population).

a. Individuals are included in one of two age groups
b. Death, disenrollment, end of data source availability, individual's 19th birthday (if aged <19 years old), or end of the study period

BMI = body mass index

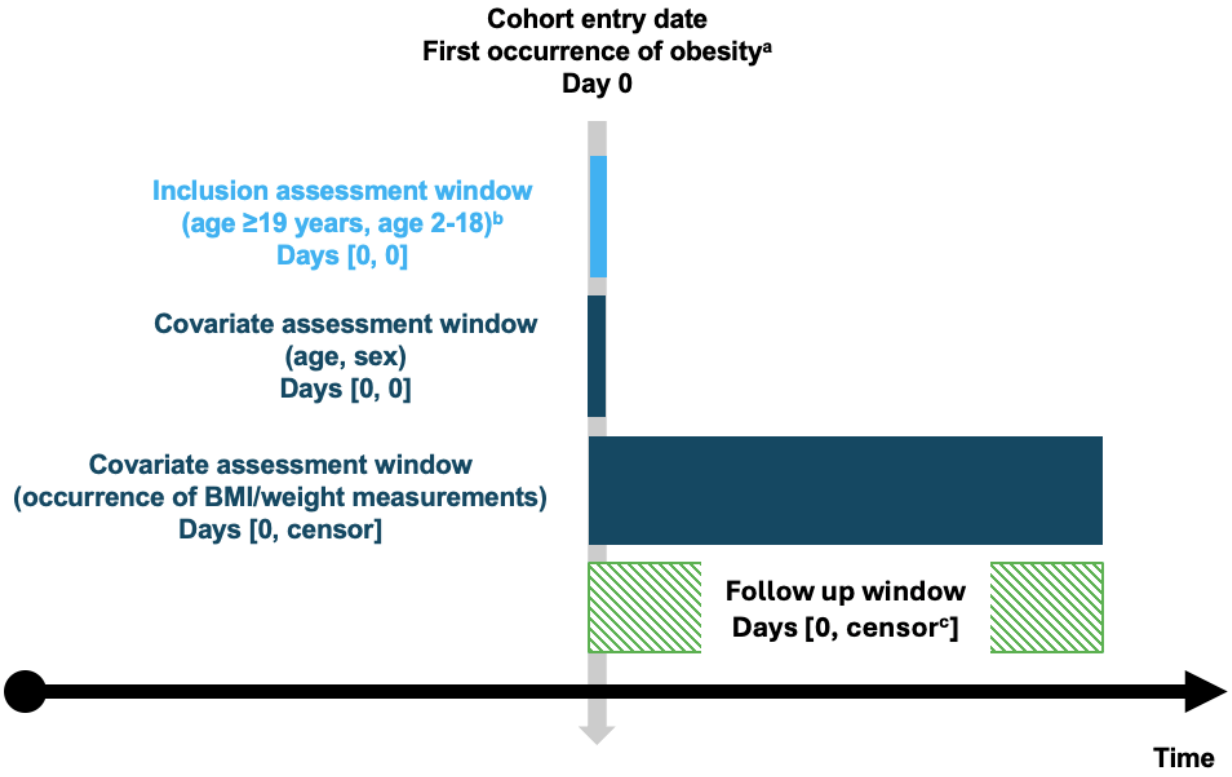


Figure 3. Graphical depiction of the study design for objectives 2 and 3 (obese population).

a. Obesity is defined as the occurrence of an obesity condition record or BMI record with a value which is determined as obese. A BMI measurement defined as obese is one with a value of ≥30 kg/m^2 for individuals ≥19 years old, and for individuals aged 2 to 18, sex-specific BMI-for-age z-score that is ≥2 standard deviations, as per the WHO growth reference curves.[6, 7]
b. Individuals are included in one of two age groups
c. Death, disenrollment, end of data source availability, individual's 19th birthday (if aged <19 years old), or end of the study period

BMI = body mass index

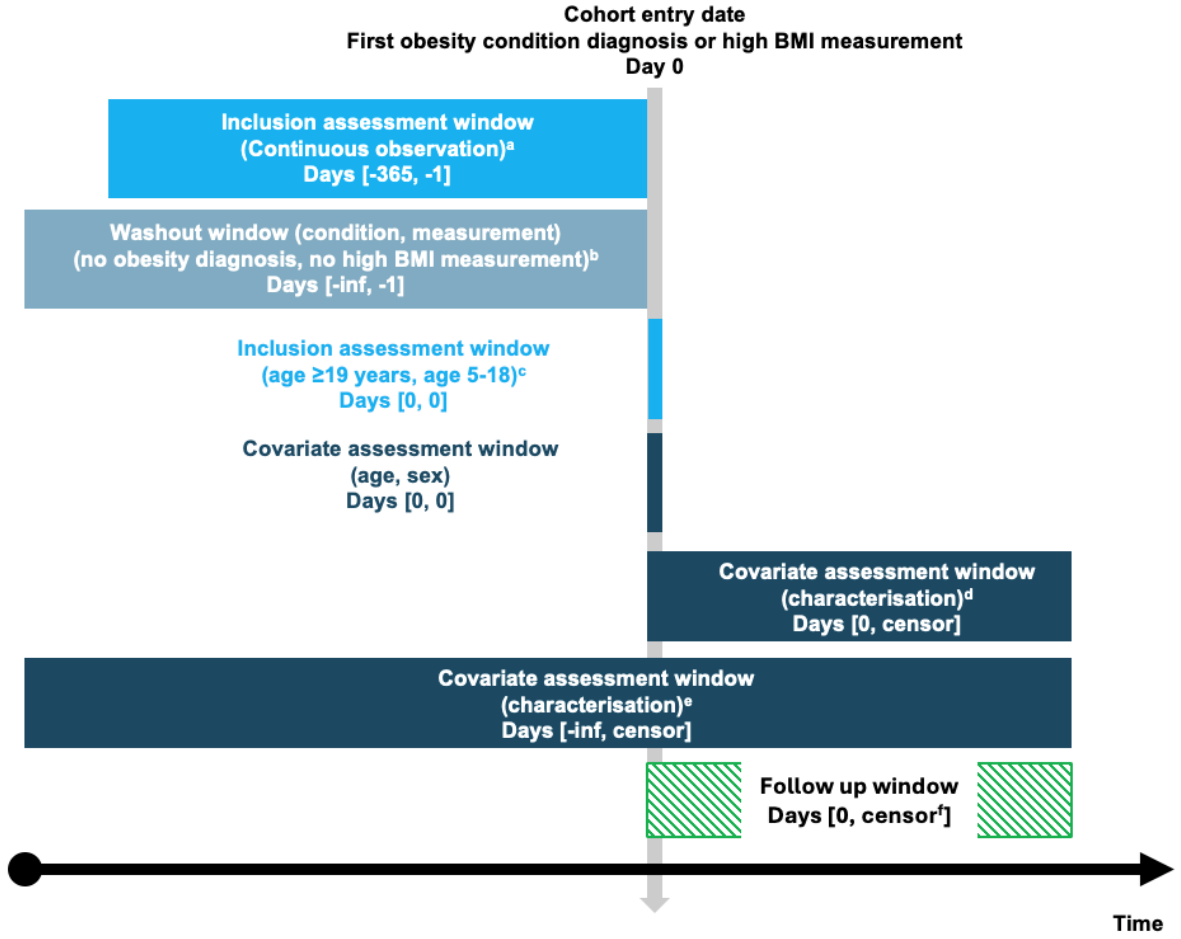


Figure 4. Graphical depiction of the study design for objective 4.

a. Does not apply to individuals in hospital data sources
b. No prior obesity diagnosis applies to the cohort of individuals defined by obesity condition record, and no prior BMI ≥30 kg/m^2 applies to the cohort of individuals defined by BMI ≥30 kg/m^2 in individuals ≥19 years old. In individuals aged 2–18 years, high BMI is defined using sex-specific BMI-for-age z-scores, where a value of ≥2 standard deviations is classified as obesity.
c. Enter one of two age-specific cohorts
d. Procedure occurrences (defined in **Section 8.6.3.**)
e. Comorbidities, lifestyle factors, drugs (all defined in **Section 8.6.3.**)
f. Death, disenrollment, end of data source availability, or end of the study period (30/06/2025)

BMI = Body mass index

8.2. Study setting and data sources

This study will be conducted using routinely collected data from 29 data sources from the DARWIN EU® network of data partners from sixteen data sources. All data were a priori mapped to the OMOP CDM.

Data sources

P4-C3-004:

1. Belgium: IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium)
2. Croatia: Croatian National Public Health Information System (NAJS)
3. France: Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)
4. Germany: InGef Research Database (InGef RDB)

5. Germany: IQVIA Disease Analyzer Germany (IQVIA DA Germany)
6. Netherlands: Integrated Primary Care Information (IPCI)

P4-C2-008:

7. Estonia: Estonian Biobank (EBB)
8. Italy: Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI)
9. Spain: Hospital Universitario 12 de Octubre (H12O)
10. Spain: Platafoma de Recerca en Informació Sanitària de les Illes Balears (PRISIB)
11. Spain: The Information System for Research on Primary Care (SIDIAP)
12. United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)
13. United Kingdom: UK BioBank (UKBB)

P4-C2-009:

14. Denmark: Danish Data Health Registries (DK-DHR)
15. Greece: Papageorgiou General Hospital (PGH)
16. Portugal: Egas Moniz Health Alliance database - Entre o Douro e Vouga (EMDB-ULSEDV)
17. Portugal: Egas Moniz Health Alliance database - Gaia E Espinho (EMDB-ULSGE)
18. Portugal: Egas Moniz Health Alliance database - Baixo Vouga (Região de Aveiro) (EMDB-ULSRA)
19. Portugal: Unidade Local de Saúde de Matosinhos Realtime Database (ULSM-RT)
20. Spain: Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público (BIFAP)
21. Spain: Institut Municipal Assistència Sanitària Information System (IMASIS)
22. Spain: Valencia Health System Integrated Dataset (VID)

P4-C2-010:

23. Finland: Auria Clinical Informatics (FinOMOP-ACI Varha)
24. Finland: Finnish Care Register for Health Care (FinOMOP-THL)
25. Finland: Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)
26. France: Assistance Publique Hôpitaux de Marseille (APHM)
27. Hungary: Semmelweis University Clinical Data (SUCD)
28. Norway: Norwegian Linked Health Registry data (NLHR)
29. Sweden: Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)

Data source justification and key characteristics

All studies share the same objectives but differ in the data partners involved due to the need to assess how obesity, obesity-related variables, and changes in weight over time are captured across the DARWIN EU® network. As a consequence, the four studies overall include data partners who constitute the majority of the DARWIN EU® network, with a few exceptions: neonatal data sources (United Kingdom: National Neonatal Research Database (NNRD)) and disease specific registries (Netherlands: Netherlands Cancer Registry (NCR), Norway: Cancer Registry Norway (CRN), Multi-country network: HARMONY - Acute Lymphoblastic Leukemia (ALL), Multi-country network HARMONY - Acute Myeloid Leukemia (AML), Multi-country network : HARMONY - Chronic Myeloid Leukemia (CML), and Multi-country network: HARMONY - Multiple Myeloma (MM)), as they can be considered less relevant and then out of scope in regards to the

study objectives. One data source is currently unavailable to execute DARWIN EU® studies (France: Système National des Données de Santé (SNDS)) and is therefore not included in any of the studies.

The included data sources for this study and the three other routine-repeated studies can be seen in **Table S1 in Annex I**. Further information on the data sources planned for use in this study is provided in **Annex I**.

8.3. Study period

The study period is from 01/01/2010 (or start of available data e.g., 01/01/2016 in InGef RDB and HI-SPEED) to 30/06/2025 or the most recent data available for each contributing data source.

8.4. Follow-up

For objective 1, the index date the first date after the study start date at which an individual has 365 days follow-up or becomes two-years old. End of follow-up will be defined as the earliest of loss to follow-up, death, end of the study (30/06/2025), or end of observation period (the latest available data), whichever occurs first.

To calculate the proportion of individuals with obesity records, an appropriate denominator population is required, for which the population is required to be observable. Study participants who contribute person time during the study period will be included in the denominator population.

An example of entry and exit into the denominator population is shown in **Figure 5**. In this example, person ID 1 already has sufficient prior history before the study start date, and the observation period ends after the study end date, so this person will contribute during the complete study period. Person IDs 2 and 4 only enter the study when they have sufficient prior history. Person ID 3 leaves when exiting the data source (the end of the observation period). Lastly, person ID 5 has two observation periods in the data source. The first period contributes time from study start until end of observation period, the second starts contributing time again once sufficient prior history is reached and exits at study end date.

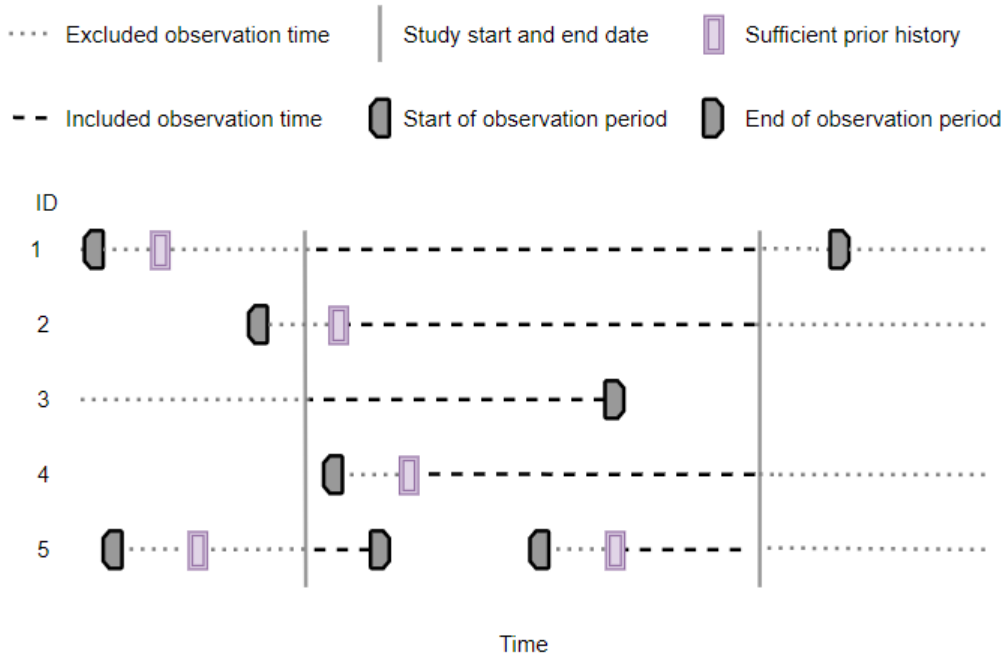


Figure 5. Included observation time for the denominator population.

Objective 2 and 3 includes two study populations, general population and the obese population. For the general population, individuals will be followed from the start of observation once the individual becomes 2- or 19-years old depending on entering the paediatric or adult cohorts. For the obese population, follow-up will begin with the first BMI measurement defined as obese or condition occurrence of obesity. A BMI measurement defined as obese is one with a value of ≥30 kg/m^2 for individuals ≥19 years old, and for individuals aged 2 to 18, sex-specific BMI-for-age z-score that is ≥2 standard deviations, as per the WHO growth reference curves.[6, 7] Individuals will be followed until loss to follow-up, death, end of the study (30/06/2025), or end of observation period (the latest available data), whichever occurs first. We perform the study in paediatric and adult populations so at the individual's 19th birthday follow-up begins in the adult cohort, and they are censored from the paediatric cohort.

For objective 4, in individuals ≥19 years old, the index date is the first condition occurrence of obesity or the first BMI measurement ≥30 kg/m^2 after the individual's 19th birthday when there is 365 days observation, with no prior BMI measurement ≥30 kg/m^2. In individuals 2–18, the index date is the first condition occurrence of obesity, or first occurrence of sex-specific BMI-for-age z-score that is ≥2 standard deviations within the study period, as per the WHO growth reference curves.[6, 7]

8.5. Study population with inclusion and exclusion criteria

For **objective 1**, the general population will constitute the study population:

Inclusion criteria

- Minimum 365 days of available history before index date (except for hospital data source CDW Bordeaux and APHM)
- Aged ≥2 years of age. In children <24 months, there is no universally accepted clinical definition of underweight/overweight/obesity. Assessment using weight-for-length (WFL) standards is recommended. Because infant growth is rapid and non-linear (especially from birth to 6 months), single cross-sectional measurements are insufficient, and would require sequential time points beyond the scope of this study

For objective 2 and 3, there are two different study populations defined. The study population consists of the general population (regardless of prior observation) and an obese population.

For **objectives 2 and 3**:

- *General population*

  Inclusion criteria

  - Aged ≥2 years of age
- *Obese population*

  Inclusion criteria

  - Aged ≥2 years of age
  - Obesity, defined as an occurrence of an obesity condition record or BMI value over the threshold that is defined as obese. For adults, we consider BMI measurement to be obese at ≥30kg/m2 (≥19 years old), and for paediatric individuals (aged 2 to 18) we consider a BMI measurement at the sex-specific BMI-for-age z-score ≥2 standard deviations within the study period as obese, as per the WHO growth reference curves.[6, 7] To implement this, we will assign an age and sex specific cut- off score to define obesity. At each measurement, we will take into account individual age and sex. For example, in all 5-year-old girls, we will use the cut-off for 5-years and six months, where obese is categorised as a

BMI measurement ≥19kg/m^2. Since it is not possible to attain age in months and z-score varies by month, we will use the value for the sixth month of each age-year.

For **objective 4**, we will explore characterising individuals aged 2–18 years defined as obese by a condition diagnosis and by high BMI:

- *Characterisation of those defined as obese by condition*

  Inclusion criteria

  - Minimum 365 days of available history before index date (except for hospital data source CDW Bordeaux)
  - Aged 2–18 years of age, since the WHO BMI for age reference values are applicable for these age groups.[6, 7]
  - Condition occurrence of obesity within the study period

  Exclusion criteria

  - Condition occurrence of obesity at any time prior to the index date

- *Characterisation of those defined as obese by BMI measurements*

  Inclusion criteria

  - Minimum 365 days of available history before index date (except for hospital data source CDW Bordeaux)
  - Aged 2–18 years of age, since the WHO BMI for age reference values are applicable for these age groups.[6, 7]
  - Sex-specific BMI-for-age z-score ≥2 standard deviations within the study period, as per the WHO growth reference curves.[6, 7]To implement this, we will assign an age and sex specific cut-off score to define obesity. At each measurement, we will take into account individual age and sex. For example, in all 5-year-old girls, we will use the cut-off for 5-years and six months, where obese is categorised as a BMI measurement ≥19kg/m^2. Since it is not possible to attain age in months and z-score varies by month, we will use the value for the sixth month of each age-year.

  Exclusion criteria

  - Sex-specific BMI-for-age z-score ≥2 standard deviations at any time prior to the index date.[6, 7]

8.6. Variables

8.6.1. Exposure

There are no medications specifically assessed under this study. Characterisation by concomitant medications is described in **Section 8.6.3.**

8.6.2. Outcomes

For Objective 1

Prevalence of the following variables will be estimated using standard concept IDs applicable to each domain of variable based on the type of data of the recording:

- Condition occurrence of obesity
- Measurements (anthropomorphic):

  - BMI
  - Weight
  - Height
  - Waist circumference
  - Hip circumference
  - Waist-to-height ratio
  - Waist-to-hip ratio
  - Abdominal skin fold thickness
  - Body composition
  - Body fat
- Measurements (lab):
  - Cholesterol
  - Glycaemia
  - Glycated haemoglobin A1c
  - Triglycerides
- Observation occurrence:
  - Diet
  - Physical activity
  - Smoking
- Procedure occurrence:
  - Bariatric surgery

The preliminary concept sets used for the identification of all outcomes are described in **Annex III**. These codes will be refined during the study execution, following the DARWIN EU® phenotyping standard processes, which involve the review of code lists by clinical experts, and the review of phenotypes after their execution in the participating data sources.

Objective 2, 3, and 4

There are no outcomes associated with these objectives since these are characterisation objectives.

8.6.3. Other covariates, including confounders, effect modifiers, and other variables

For objective 1:

Prevalence will be stratified by the following covariates:

- Sex
- Age groups: 2–11, 12–18, 19–39, 40–59, 60–79, and 80+.

For objective 2:

The frequency of recording of the following variables will be described:

- Measurements
  - BMI

  - Weight
  - Height
  - Cholesterol
  - Waist circumference
- Observations:
  - Diet
  - Physical activity
  - Smoking status

For objective 3:

The frequency and time between the recording of the following variables will be described:

- Measurements
  - BMI
  - Weight

<u>For objective 4:</u>

In this objective, individuals will be characterised by the following predefined covariates:

- To understand the overlap between the cohorts of individuals defined as obese by an obesity diagnosis and those defined as obese by BMI measurement ≥30 kg/m^2, we will characterise each cohort by the occurrence of an obesity diagnosis and by occurrence of a BMI measurement ≥30 kg/m^2 (assessment window at index date: [-inf,inf]).
- Demographics:
  - Sex
  - Age (in years) at index date
- Lifestyle factors (assessment window at index date: [-inf,inf]):
  - Diet
  - Physical activity
  - Smoking status
- Procedure (assessment window [0,inf]):
  - Bariatric surgery
- Obesity-related comorbidities [8] (assessment window [-inf,inf]):
  - Diabetes mellitus type 2
  - Hypertension
  - Ischemic heart disease
  - Chronic kidney disease
  - Hypothyroidism
  - Hypertriglyceridemia
  - Metabolic syndrome X

  - Cushing's syndrome
  - Knee arthrosis
  - Obstructive sleep apnoea
  - Mental health disorders, including depression and anxiety
  - Dyslipidaemia
  - Metabolic dysfunction-associated steatotic liver disease
  - Non-alcoholic fatty liver disease
  - Steatosis of liver
  - Cancer
- Prescription or dispensing of the following drugs during follow-up (assessment window [-inf,inf]):
  - Glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide-1 (GLP-1) receptor agonists
  - Orlistat
  - Metformin (possible off-label use for weight loss, in addition to weight loss seen when used for primary indication of diabetes mellitus type 2 [9, 10]
  - Naltrexone-bupropion

The preliminary concept sets used for the identification of all covariates are described in **Annex III**. These codes will be refined during the study execution, following the DARWIN EU® phenotyping standard processes, which involve the review of code lists by clinical experts and the review of phenotypes after their execution in the participating data sources.

8.7. Study size

No sample size has been calculated, as this is a descriptive disease epidemiology study which will not test a specific hypothesis. Thus, the sample size is driven by the availability of data for patients with obesity and obesity-related measurements, conditions, observations, and procedures.

8.8. Analysis

8.8.1. Federated network analyses

All analyses will be conducted separately for each data source, and will be carried out in a federated manner, allowing analyses to be run locally without sharing patient-level data.

Before sharing the study package, test runs of the analytics will be performed on a subset of the data sources and quality control checks will be performed. After all the tests are passed (see **Annex II**), the final package will be released in a version-controlled study repository for execution against all the participating data sources.

The data partners will locally execute the analytics against the OMOP CDM in R Studio and review and approve the default aggregated results. They will then be made available to the Principal Investigators and study team in secure online repository (Data Transfer Zone). All results will be locked and timestamped for reproducibility and transparency. The study results of all data sources are checked after which they are made available to the team, and the Study Dissemination Phase can start. All results are locked and timestamped for reproducibility and transparency.

More general-purpose diagnostic tools, *CohortDiagnostics* [11] and *DrugExposureDiagnostics* [12], have been developed. The *CohortDiagnostics* package provides additional insights into cohort characteristics,

record counts, and index event misclassification. The *DrugExposureDiagnostics* package evaluates ingredient-specific attributes and patterns in drug exposure records. Upon finalisation of the study protocol and creation of the disease and drug cohorts of interest by DARWIN EU® Coordination Centre, these packages will be executed in each data sources by each data partners.

8.8.2. Patient privacy protection

All analyses will be carried out in a federated manner, allowing analyses to be run locally for each data source without sharing patient-level data. Cell counts <5 will be suppressed when reporting results to comply with the data source's privacy protection regulations.

8.8.3. Statistical model specification and assumptions of the analytical approach considered

Objective 1

The prevalence estimation (i.e., proportion of BMI records) will be calculated based on OMOP CDM mapped data using the *IncidencePrevalence* R package, developed by DARWIN EU®.[13] Overall prevalence estimates will be calculated, as well as stratified by age group, sex, and calendar year.

Individuals enter the denominator population at the start of each calendar year from the start of the study period onwards, or once the eligibility criteria are fulfilled. Those study participants who enter the denominator population will then contribute follow-up time during each calendar year. Follow-up time subjects who die or become lost to follow up will be censored at the time of death or loss to follow-up. Subjects with data until the end of the study period without a record of death or loss to follow-up will be administratively censored at the end of the study period.

Period prevalence will be calculated by counting the number of individuals per calendar year with record of obesity (as a condition), the selected obesity-related measurements, lifestyle factors, or procedures (see **Section 8.6.2.** Outcome – Objective 1). These counts will be divided by the denominator (the number of persons contributing time at risk in the period) to calculate a proportion. Period prevalence will be reported as a percentage with 95% confidence intervals, as estimated by the Wilson Score method.

Descriptive statistics of the study population will be calculated at index date: total number of people, number of people per sex (N and percentage of the total study population), number of people per age group (N and percentage of the total study population).

Objective 2

This objective will make use of BMI measurement values, as well as condition occurrences of obesity to define obesity. All available BMI measurements will be assigned an obese/non-obese flag, depending on age and sex. A BMI measurement defined as obese is one with a value of ≥30 kg/m^2 for individuals ≥19 years old, and sex-specific BMI-for-age z-score that is ≥2 standard deviations, as per the WHO growth reference curves for individuals aged 2 to 18.[6, 7] The cohorts of obese individuals will be constructed using the *CohortConstructor* R package.[14] To estimate the denominator values in objective 2, the *IncidencePrevalence* R package will be used.[13]

To address objective 2a, the proportion of individuals who have a record of zero, one, two, three, four, or five-plus BMI and weight measurements for each calendar year, each three-year period (2010–12, 2013–15, 2016–18, 2019–21, 2022–25), and overall, within the study period will be estimated. For the paediatric population the proportion of individuals who have zero, one, or two-plus height measurements will be estimated in these time periods. In the adult population, the proportion of individuals who have zero or one-plus height measurements in these time periods will be estimated.

To estimate the number of intersections between the cohorts, the *PatientProfiles* R package will be used.[15] The output table of this analysis can be seen in **Tables 4–9**. In the second part of objective 2a, the median number of measurements (BMI, weight, height, cholesterol value, waist circumference) in the general population will be calculated, along with Q05, Q25, Q75, Q95, and min-max values. The output

table of this analysis can be seen in **Table 10**. All estimates in this sub objective will be calculated in the overall study population, adults (general population), adults (defined as obese), paediatric (general population), and paediatric (defined as obese), as specified in **Section 8.5**.

To address objective 2b, the median value of the BMI (as kg/m^2), height (in meters), and weight (in kilograms) will be estimated, along with Q05, Q25, Q75, Q95, and min-max values. The description of the values will be estimated using *MeasurementDiagnostics*.[16] For all these values, the first measurement will be selected per individual within the study period.

All estimates in objective 2 will be calculated in the overall study population, adults (general population), adults (defined as obese), paediatric (general population), and paediatric (defined as obese), as specified in **Section 8.5**

Objective 3

In objective 3a, we will first calculate the overall and annual recording rate of BMI and weight after study entry. Between the individual-level cohort start and end date, we will select all BMI measurements and all weight measurements. Summary descriptive statistics will be used to obtain an estimate of the number of recordings of weight per person-years. The output table of this analysis can be seen in **Table 11**.

For objective 3b, we will summarise the time elapsed between BMI and weight measurements up until the fifth measurement. Each individual from index date will have the time difference between each measurement calculated. The mean and median (Q05, Q25, Q50, Q75, Q95) time between first and second, second and third, third and fourth, and fourth and fifth measurement will be estimated. The output table of this analysis can be seen in **Table 12**.

In objective 3c, we will estimate mean cumulative function (MCF) of BMI and weight measurement occurrences. The MCF at time t will report the cumulative number of events (e.g., BMI recordings) per person on average up to time t. Plotting it results in a curve showing how BMI recording event accumulation evolves over time. To execute this, a table with patient ID, time (since index date in days), event indicator (event = 1 or censor = 0), and sex will be used. The MCF will be estimated using the function from the *reda* R package.[17] The parameters of the *mcf* function, the variance and estimator type, will be explored during study execution. The analysis will also be stratified by sex. The output (overall and by sex) will be plotted using *ggplot*, as seen in **Figures 12–14**.

All estimates in objective 3 will be calculated in the overall study population, adults (general population), adults (defined as obese), paediatric (general population), and paediatric (defined as obese), as specified in **Section 8.5**

Objective 4

Characterisation of individuals with incident obesity by demographics, laboratory measurements, lifestyle factors, procedures, comorbidities, and drugs around the obesity diagnosis (whether obtained via condition concept ID or by measurement record of BMI) will be conducted using the *CohortCharacteristics* R packages.[18] Two obesity cohorts will be characterised: diagnosis of incident obesity diagnosis or first BMI measurement defined as obese. All available BMI measurements will be assigned an obese/non-obese flag, depending on age and sex. A BMI measurement defined as obese is one with a value of ≥30 kg/m^2 for individuals ≥19 years old, and for individuals aged 2 to 18, sex-specific BMI-for-age z-score that is ≥2 standard deviations, as per the WHO growth reference curves.[6]

For each patient characteristic listed in **Section 8.6.3** (as defined using a list of concepts seen in **Annex III**) the number and proportion (%) of individuals with a record within each specified time window will be presented. For characterisation by conditions and drugs, all diagnoses or prescriptions/dispensing will be considered, irrespective of whether they are incident or prevalent. Sex and median age (plus IQR) will be measured at index date (i.e., date of diagnosis of incident obesity diagnosis or first BMI measurement ≥30 kg/m^2). For measurements, lifestyle factors, procedures, comorbidities, and drugs the number and

proportion of individuals with a record within each specified time window will be presented. We will estimate the standardised mean difference (SMD) between the variables in within each group. For binary categorical variables, the standardised difference is:

$$SMD = \frac{P1 - P2}{\sqrt{\frac{[P1(1-P1) + P2(1-P2)]}{2}}}$$

Where *Pi,* the proportion within each variable, varies by i = obese by condition or obese by BMI measurement.

The output table of this analysis can be seen in **Tables 13–18**.

We will calculate the number and percentage overlap between the cohorts of individuals defined as obese by condition and those defined as obese by a high BMI measurement (≥30 kg/m^2 in individuals ≥19 years old, or when BMI cut-off is defined by z-score ≥2 SD in individuals <19 years old) using the *summariseCohortOverlap* function in the *CohortCharacteristics* R package. The output table of this analysis can be seen in **Table 19**.

8.8.4. Output

Output will include a PDF report with an executive summary, and the following tables and figures. An interactive dashboard will be generated by incorporating all the results (tables and figures) included in the PDF report mentioned below.

Table 1. Distribution of study participants' characteristics (N, %, median, and IQR) in each data source.

Characteristic	Data source 1	Data source 2	Data source 3	Data source 4	Data source 5	Data source x
Overall, N						
Median age (IQR) at index date						
Mean age (SD) at index date						
Age groups in year, N (%), 2–11						
Age groups in year, N (%), 12–18						
Age groups in year, N (%), 19–39						
Age groups in year, N (%), 40–59						
Age groups in year, N (%), 60–79						
Age groups in year, N (%), 80+						
Median index year (IQR)						
Male, N (%)						
Female, N (%)						

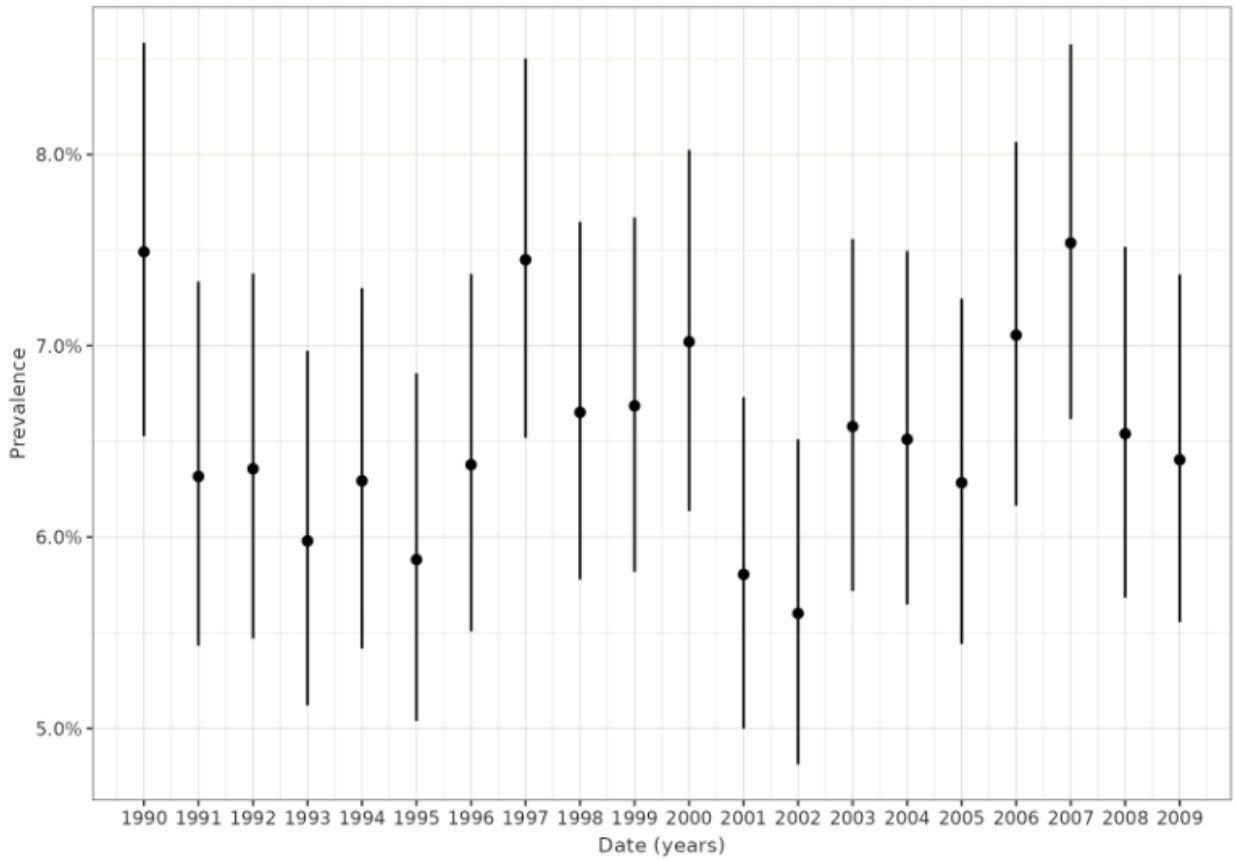


Figure 6. Period prevalence (%) of obesity defined by an obesity diagnosis record for each calendar year in each data source.

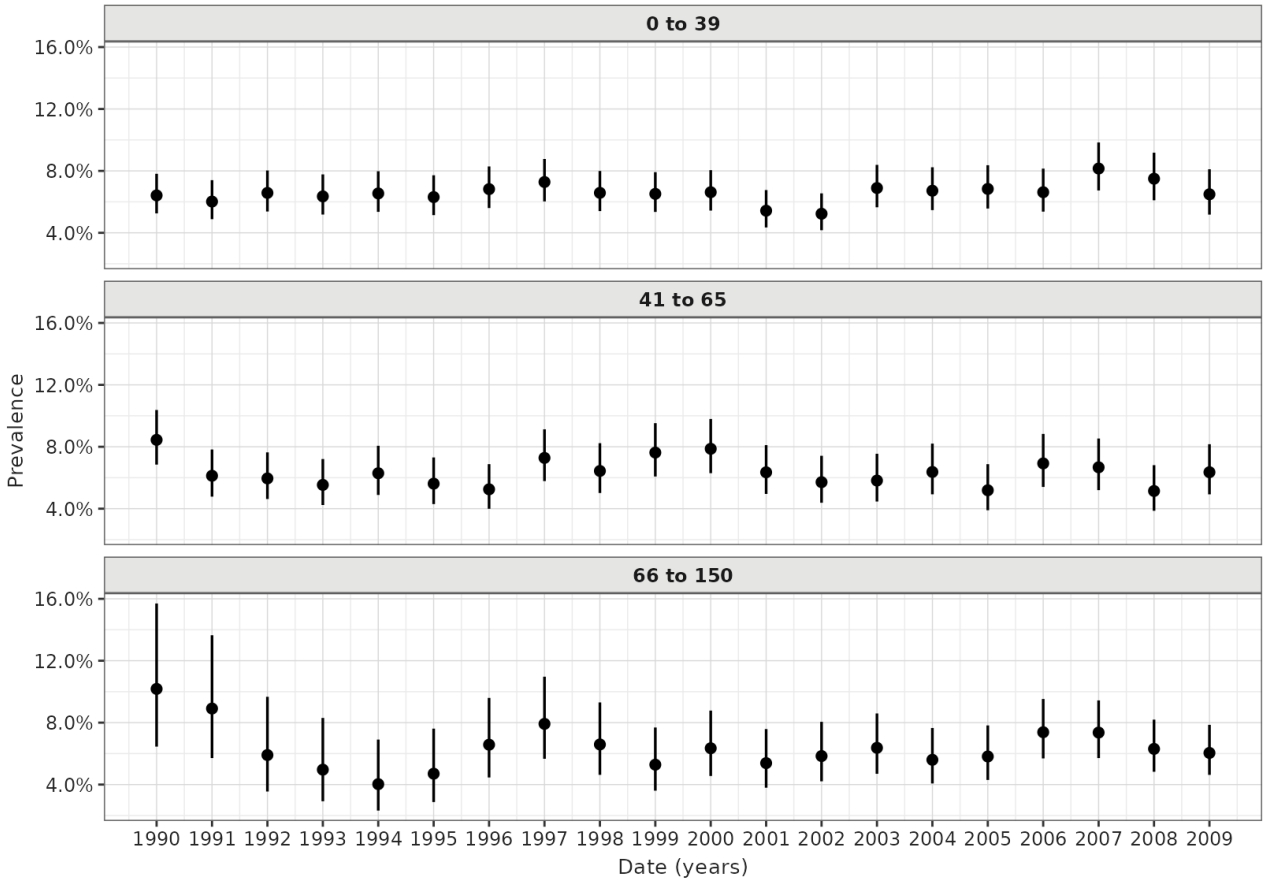


Figure 7. Period prevalence (%) of obesity defined by an obesity diagnosis for each calendar year, stratified by age group and sex in each data source.

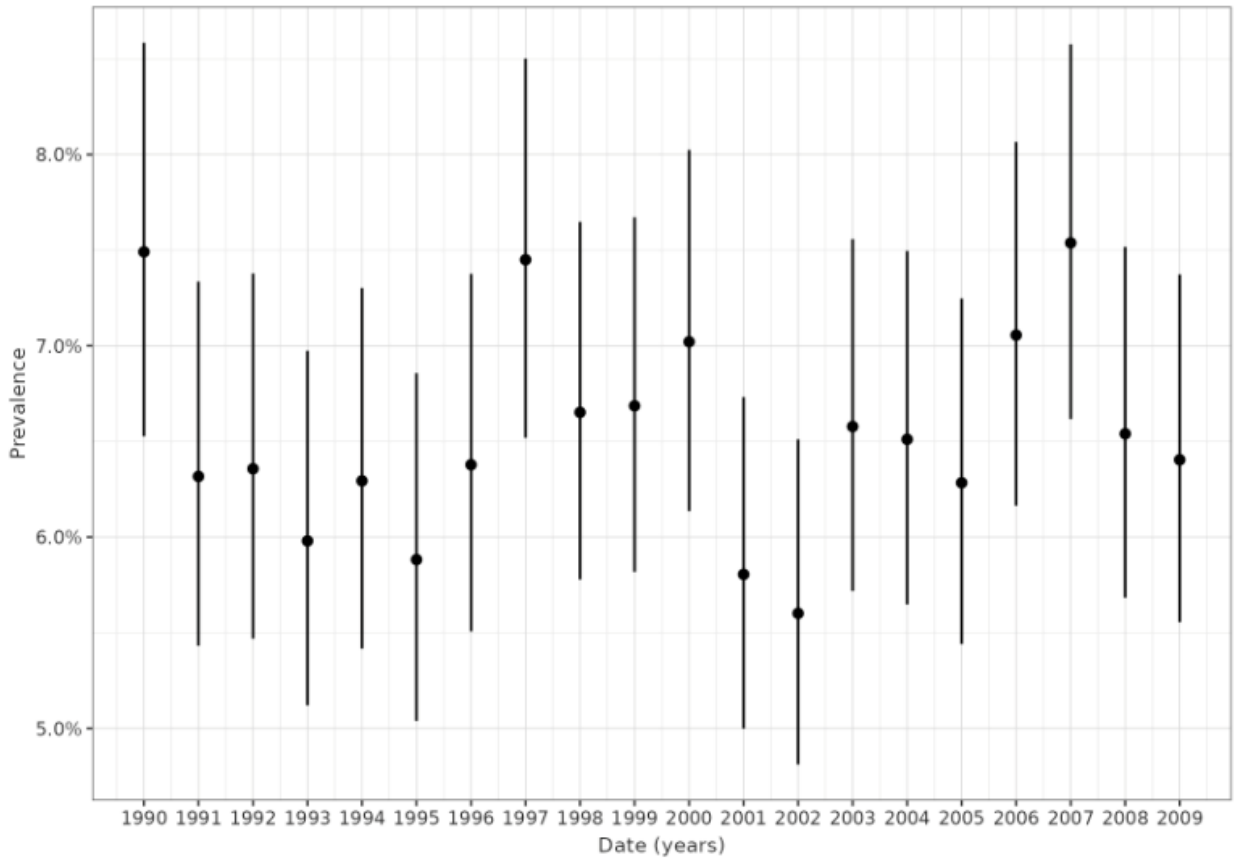


Figure 8. Period prevalence (%) of recording of BMI measurements for each calendar year in each data source.

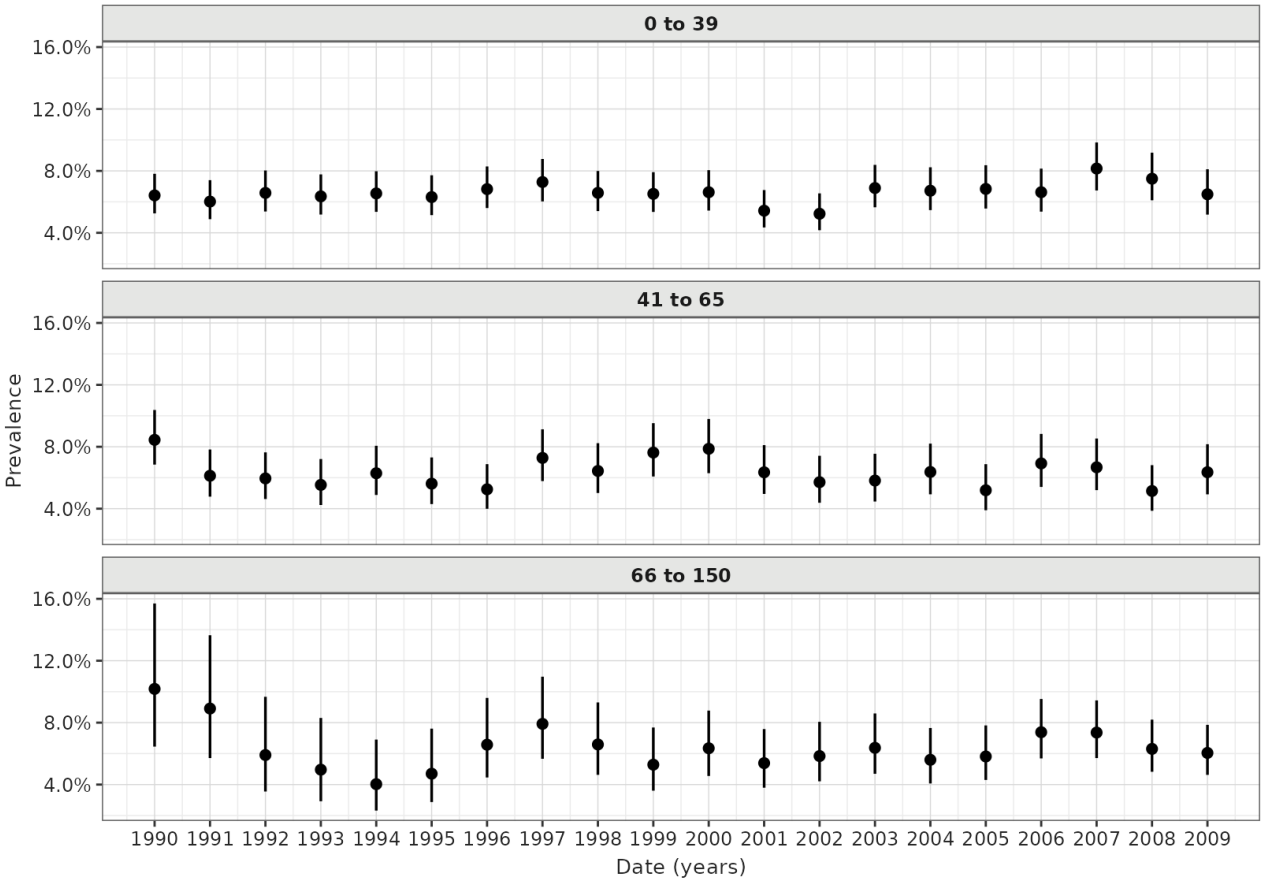


Figure 9. Period prevalence (%) of recording of BMI measurements for each calendar year, stratified by age group and sex in each data source.

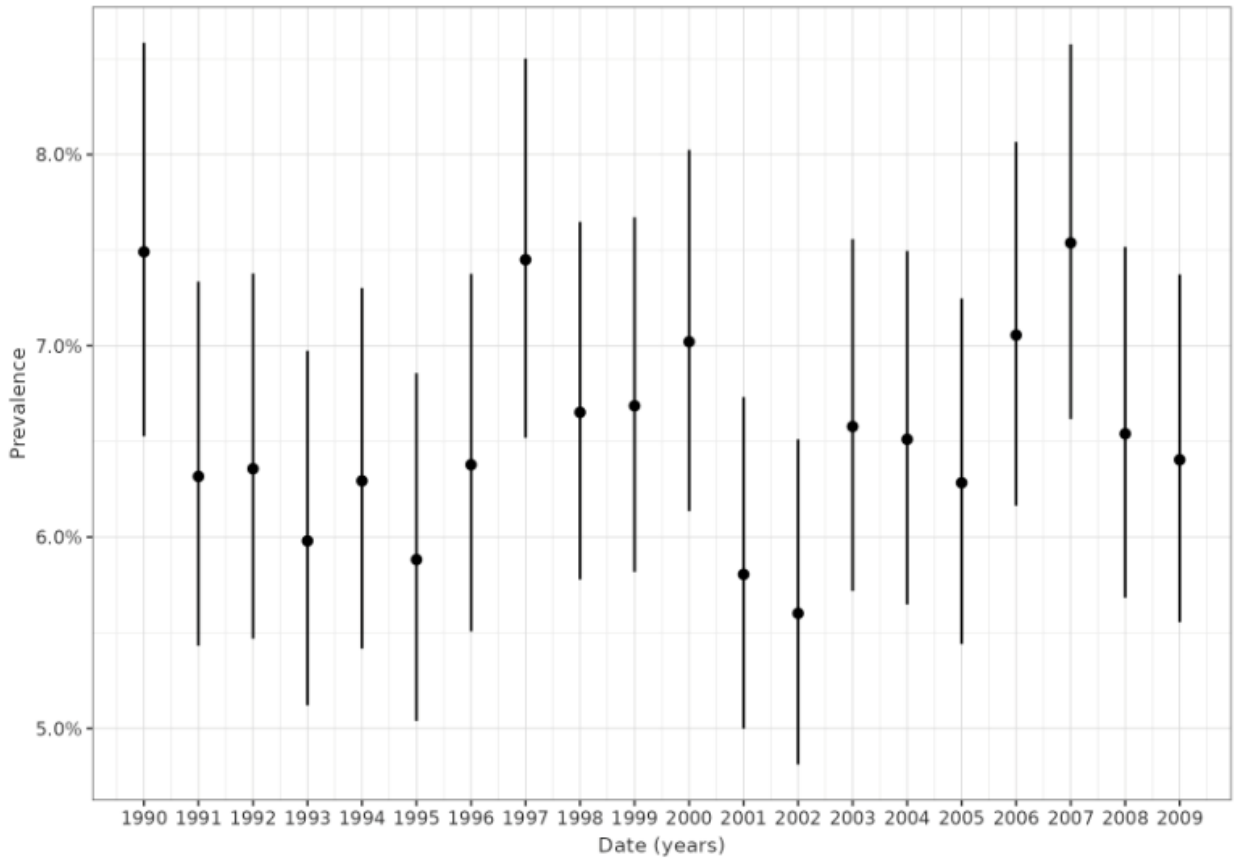


Figure 10. Period prevalence (%) of recording of weight measurements for each calendar year in each data source.

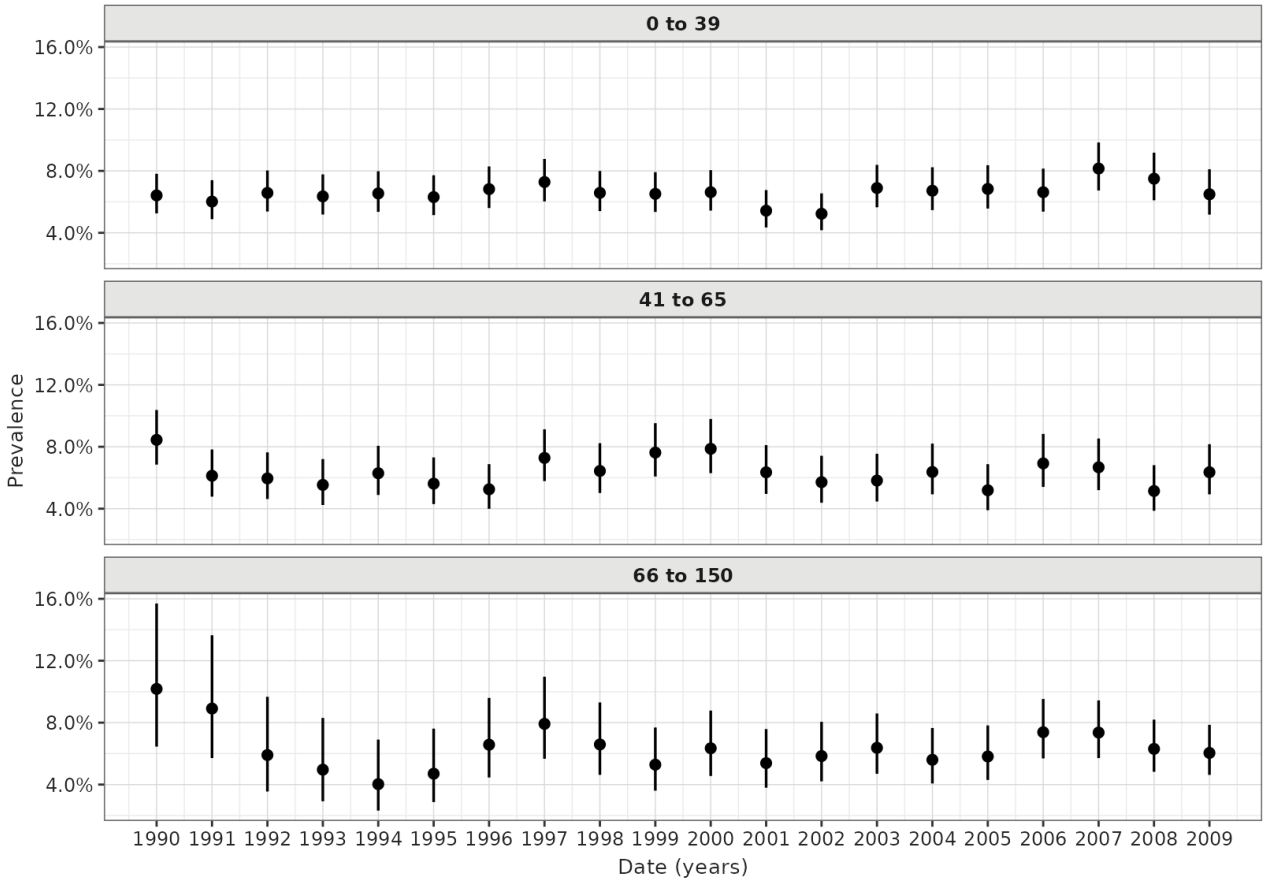


Figure 11. Period prevalence (%) of recording of weight measurements for each calendar year, stratified by age group and sex in each data source.

Table 2. Period prevalence of recording of all lifestyle measurements, lifestyle factors, and procedures per data source.

Characteristic	Data source 1	Data source 2	Data source 3	Data source 4	Data source 5	Data source x
Number subjects, N						
BMI, N (%)						
Weight, N (%)						
Height, N (%)						
Waist circumference, N (%)						
Hip circumference, N (%)						
Waist-to-height ratio, N (%)						
Waist-to-hip ratio, N (%)						
Abdominal skin fold thickness, N (%)						
Body fat, N (%)						
Cholesterol, N (%)						
Glycaemia, N (%)						
Glycated haemoglobin A1c, N (%)						
Triglycerides, N (%)						
Diet, N (%)						
Physical activity, N (%)						
Smoking, N (%)						
Bariatric surgery, N (%)						

Table 3. Median number (IQR and min-max) of the records of measurements and observations per individual per data source.

	Data source 1	Data source 2	Data source 3	Data source 4	Data source 5	Data source x
BMI, median [IQR]						
BMI, range						
Weight, median [IQR]						
Weight, range						
Height, median						
Height, range						
Waist circumference, median [IQR]						
Waist circumference, range						
Hip circumference, median [IQR]						

	Data source 1	Data source 2	Data source 3	Data source 4	Data source 5	Data source x
Hip circumference, range						
Waist-to-height ratio, median [IQR]						
Waist-to-height ratio, range						
Waist-to-hip ratio, median [IQR]						
Waist-to-hip ratio, range						
Abdominal skin fold thickness, median [IQR]						
Abdominal skin fold thickness, range						
Body fat, median [IQR]						
Body fat, range						
Cholesterol, median [IQR]						
Cholesterol, range						
Glycaemia, median [IQR]						
Glycaemia, range						
Glycated haemoglobin A1c, median [IQR]						
Glycated haemoglobin A1c, range						
Triglycerides, median [IQR]						
Triglycerides, range						
Diet, median [IQR]						
Diet, range						
Physical activity, median [IQR]						
Physical activity, range						
Smoking, median [IQR]						
Smoking, range						

Table 4. Frequency of BMI recording over the period 2010 to 2025.

Data source	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
Data source 1	Overall							
	Adults							
	Adults obese							
	Paediatric							
	Paediatric obese							
Data source 2	Overall							
	Adults							
	Adults obese							
	Paediatric							
	Paediatric obese							
Data source x	Overall							
	Adults							
	Adults obese							
	Paediatric							
	Paediatric obese							

Table 5. Frequency of weight recording over the period 2010 to 2025.

Data source	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
Data source 1	Overall							
	Adults							
	Adults obese							
	Paediatric							
	Paediatric obese							
Data source 2	Overall							
	Adults							
	Adults obese							
	Paediatric							
	Paediatric obese							
Data source x	Overall							
	Adults							
	Adults obese							
	Paediatric							
	Paediatric obese							

Table 6. Frequency of height recordings over the period 2010 to 2025.

Data source	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2+ recordings (N individuals, %)
Data source 1	Overall				
	Adults				
	Adults obese				
	Paediatric				
	Paediatric obese				
Data source 2	Overall				
	Adults				
	Adults obese				
	Paediatric				
	Paediatric obese				
Data source x	Overall				
	Adults				
	Adults obese				
	Paediatric				
	Paediatric obese				

Table 7. Frequency of BMI recording over the period 2010 to 2025, with 3-year periods (2010 to 2012, 2013 to 15, 2016 to 2018, 2019 to 2021, and 2022 to 2025).

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
Data source 1	2010 to 2012	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2013 to 2015	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2016 to 2018	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2019 to 2021	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2022 to 2015	Overall							
		Adults							
		Adults obese							

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
		Paediatric							
		Paediatric obese							
Data source 2	2010 to 2012	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2013 to 2015	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2016 to 2018	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2019 to 2021	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2022 to 2015	Overall							
		Adults							

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
		Adults obese							
		Paediatric							
		Paediatric obese							
Data source x	2010 to 2012	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2013 to 2015	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2016 to 2018	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2019 to 2021	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2022 to 2015	Overall							

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							

Table 8. Frequency of weight recording over the period 2010 to 2025, with 3-year periods (2010 to 2012, 2013 to 15, 2016 to 2018, 2019 to 2021, and 2022 to 2025).

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
Data source 1	2010 to 2012	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2013 to 2015	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2016 to 2018	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2019 to 2021	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2022 to 2015	Overall							
		Adults							
		Adults obese							

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
		Paediatric							
		Paediatric obese							
Data source 2	2010 to 2012	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2013 to 2015	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2016 to 2018	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2019 to 2021	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2022 to 2015	Overall							
		Adults							

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
		Adults obese							
		Paediatric							
		Paediatric obese							
Data source x	2010 to 2012	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2013 to 2015	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2016 to 2018	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2019 to 2021	Overall							
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							
	2022 to 2015	Overall							

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2 recording (N individuals, %)	3 recording (N individuals, %)	4 recording (N individuals, %)	5+ recording (N individuals, %)
		Adults							
		Adults obese							
		Paediatric							
		Paediatric obese							

Table 9. Frequency of height recordings over the period 2010 to 2025, with 3-year periods (2010 to 2012, 2013 to 15, 2016 to 2018, 2019 to 2021, and 2022 to 2025).

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2+ recording (N individuals, %)
Data source 1	2010 to 2012	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2013 to 2015	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2016 to 2018	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2019 to 2021	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2022 to 2015	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
Data source 2	2010 to 2012	Overall				
		Adults				
		Adults obese				

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2+ recording (N individuals, %)
		Paediatric				
		Paediatric obese				
	2013 to 2015	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2016 to 2018	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2019 to 2021	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2022 to 2015	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
Data source x	2010 to 2012	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2013 to 2015	Overall				
		Adults				

Data source	Time period	Population	N individuals in data source	No recording (N individuals, %)	1 recording (N individuals, %)	2+ recording (N individuals, %)
		Adults obese				
		Paediatric				
		Paediatric obese				
	2016 to 2018	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2019 to 2021	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				
	2022 to 2015	Overall				
		Adults				
		Adults obese				
		Paediatric				
		Paediatric obese				

Table 10. Q10, Q25, Q50 (median), Q75, Q90, and range of the values of BMI, weight, and height measurements per data source.

	Data source 1	Data source 2	Data source 3	Data source 4	Data source 5	Data source x
BMI, N						
BMI, Q10						
BMI, Q25						
BMI, Q50						
BMI, Q75						
BMI, Q90						
BMI, range						
Weight, N						
Weight (kg), Q10						
Weight (kg), Q25						
Weight (kg), Q50						
Weight (kg), Q75						
Weight (kg), Q90						
Weight (kg), range						
Height, N						
Height (m), Q10						
Height (m), Q25						
Height (m), Q50						
Height (m), Q75						
Height (m), Q90						
Height (m), range						

Table 11. Overall and annual recording rate of BMI measurements after study entry.

Data source	Population	N individuals	Total BMI measurements	Total follow-up months	Mean rate (95% CI)	Median rate (Q10, Q25, Q75, Q90)
Data source 1	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source 2	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source x	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					

Table 12. Overall and annual recording rate of weight measurements after study entry.

Data source	Population	N individuals	Total weight measurements	Total follow-up months	Mean rate (95% CI)	Median rate (Q10, Q25, Q75, Q90)
Data source 1	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source 2	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source x	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					

Table 13. Summarised time elapsed between BMI measurements.

Data source	Population	N individuals in data source	Mean/ Median (Q10, Q25, Q75, Q90) time between first and second record	Mean/ Median (Q10, Q25, Q75, Q90) time between second and third record	Mean/ Median (Q10, Q25, Q75, Q90) time between third and fourth record	Mean/ Median (Q10, Q25, Q75, Q90) time between fourth and fifth record
Data source 1	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source 2	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source x	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					

Table 14. Summarised time elapsed between weight measurements.

Data source	Population	N individuals in data source	Mean/ Median (Q10, Q25, Q75, Q90) time between first and second record	Mean/ Median (Q10, Q25, Q75, Q90) time between second and third record	Mean/ Median (Q10, Q25, Q75, Q90) time between third and fourth record	Mean/ Median (Q10, Q25, Q75, Q90) time between fourth and fifth record
Data source 1	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source 2	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					
Data source x	Overall					
	Adults					
	Adults obese					
	Paediatric					
	Paediatric obese					

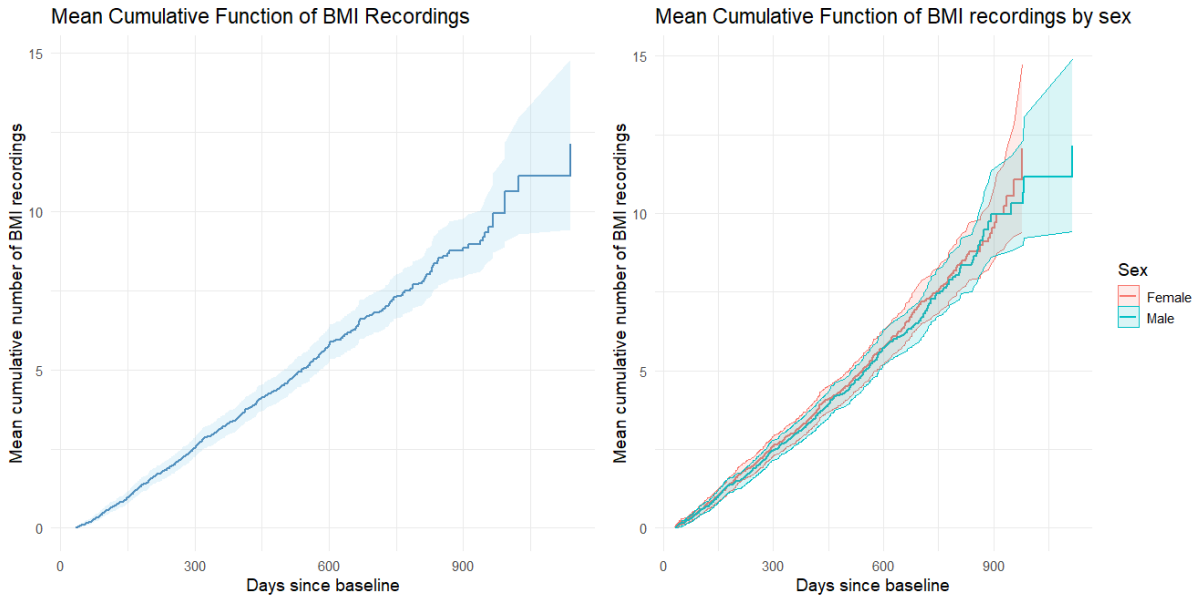


Figure 12. Mean cumulative function in each data source. Overall (left) and stratified by sex (right).

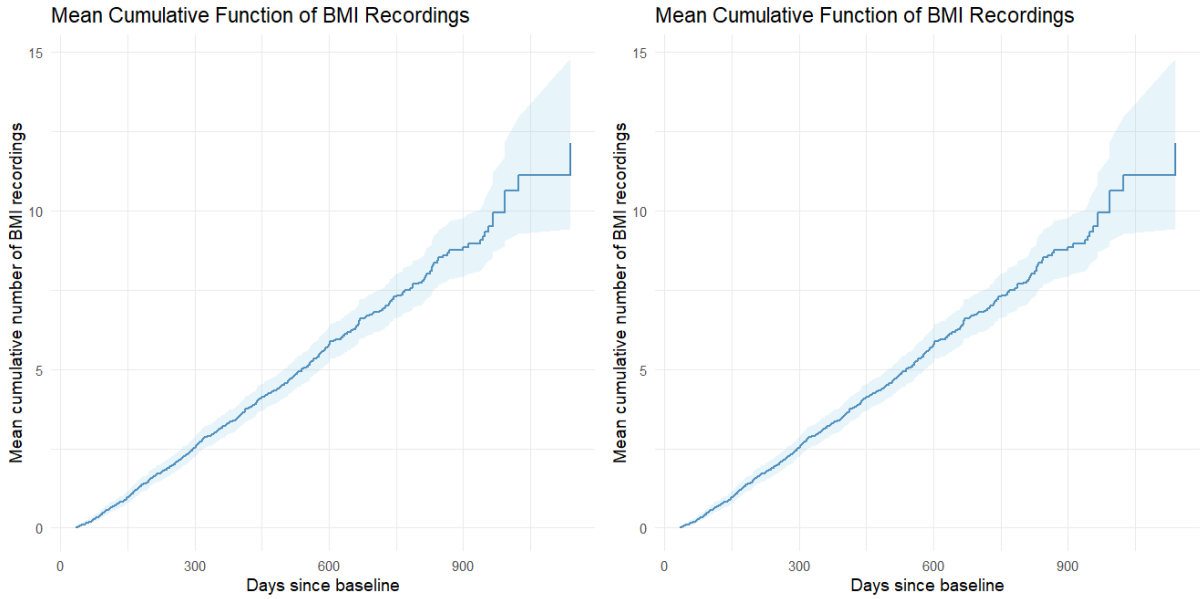


Figure 13. Mean cumulative function in a paediatric population in each data source. General (left) and in an obese population (right).

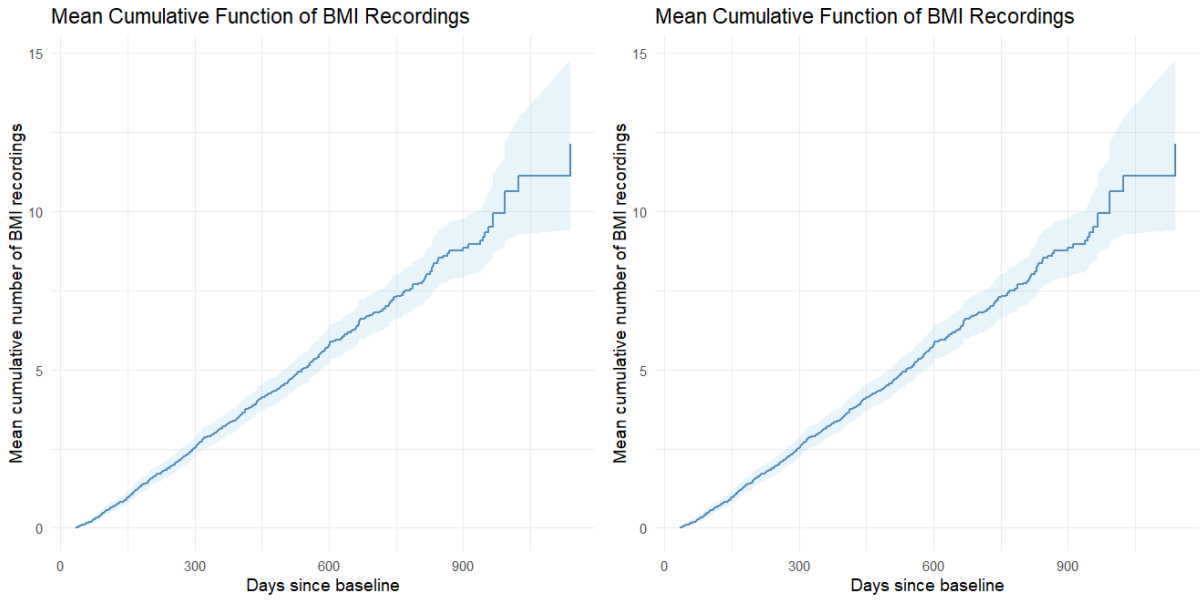


Figure 14. Mean cumulative function in a adult population in each data source. General (left) and in an obese population (right).

Table 15. Demographic characteristics of the study population on index date within individuals with an incident obesity condition diagnosis and within individuals with a BMI measurement ≥30 kg/m^2.

	Data source 1		Data source 2		Data source 3		Data source 4		Data source 5		Data source x	
Characteristic	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2
Overall, N												
Median age (IQR) at index date												
Mean age (SD) at index date												
Age groups in year, N (%), 19–39												
Age groups in year, N (%), 40–59												
Age groups in year, N (%), 60–79												
Age groups in year, N (%), 80+												
Male, N (%)												
Female, N (%)												

Table 16. The number and proportion of individuals with a record of each of the prespecified characteristics within individuals with an incident obesity condition diagnosis and within individuals with a BMI measurement ≥30 kg/m^2.

	Data source 1		Data source 2		Data source 3		Data source 4		Data source 5		Data source x	
Characteristic	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2	Obesity condition	≥30 kg/m^2
Number subjects, N												
Diet, N (%)												
Physical activity, N (%)												
Smoking status, N (%)												
Bariatric surgery, N (%)												
Diabetes mellitus type 2, N (%)												
Hypertension, N (%)												
Ischemic heart disease, N (%)												
Chronic kidney disease, N (%)												
Hypothyroidism, N (%)												
Hypertriglyceridemia, N (%)												
Metabolic syndrome X, N (%)												
Cushing's syndrome, N (%)												
Knee arthrosis, N (%)												
Obstructive sleep apnoea, N (%)												
Mental health disorders, including												

	Data source 1		Data source 2		Data source 3		Data source 4		Data source 5		Data source x	
depression and anxiety, N (%)												
Dyslipidaemia, N (%)												
Metabolic dysfunction-associated steatotic liver disease, N (%)												
Non-alcoholic fatty liver disease, N (%)												
Steatosis of liver, N (%)												
Cancer, N (%)												
Glucagon-like peptide-1 (GLP-1) receptor agonists, N (%)												
Orlistat, N (%)												
Metformin, N (%)												
Naltrexone-bupropion, N (%)												

Table 17. Demographic characteristics of the study population on index date within individuals aged 2–18 years old with an incident obesity condition diagnosis and within individuals 2–18 years old with a sex-specific BMI-for-age z-score ≥2 standard deviations.

	Data source 1		**Data source 2**		**Data source 3**		**Data source 4**		**Data source 5**		**Data source x**	
Characteristic	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD
Overall, N												
Median age (IQR) at index date												
Mean age (SD) at index date												
Male, N (%)												
Female, N (%)												

Table 18. The number and proportion of individuals within individuals aged 2–18 years old with an incident obesity condition diagnosis and within individuals 2–18 years old with a sex-specific BMI-for-age z-score ≥2 standard deviations.

	Data source 1		**Data source 2**		**Data source 3**		**Data source 4**		**Data source 5**		**Data source x**	
Characteristic	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD	Obesity condition	z-score ≥2 SD
Number subjects, N												
Diet, N (%)												
Physical activity, N (%)												
Smoking status, N (%)												
Bariatric surgery, N (%)												
Diabetes mellitus type 2, N (%)												
Hypertension, N (%)												
Ischemic heart disease, N (%)												
Chronic kidney disease, N (%)												

	Data source 1		Data source 2		Data source 3		Data source 4		Data source 5		Data source x	
Hypothyroidism, N (%)												
Hypertriglyceridemia, N (%)												
Metabolic syndrome X, N (%)												
Cushing's syndrome, N (%)												
Knee arthrosis, N (%)												
Obstructive sleep apnoea, N (%)												
Mental health disorders, including depression and anxiety, N (%)												
Dyslipidaemia, N (%)												
Metabolic dysfunction-associated steatotic liver												

	Data source 1		Data source 2		Data source 3		Data source 4		Data source 5		Data source x	
disease, N (%)												
Non-alcoholic fatty liver disease, N (%)												
Steatosis of liver, N (%)												
Cancer, N (%)												
Glucagon-like peptide-1 (GLP-1) receptor agonists, N (%)												
Orlistat, N (%)												
Metformin, N (%)												
Naltrexone - bupropion, N (%)												

Table 19. The number and percentage overlap between the cohorts of individuals defined as obese by condition and those defined as obese by a BMI measurement (≥30 kg/m2 in individuals ≥19 years old, z-score ≥2 SD in individuals <19 years old).

Cohort name reference	Cohort name comparator	Estimate name	Only in reference cohort	In both cohorts	Only in comparator cohort
Obesity by condition (≥19 years old)	Obesity by measurement (≥19 years old)	N (%)			
Obesity by condition (<19 years old)	Obesity by measurement (<19 years old)	N (%)			

8.9. Evidence synthesis

Results from analyses described in **Section 8.8.4.** will be presented separately for each data source. No meta-analysis of results will be conducted. The results of this protocol will be reported together in a final report. Once all four studies yield results, tables and figures will be made according to data source type: primary care data source, secondary care data source, or data sources encompassing information from both.

9. STRENGTHS AND LIMITATIONS

This study accounts for the majority of the DARWIN EU® data partner network. This allows for broad analysis of the use of obesity as a covariate in a large number of RWD sources across Europe. Furthermore, this study classifies obesity not only by a condition record, but also by anthropometric measurement over the cutoff value, which increases sensitivity. This study will not comprehensively address all five of the data quality (DQ) dimensions, as outlined in the DQ Framework: reliability, relevance, timeliness, coherence, and extensiveness.[19] It would not be feasible to validate each component of these dimensions, some of which are beyond the scope of this study. Instead, this study considers extensiveness and relevance as the DQ dimensions that are to be most explored.

Classification of obesity through anthropomorphic measurements, such as waist circumference, waist-to-hip ratio, waist-to-height ratio, abdominal skin fold thickness, body composition, and body fat percentage, will not be possible. This study will estimate the prevalence of these measures to understand the scope of their capture, since clinically these can be seen as reliable diagnostic measures, but in RWD, they are likely missing or not reported.[8]

Pregnant individuals are not accounted for in the study design and analysis, as these individuals will likely have BMI measurements where the same assumptions as for non-pregnant individuals will not hold. Some data partners involved use EHR data where BMI is more likely to be recorded for pregnant individuals. In the stratification by age and sex, the results in younger adult and female populations are likely to be impacted by this. This study does not account for pregnant individuals due to the complexity involved in identifying and estimating the timing of a pregnancy episode in RWD. The results estimated from this study will only reflect the populations from the included data sources; in some cases, these will be selected populations and not necessarily representative of the entire country or region it is within.

10. REFERENCES

1. Huisman MV, Rothman KJ, Paquette M, Teutsch C, Diener HC, Dubner SJ, et al. The Changing Landscape for Stroke Prevention in AF: Findings From the GLORIA-AF Registry Phase 2. Journal of the American College of Cardiology. 2017;69(7):777-85.
2. Stevens GA, Singh GM, Lu Y, Danaei G, Lin JK, Finucane MM, et al. National, regional, and global trends in adult overweight and obesity prevalences. Popul Health Metr. 2012;10(1):22.
3. Blüher M. Obesity: global epidemiology and pathogenesis. Nat Rev Endocrinol. 2019;15(5):288-98.
4. Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, Lee A, et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N Engl J Med. 2017;377(1):13-27.
5. WHO. Global status report on noncommunicable diseases 2014. WHO; 2014.
6. WHO. BMI-for-age (5-19 years) 2025 [Available from: https://www.who.int/tools/growth-reference-data-for-5to19-years/indicators/bmi-for-age.
7. WHO. BMI-for-age (birth-5 years) 2025 [Available from: https://www.who.int/toolkits/child-growth-standards/standards/body-mass-index-for-age-bmi-for-age.
8. Rubino F, Cummings DE, Eckel RH, Cohen RV, Wilding JPH, Brown WA, et al. Definition and diagnostic criteria of clinical obesity. Lancet Diabetes Endocrinol. 2025;13(3):221-62.
9. Lund SS, Tarnow L, Stehouwer CD, Schalkwijk CG, Frandsen M, Smidt UM, et al. Targeting hyperglycaemia with either metformin or repaglinide in non-obese patients with type 2 diabetes: results from a randomized crossover trial. Diabetes Obes Metab. 2007;9(3):394-407.
10. Long-term safety, tolerability, and weight loss associated with metformin in the Diabetes Prevention Program Outcomes Study. Diabetes Care. 2012;35(4):731-7.
11. Gilbert J, Rao G, Schuemie M, Ryan P, Weaver J. CohortDiagnostics: Diagnostics for OHDSI Cohorts. R package version 3.3.0, https://github.com/OHDSI/CohortDiagnostics, https://ohdsi.github.io/CohortDiagnostics. 2024.
12. Inberg G, Burn E, Burkard T. DrugExposureDiagnostics: Diagnostics for OMOP Common Data Model Drug Records. R package version 1.0.9, https://github.com/darwin-eu/DrugExposureDiagnostics, https://darwin-eu.github.io/DrugExposureDiagnostics/. 2024.
13. Raventós B, Català M, Du M, Guo Y, Black A, Inberg G, et al. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. Pharmacoepidemiology and drug safety. 2024;33(1).
14. Burn E CM, Mercade-Besora N, Alcalde-Herraiz M, Du M, Guo Y, Chen X, Lopez-Guell K, Rowlands E CohortConstructor: Build and Manipulate Study Cohorts Using a Common Data Model. R package version 0.4.0 2025 [Available from: https://ohdsi.github.io/CohortConstructor/.
15. Català M GY, Du M, Lopez-Guell K, Burn E, Mercade-Besora N. PatientProfiles: Identify Characteristics of Patients in the OMOP Common Data Model. R package version 1.4.3 ed2025.
16. Burn E M-BN, Alcalde-Herraiz M. MeasurementDiagnostics: Diagnostics for Lists of Codes Based on Measurements. R package version 0.1.0999 ed2025.
17. Wang W FH, Yan J. reda: Recurrent Event Data Analysis. R package. 0.5.6 ed2025.
18. Catala M GY, Lopez-Guell K, Burn E, Mercade-Besora N, Alcalde M CohortCharacteristics: Summarise and Visualise Characteristics of Patients in the OMOP CDM. R package version 0.4.0 2024 [Available from: https://darwin-eu.github.io/CohortCharacteristics/.
19. Data Analytics and Methods Task Force. Data Quality Framework for EU medicines regulation. EMA: EMA; 2023.

11. ANNEXES

ANNEX I: Description of data sources

Table S1. Overview of data sources included in all studies: P4-C3-004, P4-C2-008, P4-C2-009, and P4-C2-010.

Country	Data partner	Database type	Study	Counts of obesity (condition)	Counts of BMI (measurements)
Belgium	IQVIA LPD Belgium	Outpatient General Practitioner Care	P4-C3-004	Yes	Yes
Croatia	NAJS	Registry	P4-C3-004	Yes	No
Denmark	DK-DHR	Registry	P4-C2-009	Yes	No
Estonia	EBB	Registry	P4-C2-008	Yes	No
Finland	FinOMOP-ACI Varha	Inpatient (Hospital) Care	P4-C2-010	Yes	No
Finland	FinOMOP-THL	Registry	P4-C2-010	Yes	No
Finland	FinOMOP-TaUH Pirha	Inpatient (Hospital) Care	P4-C2-010	Yes	No
France	APHM	Inpatient And Outpatient Hospital Care	P4-C2-010	Yes	No
France	CDW Bordeaux	Inpatient (Hospital) Care	P4-C3-004	Yes	No
Germany	InGef RDB	Claims	P4-C3-004	Yes	No
Germany	IQVIA DA Germany	Outpatient General Practitioner Care	P4-C3-004	Yes	Yes
Greece	PGH	Inpatient (Hospital) Care	P4-C2-009	Yes	No
Hungary	SUCD	Inpatient (Hospital) Care	P4-C2-010	Yes	No
Italy	POLIMI	Inpatient (Hospital) Care	P4-C2-008	Yes	No
Netherlands	IPCI	Outpatient General Practitioner Care	P4-C3-004	Yes	Yes
Norway	NLHR	Registry	P4-C2-010	Yes	No
Portugal	EMDB-ULSEDV	Inpatient (Hospital) Care	P4-C2-009	Yes	No
Portugal	EMDB-ULSGE	Inpatient (Hospital) Care	P4-C2-009	Yes	No
Portugal	EMDB-ULSRA	Inpatient (Hospital) Care	P4-C2-009	Yes	No
Portugal	ULSM-RT	Inpatient (Hospital) Care	P4-C2-009	Yes	No
Spain	BIFAP	Outpatient general	P4-C2-009	Yes	Yes

Country	Data partner	Database type	Study	Counts of obesity (condition)	Counts of BMI (measurements)
		practitioner And Hospital Care			
Spain	H12O	Inpatient And Outpatient Hospital Care	P4-C2-008	Yes	Yes
Spain	IMASIS	Inpatient (Hospital) Care	P4-C2-009	Yes	Yes
Spain	PRISIB	Outpatient General Practitioner Care, Inpatient (Hospital) Care	P4-C2-008	Yes	Yes
Spain	SIDIAP	Outpatient General Practitioner Care	P4-C2-008	Yes	Yes
Spain	VID	Outpatient General Practitioner Care	P4-C2-009	Yes	Yes
Sweden	HI-SPEED	Registry, Outpatient General Practitioner Care, Inpatient (Hospital) Care	P4-C2-010	Yes	No
United Kingdom	CPRD GOLD	Outpatient General Practitioner Care	P4-C2-008	Yes	Yes
United Kingdom	UKBB	Registry + Outpatient General Practitioner Care, Inpatient (Hospital) Care	P4-C2-008	Yes	Yes

P4-C3-004:

Belgium: IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium)

Belgium Longitudinal patient data (LPD) is collected from GP prescribing systems and contains patient records on all signs and symptoms, diagnoses, and prescribed medications. The information recorded allows patients and doctors to be monitored longitudinally. Data are recorded directly in the LPD from doctors' surgeries in real-time during patient consultations via a practice management software system. It is used in studies to provide various market insights such as treatment trends, patient pathway analysis and treatment compliance. The panel of contributing physicians (a stable 300 GPs) is maintained as a representative sample of the primary care physician population in Belgium according to three criteria known to influence prescribing: age, sex, and geographical distribution. Currently, the database is covering 1.1 M cumulative patients and covers from 2012 through to the present. The panel consists of a stable 300 GPs that are geographically well spread. The total number of active GPs in Belgium is 15.602. The regional geographical spread of physicians in the LPD data is also representative of the distribution across the country: 57% GPs in the North (compared to 54% nationally), 31% in the South (33% nationally) and 12% in Brussels (13%).The provider of the data has more than 2,250 GPs under contract so in case of a drop out a

replacement is easily found. Drugs obtained over the counter by the patient outside the prescription system are not reported. No explicit registration or approval is necessary for drug utilization studies.

Croatia: Croatian National Public Health Information System (NAJS)

The National Public Health Information System (Nacionalni javnozdravstveni informacijski sustav - NAJS) is an organised system of information services by Croatian Institute of Public Health (CIPH). NAJS enables data collecting, processing, recording, managing, and storing of health-related data from health care providers as well as production and management of health information. NAJS contains medical and public health data collected and stored in health registries and other health data collections including cancer registry, mortality, work injuries, occupational diseases, communicable and non-communicable diseases, health events, disabilities, psychosis and suicide, diabetes, drug abuse and others.

France: Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)

The clinical data warehouse of the Bordeaux University Hospital comprises electronic health records on more than 2 million patients with data collection starting in 2005. The hospital complex is made up of three main sites and comprises a total of 3,041 beds (2021 figures). The database currently holds information about the person (demographics), visits (inpatient and outpatient), conditions and procedures (billing codes), drugs (outpatient prescriptions and inpatient orders and administrations), measurements (laboratory tests and vital signs) and dates of death (in or out-hospital death). The data from the hospital production information system are loaded daily into a CDW in i2b2 format. A specific ETL from i2b2 to OMOP has been set up to standardize the data in OMOP-CDM format. Currently, this ETL from i2b2 to OMOP is launched manually when needed (this ETL can be scheduled regularly if necessary). The data is integrated in the OMOP Common Data Model version 5.3.1 and is stored in Oracle version 19c.

Germany: InGef Research Database (InGef RDB)

The InGef database comprises anonymized longitudinal claims data of about 10 million individuals across more than 50 statutory health insurance providers (SHIs) throughout Germany. Data are longitudinally linked over a period of currently ten years. Patients can be traced across health care sectors. All patient-level and provider-level data in the InGef research database are anonymised to comply with German data protection regulations and German federal law. German SHI claims data available in the InGef database includes information on demographics (year and quarter of birth, gender, death date if applicable, region of residence on administrative district level); hospitalizations; outpatient services (diagnoses, treatments; specialities of physicians); dispensing of drugs; dispensing of remedies and aids; and sick leave and sickness allowance times. In addition, costs or cost estimates from SHI perspective are available for all important cost elements. All diagnoses in Germany are coded using the International Classification of Diseases, version 10 in the German Modification (ICD-10-GM). The persistence (membership over time) is rather high in the InGef database: During a time period of 5 years (2009 to 2013), 70.6% of insurance members survived and remained insured with the same SHI without any gap in their observational time. Persons leaving one of the participating SHIs and entering another participating SHI, can be linked during yearly database consistency updates and are thus not lost over time. The InGef database is dynamic in nature, i.e. claims data are updated in an ongoing process and new SHIs may join or leave the database. By law, only the last 10 years of data are allowed to be used. At every new release this window shifts, dropping older data and adding new data.

Germany: IQVIA Disease Analyzer Germany (IQVIA DA Germany)

Germany DA is collected from extracts of patient management software used by GPs and specialists practicing in ambulatory care settings. Data coverage includes 39.6 M cumulative person. Patient visiting more than one provider are not cross identified for data protection reasons and therefore recorded as separate in the system. Dates of service include from 1992 through present. Observation time is defined by the first and last consultation dates. Germany has no mandatory GP system and patient have free choice of

specialist. Drugs are recorded as prescriptions of marketed products. No registration or approval is required for drug utilization studies.

Netherlands: Integrated Primary Care Information (IPCI)

The Integrated Primary Care Information (IPCI) database is a longitudinal observational database containing routinely collected data from computer-based patient records of a selected group of GPs throughout the Netherlands (N=723). IPCI was started in 1992 by the department of Medical Informatics of the Erasmus University Medical Center in Rotterdam with the objective to enable better post marketing surveillance of drugs. The current database includes patient records from 2006 on, when the size of the database started to increase significantly. In 2016, IPCI was certified as Regional Data Center. Since 2019 the data is also standardized to the Observational Medical Outcomes Partnership common data model (OMOP CDM), enabling collaborative research in a large network of databases within the Observational Health Data Sciences and Informatics (OHDSI) community. The primary goal of IPCI is to enable medical research. In addition, reports are generated to inform GPs and their organizations about the provided care. Contributing GPs are encouraged to use this information for their internal quality evaluation. The IPCI database is registered on the European Medicines Agency (EMA) ENCePP resources database (http://www.encepp.eu).

P4-C2-008:

Estonia: Estonian Biobank (EBB)

The Estonian Biobank (EBB) is a population-based biobank of the Estonian Genome Center at the University of Tartu (EGCUT). Its cohort size is currently close to 200,000 participants (“gene donors” ≥ 18 years of age), which closely reflects the age, sex, and geographical distribution of the Estonian population. Estonians represent 83%, Russians 14%, and other nationalities 3% of all participants. Genomic GWAS analyses have been performed on all gene donors. The database also covers health insurance claims, digital prescriptions, discharge reports, information about incident cancer cases, and causes of death from national sources for each donor.

Italy: Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI)

Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, known simply as Policlinico of Milan, is a general hospital that can count on important excellence in different areas of care, with a strong interdisciplinary focus. Given its nature as IRCCS—Institute for Research, Hospitalization, and Health Care—in addition to care, it carries out biomedical and health research activities of a clinical and translational nature, involving the rapid transfer of therapies from the laboratories to the bedside of the sick person. The research activity is conducted in the different fields of medicine, from neurology to cardiology, from transplantation to haematology, to excellence of care in gynaecology, neonatology, geriatrics, and rare diseases. Our DWH was born a few years ago, with the aim of helping researchers in identifying patient cohorts and in obtaining large amounts of data for their studies more easily. A few years later, thanks to the EHDEN Project, we were also able to introduce the OMOP CDM. Currently the DWH contains data from Hospitalization, Outpatients visits, Laboratory test, Therapies, Radiology, Anatomic Pathology, and a REDCap instance for non-profit studies.

Spain: Hospital Universitario 12 de Octubre (H12O)

The data source is mainly the Electronic Health Record of the Hospital Universitario 12 de Octubre. It contains information from the different health domains (laboratory, prescriptions, treatments, administrative, diagnoses, etc.). In addition, information is also obtained from other data sources, such as the pathological anatomy system, which provides information about sample analysis, and the cost system, containing information on the cost associated with a contact with the hospital. Work for the inclusion of further data is ongoing, among others, radiological information, or PROMs.

Spain: Platafoma de Recerca en Informació Sanitària de les Illes Balears (PRISIB)

The PRISIB database is the result of the harmonization and merging of different data sources across the Public Health System of the Balearic Islands (IBSalut), including data from 7 hospitals, 61 primary care areas, and around 1,300,000 active patients, which encompasses the whole of the population of the archipelago. Hospital data includes diagnostic and procedure codes from the discharge reports. Primary care data includes visits, measurements, diagnoses, procedures, prescriptions, and laboratory results. PRISIB was established in 2018 and contains data from 2011 onwards, with full coverage of the population of the Balearic Islands. The database also contains many visitors, i.e., tourists visiting the islands, mainly in the summer period.

Spain: The Information System for Research on Primary Care (SIDIAP)

The Information System for Research in Primary Care (SIDIAP) is a clinical database of anonymized patient records in Catalonia, Spain. The Spanish public healthcare system covers more than 98% of the population, and more than two thirds of the Catalan population see their GP at least once a year. The computerisation of the primary care patient records of the Catalan Health Institute (CHI) was completed in 2005. SIDIAP was designed to provide a valid and reliable database of information from clinical records of patients registered in primary care centres for use in biomedical research. SIDIAP contains data of anonymized patients' healthcare records for nearly six million people (approximately 80% of the Catalan population), registered in 287 primary care practices throughout Catalonia since 2005. It includes data collected by health professionals during routine visits in primary care, including anthropometric measurements, clinical diagnoses (International Classification of Diseases 10th revision ICD-10), laboratory tests, prescribed and dispensed drugs, hospital referrals, and demographic and lifestyle information. It was previously shown that the SIDIAP population is highly representative of the entire Catalan region in terms of geographic, age, and sex distributions. The high quality of these data has been previously documented, and SIDIAP has been successfully applied to epidemiological studies of key exposures and outcomes. Quality checks to identify duplicate patient IDs are performed centrally at each SIDIAP database update. Checks for logical values and data harmonisation are performed. For biochemistry data, consistency for measurements taken in different laboratories is assessed, and unit conversion is undertaken, when needed.

United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)

The Clinical Practice Research Datalink (CPRD) GOLD is a database of anonymised electronic health records (EHR) from General Practitioner (GP) clinics in the UK that use the Vision® software system for their management. 98% of the UK population is registered with a GP who is primarily responsible for non-emergency care and referrals to secondary care as needed. Participating GPs provide CPRD EHR for all registered patients who did not specifically request to opt out of data sharing. GOLD currently contains data from 985 up-to-standard GP practices, and for nearly 21-million patients whose data quality is routinely assessed by CPRD as acceptable for clinical research. More than 3 million of these patients are alive and currently registered in 401 contributing practices. Based on latest UK population estimates from the UK Office of National Statistics, GOLD covers 4.6% of the current UK population and includes 4.9% of currently contributing GP practices. GOLD contains data from all four UK constituent countries, and the current regional distribution of its GP practices is 5.7% in England, 55.6% in Scotland, 28.4% in Wales, and 10.2% in Northern Ireland (May 2022). GOLD data include patient's demographic, biological measurements, clinical symptoms and diagnoses, referrals to specialist/hospital and their outcome, laboratory tests/results, and prescribed medications. GOLD has been assessed and found broadly representative of the UK general population in terms of age, gender, and ethnicity. GOLD has been widely used internationally for observational research to produce nearly 3,000 peer-reviewed publications, making GOLD the most influential UK clinical database so far. In 2019, CPRD launched AURUM and since then has encouraged practices from England to move from the software that feeds GOLD (Vision) to the one that feeds AURUM (Emis). GOLD data from 2019 therefore mainly represents Wales/Scotland/NI and AURUM represents England. However, GOLD data collected before 2019 fully represent the UK. CPRD provides an updated list

of practices which moved from GOLD to AURUM for each build release. An overlap between GOLD and AURUM can occur because historical data for these practices have been transferred from Vision/GOLD to Emis/AURUM. When DARWIN EU® uses both databases, the safest and easiest solution would be to disregard these practices in GOLD. The licence also covers HES/ONS data, which can be requested on a study-by-study basis as linked data. This data only covers England and is planned to be mapped to OMOP in the future.

United Kingdom: UK BioBank (UKBB)

UK Biobank is a powerful biomedical database that can be accessed globally to enable new discoveries to improve public health. UK Biobank contains in-depth genetic, biomarker, imaging, and health information from over half a million volunteers living in the UK aged 40–69 years at the time of recruitment (2006–2010). UK Biobank has collected an unprecedented amount of biological and medical data as part of a large-scale long term prospective study. With their consent, they regularly provide blood, urine, and saliva samples, as well as detailed information about their lifestyle, which is then linked to their health-related records (e.g. primary care data, hospital data, cancer registry) to provide a deeper understanding of how individuals experience diseases. Since 2012 UK Biobank database, the largest and richest of its kind, is opened to applications from researchers. The resource is available in a strictly anonymised format to scientists from the UK and around the world, subject to verification that the research is health-related and in the public interest. Researchers are required to publish their results in an open source publication site or in an academic journal and return their findings to the UK Biobank. At the time of writing nearly 4,056 research applications have been approved for the usage of UK Biobank data and 8,553 peer-reviewed articles based on them have been published.

P4-C2-009:

Denmark: Danish Data Health Registries (DK-DHR)

Danish health data is collected, stored, and managed in national health registers at the Danish Health Data Authority, and covers the entire population, which makes it possible to study the development of diseases and their treatment over time. There are no gaps in terms of gender, age, and geography in Danish health data due to mandatory reporting on all patients from cradle to grave, in all hospitals and medical clinics. Personal identification numbers enable linking of data across registers, so we have data on all Danes throughout their lives, regardless of whether they have moved around the country. High data quality due to standardization, digitization, and documentation means that Danish health data is not based on interpretation. The Danish Health Data Authority is responsible for the national health registers and for maintaining and developing standards and classifications in the Danish healthcare system. Legislation ensures balance between personal data protection and use. In the present data base, we have access to the following registries for the entire Danish population of 5.9 million persons from 1/1/1995: The central Person Registry (CPR), The National Patient Registry (LPR), The Register of Pharmaceutical Sales (LSR), The National Cancer Register (CAR), The Cause of Death registry (DAR), The Clinical Laboratory Information Register (LAB), COVID-19 test and vaccination Registries (SSI-OVD, SSI-DDV), and The complete Vaccination registry (DDV_all). All data registered from 1/1/1995 will be included.

Greece: Papageorgiou General Hospital (PGH)

PGH, situated in Thessaloniki—the second-largest city in Greece—began operations in 1999. By 2004, it had formed a partnership with Aristotle University of Thessaloniki's School of Medicine, hosting its Teaching Clinics. The hospital boasts 30 clinics, including 2 Internal Medicine clinics, 2 Surgery clinics, 10 collaborative departments, and 8 laboratory centers. With a capacity of 745 beds, PGH employs 1,939 staff members and supports over 200,000 hospitalization days annually, conducts 20,000 surgeries, and manages approximately 1,000 daily visits to outpatient departments. The eHealth Lab at the Institute of Applied Biosciences, part of the Centre for Research and Technology Hellas (INAB|CERTH), focuses on developing

software for medical informatics applications. INAB|CERTH, certified by EHDEN for its proficiency in Extract-Transform-Load (ETL) operations for OMOP CDM, serves as a subcontractor managing PGH's OMOP CDM instance on its behalf. PGH's information system integrates multiple databases, including Electronic Healthcare Records and Laboratory Information Systems, and aligns with international medical vocabularies and standards. It also encompasses an imaging system (PACS) to handle the extensive daily diagnostic imaging. Furthermore, PGH utilizes specialized software for logistical management, blood transfusion services, and more. The hospital's significant daily patient influx results in the production of a vast and diverse array of medical data.

Portugal: Egas Moniz Health Alliance database - Entre o Douro e Vouga (EMDB-ULSEDV)

The Clinical Academic Center Egas Moniz Health Alliance (CAC-EMHA) integrates several Portuguese institutions—the University of Aveiro and 4 Local Health Units (Aveiro, Entre Douro e Vouga, Vila Nova de Gaia/Espinho, and Matosinhos). More than 1 million clinical records of patients are included. The CAC-EMHA has defined main problems for intervention, aligned with the needs of public health and considering the clinical and scientific differentiation of its professionals, in the following areas: a) cardiovascular and respiratory; b) muscle and bone; c) infection and resistance; d) neurosciences. Unidade Local de Saúde de Entre Douro e Vouga (ULSEDV) is an integrated public medical care centre comprising of primary, secondary, and tertiary healthcare. It fully serves approximately 274,000 patients of the municipalities of Santa Maria da Feira, Arouca, São João da Madeira, Oliveira de Azeméis, Vale de Cambra, Ovar, and Castelo de Paiva. The ULSEDV includes 32 primary care centres assisted by three hospitals (Hospital de São Sebastião, Hospital São João da Madeira, and Hospital São Miguel).

Portugal: Egas Moniz Health Alliance database - Gaia E Espinho (EMDB-ULSGE)

The Clinical Academic Center Egas Moniz Health Alliance (CAC-EMHA) integrates several Portuguese institutions—the University of Aveiro and 4 Local Health Units (Aveiro, Entre Douro e Vouga, Vila Nova de Gaia/Espinho, and Matosinhos). More than 1 million clinical records of patients are included. The CAC-EMHA has defined main problems for intervention, aligned with the needs of public health and considering the clinical and scientific differentiation of its professionals, in the following areas: a) cardiovascular and respiratory; b) muscle and bone; c) infection and resistance; d) neurosciences. Unidade Local de Saúde de Gaia e Espinho (ULSGE) is an integrated public medical care centre, comprising of primary, secondary, and tertiary healthcare. It fully serves the population of the municipalities of Gaia and Espinho, which amounts to approximately 350,000 patients. The ULSGE includes 32 primary care centres, assisted by three hospitals (Hospital Eduardo Santos Silva, Hospital Distrital Vila Nova de Gaia, Hospital Nossa Senhora da Ajuda) and one specialized rehabilitation centre (Centro de Reabilitação do Norte).

Portugal: Egas Moniz Health Alliance database - Baixo Vouga (Região de Aveiro) (EMDB-ULSRA)

The Clinical Academic Center Egas Moniz Health Alliance (CAC-EMHA) integrates several Portuguese institutions—the University of Aveiro and 4 Local Health Units (Aveiro, Entre Douro e Vouga, Vila Nova de Gaia/Espinho, and Matosinhos). More than 1 million clinical records of patients are included. The CAC-EMHA has defined main problems for intervention, aligned with the needs of public health and considering the clinical and scientific differentiation of its professionals, in the following areas: a) cardiovascular and respiratory; b) muscle and bone; c) infection and resistance; d) neurosciences. Unidade Local da Região de Aveiro (ULSRA) is an integrated public medical care centre, comprising of primary, secondary, and tertiary healthcare. It serves approximately 390,000 patients from most of the municipalities of Aveiro. The ULSRA includes 41 primary care centres, assisted by four hospitals (Hospital Dr. Francisco Zagalo, Hospital Visconde de Salreu, Hospital Distrital de Águeda, Hospital Infante D. Pedro).

Portugal: Unidade Local de Saúde de Matosinhos Realtime Database (ULSM-RT)

The database is comprised of clinical information of patients admitted at 1 public hospital centre, located in Portugal, and 14 connected primary healthcare centres providing full coverage to the municipality of Matosinhos. It includes administrative, sociodemographic, and clinical data (visits, measurements, drugs,

procedures, observations) of over 700,000 patients with ages ranging from 0 to 100 years. Of these, 200,000 patients have complete follow-up in primary care with complete data in the mentioned domains. The hospital centre includes departments of most medical specialties, comprising data from surgery, outpatient, ward, accident and emergency, and ICU areas. Data spans back to 1998, with the highest density starting from 2014.

Spain: Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público (BIFAP)

BIFAP (http://www.bifap.org/index_EN.html) is a longitudinal population-based data source of medical patient records of the Spanish National Health Service (SNS). It includes data from 9 of the 17 regions in Spain. The population currently included represents 36% of the total Spanish population. Spain has a SNS that provides universal access to health services through the Regional Healthcare Services. Primary care physicians (PCPs), both general practitioners and paediatricians, have a central role. They act as gatekeepers of the system and exchange information with other levels of care to ensure the continuity of care. Most of the population (98.9%) is registered with a PCP and, in addition, most drug prescriptions are written at the primary care level. BIFAP includes a collection of databases linked at individual patient level. The main one is the Primary care Database, given the central role of PCPs in the SNS. Linked, there are additional important structural databases, like the medicines dispensed at community pharmacies and the patients' hospital diagnosis at discharge. 7 out of the 9 regions have linkage to hospital data. However, hospital data is available for different time periods for each region. From 2014 onwards, linkage to hospital data is available for >68% of patients. Linkage to SARS-CoV-2 diagnostics tests and COVID-19 vaccination registries is also included. Additional databases are also linked for a subset of patients (hospital pharmacy, cause of death registry). The BIFAP program is a non-profit program financed by the Spanish Agency of Medicines and Medical Devices (AEMPS), a government agency belonging to the Ministry of Health, in collaboration with the Regional health authorities. The main use of BIFAP is for research purposes in order to evaluate the adverse and beneficial effects of drugs and drug utilization patterns in the general population under real use conditions.

Spain: Institut Municipal Assistència Sanitària Information System (IMASIS)

The Institut Municipal Assistència Sanitària Information System (IMASIS) is the Electronic Health Record (EHR) system of Parc de Salut Mar Barcelona (PSMar), which is a complete healthcare services organisation. Currently, this information system includes and shares the clinical information of two general hospitals (Hospital del Mar and Hospital de l'Esperança), one mental health care centre (Centre Dr. Emili Mira), and one social-healthcare centre (Centre Fòrum), including emergency room settings, that are offering specific and different services in the Barcelona city area (Spain). At present, IMASIS includes clinical information from around 1 million patients with at least one diagnosis, and who have used the services of this healthcare system since 1990, and from different settings such as admissions, outpatients, emergency room, and major ambulatory surgery. The diagnoses are coded using The International Classification of Diseases ICD-9-CM and ICD-10-CM. The average follow-up period per patient in years is 6.37 (SD±6.82). IMASIS-2 is the anonymized relational database of IMASIS which is used for mapping to OMOP, including additional sources of information such as the Tumours Registry

Spain: Valencia Health System Integrated Dataset (VID)

The Valencia Health System Integrated Dataset (VID) is a set of multiple, public, population-wide electronic databases for the Valencia Region, the fourth most populated Spanish region, with about 5 million inhabitants and an annual birth cohort of 48,000 new-borns, representing 10.7% of the Spanish population and around 1% of the European population. The VID provides exhaustive longitudinal information, including sociodemographic and administrative data (sex, age, nationality, etc.), clinical (diagnoses, procedures, diagnostic tests, imaging, etc.), pharmaceutical (prescription, dispensing), and healthcare utilization data from hospital care, emergency departments, specialized care (including mental and obstetrics care), primary care, and other public health services. It also includes a set of associated population databases and registries of significant care areas, such as cancer, rare diseases, vaccines,

congenital anomalies, microbiology (including COVID-19 test results registry), and others, as well as public health databases from the population screening programmes. All the information in the VID databases can be linked at the individual level through a single personal identification code. The databases were initiated at different moments in time, but all in all the VID provides comprehensive individual-level data fed by all the databases since 2008 to date. IMPORTANT: The OMOP instance has been created using CONSIGN project data, where 1.96 Million of females in fertile age are studied from the start of 2018 to the end of 2021.

P4-C2-010:

Finland: Auria Clinical Informatics (FinOMOP-ACI Varha)

The data covers the patient register at the Hospital District of Southwest Finland (HDSF), containing Turku University Hospital, which is one of the five university hospitals in Finland. It covers the public specialist health care and most emergency health care in the area of Southwest Finland (Varsinais-Suomi) for all demographic groups. The data is utilized for scientific research from the data lake in the HDSF under the Finnish legislation (The Act on Secondary Use of Health and Social Data). The most relevant data domains are patients, visits, inpatient episodes, diagnoses, laboratory results, procedures, medication, pathology, radiology, radiotherapy, chemotherapy, obstetrics, and narrative patient reports, however there are also other data domains available.

Finland: Finnish Care Register for Health Care (FinOMOP-THL)

The THL database covers both public and private, primary and specialised inpatient and outpatient health care encounters in Finland, starting from 2011. The entire public sector and private inpatient encounters have been included since 2011, while private outpatient encounters, including occupational care, are included since 2020. The main content of the THL CDM is The Finnish Care Register for Health Care (fi:Hoitoilmoitusrekisteri, HILMO). It is a continuation of the former Hospital Discharge Register, which originally gathered data on patients discharged from hospitals. The Care Register has comprehensive data on the use of services and service users from Finnish public inpatient and outpatient primary and specialised care nationwide. Since 1998, the register has covered both public outpatient and inpatient specialized care and private inpatient care (TerveysHilmo). Since 2011, the register has covered public primary care (AvoHilmo). Since 2020, the register has covered private outpatient care and occupational care. In addition, the CDM also contains the vaccination data from the Finnish National Vaccination Register, and positive COVID-19 test results from the Finnish National Infectious Diseases Register, which is maintained by THL. The CDM is currently produced from the above-mentioned, and limited to observation periods commencing after 1/1/2011. The National Population registry is also used as a source for the CDM database. The National Population registry data forms the basis for forming the patient population. This ensures an up-to-date location (municipality of residence) of patients, as well as complete death occurrences (although not the cause of death). Using the complete population as a basis for the person table also serves to facilitate calculations on a population level, e.g. incidence rates. The current CDM population comprises all persons having been alive and residing in Finland since the beginning of 2011.

Finland: Hospital District of Helsinki and Uusimaa (FinOMOP-HUS)

The HUS data lake is a comprehensive, integrated data source derived in real-time from all patients who visit the HUS hospitals and receive treatment. HUS is responsible for specialized healthcare in Finland's Uusimaa region and treatment of many rare and severe diseases that are nationally centralized to HUS. HUS' catchment area covers a population of about 2.2 million people: 2.4 million outpatient clinic visits, 670,000 patients treated, 84,000 surgical procedures, 710 specialist care emergency clinic visits per day. All visits, examinations, laboratory test, procedures, and treatments are recoded in the HUS IT systems and integrated into the data lake. The data lake stores decades of clinical information in digital format, and data from both past and current source systems are available. Systems providing data into the data lake include:

CGI Uranus and Epic Apotti (EHR: visits, diagnoses, medication, etc.), Opera (procedure records), Kemokur and Beacon (cancer-specific medications), Marela (hospital pharmacy), Multilab/Mylab+ (laboratory system), Qpati/Mylab+ (pathology records system). HUS Acamedic is a dedicated, accredited, secure analytics environment for research utilizing these data.

Finland: Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)

TaUH Research Database includes all specialties/all patient groups treated in the Tampere University Hospital, secondary, and tertiary care given in the region, including given clinical and pathology diagnoses, diagnostic and therapeutic procedures, laboratory findings, radiology and pathology reports, medication given in the hospital and electronic prescriptions, and continuous medical records (free text), including discharge letters since 2007.

France: Assistance Publique Hôpitaux de Marseille (APHM)

The data source used in this study includes all hospital stays across various care settings—acute care, psychiatric care, rehabilitation care, and home hospitalization—capturing approximately 300,000 stays annually. Diagnoses are coded using ICD-10 and procedures are recorded using CCAM, in line with the French DRG system, managed via the CORA software. The database also captures comprehensive drug prescription and administration data, including UCD drug codes, ATC classifications, quantities, and dosages, managed through the PHARMA software. Additionally, medical and paramedical notes, such as hospitalization reports, radiology, EEG, endoscopy, and consultation summaries are recorded using the AXIGATE software. Laboratory data, covering both prescriptions and test results, is also included.

Hungary: Semmelweis University Clinical Data (SUCD)

Semmelweis University is the largest provider of health care services in Hungary. Most of the departments cater to the most serious cases and patients requiring complex treatment, thus making the university a national health care provider. The overwhelming majority of patient data originates from Hungary, mainly from the central region of the country: Budapest and Pest County. The database contains approximately 2 million individual patients across all care settings of the University since 2011. The hospital information system (MedSolution) is an integrated IT system, which provides functional support for inpatient and outpatient care processes and serves as an integrated platform for different diagnostic areas, as well as, in some specific areas, supporting the registration of medications. It supports all kinds of hospital work processes from admission to discharge. The outpatient module serves as a platform for the registration of activities related to a care episode within the outpatient specialist care. During the care provision, data related to health state of the patient, the diagnosis, the documentation of requested examinations and medical consultations, prescribed medication, final reports, and performed interventions are recorded. The functions of the inpatient module assist the care provision within the inpatient settings. It documents the health state of the patient at admission and during the hospital stay, along with the anamnesis, diagnosis, the performed examinations and interventions, hospital final reports, and provided medication in some areas of care provision, such as chemotherapy. Among other modules, the diagnostic module registers the requested laboratory and imaging examinations and records the laboratory results.

Norway: Norwegian Linked Health Registry data (NLHR)

Norway has a universal public health care system, consisting of primary and specialist health care services covering a population of approximately 5.4 million inhabitants. Many population-based health registries were established in the 1960s, with use of unique personal identifiers facilitating linkage between registries. Data in these health registries are used for health analysis, health statistics, improving the quality of healthcare, research, administration, and emergency preparedness. We harmonized data from the following registries: the Medical Birth Registry of Norway (MBRN), the Norwegian Prescription Registry (NorPD), the Norwegian Patient Registry (NPR), Norway Control and Payment of Health Reimbursement (KUHR), the Norwegian Surveillance System for Communicable Diseases (MSIS), the Norwegian Immunisation Registry (SYSVAK), the National Death Registry, and the National Registry (NR). Linkage

between the registries was facilitated using project-specific person IDs generated from unique personal identification assigned at birth or immigration for all legal residents in Norway. Briefly: MBRN stores information about the pregnancy, the mother, father, and child; NPR records diagnosis in secondary care (e.g., hospital); KUHR contains information about diagnosis and contact in primary care (e.g., GPs and outpatient specialists) – to be included in third release; NorPD recorded all medications dispensed outside of hospitals; MSIS collects test results of communicable diseases (e.g., Sars-Cov-2); and SYSVAK recorded vaccinations.

Sweden: Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)

The Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED) study is a nationwide linked multi-register, regularly updated, observational study for a timely response over time to scientific questions around effectiveness and safety of approved drugs that can arise suddenly, requiring rapid evidence for timely regulatory action to protect patients' health and lives. The study data covers the entire Swedish population (about 10 million), with data on specialist care (National Patient Register), drug use (Prescribed Drug Register), cause of death (Cause-of-Death Register), sociodemographic data, and selected clinical data. Primary care visit diagnoses and procedures are available for 40% of the population (two largest Swedish regions). Most data start from 2015; prescription drug data on all prescriptions filled nationally are available from 2018. Data on hospital-administered drugs is not available if the prescription was not individually filled by the patient. The study population and all data are updated quarter-yearly. HI-SPEED builds on the predecessor project SCIFI-PEARL (Swedish COVID-19 Investigation for Future Insights - a Population Epidemiology Approach using Register Linkage) that was initiated in 2020 to conduct research on Covid-19 and pandemic-relations (https://www.gu.se/en/research/scifi-pearl).

ANNEX II: Additional information

DATA MANAGEMENT

Data management

All data sources have previously mapped their data to the OMOP common data model. This enables the use of standardised analytics and using DARWIN EU® tools across the network since the structure of the data and the terminology system is harmonised. The OMOP CDM was developed and maintained by the Observational Health Data Sciences and Informatics (OHDSI) initiative and is described in detail on the wiki page of the CDM: https://ohdsi.github.io/CommonDataModel and in The Book of OHDSI: http://book.ohdsi.org.

The analytic code for this study will be written in R and will use standardized analytics wherever possible. Each data partner will execute the study code against their data source containing patient-level data and then return the results (csv files) which will only contain aggregated data. The results from each of the contributing data sites will then be combined in tables and figures for the study report.

Data storage and protection

For this study, participants from various EU member states will process personal data from individuals which is collected in national/regional electronic health record data sources. Due to the sensitive nature of this personal medical data, it is important to be fully aware of ethical and regulatory aspects and to strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy.

All data sources used in this study are already used for pharmaco-epidemiological research and have a well-developed mechanism to ensure that European and local regulations dealing with ethical use of the data and adequate privacy control are adhered to. In agreement with these regulations, rather than combining person level data and performing only a central analysis, local analyses will be run, which generate non-identifiable aggregate summary results.

The output files are stored in the DARWIN EU® Remote Research Environment (RRE). These output files do not contain any data that allow identification of subjects included in the study. The RRE implements further security measures to ensure a high level of stored data protection to comply with the local implementation of the General Data Protection Regulation (GDPR) (EU) 679/20161 in the various member states.

QUALITY CONTROL

General data source quality control

A number of open-source quality control mechanisms for the OMOP CDM have been developed (see Chapter 15 of The Book of OHDSI http://book.ohdsi.org/DataQuality.html). In particular, it is expected that data partners will have run the OHDSI Data Quality Dashboard tool (https://github.com/OHDSI/DataQualityDashboard). This tool provides numerous checks relating to the conformance, completeness, and plausibility of the mapped data. Conformance focuses on checks that describe the compliance of the representation of data against internal or external formatting, relational, or computational definitions, completeness in the sense of data quality is solely focused on quantifying missingness, or the absence of data, while plausibility seeks to determine the believability or truthfulness of data values. Each of these categories has one or more subcategories and are evaluated in two contexts: validation and verification. Validation relates to how well data align with external benchmarks with expectations derived from known true standards, while verification relates to how well data conform to local knowledge, metadata descriptions, and system assumptions.

Study specific quality control

When defining drug cohorts, non-systemic products will be excluded from the list of included codes summarised on the ingredient level.

When defining cohorts for indications, a systematic search of possible codes for inclusion will be identified using the *CodelistGenerator* R package (https://github.com/darwin-eu/CodelistGenerator). This software allows the user to define a search strategy and using this will then query the vocabulary tables of the OMOP common data model so as to find potentially relevant codes. In addition, the *CohortDiagnostics* R package (https://github.com/OHDSI/CohortDiagnostics) will be run if needed to assess the use of different codes across the data sources contributing to the study and identify any codes potentially omitted in error.

The study code will be based on two R packages currently being developed to estimate prevalence and characterise individuals with obesity and obesity-related measurements using the OMOP common data model. These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via GitHub.

PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

A PDF report including an executive summary, and the specified tables and/or figures will be submitted to EMA by the DARWIN EU® CC upon completion of the study. An interactive dashboard incorporating all the results (tables and figures) will be provided alongside the PDF report. The full set of underlying aggregated data used in the dashboard will also be made available if requested. Once the report has been submitted, a manuscript will be submitted to a peer-review journal.

ANNEX III: List of stand-alone documents

Table S2. Preliminary list of conditions definitions.

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
Obesity	Obese	4215968	Including descendants: 380500	SNOMED
Diabetes mellitus type 2	Type 2 diabetes mellitus	201826	-	SNOMED
Hypertension	Hypertensive disorder	316866	-	SNOMED
Ischemic heart disease	Ischemic heart disease	4185932	-	SNOMED
Chronic kidney disease	Chronic kidney disease	46271022	Including descendants: 443961, 37019193, 45757392, 45757393, 45768813, 45771064, 45771067, 45772751	SNOMED
Hypothyroidism	Hypothyroidism	140673	-	SNOMED
Hypertriglyceridemia	Hypertriglyceridemia	4120314	-	SNOMED
Metabolic syndrome X	Metabolic syndrome X	436670	-	SNOMED
Cushing's syndrome	Hypercortisolism	195212	Including descendants: 4028805, 4029451, 4030207, 4030208, 4182585	SNOMED
Knee arthrosis	Osteoarthritis of knee	4079750	Including descendants: 74444, 761407, 4264472, 40479260	SNOMED
	Primary gonarthrosis, bilateral	4114585		
Obstructive sleep apnoea	Obstructive sleep apnea syndrome	442588	-	SNOMED
Mental health disorders, including depression and anxiety	Depressive disorder	440383	Including descendants: 35622958, 44782943, 44813499	SNOMED
	Depressed mood	40546087		
	Anxiety	441542		
Dyslipidaemia	Dyslipidemia	4159131	-	SNOMED
Metabolic dysfunction-associated steatotic liver disease	Steatosis of liver	4059290	-	SNOMED
Non-alcoholic fatty liver disease				
Steatosis of liver				

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
Cancer	Cancer	141232,443392,4189640	4155297, 433435	SNOMED

Table S3. Preliminary list of medicines definitions.

Substance Name	Concept name	Class	ATC code	Ingredient Concept ID	Include descendants
Glucagon-like peptide-1 (GLP-1) receptor agonists	Glucagon-like peptide-1 (GLP-1) analogues (	ATC 4th	A10BJ	1123618	Yes
	semaglutide	Ingredient	-	793143	Yes
	Tirzepatide	Ingredient	-	36860244	Yes
	liraglutide	Ingredient	-	40170911	Yes
	dulaglutide	Ingredient	-	45774435	Yes
	exenatide	Ingredient	-	1583722	Yes
	lixisenatide	Ingredient	-	44506754	Yes
	albiglutide	Ingredient	-	44816332	Yes
Orlistat	Orlistat	Ingredient	-	741530	Yes
Metformin	Metformin	Ingredient	-	1503297	Yes
Naltrexone-bupropion	Naltrexone-bupropion	ATC 5th	A08AA62	1588704	Yes, but exclude without descendants: 750982,1714319,45774483,45774484

Table S4. Preliminary list of observation definitions.

Phenotype	Concept id (including descendants)	Exclude concept id	Vocabulary
Diet	443359	With descendants: 4103471	LOINC
Physical activity	40654125	With descendants: 1761716, 1988912, 1989396, 3004249, 3004691, 3005629, 3012888, 3015514, 3022007, 3022304, 3025315, 3036277, 3038553, 3040891, 3042292, 3042888, 3042942, 3044062, 3045837, 3046098, 3046810, 3046965, 3050959, 21493624, 21493625, 21493626, 21493628, 21493629, 21493630, 21493632, 21493633, 21493635, 21493637, 21493638, 21493639, 36305205, 36305295, 36305415, 40762503, 40762508, 40763897, 40766813, 40766817, 40766818, 40766821, 40766822, 40766825, 40766828, 40767145, 40767146, 40767206, 40767207, 40768893, 40768894, 40768899, 40768903, 40768904, 40768905, 40768907, 40768909, 40768910, 40768913, 40768916, 40768921, 40768923, 40768929, 40768948, 40768950, 40768957, 40768973, 40769011, 40769016, 40769017, 40769019, 40769020, 40769021, 40769024, 40769025, 40769026, 40769027, 40769029, 40769030, 40769031, 40769032, 40769033, 40769034, 40769035, 40769110, 40769763	LOINC
Smoking status	40654161	-	LOINC

Table S5. Preliminary list of measurement definitions.

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
BMI	Body mass index (BMI	3038553	-	LOINC
	Body mass index (BMI) [Percentile] Per age and sex	40762638		
	Body mass index	4245997		
'Weight	Height and weight Set	3045269	With descendants: 3023540, 3038553, 3019204	LOINC
	Weight and Height tracking panel	40758547		
	Body weight	3025315		
	Height and weight	40757698		
	Height and weight	4061893		
Height	Height / growth measure	4154781	607563, 4061893, 4090586 , 4200189, 45770285, 45770287	SNOMED
Waist circumference	Waist circumference	4172830	-	SNOMED
	Abdominal circumference	4245376		
Hip circumference	Hip circumference	4111665	-	SNOMED
Waist-to-height ratio	Waist to height ratio	44809433	-	SNOMED
Waist-to-hip ratio	Waist/hip ratio	4087501	-	SNOMED
Abdominal skin fold thickness	Skin-fold thickness	4013687	-	SNOMED
Body composition	Body composition measure	4178503	4087498, 4245997	
Body fat	Total body fat	4087498	-	SNOMED
Cholesterol	Calculus cholesterol content measurement	4210880	-	SNOMED
	Serum cholesterol/HDL ratio measurement	4195490		
	Serum cholesterol/LDL ratio measurement	4195491		
	Plasma cholesterol/VLDL ratio measurement	4195492		
	Cholesterol measurement	4299360		

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
	Plasma cholesterol/LDL ratio measurement	4199032		
	Calculated LDL cholesterol level	4191837		
	Serum cholesterol/VLDL ratio measurement	4198117		
Glycaemia	Glucose measurement, fasting	4182052	-	SNOMED
	Glucose measurement, quantitative	4018317		
	Glucose measurement, body fluid	4151548		
	Glucose measurement estimated from glycated hemoglobin	4036846		
	Glucose concentration, test strip measurement	4094447		
	Glucose measurement, post glucose dose	4017078		
	Glucose measurement, blood	4144235		
	Glucose measurement by monitoring device	4230393		
	Glucose measurement, random	4249006		
Glycated haemoglobin A1c	Glycated hemoglobin-A1c	37116827	-	SNOMED
	Glycated haemoglobin (HbA1c) correction level	35814471		
	Glycated haemoglobin (HbA1c) reportability	35814474		
	Glycated haemoglobin	35814472		

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
	(HbA1c) correction reason			
	Glycated haemoglobin (HbA1c) missing reason	35814473		
Triglycerides	Fasting blood lipids	4150326	-	SNOMED
	Fasting lipid profile	4090059		
	Triglycerides measurement	4032789		
	Lipids, triglycerides measurement	4017787		

Table S6. Preliminary list of procedure definitions.

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
Bariatric surgery	Gastric banding for obesity	3170866	-	Nebraska Lexicon
	Jejunocecostomy for obesity	4135346		SNOMED
	Sleeve resection of stomach	4227605		
	Jejunocolostomy for obesity	4232489		
	Jejunoileostomy bypass shunt for obesity	4234135		
	Intestinal bypass for morbid obesity	4245767		
	Gastric stapling for obesity	4307794		
	Bypass of stomach	Excluding descendants: 40483096		
	Laparoscopic bypass of stomach	46270974		
	Duodenal switch	4145504		

ANNEX IV: ENCePP checklist for study protocols

Doc.Ref. EMA/540136/2009

ENCePP Checklist for Study Protocols (Revision 4)

Adopted by the ENCePP Steering Group on 15/10/2018

Study title: DARWIN EU® - Capturing obesity, obesity-related variables, and changes in weight over time across the DARWIN EU® network

EU PAS Register® number: EUPAS1000000820 **Study reference number:** P4-C3-004/P4-C2-008/P4-C2-009/P4-C2-010

Section 1: Milestones		**Yes**	**No**	**N/A**	**Section Number**
1.1	Does the protocol specify timelines for				
	1.1.1 Start of data collection[1]	☒	☐	☐	
	1.1.2 End of data collection[2]	☒	☐	☐	
	1.1.3 Progress report(s)	☐	☐	☒	5
	1.1.4 Interim report(s)	☒	☐	☐	
	1.1.5 Registration in the EU PAS Register®	☒	☐	☐	
	1.1.6 Final report of study results.	☒	☐	☐	

Comments:

Section 2: Research question		**Yes**	**No**	**N/A**	**Section Number**
2.1	Does the formulation of the research question and objectives clearly explain:	☒	☐	☐	
	2.1.1 Why the study is conducted? (e.g. to address an important public health concern, a risk identified in the risk management plan, an emerging safety issue)	☒	☐	☐	
	2.1.2 The objective(s) of the study?	☒	☐	☐	7
	2.1.3 The target population? (i.e. population or subgroup to whom the study results are intended to be generalised)	☒	☐	☐	
	2.1.4 Which hypothesis(-es) is (are) to be tested?	☐	☐	☒	
	2.1.5 If applicable, that there is no *a priori* hypothesis?	☐	☐	☒	

Comments:

[1] Date from which information on the first study is first recorded in the study dataset or, in the case of secondary use of data, the date from which data extraction starts.
[2] Date from which the analytical dataset is completely available.

Section 3: Study design		**Yes**	**No**	**N/A**	**Section Number**
3.1	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	☒	☐	☐	8.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	8.1
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.8
3.4	Does the protocol specify measure(s) of association? (e.g. risk, odds ratio, excess risk, rate ratio, hazard ratio, risk/rate difference, number needed to harm (NNH))	☐	☐	☒	
3.5	Does the protocol describe the approach for the collection and reporting of adverse events/adverse reactions? (e.g. adverse events that will not be collected in case of primary data collection)	☐	☐	☒	

Comments:

Section 4: Source and study populations		**Yes**	**No**	**N/A**	**Section Number**
4.1	Is the source population described?	☒	☐	☐	8.5
4.2	Is the planned study population defined in terms of:				8.5
	4.2.1 Study time period	☒	☐	☐	
	4.2.2 Age and sex	☒	☐	☐	
	4.2.3 Country of origin	☒	☐	☐	
	4.2.4 Disease/indication	☒	☐	☐	
	4.2.5 Duration of follow-up	☐	☐	☒	
4.3	Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	☒	☐	☐	8.5

Comments:

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.1	Does the protocol describe how the study exposure is defined and measured? (e.g. operational details for defining and categorising exposure, measurement of dose and duration of drug exposure)	☒	☐	☐	8.6
5.2	Does the protocol address the validity of the exposure measurement? (e.g. precision, accuracy, use of validation sub-study)	☐	☐	☒	
5.3	Is exposure categorised according to time windows?	☐	☐	☒	
5.4	Is intensity of exposure addressed? (e.g. dose, duration)	☐	☐	☒	
5.5	Is exposure categorised based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?	☐	☐	☒	
5.6	Is (are) (an) appropriate comparator(s) identified?	☐	☐	☒	

Comments:

Section 6: Outcome definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
6.1	Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated?	☒	☐	☐	8.6
6.2	Does the protocol describe how the outcomes are defined and measured?	☒	☐	☐	8.6
6.3	Does the protocol address the validity of outcome measurement? (e.g. precision, accuracy, sensitivity, specificity, positive predictive value, use of validation sub-study)	☐	☐	☒	
6.4	Does the protocol describe specific outcomes relevant for Health Technology Assessment? (e.g. HRQoL, QALYs, DALYS, health care services utilisation, burden of disease or treatment, compliance, disease management)	☐	☐	☒	

Comments:

Section 7: Bias		**Yes**	**No**	**N/A**	**Section Number**
7.1	Does the protocol address ways to measure confounding? (e.g. confounding by indication)	☐	☐	☒	
7.2	Does the protocol address selection bias? (e.g. healthy user/adherer bias)	☒	☐	☐	8, 9
7.3	Does the protocol address information bias? (e.g. misclassification of exposure and outcomes, time-related bias)	☒	☐	☐	8, 9

Comments:

Section 8: Effect measure modification		**Yes**	**No**	**N/A**	**Section Number**
8.1	Does the protocol address effect modifiers? (e.g. collection of data on known effect modifiers, sub-group analyses, anticipated direction of effect)	☒	☐	☐	8, 9

Comments:

Section 9: Data sources		**Yes**	**No**	**N/A**	**Section Number**
9.1	Does the protocol describe the data source(s) used in the study for the ascertainment of:				
	9.1.1 Exposure? (e.g. pharmacy dispensing, general practice prescribing, claims data, self-report, face-to-face interview)	☒	☐	☐	8.2
	9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	☒	☐	☐	8.2
	9.1.3 Covariates and other characteristics?	☒	☐	☐	8.2

Section 9: Data sources		**Yes**	**No**	**N/A**	**Section Number**
9.2	Does the protocol describe the information available from the data source(s) on:				
	9.2.1 Exposure? (e.g. date of dispensing, drug quantity, dose, number of days of supply prescription, daily dosage, prescriber)	☒	☐	☐	8.2
	9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	☒	☐	☐	8.2
	9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	☒	☐	☐	8.2
9.3	Is a coding system described for:				
	9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	☒	☐	☐	8.2
	9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	☒	☐	☐	8.2
	9.3.3 Covariates and other characteristics?	☒	☐	☐	8.2
9.4	Is a linkage method between data sources described? (e.g. based on a unique identifier or other)	☒	☐	☐	8.2

Comments:

Section 10: Analysis plan		**Yes**	**No**	**N/A**	**Section Number**
10.1	Are the statistical methods and the reason for their choice described?	☒	☐	☐	8.8
10.2	Is study size and/or statistical precision estimated?	☐	☐	☒	
10.3	Are descriptive analyses included?	☒	☐	☐	8.8
10.4	Are stratified analyses included?	☒	☐	☐	8.8
10.5	Does the plan describe methods for analytic control of confounding?	☐	☐	☒	
10.6	Does the plan describe methods for analytic control of outcome misclassification?	☐	☐	☒	
10.7	Does the plan describe methods for handling missing data?	☐	☐	☒	
10.8	Are relevant sensitivity analyses described?	☐	☐	☒	

Comments:

Section 11: Data management and quality control		**Yes**	**No**	**N/A**	**Section Number**
11.1	Does the protocol provide information on data storage? (e.g. software and IT environment, database maintenance and anti-fraud protection, archiving)	☒	☐	☐	Annex II
11.2	Are methods of quality assurance described?	☒	☐	☐	Annex II
11.3	Is there a system in place for independent review of study results?	☒	☐	☐	Annex II

Comments:

Section 12: Limitations	**Yes**	**No**	**N/A**	**Section Number**
12.1 Does the protocol discuss the impact on the study results of:				
12.1.1 Selection bias?	☒	☐	☐	
12.1.2 Information bias?	☒	☐	☐	9
12.1.3 Residual/unmeasured confounding? (e.g. anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).	☐	☐	☒	
12.2 Does the protocol discuss study feasibility? (e.g. study size, anticipated exposure uptake, duration of follow-up in a cohort study, patient recruitment, precision of the estimates)	☐	☐	☒	

Comments:

Section 13: Ethical/data protection issues	**Yes**	**No**	**N/A**	**Section Number**
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	☒	☐	☐	Annex II
13.2 Has any outcome of an ethical review procedure been addressed?	☐	☐	☒	
13.3 Have data protection requirements been described?	☒	☐	☐	Annex II

Comments:

Section 14: Amendments and deviations	**Yes**	**No**	**N/A**	**Section Number**
14.1 Does the protocol include a section to document amendments and deviations?	☒	☐	☐	4

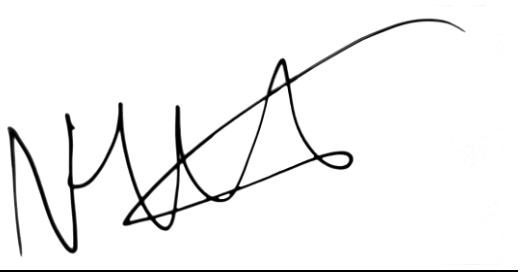
Comments:

Section 15: Plans for communication of study results	**Yes**	**No**	**N/A**	**Section Number**
15.1 Are plans described for communicating study results (e.g. to regulatory authorities)?	☒	☐	☐	Annex II
15.2 Are plans described for disseminating study results externally, including publication?	☒	☐	☐	Annex II

Comments:

Name of the main author of the protocol: Nicholas Hunt

Date: 27/08/2025

Signature: 

ANNEX V: Glossary

Additional definitions are available in the EMA Glossary of terms https://www.ema.europa.eu/en/about-us/glossaries.

Aggregated Data

Data collected and combined from multiple sources to generate summary information, typically anonymised.

Benefit-Risk Assessment

Evaluation of the positive therapeutic effects of a medicine compared to its risks (e.g., side effects).

Common Data Model (CDM)

A standardized data structure that enables data from multiple sources to be harmonized, making analysis consistent and reproducible. DARWIN EU® utilises the OMOP CDM maintained by the OHDSI community.

Complex Studies (C3)

Studies requiring the development or customisation of specific study designs, protocols, and Statistical Analysis Plans (SAPs), with extensive collection or extraction of data. Examples include etiological studies measuring the strength and determinants of an association between an exposure and the occurrence of a health outcome in a defined population considering sources of bias, potential confounding factors, and effect modifiers.

Coordination Centre (CC)

The central hub responsible for managing and overseeing the activities within DARWIN EU®. It is based at Erasmus University Medical Centre in Rotterdam, the Netherlands.

Data Access

The process of obtaining permission to use specific datasets for regulatory or scientific studies.

Data Quality Framework

A set of standards and procedures to ensure accuracy, completeness, timeliness, and consistency of data used in DARWIN EU®.

Data Source

A data source or repository of structured health-related data, such as electronic health records (EHRs), insurance claims, or registries.

DARWIN EU®

The European Medicines Agency's (EMA) federated network of real-world data sources designed to generate evidence to support regulatory decision-making.

EMA (European Medicines Agency)

The regulatory body responsible for the evaluation and supervision of medicinal products in the EU, overseeing DARWIN EU®.

Evidence Generation

The process of analysing real-world data to produce scientific information that can inform healthcare or regulatory decisions.

Federated Network

A data infrastructure where data remain at their original location but can be analysed in a harmonised way across multiple partners using a common model and tools.

GDPR (General Data Protection Regulation)

The EU regulation governing the protection of personal data and privacy, crucial to how DARWIN EU® handles health data.

Health Technology Assessment (HTA)

A systematic evaluation of properties and impacts of health technology, often using DARWIN EU® data to support assessments.

Metadata

Descriptive information about a data source (e.g., its content, quality, and structure), essential for identifying relevant data sources in DARWIN EU® studies.

Off-the-Shelf Studies (OTS)

Studies for which a standard protocol per study/analysis type and standardised analytics may be developed and applied or adapted, typically relating to a descriptive research question. This includes studies on disease epidemiology, for example, the estimation of the prevalence or incidence of health outcomes in defined time periods and population groups, or drug utilisation studies at the population or patient level.

OHDSI (Observational Health Data Sciences and Informatics)

An open-science collaborative community that develops tools and standards (including the OMOP CDM) to enable large-scale analytics of observational health data. OHDSI provides the technical and scientific foundation for DARWIN EU®'s analytical ecosystem.

Patient-Level Data

Data related to individuals, de-identified, used for longitudinal or detailed analyses.

OMOP (Observational Medical Outcomes Partnership)

A common data model (CDM) that standardises the structure and content of observational healthcare data, enabling systematic analysis across disparate datasets. DARWIN EU® uses the OMOP CDM to ensure interoperability and consistency in real-world evidence generation.

Real-World Data (RWD)

Data relating to individual health status or healthcare delivery that is collected from routine clinical practice rather than from randomised controlled trials.

Real-World Evidence (RWE)

Clinical evidence derived from the analysis of RWD, used to inform decisions by regulators, payers, or clinicians.

Regulatory Decision-Making

The process by which authorities like EMA assess data to authorise, monitor, or modify the use of medicines in the EU.

Routine Repeated Studies (RR)

Studies that are either Off-the-Shelf or Complex studies repeated on a regular basis, following the same protocol and study code, but with updated data and/or different data partners.

Study Protocol

A detailed plan describing how a specific real-world study will be conducted, including objectives, design, data sources, and analyses.

Very Complex Studies (C4)

Studies which cannot rely only on electronic health care data sources, or which would require complex methodological work, for example, due to the occurrence of events that cannot be defined by existing diagnosis codes, including events that do not yet have a diagnosis code, where it may be necessary to combine a diagnosis code with other data such as results of laboratory investigations. These studies might require the collection of data prospectively, or the inclusion of new (not previously onboarded) data sources.