Study Protocol

P4-C3-006, P4-C2-019, and P4-C2-020

DARWIN EU® -

Population demographics and disease frequency across the DARWIN EU® network

23/03/2026

Version 3.0

Authors: Anna Saura-Lazaro, Albert Prats-Uribe

Public

CONTENTS

Study title	DARWIN EU® - Population demographics and disease frequency across the DARWIN EU® network
Outline version	V3.0
Date	23/03/2026
EUPAS number	EUPAS1000000962
Active Substance	NA
Medicinal Product	NA
Research question and objectives	The aim of this study is to characterise the Data Analysis and Real-World Interrogation Network (DARWIN EU®) in terms of demographics and common clinical conditions, and to develop, document, and validate reusable standardised phenotypes for the description of disease prevalence and the characterisation of study populations within the network. The specific objectives of this study are: 1. To describe the demographic characteristics (age and sex) of the populations under observation in the DARWIN EU® data sources and to compare these characteristics with relevant national or regional reference data. 2. To develop reusable standardised phenotypes for selected common clinical conditions. 3. To estimate the period prevalence of selected common clinical conditions within the population under observation. 4. To describe the median age at first record and sex distribution of selected common clinical conditions within the population under observation.
Countries of study	P4-C3-006: Belgium, Germany, the Netherlands, Norway, Spain, Sweden, the United Kingdom P4-C2-019: France, Greece, Hungary, Portugal P4-C2-020: Croatia, Estonia, Germany, Spain
Authors	Anna Saura-Lazaro (a.sauralazaro@darwin-eu.org) Albert Prats-Uribe (a.prats-uribe@darwin-eu.org)

LIST OF ABBREVIATIONS

Acronyms/terms	Description
APHM	Assistance Publique Hôpitaux de Marseille
ATC	Anatomical Therapeutic Chemical
BIFAP	Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público
CDM	Common Data Model
CI	Confidence Interval
CPRD GOLD	Clinical Practice Research Datalink GOLD
DARWIN EU®	Data Analysis and Real-World Interrogation Network
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	Portugal: Egas Moniz Health Alliance database - Gaia e Espinho
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EUPAS	EU Post-Authorisation Studies Register
GP	General Practitioner
HI-SPEED	Swedish Population Evidence Enabling Data-linkage
InGef RDB	InGef Research Database
IRB	Institutional Review Board
IPCI	Integrated Primary Care Information
IQVIA DA Germany	IQVIA Disease Analyzer Germany
IQVIA LPD Belgium	IQVIA Longitudinal Patient Database Belgium
LOINC	Logical Observation Identifiers Names and Codes
NA	Not applicable
NAJS	Croatian National Public Health Information System
NLHR	Norwegian Linked Health Registry data
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PGH	Papageorgiou General Hospital
PRISIB	Plataforma de Recerca en Informació Sanitària de les Illes Balears
RxNorm	Medical prescription normalised
SIDIAP	The Information System for Research on Primary Care
SNOMED	Systemised Nomenclature of Medicine
SUCD	Semmelweis University Clinical Data
UK	The United Kingdom
ULSM-RT	Unidade Local de Saúde de Matosinhos Realtime Database
VID	Valencia Health System Integrated Dataset

1. TITLE

DARWIN EU® - Population demographics and disease frequency across the DARWIN EU® network

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator	Anna Saura-Lazaro Albert Prats-Uribe	University of Oxford
Data Scientist	Kim Lopez-Guell Marti Catala-Sabate	University of Oxford
Clinical Domain Expert	Anna Saura-Lazaro Albert Prats-Uribe Annika Jodicke	University of Oxford
Study Manager	Natasha Yefimenko Nosova	Erasmus MC
Data source **P4-C3-006**	**Names**	**Data Partner Organisation***
IQVIA LPD Belgium	Dina Vojinovic Ellen Gerritsen Akram Mendez Gargi Jadhav	IQVIA Longitudinal Patient Database Belgium
IQVIA DA Germany	Dina Vojinovic Ellen Gerritsen Akram Mendez Gargi Jadhav	IQVIA Disease Analyzer Germany
IPCI	Katia Verhamme Guido van Leeuwen Mees Mosseveld	Erasmus MC
NLHR	Hedvig Marie Egeland Nordeng Saeed Hayati Maren Mackenzie Olson Peter (Petrica-Ioan) Olteanu	Norwegian Linked Health Registry data
SIDIAP	Anna Palomar Cros Laura Granés González Agustina Giuliodori Picco Talita Duarte-Salles	IDIAPJGol
HI-SPEED	Fredrik Nyberg Huiqi Li Rickard Ljung Marcel Ballin Mats Talbäck	Health Impact - Swedish Population Evidence Enabling Data-linkage

CPRD GOLD	Antonella Delmestri	Clinical Practice Research Datalink GOLD
P4-C2-019		
APHM	Vanessa Pauly Laurent Boyer Dorian Grousset	Assistance publique Hôpitaux de Marseille
PGH	Alexandros Rekkas Anastasia Farmaki Achilleas Chytas Pantelis Natsiavas Antonia Sipaki	Papageorgiou General Hospital
SUCD	Dr. Loretta Kiss Dr. Zsolt Bagyura András Sallai Csaba Nemes Orsolya Székely	Semmelweis University Clinical Data
EMDB-ULSEDV	Luís Ruano Ana Pinto Teresa Monjardino Tiago Taveira Gomes	Egas Moniz Health Alliance database - Unidade Local de Saúde de Entre Douro e Vouga
EMDB-ULSGE	Firmino Machado Ana Pinto Teresa Monjardino Tiago Taveira Gomes	Egas Moniz Health Alliance database - Unidade Local de Saúde de Gaia e Espinho
ULSM-RT	Nuno Silva Fernando Montenegro Sá Tiago Taveira-Gomes	Unidade Local de Saúde de Matosinhos Realtime Database
P4-C2-020		
NAJS	Karlo Pintaric Antea Jezidzic Marko Cavlina Anamaria Jurcevic Jakov Vukovic	Croatian National Public Health Information System
EBB	Marek Oja Raivo Kolde Ami Sild	Estonian Biobank
InGef RDB	Raeleesha Norris Alexander Harms Annika Vivirito	InGef – Institute for Applied Health Research Berlin GmbH

BIFAP	Cristina Justo Astorgano Hermenegildo Carlos Martínez-Alcalá García Virginia Arroyo Nebreda Alicia Peñaranda Navazo Miguel Angel Macia Martinez Ana Llorente Garcia	Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público (Pharmacoepidemiological Research Database for Public Health Systems)
PRISIB	Pau Pericas Pulido Joan Vicenç Cladera Salva	Platafoma de Recerca en Informació Sanitària de les Illes Balears
VID	Gabriel Sanfélix Gimeno Celia Robles Cabaniñas Fran Llopis Cardona	Valencia Health System Integrated Dataset

*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® - Population demographics and disease frequency across the DARWIN EU® network

Rationale and background

The DARWIN EU® network consists of heterogeneous real-world data sources with varying population characteristics, which may affect study feasibility, analytical approaches, and the interpretation and generalisability of results. Understanding clinical heterogeneity is important when conducting studies in the network.

Research question and objectives

Research questions

The aim of this study is to characterise the DARWIN EU® network in terms of demographics and common clinical conditions.

Objectives

1. To describe age and sex of the populations under observation in the DARWIN EU® data sources.
2. To develop reusable standardised phenotypes for selected common clinical conditions.
3. To estimate the period prevalence of selected common clinical conditions.
4. To describe the median age at first record and sex distribution of selected common clinical conditions.

Methods

Study design

A descriptive cohort study will be conducted.

The index date will be the start of observation during the study period (1 January–31 December 2023) for Objectives 1 and 3 or the date of first record of the clinical condition for Objective 4. Individuals will be followed until the end of the study period for Objective 3.

Population

Inclusion criteria:

- Under observation at any time during the study period.
- At least 365 days of available history before index date, except for children aged <1 year at the start of observation during the study period and for hospital-based data sources (APHM, PGH, SUCD, EMDB-ULSEDV, EMDB-ULSGE, and ULSM-RT).

Exclusion criteria:

- Missing information on age or sex.

Variables

Outcomes:

- Selected cancers and conditions across diseases of the circulatory, digestive, endocrine, genitourinary, and respiratory systems, haematological, infectious, mental health, musculoskeletal, neurological, and skin conditions.

Relevant covariates:

- Age
- Sex

Data sources

This study will be conducted using routinely collected data from 19 data sources within the DARWIN EU® network, covering 13 European countries. These include: i) four nation-wide primary care data sources; ii) four nation-wide data sources covering primary and secondary outpatient care, as well as hospital inpatient care; iii) three regional data sources covering primary and secondary outpatient care, as well as hospital inpatient care; iv) six hospital-based data sources; v) one regional primary care data source linked to hospital discharge data; and vi) one biobank including primary and secondary outpatient care, as well as hospital inpatient care data. All data were *a priori* mapped to the OMOP CDM.

Study size

No sample size has been calculated. The study aims to describe the entire available eligible population in each data-source irrespective of size.

Statistical analysis

All analyses will be conducted using Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) mapped data. A minimum cell count of 5 will be used when reporting results, with any smaller count reported as "<5".

Objectives 1 and 4:

Age will be summarised using medians and interquartile ranges at the index date, and sex using proportions. The age distribution will be presented as proportions across predefined age categories and stratified by sex.

Objective 3:

Annual period prevalence will be defined as the total number of individuals in observation during the study period with the clinical condition of interest, divided by the entire eligible population in observation during the same period. Binomial 95% confidence intervals will be calculated.

Sensitivity analysis:

Observation period definitions vary across the participating data sources. All study estimates will be re-estimated using a harmonised observation period definition across all data sources, defined from first recorded event to end of data collection or death.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

P4-C3-006/P4-C2-019/P4-C2-020

Study milestones and deliverables	Planned dates*
Final Study Protocol	February 2026
Phenotype Review and Approval (in Batches)	April–May 2026
Creation of Analytical code	May 2026
Execution of Analytical Code on the data	June 2026
Interim Preliminary Results Review (Shiny App)	June 2026
Draft Study Report	August 2026
Final Study Report	October 2026

*Planned dates are dependent on obtaining approvals from the internal review boards of the data sources, as well as on deliverables review cycles.

6. RATIONALE AND BACKGROUND

The DARWIN EU® network comprises multiple real-world data sources that differ in data types, healthcare settings, coding practices, and population coverage. Population characteristics have been described within individual DARWIN EU® studies to support specific research questions. However, these descriptions have typically been study-specific and tailored to particular diseases or treatments, limiting their reuse and comparability across studies. Differences in age distributions, sex composition, and comorbidity profiles across data sources may influence study feasibility, analytical choices, and the interpretation and generalisability of results. It is therefore essential to establish a clear understanding of the populations captured within each data source.

This study represents a foundational step in characterising the DARWIN EU® network by systematically describing and visualising its demographic composition and clinical profile, including the prevalence of selected common conditions across data sources within the network. Through this work, the study will support the interpretation of results and their heterogeneity, facilitate the contextualisation of findings, inform study design and analytical strategies for future research, support feasibility assessments, and evaluations of data fitness for purpose, and contribute to harmonisation of phenotyping across the DARWIN EU® network. Ultimately, by providing a transparent and well-characterised population reference, this work will enhance the utility of DARWIN EU® for regulatory science.

In addition, this study will support the development, documentation, and validation of standardised, fully reusable phenotypes for common clinical conditions used to characterise study populations. These phenotypes will be developed for reuse in future studies within the DARWIN EU® network, thereby promoting consistency, transparency, and efficiency in population characterisation and prevalence estimation across the network.

7. RESEARCH QUESTION AND OBJECTIVES

Research questions

The aim of this study is to characterise the DARWIN EU® network in terms of demographics and common clinical conditions, and to develop, document, and validate reusable standardised phenotypes for the description of disease prevalence and the characterisation of study populations within the network.

Research objectives

The specific objectives of this study are:

1. To describe the demographic characteristics (age and sex) of the populations under observation in the DARWIN EU® data sources and to compare these characteristics with relevant national or regional reference data.
2. To develop reusable standardised phenotypes for selected common clinical conditions.
3. To estimate the period prevalence of selected common clinical conditions within the population under observation.
4. To describe the median age at first record and sex distribution of selected common clinical conditions within the population under observation.

8. RESEARCH METHODS

8.1. Study design

A descriptive cohort study will be conducted using routinely collected health data from 19 data sources from 13 countries across Europe.

The index date will be the start of observation during the study period (1 January–31 December 2023) for Objectives 1 and 3 or the date of first record of the clinical condition for Objective 4.

8.2. Follow-up

Objective 3: Descriptive epidemiology of selected clinical conditions

Follow-up for prevalence estimation will begin on 1 January 2023 (or later if an individual's eligible observation period starts after this date) and will end on 31 December 2023 (or earlier if the observation ends before that date).

Estimating prevalence requires first defining an appropriate denominator population and the contributed observation time. Study participants in the denominator population will begin contributing observation time as described above. For prevalence estimation, data sources will be required to capture the complete interval of interest. For data sources which do not yet provide complete coverage of the interval of interest, an updated Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) release will be required prior to study execution. Individuals will not be required to be present for the entire follow-up period to contribute to the prevalence estimates.

An example of entry into and exit from the denominator population is shown in **Figure 1**. In this example, person ID 1 has the observation period that starts before the beginning of the study period and ends after the end of the study period. Therefore, this individual will contribute the complete study period and will be included in the denominator. Person IDs 2, 3, and 4 either start the observation period after the beginning of the study period or end the observation period before the end of the study period. These three individuals will contribute to the denominator cohort only for the time they were in observation within the study period. Lastly, person ID 5 has two observation periods in the data source, neither of which covers the complete study period. Person ID 5 contributes to the denominator cohort from the beginning of the

study period until the end of their first observation period and from the start of the second observation period until the end of the study period.

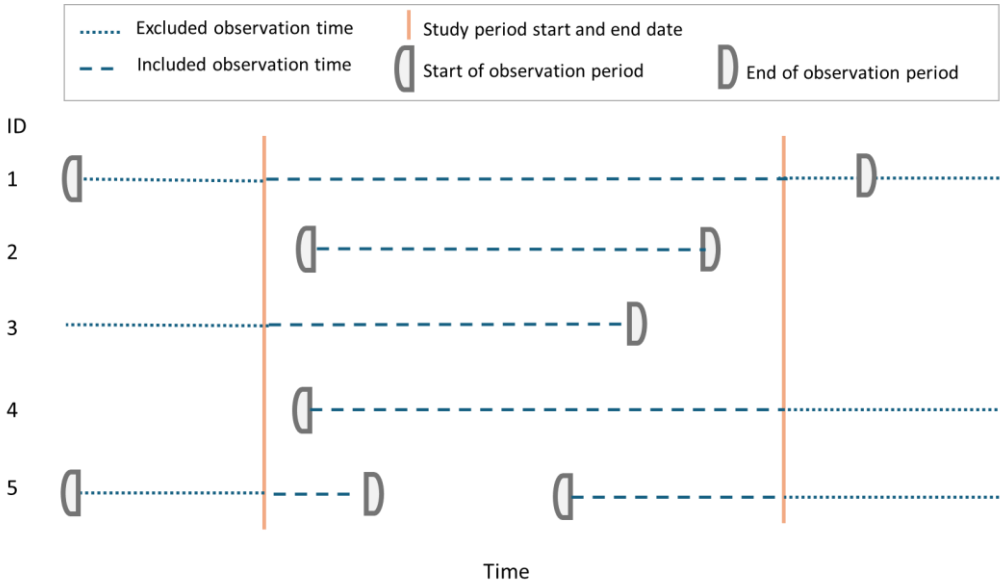

Figure 1. Included observation time for the denominator population for descriptive epidemiology of selected clinical conditions (Objective 3).

8.3. Study population with inclusion and exclusion criteria

Objectives 1 and 3 - General population cohort

Inclusion criteria

- Under observation at any time during the study period (1 January 2023 to 31 December 2023).
- At least 365 days of available history before index date, except for children aged <1 year at the start of observation during the study period and for hospital-based data sources (APHM, PGH, SUCD, EMDB-ULSEDV, EMDB-ULSGE, and ULSM-RT).

Exclusion criteria

- Missing information on age or sex.

Objective 4 - Clinical condition cohorts

Inclusion criteria

- Under observation at any time during the study period (1 January 2023 to 31 December 2023).
- At least 365 days of available history before index date except for children aged <1 year at the start of observation during the study period and for hospital-based data sources (APHM, PGH, SUCD, EMDB-ULSEDV, EMDB-ULSGE, and ULSM-RT).
- Has or has ever had a clinical condition of interest (listed in **Section 8.6.1.**)

Exclusion criteria

- Missing information on age or sex.

8.4. Study setting and data sources

This study will be conducted using routinely collected data from 19 data sources within the DARWIN EU® network, covering 13 European countries. These include: i) four nation-wide primary care data sources; ii)

four nation-wide data sources covering primary and secondary outpatient care, as well as hospital inpatient care; iii) three regional data sources covering primary and secondary outpatient care, as well as hospital inpatient care; iv) six hospital-based data sources; v) one regional primary care data source linked to hospital discharge data; and vi) one biobank including primary and secondary outpatient care, as well as hospital inpatient care data. All data were *a priori* mapped to the OMOP CDM.

Table 1. Data sources grouped according to healthcare setting or type of data.

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals	Calendar period covered by each data source	Contributing to	Observation period definition
Primary care							
Belgium	IQVIA LPD Belgium	Primary care	EHR	189k	2005–07/05/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit
Germany	IQVIA DA Germany	Primary care	EHR	4.72M	1989–17/10/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit
The Netherlands	IPCI	Primary care	EHR	1.32M	2006–22/10/2025	Objectives 1, 3, 4	From registration to deregistration or end of data source coverage
The United Kingdom	CPRD GOLD	Primary care	EHR	2.63M	1987–15/08/2025	Objectives 1, 3, 4	From registration to deregistration or end of data source coverage
Primary, secondary, and hospital inpatient care							
Norway	NLHR	Primary, secondary, and hospital inpatient care	Registry	5.55M	2008–31/12/2023	Objectives 1, 3, 4	Spanning birth, start of data source coverage, or registration to death, end of data source coverage, or deregistration
Spain	BIFAP	Primary, secondary, and hospital inpatient care	EHR	23.2M	2001–31/10/2025	Objectives 1, 3, 4	From first recorded event or registration to death, deregistration, or end of

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals	Calendar period covered by each data source	Contributing to	Observation period definition
							data source coverage
Spain	PRISIB	Primary, secondary, and hospital inpatient care	EHR	1.79M	2008–01/12/2023	Objectives 1, 3, 4	From first recorded event or registration to death, deregistration, or end of data source coverage
Spain	SIDIAP	Primary care (linked to hospital discharged records)	EHR	6.07M	2006–26/11/2025	Objectives 1, 3, 4	From registration to deregistration or end of data source coverage
Spain	VID	Primary, secondary, and hospital inpatient care	EHR	5.58M	2009–01/10/2025	Objectives 1, 3, 4	From registration to deregistration or end of data source coverage
Sweden	HI-SPEED	Primary, secondary, and hospital inpatient care	Registry	10.6M	2015–31/08/2025	Objectives 1, 3, 4	Spanning birth, start of data source coverage, or registration to death, end of data source coverage, or deregistration
Secondary and hospital inpatient care							
France	APHM	Secondary and hospital inpatient care	EHR	250k	2014–11/01/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit
Greece	PGH	Secondary and hospital inpatient care	EHR	55.7k	1999–03/05/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit
Hungary	SUCD	Secondary and hospital	EHR	227k	2010–28/07/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals	Calendar period covered by each data source	Contributing to	Observation period definition
		inpatient care					
Portugal	EMDB-ULSEDV	Secondary and hospital inpatient care	EHR	101k	1999–20/10/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit
Portugal	EMDB-ULSGE	Secondary and hospital inpatient care	EHR	130k	2004–01/11/2023	Objectives 1, 3, 4	From first to last recorded clinical event/visit
Portugal	ULSM-RT	Secondary and hospital inpatient care	EHR	76k	1998–22/10/2025	Objectives 1, 3, 4	From first to last recorded clinical event/visit
Biobank							
Estonia	EBB	Primary, secondary, and hospital inpatient care	Biobank	212k	2004–27/02/2025	Objectives 1, 3, 4	Spanning birth, start of data source coverage, or registration to death, end of data source coverage, or deregistration
Claims data							
Croatia	NAJS	Primary, secondary, and hospital inpatient care	Claims	4.3M	1998–14/08/2025	Objectives 1, 3, 4	From registration to deregistration or end of data source coverage
Germany	InGef RDB	Primary, secondary, and hospital inpatient care	Claims	7.43M	2014–14/11/2025	Objectives 1, 3, 4	From registration to deregistration or end of data source coverage

APHM = Assistance publique Hôpitaux de Marseille, BIFAP = Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público, CPRD GOLD = Clinical Practice Research Datalink GOLD, EBB = Estonian Biobank, EHR = Electronic Health Record, EMDB-ULSEDV = Egas Moniz Health Alliance database - Unidade Local de Saúde de Entre Douro e Vouga, EMDB-ULSGE = Egas Moniz Health Alliance database - Unidade Local de Saúde de Gaia e Espinho, HI-SPEED = Health Impact – Swedish Population Evidence Enabling Data-linkage, InGef RDB = InGef Research Database, IPCI = Integrated Primary Care Information, IQVIA DA Germany = IQVIA Disease Analyzer Germany, IQVIA LPD Belgium = IQVIA Longitudinal Patient Database Belgium, M= Million, NAJS =

Croatian National Public Health Information System, NLHR = Norwegian Linked Health Registry data, PGH = Papageorgiou General Hospital, PRISIB = Plataforma de Recerca en Informació Sanitària de les Illes Balears, SIDIAP = The Information System for the Development of Research in Primary Care, SUCD = Semmelweis University Clinical Data, ULSM-RT = Unidade Local de Saúde de Matosinhos Realtime Database, VID = Valencia Health System Integrated Dataset

Data sources selection

Data sources were included using an inclusive approach to enable a comprehensive characterisation of the DARWIN EU® network across different European regions (**Annex II**). No specific selection criteria were applied beyond the exclusion of highly specialised data sources that would not be appropriate for general population characterisation.

8.5. Study period

The study period will be defined as from 1 January to 31 December 2023. This study period was selected to ensure recent, complete, and accurate data capture across the entire interval for the estimation of period prevalence across all participating data sources.

8.6. Variables

8.6.1. Outcomes

Objective 3: Descriptive epidemiology of selected clinical conditions

The outcomes for this objective are summarised in **Table 2.**

All outcomes will be defined using a prevalent disease definition, in line with the objective of characterising study populations. Accordingly, for each condition, disease onset will be considered to be at individual's first record of the condition, including diagnoses prior to the study period. The disease duration will be considered to extend all the way until the end of the individual's observation period. This approach will be applied only to chronic and lifelong conditions. For conditions that are acute or potentially curable, an adapted lookback window will be applied. An individual would be considered as having the conditions of interest (i.e., being part of the prevalence numerator) if they had any record of the conditions within the lookback window or afterwards, during follow-up.

Table 2 describes how conditions are categorised and what would be the lookback window.

Table 2. Outcomes and lookback window for Objective 3.

Outcome	Lookback window
Cancers	
Colorectal and anus cancer	5 years
Breast cancer	5 years
Prostate cancer	5 years
Lung cancer	5 years
Diseases of the circulatory system	
Atrial fibrillation	Entire available time
Cardiac arrhythmia	Entire available time
Coronary heart disease (composite phenotype including coronary heart disease not otherwise specified, myocardial infarction, stable angina, or unstable angina)	Entire available time (1 year for myocardial infarction)
Coronary heart diseases not otherwise specified	Entire available time
Myocardial infarction	1 year
Stable angina	Entire available time

Outcome	Lookback window
Unstable angina	Entire available time
Heart failure	Entire available time
Hypertension	Entire available time
Peripheral arterial disease	Entire available time
Stroke (composite phenotype, including ischaemic stroke, haemorrhagic stroke, or transient ischaemic attack)	1 year
Haemorrhagic stroke	1 year
Ischaemic stroke	1 year
Transient ischaemic attack	1 year
Non-fatal major adverse cardiovascular events (composite phenotype including myocardial infarction or stroke)	1 year
Diseases of the digestive system	
Cholecystitis	1 year
Chronic liver disease (excluding chronic viral hepatitis)	Entire available time
Inflammatory bowel disease (composite phenotype, including Crohn's disease or ulcerative colitis)	Entire available time
Crohn's disease	Entire available time
Ulcerative colitis	Entire available time
Gastro-oesophageal reflux disease	Entire available time
Metabolic dysfunction-associated steatohepatitis	Entire available time
Diseases of the endocrine system	
Hypercholesterolaemia	Entire available time
Hyperlipidaemia	Entire available time
Hypothyroidism	Entire available time
Obesity	Entire available time
Type 1 diabetes mellitus	Entire available time
Type 2 diabetes mellitus	Entire available time
Diseases of the genitourinary system	
Acute kidney injury	1 year
Chronic kidney disease	Entire available time
Benign prostatic hyperplasia	Entire available time
Diseases of the respiratory system	
Asthma	Entire available time
Chronic obstructive pulmonary disease	Entire available time
Anaemia (limited to nutritional and metabolic anaemias, including iron, vitamin B12, and folate deficiency)	1 year
Chronic infectious diseases	
Chronic viral hepatitis	Entire available time
Human immunodeficiency virus	Entire available time

Outcome	Lookback window
Mental health disorders	
Alcohol use-related disorders	5 years
Anxiety disorders	5 years
Bipolar affective disorder and mania	5 years
Depression	5 years
Schizophrenia, schizotypal, and delusional disorders	5 years
Musculoskeletal conditions	
Osteoporosis	Entire available time
Rheumatoid arthritis	Entire available time
Neurological conditions	
Alzheimer's dementia	Entire available time
Dementia	Entire available time
Epilepsy	Entire available time
Migraine	Entire available time
Parkinson's disease	Entire available time
Skin conditions	
Acne	5 years
Atopic dermatitis	Entire available time

Most outcomes will be defined using condition and/or observation codes. For more complex phenotypes, such as type 1 diabetes mellitus, phenotyping algorithms will incorporate additional elements, for example excluding individuals treated with oral antidiabetic medications without evidence of insulin use. Other potentially complex phenotypes may include type 2 diabetes mellitus, where the algorithm may combine condition codes with antidiabetic treatment codes (oral and/or insulin) and apply exclusions for alternative indications, such as polycystic ovary syndrome. Chronic kidney disease may require combining condition codes with estimated glomerular filtration rate measurement data. Obesity may also require combining condition codes with body mass index measurements. Unstable angina may require additional refinement to distinguish it from stable angina and acute myocardial infarction, for example by excluding individuals with concurrent myocardial infarction codes during the same episode of care or within a predefined time window in order to improve specificity. The development and validation of these algorithms for each selected clinical condition will be undertaken as part of the study execution.

The preliminary concept sets used for the identification of outcomes are described in **Annex IV** and will be refined during study execution in accordance with the DARWIN EU® phenotyping standard processes. These processes include review of code lists by clinical experts and post-execution review of phenotype performance across participating data sources using the *PhenotypeR* R package, which provides insights into cohort characteristics, record counts, and potential index event misclassification, supporting quality assurance of the developed phenotypes. Phenotypes will be reviewed and approved by the European Medicines Agency (EMA) team prior to study execution.[1]

Furthermore, phenotypes will be adapted (if needed) after a first run in the data sources if the clinical characteristics show a possible misclassification.

For hospital data sources, the availability of information to distinguish primary conditions (reason for admission) from secondary conditions will be explored. Any characterisation based on this distinction will be contingent on data availability within each participating data source.

8.6.2. Covariates, including confounders, effect modifiers, and other variables

Objectives 1 and 4: Characterisation of the population under observation

The following covariates will be used to characterise the study population at the index date [0, 0], defined as the date on which individuals first contribute observation time to the denominator during the study period for Objective 1 and as the date of the first record of the clinical condition for Objective 4:

- Age
- Sex

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. In addition, we will use already collected available data to estimate the period prevalence of selected clinical conditions. Thus, the sample size is driven by the availability of data for individuals with the clinical conditions of interest.

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 individuals' 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 III**), the final package will be released in a version-controlled study repository for execution against all the participating data sources.

8.8.2. Data privacy protection

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 a secure online repository of the Data Transfer Zone (DTZ). All results will be locked and timestamped for reproducibility and transparency. The study results from all data sources will be checked, after which they are made available to the team, and the Study Dissemination Phase can start. 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. 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

Objectives 1 and 4

Individuals' characterisation will be done using OMOP CDM mapped data using the *CohortCharacteristics* R package, developed by DARWIN EU®.[2]

For Objective 1, characterisation in terms of age and sex will be provided for the population under observation during the study period. Age and sex will be assessed at the index date, defined as the date on which individuals first contribute observation time to the denominator during the study period. Age will be summarised using the median and interquartile range, and sex will be summarised using proportions. In addition, the age distribution will be presented as proportions within predefined 5-age categories and stratified by sex.

For Objective 4, characterisation will be conducted among individuals with a selected clinical condition. Age will be assessed at the date of the first ever record of the clinical condition within the data source. Age at first record will be summarised using the median and interquartile range, and the sex distribution will be summarised using proportions.

Objective 3

Prevalence of selected clinical conditions will be calculated based on OMOP CDM mapped data using the *IncidencePrevalence* R package, developed by DARWIN EU®.[3]

Prevalence of selected clinical conditions will be estimated as annual period prevalence, defined as the total number of individuals in observation during the study period with the clinical condition of interest, divided by the population in observation during the same period. Binomial 95% confidence intervals will be calculated for all prevalence estimates. For chronic conditions, disease onset will be defined as the individual's first recorded diagnosis of the condition, including records prior to the study period. For acute or potentially curable conditions, an adapted look-back window will be applied, as specified in **Table 2**.

Sensitivity analysis

Observation period definitions vary across the participating data sources. Specifically, four definitions are identified:

1. Observation time defined from registration to deregistration or end of data source coverage: IPCI, NAJS, InGef RDB, SIDIAP, VID, and CPRD GOLD.
2. Observation time defined from first to last recorded clinical event/visit: APHM, IQVIA LPD Belgium, IQVIA DA Germany, PGH, EMDB-ULSEDV, EMDB-ULSGE, SUCD, and ULSM-RT.
3. Observation time spanning birth, start of data source coverage, or registration to death, end of data source coverage, or deregistration: EBB, NLHR, and HI-SPEED.
4. Observation time defined from first recorded event or registration to death, deregistration, or end of data source coverage: BIFAP and PRISIB.

These differences in observation period definitions may influence the size and composition of the population considered under observation and, consequently, both the demographic characterisation and the estimated period prevalence. In particular, observation period definitions based on first and last recorded events or visits, most commonly used in hospital-based data sources, may be more restrictive, potentially underestimating follow-up time and reducing the population under observation. However, as most included data sources contain data extending into 2025, nearly two additional years of data are available after the end of the study period, which may mitigate potential truncation effects towards the end of the study period.

To assess the robustness of the findings, a sensitivity analysis will be conducted in which all study estimates, including demographic characterisation (Objective 1), period prevalence (Objective 3), and median age at first record and sex distribution (Objective 4), will be re-estimated using a harmonised observation period definition applied across all participating data sources. This harmonised definition will align with the observation time defined from first recorded event to death, deregistration, or end of data source coverage.

For this sensitivity analysis, we will first characterise the observation periods under both the original and harmonised definitions by calculating the number of records, number of subjects, number of records per person, and the length of observation periods, as well as estimating the number of person-years under observation per calendar year and the median age under observation over time (2019–2023). This characterisation will help describe the differences between observation period definitions.

Prevalence estimates from the sensitivity analysis will be compared with those from the primary analysis to evaluate the impact of differing observation period definitions on the study findings.

Description of sensitivity analyses are presented by means of **Table 3**.

Table 3. Sensitivity analyses – rationale, strengths, and limitations.

	What is being varied? How?	Why? (What do you expect to learn?)	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
Objectives 1, 3, 4: definition of observation period	Primary analysis uses the original observation period definition specific to each data source. Sensitivity analysis applies a harmonised observation period definition across all data sources: from first recorded event to end of data source coverage or death.	Differences in observation period definitions across data sources may influence the size and composition of the population considered under observation and, consequently, affect demographic characterisation, period prevalence estimates, and age at first record of a clinical condition. This sensitivity analysis aims to assess the robustness of the results to these differences.	- Reduces the risk of underestimating follow-up time and restricting the population under observation, particularly in hospital-based data sources where observation periods are often defined by first and last recorded events or visits. - Improves comparability of results across data sources by applying a consistent definition of observation period.	- Results from the sensitivity analysis may be less aligned with how specific data sources are typically used or interpreted in standalone analyses.

8.8.4. Output

Output will include the following:

A PDF report including an executive summary and the following tables and figures. All tables and figures have been generated using synthetic data and do not represent real results.

Table 1. Number of individuals under observation during the study period, by data source.

Eligibility criterion	Number of subjects	
	Data source 1	Data source X
Initial population	213,953	10,678
Under observation in the study period	205,771	10,257
Prior observation requirement: 365 days	27,912	1,416
Sex requirement: Both	27,912	1,416
Age requirement: 0 to 150	27,912	1,416

Objective 1: Characterisation of the population under observation

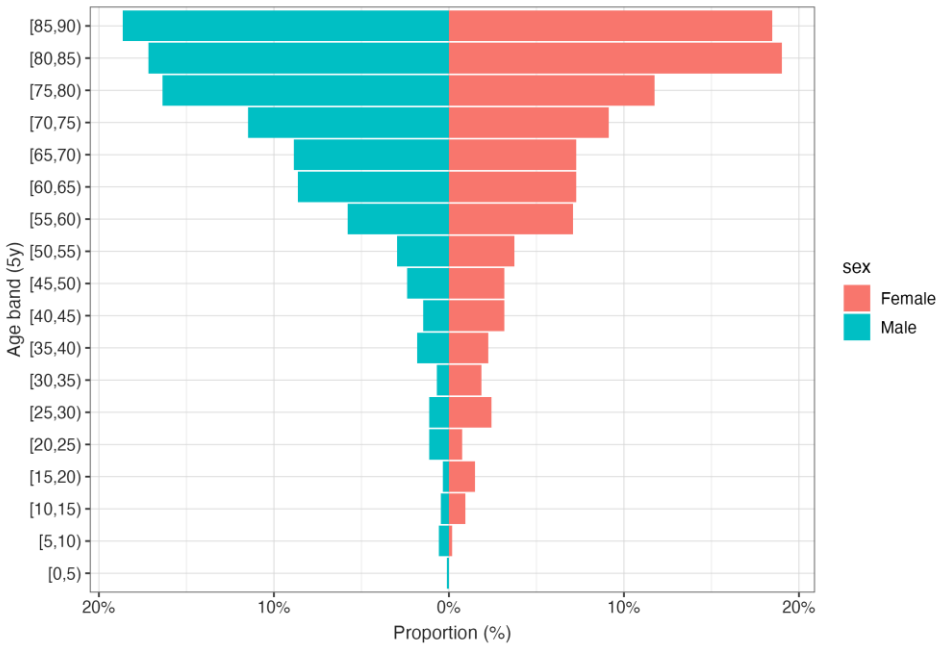

Figure 1a–1x. Population pyramid displaying the age distribution of the study population stratified by sex at the index date.

One figure will be generated for each data source.

Table S1a–1x. Demographic characteristics of the study population by data source.

Variable name	Data source 1	Data source X
Number of subjects	27,912	1,416
Age (Median [Q25, Q75])	75 [61, 83]	75 [61, 83]
Female (N (%))	11,479 (41.13%)	536 (37.85%)

Tables will be generated for each group of data sources, as described in **Section 8.9**.

Objective 3: Descriptive epidemiology of selected clinical conditions

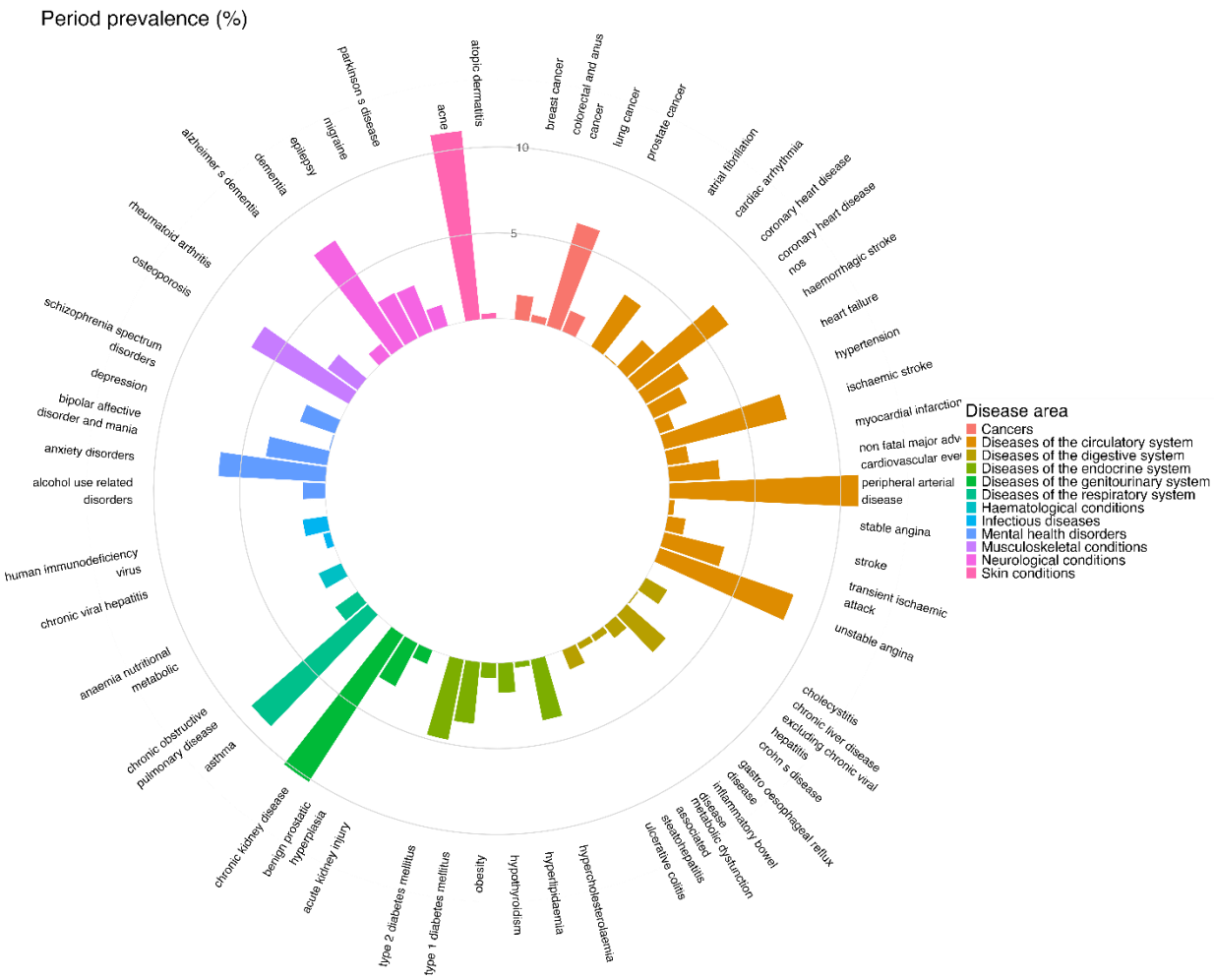

Figure 2a–2x. Radial graphs of period prevalence of selected clinical conditions, with conditions colour-coded by disease area.

One figure will be generated for each data source.

Table S2a–2x. Period prevalence of selected clinical conditions by data source.

Outcome cohort name	Data source 1			Data source X		
	Denominator count	Outcome count	Prevalence (95% CI) (%)	Denominator count	Outcome count	Prevalence (95% CI) (%)
Cancers						
breast cancer	27912	417	1.49 (1.36, 1.64)	1416	0	0 (0, 0.27)
colorectal and anus cancer	27912	134	0.48 (0.41, 0.57)	1416	0	0 (0, 0.27)
lung cancer	27912	1754	6.28 (6, 6.58)	1416	0	0 (0, 0.27)
prostate cancer	27912	376	1.35 (1.22, 1.49)	1416	0	0 (0, 0.27)
Diseases of the circulatory system						
atrial fibrillation	27912	1019	3.65 (3.44, 3.88)	1416	0	0 (0, 0.27)
cardiac arrhythmia	27912	33	0.12 (0.08, 0.17)	1416	0	0 (0, 0.27)
coronary heart disease	27912	1844	6.61 (6.32, 6.9)	1416	100	7.06 (5.84, 8.52)
coronary heart disease not otherwise specified	27912	615	2.2 (2.04, 2.38)	1416	0	0 (0, 0.27)
haemorrhagic stroke	27912	369	1.32 (1.2, 1.46)	1416	0	0 (0, 0.27)
heart failure	27912	265	0.95 (0.84, 1.07)	1416	0	0 (0, 0.27)
hypertension	27912	2063	7.39 (7.09, 7.7)	1416	92	6.5 (5.33, 7.9)
ischaemic stroke	27912	840	3.01 (2.82, 3.22)	1416	0	0 (0, 0.27)
non-fatal major adverse cardiovascular events	27912	819	2.93 (2.74, 3.14)	1416	49	3.46 (2.63, 4.54)
myocardial infarction	27912	3088	11.06 (10.7, 11.44)	1416	167	11.79 (10.22, 13.58)
peripheral arterial disease	27912	103	0.37 (0.3, 0.45)	1416	0	0 (0, 0.27)
stable angina	27912	313	1.12 (1, 1.25)	1416	0	0 (0, 0.27)
stroke	27912	1019	3.65 (3.44, 3.88)	1416	0	0 (0, 0.27)
transient ischaemic attack	27912	2317	8.3 (7.98, 8.63)	1416	0	0 (0, 0.27)
unstable angina	27912	417	1.49 (1.36, 1.64)	1416	0	0 (0, 0.27)
Diseases of the digestive system						
cholecystitis	27912	39	0.14 (0.1, 0.19)	1416	0	0 (0, 0.27)
chronic liver disease excluding chronic viral hepatitis	27912	858	3.07 (2.88, 3.28)	1416	0	0 (0, 0.27)
Crohn's disease	27912	134	0.48 (0.41, 0.57)	1416	0	0 (0, 0.27)
gastro oesophageal reflux disease	27912	123	0.44 (0.37, 0.52)	1416	0	0 (0, 0.27)
inflammatory bowel disease	27912	277	0.99 (0.88, 1.12)	1416	0	0 (0, 0.27)
metabolic dysfunction associated steatohepatitis	27912	352	1.26 (1.14, 1.4)	1416	0	0 (0, 0.27)
ulcerative colitis	27912	1019	3.65 (3.44, 3.88)	1416	0	0 (0, 0.27)

Outcome cohort name	Data source 1			Data source X		
	Denominator count	Outcome count	Prevalence (95% CI) (%)	Denominator count	Outcome count	Prevalence (95% CI) (%)
Diseases of the endocrine system						
hypercholesterolaemia	27912	107	0.38 (0.32, 0.46)	1416	0	0 (0, 0.27)
hyperlipidaemia	27912	498	1.78 (1.64, 1.95)	1416	0	0 (0, 0.27)
hypothyroidism	27912	261	0.94 (0.83, 1.05)	1416	0	0 (0, 0.27)
obesity	27912	1021	3.66 (3.44, 3.89)	1416	0	0 (0, 0.27)
type 1 diabetes mellitus	27912	1338	4.79 (4.55, 5.05)	1416	0	0 (0, 0.27)
type 2 diabetes mellitus	27912	254	0.91 (0.8, 1.03)	1416	0	0 (0, 0.27)
Diseases of the genitourinary system						
acute kidney injury	27912	2849	10.21 (9.86, 10.57)	1416	127	8.97 (7.59, 10.57)
benign prostatic hyperplasia	27912	2519	9.03 (8.69, 9.37)	1416	0	0 (0, 0.27)
chronic kidney disease	27912	800	2.87 (2.68, 3.07)	1416	0	0 (0, 0.27)
Diseases of the respiratory system						
asthma	27912	459	1.64 (1.5, 1.8)	1416	0	0 (0, 0.27)
chronic obstructive pulmonary disease	27912	418	1.5 (1.36, 1.65)	1416	0	0 (0, 0.27)
Haematological conditions						
anaemia nutritional metabolic	27912	134	0.48 (0.41, 0.57)	1416	0	0 (0, 0.27)
Infectious diseases						
chronic viral hepatitis	27912	417	1.49 (1.36, 1.64)	1416	0	0 (0, 0.27)
human immunodeficiency virus	27912	376	1.35 (1.22, 1.49)	1416	0	0 (0, 0.27)
Mental health disorders						
alcohol use related disorders	27912	1754	6.28 (6, 6.58)	1416	0	0 (0, 0.27)
anxiety disorders	27912	1019	3.65 (3.44, 3.88)	1416	0	0 (0, 0.27)
bipolar affective disorder and mania	27912	33	0.12 (0.08, 0.17)	1416	0	0 (0, 0.27)
depression	27912	615	2.2 (2.04, 2.38)	1416	0	0 (0, 0.27)
schizophrenia spectrum disorders	27912	1844	6.61 (6.32, 6.9)	1416	100	7.06 (5.84, 8.52)
Musculoskeletal conditions						
osteoporosis	27912	615	2.2 (2.04, 2.38)	1416	0	0 (0, 0.27)
rheumatoid arthritis	27912	265	0.95 (0.84, 1.07)	1416	0	0 (0, 0.27)
Neurological conditions						
Alzheimer's dementia	27912	2063	7.39 (7.09, 7.7)	1416	92	6.5 (5.33, 7.9)
dementia	27912	840	3.01 (2.82, 3.22)	1416	0	0 (0, 0.27)

Outcome cohort name	Data source 1			Data source X		
	Denominator count	Outcome count	Prevalence (95% CI) (%)	Denominator count	Outcome count	Prevalence (95% CI) (%)
epilepsy	27912	819	2.93 (2.74, 3.14)	1416	49	3.46 (2.63, 4.54)
migraine	27912	369	1.32 (1.2, 1.46)	1416	0	0 (0, 0.27)
Parkinson's disease	27912	3088	11.06 (10.7, 11.44)	1416	167	11.79 (10.22, 13.58)
Skin conditions						
acne	27912	103	0.37 (0.3, 0.45)	1416	0	0 (0, 0.27)
atopic dermatitis	27912	313	1.12 (1, 1.25)	1416	0	0 (0, 0.27)

CI = Confidence Interval
Tables will be generated for each group of data sources, as described in **Section 8.9**.

Objective 4: Characterisation of median age of first record and sex distribution of selected clinical conditions

Figure 3a–3x. Radial graphs of the median age and interquartile rage at first record of selected clinical conditions, with conditions colour-coded by disease area.

One figure will be generated for each data source.

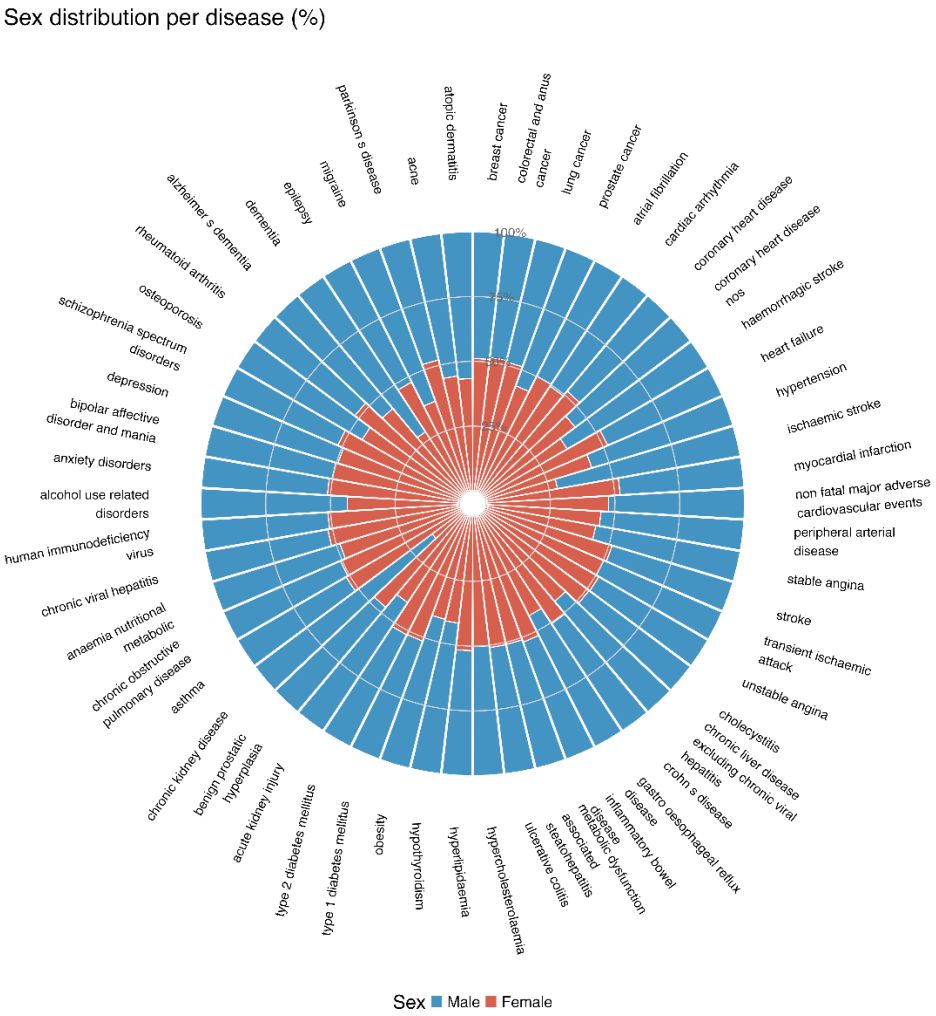

Figure 4a–4x. Radial graphs of the sex distribution of selected clinical conditions.

One figure will be generated for each data source.

Table S3a–3x. Median age and interquartile range at first diagnosis.

	Age (Median [Q25, Q75])	
Outcome cohort name	**Data source 1**	**Data source X**
Cancers		
breast cancer	47.0 (23.0, 72.0)	–
colorectal and anus cancer	48.0 (23.0, 71.0)	–
lung cancer	45.0 (21.0, 70.0)	–
prostate cancer	82.0 (72.0, 90.0)	–
Diseases of the circulatory system		
atrial fibrillation	62.0 (41.0, 80.0)	–
cardiac arrhythmia	53.0 (29.0, 76.0)	–
coronary heart disease	81.0 (72.0, 88.0)	82.0 (71.0, 88.0)
coronary heart disease not otherwise specified	50.0 (24.0, 73.0)	–

	Age (Median [Q25, Q75])	
Outcome cohort name	**Data source 1**	**Data source X**
haemorrhagic stroke	45.0 (20.0, 70.5)	–
heart failure	82.0 (71.0, 90.0)	–
hypertension	75.0 (62.0, 84.0)	75.0 (62.0, 84.0)
ischaemic stroke	71.0 (52.0, 85.0)	–
non-fatal major adverse cardiovascular events	62.0 (32.0, 80.0)	61.0 (29.0, 83.5)
myocardial infarction	81.0 (69.0, 88.0)	81.0 (69.0, 88.0)
peripheral arterial disease	77.0 (59.0, 87.5)	–
stable angina	46.0 (21.0, 71.0)	–
stroke	62.0 (41.0, 80.0)	–
transient ischaemic attack	46.0 (22.0, 70.0)	–
unstable angina	47.0 (23.0, 72.0)	–
Diseases of the digestive system		
cholecystitis	58.5 (31.0, 79.8)	–
chronic liver disease excluding chronic viral hepatitis	46.0 (22.0, 70.0)	–
Crohn's disease	48.0 (23.0, 71.0)	–
gastro oesophageal reflux disease	47.0 (22.5, 73.0)	–
inflammatory bowel disease	82.0 (72.0, 90.0)	–
metabolic dysfunction associated steatohepatitis	45.0 (21.0, 70.0)	–
ulcerative colitis	62.0 (41.0, 80.0)	–
Diseases of the endocrine system		
hypercholesterolaemia	45.0 (21.0, 70.0)	–
hyperlipidaemia	82.0 (71.0, 90.0)	–
hypothyroidism	83.0 (72.0, 90.0)	–
obesity	46.0 (22.0, 71.0)	–
type 1 diabetes mellitus	45.0 (21.0, 70.0)	–
type 2 diabetes mellitus	84.0 (71.8, 90.0)	–
Diseases of the genitourinary system		
acute kidney injury	79.0 (66.0, 87.0)	77.5 (67.0, 86.0)
benign prostatic hyperplasia	46.0 (22.0, 71.0)	–
chronic kidney disease	75.0 (55.0, 87.0)	–
Diseases of the respiratory system		
asthma	65.0 (45.2, 82.0)	–
chronic obstructive pulmonary disease	65.0 (42.0, 81.0)	–
Haematological conditions		
anaemia nutritional metabolic	48.0 (23.0, 71.0)	–
Infectious diseases		
chronic viral hepatitis	47.0 (23.0, 72.0)	–

Outcome cohort name	Age (Median [Q25, Q75])	
	Data source 1	Data source X
human immunodeficiency virus	82.0 (72.0, 90.0)	–
Mental health disorders		
alcohol use related disorders	45.0 (21.0, 70.0)	–
anxiety disorders	62.0 (41.0, 80.0)	–
bipolar affective disorder and mania	53.0 (29.0, 76.0)	–
depression	50.0 (24.0, 73.0)	–
schizophrenia spectrum disorders	81.0 (72.0, 88.0)	82.0 (71.0, 88.0)
Musculoskeletal conditions		
osteoporosis	50.0 (24.0, 73.0)	–
rheumatoid arthritis	82.0 (71.0, 90.0)	–
Neurological conditions		
Alzheimer's dementia	75.0 (62.0, 84.0)	75.0 (62.0, 84.0)
dementia	71.0 (52.0, 85.0)	–
epilepsy	62.0 (32.0, 80.0)	61.0 (29.0, 83.5)
migraine	45.0 (20.0, 70.5)	–
Parkinson's disease	81.0 (69.0, 88.0)	81.0 (69.0, 88.0)
Skin conditions		
acne	77.0 (59.0, 87.5)	–
atopic dermatitis	46.0 (21.0, 71.0)	–

Tables will be generated for each group of data sources, as described in **Section 8.9**.

Sensitivity analyses

Table S4a–4x. Characterisation of the observation periods under both the original and harmonised definitions.

Observation period definition	Variable name	Data source 1	Data source X
Original	**Number subjects (N)**	213,953	10,678
	Length of observation time (days) Median [Q25, Q75]	3,312 [2,920, 3,340]	3,319 [2,920, 3,340]
Harmonised	**Number subjects (N)**	186,139	9,290
	Length of observation time (days) Median [Q25, Q75]	3,312 [2,920, 3,340]	3,319 [2,920, 3,340]

Tables will be generated for each group of data sources, as described in **Section 8.9**.

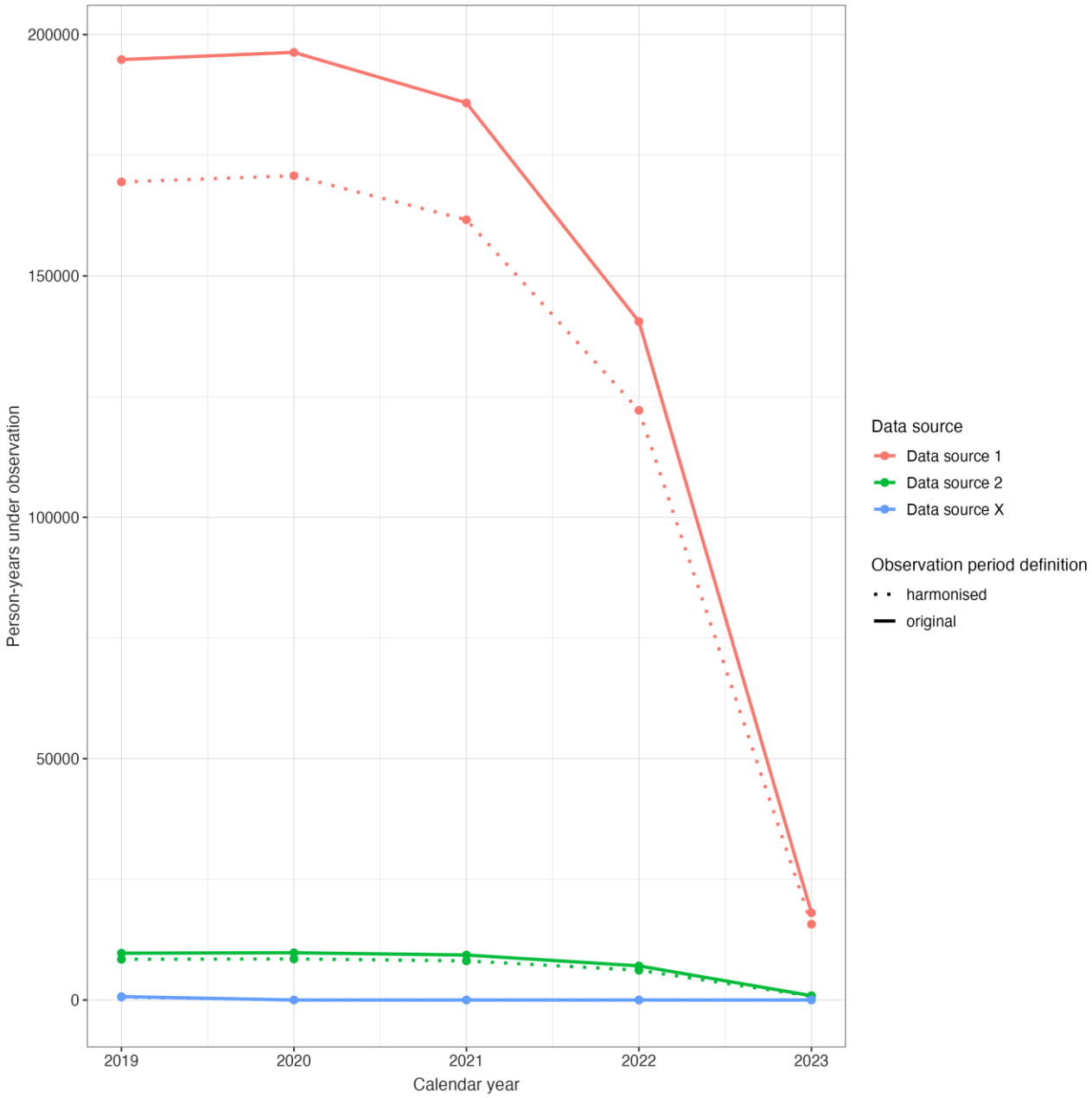

Figure S1a–1x. Number of person-years under observation per year (2019–2023), by data source and observation period definition (original and harmonised).

Figures will be generated for each group of data sources, as described in **Section 8.9**.

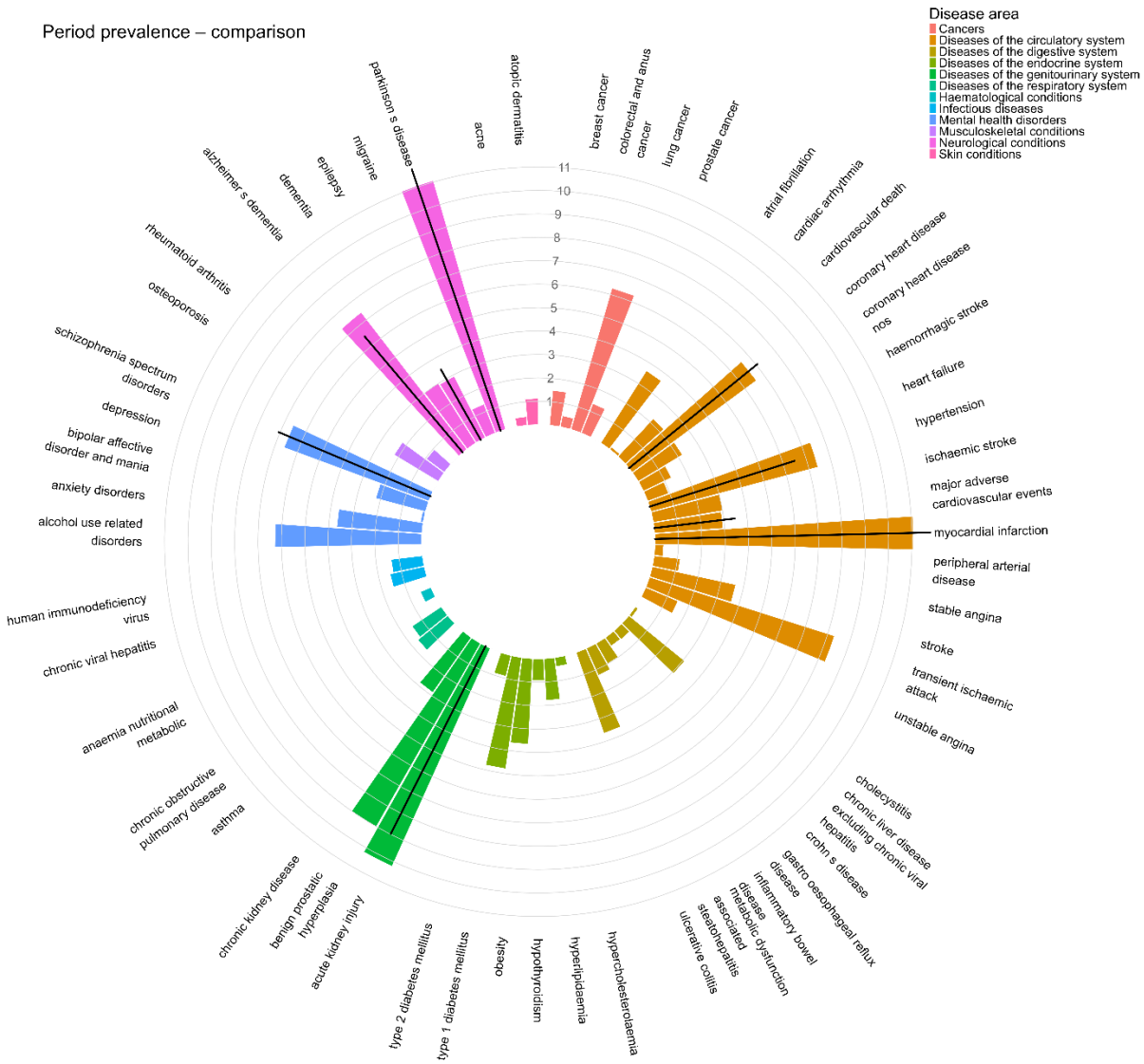

Figure S2. Radial graphs of period prevalence of selected clinical conditions by observation period definition (original and harmonised), with conditions colour-coded by disease area.

The coloured bars represent estimates calculated using the original observation period definition, while the black line indicates period prevalence estimated using the harmonised observation period definition. In this illustrative figure, estimates based on the harmonised observation period are shown for only some conditions; in the final report, harmonised estimates will be generated and displayed for all conditions.
One figure will be generated for each data source.

In addition, phenotyping algorithms will be documented for each clinical condition, including the corresponding code lists and cohort definitions. These materials will be stored and made available to support transparency and reuse in future studies within the DARWIN EU® network for study population characterisation and prevalence estimation.

8.9. Evidence synthesis

Results from the analyses described in **Section 8.8.** will be presented separately for each data source. No meta-analysis will be conducted. To support interpretation, results will be reported by grouping data sources according to healthcare setting and type of data: i) primary care; ii) primary, secondary, and hospital inpatient care; iii) secondary and hospital inpatient care; iv) biobank; and v) claims data.

For Objective 1, demographic characteristics (age and sex) of the populations under observation in the DARWIN EU® data sources will be compared with relevant national or regional reference population statistics. These comparisons are intended to contextualise observed population structures and support interpretation of results. A preliminary list of external reference population statistics to be used for each data source is provided in **Table 4.** The most recent available reference data overlapping the study period will be used where possible.

Table 4. External reference population statistics used for demographic contextualisation by country (Objective 1).

Country	External reference population statistics
Belgium	Belgian national population statistics: https://statbel.fgov.be/en/themes/population
Croatia	Croatian Bureau of Statistics (DZS – population statistics): https://web.dzs.hr/PxWeb/pxweb/en/Stanovni%C5%A1tvo/
Estonia	Statistics Estonia: https://andmed.stat.ee/en/stat Estonian National Institute for Health Development: https://statistika.tai.ee/pxweb/en/Andmebaas/#/
France	INSEE (French National Institute of Statistics and Economic Studies): https://www.insee.fr/fr/accueil
Germany	German Federal Statistical Office (Destatis): https://www-genesis.destatis.de/datenbank/online/statistic/12411/table/12411-0001
Greece	The National Statistical Authority of Greece: https://www.statistics.gr/en/home
Hungary	Hungarian Central Statistical Office (KSH): https://www.ksh.hu/stadat_files/nep/hu/nep0003.html
Norway	Statistics Norway (SSB): https://www.ssb.no/en/befolkning Norwegian Institute of Public Health: https://statistikk.fhi.no/
Portugal	Instituto Nacional de Estatística (INE – national population statistics): https://www.ine.pt/
The United Kingdom	Office for National Statistics (ONS) – population statistics: https://www.ons.gov.uk/peoplepopulationandcommunity
The Netherlands	Statistics Netherlands (CBS): https://www.cbs.nl/
Spain	Spanish national demographic statistics: https://www.ine.es/ Catalonia regional statistics (IDESCAT): https://www.idescat.cat/pub/?id=inddt&lang=en Atlas VPM (Atlas of Variations in Medical Practice in the Spanish National Health System – descriptive population and healthcare indicators): https://cienciadedatosysalud.org/en/atlas-vpm/
Sweden	Statistics Sweden: https://www.scb.se/en/

9. STRENGTHS AND LIMITATIONS

The study will be informed by routinely collected health care data and so data quality issues must be considered. Electronic health records and other routinely collected data have certain inherent limitations because they were collected for clinical purposes rather than primarily for research use. In particular, the completeness and accuracy of recorded clinical conditions may vary across data sources due to differences in healthcare settings, clinical practice patterns, and coding practices. As disease prevalence will be estimated using recorded clinical history, undiagnosed or unrecorded conditions may not be captured in some data sources. These limitations will be identified and mitigated through the use of a standardised, transparent, and reproducible phenotyping process, which will support systematic assessment and validation of phenotype algorithms for each clinical condition of interest across data sources[4].

It is also important to note that the phenotypes developed in this study will be explicitly defined to capture history of disease and will not distinguish incident from prevalent events. Consequently, these phenotype

algorithms will not be suitable for incidence estimation or analyses requiring incident outcomes (e.g., time-to-event or causal inference analyses) and would need to be further adapted. Nevertheless, these standardized phenotypes may serve as a reference standard, provided their suitability is assessed in relation to the specific research objective.

In addition, the results generated in this study will reflect only the populations from the included data sources and may not be always fully representative of national populations, particularly for data sources that are not based on national registries. However, as the primary objective of this study is to characterise the DARWIN EU® network as it is, rather than to produce population-level estimates, this limitation does not affect the validity of results (in the context of the study objectives).

Despite these limitations, the study will have several important strengths. It will develop, document, and validate reusable standardised phenotypes for the description of disease prevalence and population characterisation across the DARWIN EU® network, supporting consistent study population description in future research. The study will also improve understanding of the data sources included in the network, supporting feasibility assessment and study planning. Finally, the inclusion of data partners from both Northern and Southern Europe, spanning a wide range of healthcare settings and data types, including electronic health records, claims data, registries, and biobanks, supports broad generalisability within the network and provides a diverse overview of populations captured by DARWIN EU®.

10. REFERENCES

1. Burn E, Catala M, Chen X, Alcalde-Herraiz M, Prats-Uribe A. PhenotypeR: Assess Study Cohorts Using a Common Data Model. [cited 2025 Dec 22]; https://ohdsi.github.io/PhenotypeR/. Accessed 22 Dec 2025

2. Catala M, Guo Y, Lopez-Guell K, Burn E, Mercade-Besora N, Alcalde M. CohortCharacteristics: Summarise and Visualise Characteristics of Patients in the OMOP CDM. [cited 2025 Dec 22]; https://darwin-eu.github.io/CohortCharacteristics/. Accessed 22 Dec 2025

3. Burn E, Raventos B, Catala M. IncidencePrevalence: Estimate Incidence and Prevalence using the OMOP Common Data Model. [cited 2025 Dec 22]; https://darwin-eu.github.io/IncidencePrevalence/. Accessed 22 Dec 2025

4. Dernie F, Corby G, Robinson A, Bezer J, Mercade-Besora N, Griffier R, et al. Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU). Pharmacoepidemiol Drug Saf [Internet]. Pharmacoepidemiol Drug Saf; 2024 [cited 2026 Jan 12];33. https://doi.org/10.1002/PDS.70042

11. ANNEXES

ANNEX I. Description of data sources

IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium)

#	Section	Description
1	Data source identification and country	IQVIA LPD Belgium (IQVIA Longitudinal Patient Database Belgium) Belgium
2	Data partner information section	IQVIA IQVIA Europe
3	Coverage and timespan	Data collection since: 2005 Extent: Nation-wide. Panel of 300 GPs in Belgium. The panel 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.
4	Healthcare setting / type of data	Primary care – General Practitioner. Ambulatory visits, with diagnosis, prescriptions, procedures, and laboratory tests.
5	Data collection process	Outpatient electronic health records. Records are entered by GPs at the healthcare encounter.
6	General representativeness	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. 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.
7	Data content /source coding	No information on source coding.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (Systemised Nomenclature of Medicine (SNOMED), Medical prescription normalised (RxNorm), Logical Observation Identifiers Names and Codes (LOINC)). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. The patient ID is per practice. So a patient can have different IDs in the DB, one per practice. In Belgium, patients are typically registered at only one general practitioner (GP) practice, so duplication should be minimal.
9	Quality control (data source specific)	No QC. Integrity constraints only.
10	Linkage	No linkage.
11	Vital status	Death information is derived from healthcare events.
12	Limitations	Observation period is defined from first to last visit, causing a drop in 'healthy' persons at the end of the interval. This creates an artefactual increase in incidence and prevalence ratios. There is no sociodemographic or over-the-counter medication information.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111116 Website: https://iqvia.com

Croatian National Public Health Information System (NAJS)

#	Section	Description
1	Data source identification and country	NAJS (Croatian National Public Health Information System) Croatia
2	Data partner information section	Croatian Institute of Public Health Department of Data Science and Analytics
3	Coverage and timespan	Data collection since: 1998 Extent: Nation-wide. Geographic coverage covers whole Croatia, with various levels of resolution for different registries. Current estimates for the population in Croatia will be available at: https://podaci.dzs.hr/hr/podaci/stanovnistvo/procjena-stanovnistva/ for each year. The total and active person count in the NAJS data is larger than the current population of Croatia. This explained by: a) the person table included deceased and all previously insured people and b) there is no information about insurance ending, c) healthcare is also used by people with dual citizenship from neighbouring countries It is known that a lot of people emigrated (300k-400k) and weren't included in the last population census but still are in the NAJS database. There is also an influx of immigrant workers that are insured and registered but weren't included in the census.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Primary care – gps, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. For both inpatient and outpatient setting diagnoses, medication, procedures, and measurements are captured. The year of availability of information depends on the setting • 2014-2025 for biochemical lab tests in primary care from EHR patients records (measurements with results) • 2015-2025 for primary care data from EHR patient records (conditions, procedures, and drug prescriptions) • 2015- 2024 for inpatient hospital data from EHR administrative records (conditions, procedures, measurements without results and drug administrations) • 2016-2025 for health risk assessment data entered by GPs (measurements with results - height, weight…) • 2016-2022 for secondary conciliatory care data from EHR administrative records (conditions, procedures, measurements without results and drug administrations) • 2016-2022 for emergency care data from EHR patient records (conditions) • 2017-2025 for hospital records from registry data (conditions and procedures) • 2020-2025 for vaccination data from EHR patient records
5	Data collection process	Inpatient hospital billing systems, and Other. Data is entered by clinicians at healthcare contact, then combined by CIPH into the NAJS database and integrated with registries for public health purposes.
6	General representativeness	The data is collected from the evidence of public health records collected for public health purposes, as the majority of health care in Croatia is public and under single health insurance provider. Personal details are collected to a better extent for insured individuals compared to uninsured patients, who are excluded in the ETL process.
7	Data content /source coding	Medication prescriptions are recorded with Anatomical Therapeutic Chemical (ATC) codes with an additional 3 digit code denoting the package. Diagnoses with ICD10 codes (Australian modification). Procedures with local source codes. Lab results with local source codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Records from 2017 include insured patients with reliable IDs. Uninsured patients do not have reliable IDs. For example, if a patient changed her status from insured to uninsured, or vice versa, she could be counted several times, as could tracking records from before 2017 and after. By using the unique personal identifier for Croatian citizens, it can be checked and verified.
9	Quality control (data source specific)	There is a network of registry personnel (leaders, administrators, coders, sources) working on data coverage and other quality dimensions. An analytical team routinely checks for erroneous entries in hospital records, removing double entries, false dates, and overlapping stays. Entries

#	Section	Description
		without enough data or with obviously erroneous dates from primary care analysis are being excluded.
10	Linkage	The national death registry is updated yearly, with one year lag, but the fact of someone's death (just the date) is updated daily, without the cause of death or any other additional details. Primary care is updated weekly and hospital level care monthly. Specific registries are included in NAJS (e.g. diabetes registry), where inclusion criteria vary across these registries.
11	Vital status	NAJS is linked to the national death registry.
12	Limitations	Hospital data is available from 2017 onwards. This is often used as start of data collection, while laboratory and GP data is captured before that (since 2014 and 2015 respectively). Drug duration is often not available and set to 1 day for administration and 30 days for prescription. Hospital discharge summaries are currently not captured in NAJS. Hospital drug administration data is less reliable than prescription data from primary care, with some drugs (monoclonal antibodies / precision medicine drugs) that require additional approval not being recorded at all.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111155 Website: https://www.hzjz.hr/nacionalni-javnozdravstveni-informacijski-sustav-najs/

Estonian Biobank (EBB)

#	Section	Description
1	Data source identification and country	EBB (Estonian Biobank) Estonia
2	Data partner information section	University of Tartu Institute of Computer Science
3	Coverage and timespan	Data collection since: 2004 Extent: Nation-wide. EBB is a nation-wide database containing records from 2004 onwards. Estonian population-based cohort size of 211,800 participants (01/01/2024) aged 18 years and older recruited at GP offices, private practices, and hospitals or in the recruitment offices of the Estonian Genome Center.
4	Healthcare setting / type of data	Primary care – General Practitioner, and community pharmacists, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Healthcare providers must report the data to national registries. The data includes claims, prescriptions, EHR and two registries - the cause of death and cancer registry. The claims data covers diagnosis, procedures and services performed. Claims data contains information about services paid by the Estonian Health Insurance Fund and not the services paid by the person themself. The prescription data contains information prescribed and dispensed prescription drugs. It does not contain information about the over-the-counter drugs. The EHR data includes diagnosis, results of the procedures and lab measurements and case summaries. The cause of death and cancer registry data is checked by the registry before it is entered to the registry. Also, Estonian Biobank, has collected genetic and questionnaire data when the person gives consent. Questionnaire data includes personal data (e.g. nationality), genealogical data (family history of medical conditions), educational history, lifestyle data (e.g. smoking, alcohol consumption etc).
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and Biobank, and Other. Data is retrieved by Estonian Biobank once a year from national registries. The insurance claims are requested from Estonian Health Insurance Fund. The inpatient and outpatient electronic health records are requested from National Health Information System. The cancer registry and

#	Section	Description
		cause of death registry information is requested from The National Institute for Health Development. The data is sent to the national registry by the healthcare providers.
6	General representativeness	The age, sex, and geographical distribution closely reflect those of the Estonian adult population and encompass approximately 20% of adult population. Female participants are over-represented in EBB. Older people tend to participate less frequently; however, all age groups are well represented.
7	Data content /source coding	All participants have undergone a standardized health assessment, including provision of blood samples for purification of DNA, white blood cells, and plasma, and completed a questionnaire covering various health-related topics, such as lifestyle, diet, and clinical diagnoses. Diseases and health problems are recorded as ICD-10 codes and prescribed medicine according to the ATC classification and local package codes. Procedures and services are coded with NOMESCO classifier and local service codes. Lab measurements are coded with local codes and in later years with LOINC codes. In cancer registry also ICD-O-3 codes are used. The data is available with different granularity. For claims data we only now that procedure or service is done. The result of the procedure or measurement is recorded in EHR. Usually, the result is in unstructured text (except lab measurements) and needs specific pipelines for extraction.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. There is one national identifier that allows linking together all encounters across databases.
9	Quality control (data source specific)	The quality control procedures in the Estonian Biobank aim to remove the most obvious mistakes in the data, misspellings, impossible dates, duplicates. Before performing the ETL, several problems are fixed on the source data. Since the ETL procedures are used for a number of different datasets (from the same national sources), we have a growing number of pre-processing steps that correspond to the issues we have discovered previously in the data, such as checking for the presence of critical values, harmonizing date and unit of measurement formats, checking the validity of certain entries against classifiers, etc.
10	Linkage	Follow-up data are available via linkage with national health-related registries and via re-examination of participants. Furthermore, electronic health records are updated for phenotypic outcome information every year. The EBB database is regularly linked with national registries, hospital databases, and the databases of the Estonian Health Insurance Fund (EHIF) and the National Health Information System (NHIS).
11	Vital status	Vital status (death date and causes of death) are obtained from the Causes of Death Registry.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	Milani L, Alver M,Laur S,Reisberg S,Haller T,Aasmets O,Abner E,Alavere H,Allik A,Annilo T,Fischer K,Hofmeister R,Hudjashov G,Jõeloo M,Kals M,Karo-Astover L,Kasela S,Kolde A,Krebs K,Krigul KL,Kronberg J,Kruusmaa K,Kukuškina V,Kõiv K,Lehto K,Leitsalu L,Lind S,Luitva LB,Läll K,Lüll K,Metsalu K,Metspalu M,Mõttus R,Nelis M,Nikopensius T,Nurm M,Nõukas M,Oja M,Org E,Palover M,Palta P,Pankratov V,Pantiukh K,Pervjakova N,Pujol-Gualdo N,Reigo A,Reimann E,Smit S,Rogozina D,Särg D,Taba N,Talvik HA,Teder-Laving M,Tõnisson N,Vaht M,Vainik U,Võsa U,Yelmen B,Esko T,Kolde R,Mägi R,Vilo J,Laisk T,Metspalu A "The Estonian Biobank's journey from biobanking to personalized medicine." Nature communications (2025): 40188112
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111114 Website: https://genomics.ut.ee/en/content/estonian-biobank

Assistance publique Hôpitaux de Marseille (APHM)

#	Section	Description
1	Data source identification and country	APHM (Assistance publique Hôpitaux de Marseille) France
2	Data partner information section	Assistance Publique Hôpitaux de Marseille Public Health
3	Coverage and timespan	Data collection since: 2014 Extent: Regional. The data covers all inpatients and outpatients treated at the Assistance Publique – Hôpitaux de Marseille (APHM), which includes five public university hospitals, all located in Marseille, France. The APHM serves not only the local population of Marseille and the surrounding Bouches-du-Rhône department but also attracts patients from the broader Provence-Alpes-Côte d'Azur (PACA) region and Corsica, representing a combined population of over 5 million inhabitants. Additionally, the hospital's specialized services attract patients from across France and abroad, including foreign nationals, all of whom are integrated into the OMOP CDM.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The data source used in this study includes all hospital stays across various care settings—acute care, psychiatric care, rehabilitation care, and home hospitalization. The EHR system covers diagnoses, procedures, drug prescription and administration, medical and paramedical notes, such as hospitalization reports, radiology, EEG, endoscopy, and consultation summaries, and laboratory data.
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries, and Biobank. Data is entered by clinicians into the hospitals EHR system, consisting of several pieces of software. Diagnoses and procedures are managed via the CORA software; drug prescription and administration data, through the PHARMA software. Additionally, reports, radiology and consultation summaries are recorded using the AXIGATE software. These systems are integrated in the IATROS database for secondary use.
6	General representativeness	The database population are limited to patients visiting a specialised hospital.
7	Data content /source coding	Diagnoses are coded using ICD-10 and procedures are recorded using CCAM, in line with the French DRG system. Drug prescription and administration use UCD drug codes, ATC classifications, quantities, and dosages. 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.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Each patient is assigned a unique patient number, which remains consistent throughout their care. In the rare event of duplicate records, a dedicated team is responsible for merging these records and ensuring quality control. Additionally, in France, patients have a unique national identifier (INS), which will eventually allow seamless linkage between hospital data and the SNDS (national health data system).
9	Quality control (data source specific)	Each software used for data collection undergoes quality verification before allowing validation and integration into the databases. These processes are managed by the hospital's IT department. Rigorous quality control is performed at multiple stages by various stakeholders, including the IT department, the medical information department, and internal controls. Quality assurance in the source systems is managed through a series of checks. These include validation loops when studies are conducted, ensuring that data is research-ready and meets required standards. Additionally, these controls help in identifying and resolving any data inconsistencies or errors before the data is made available for research purposes.

#	Section	Description
10	Linkage	Patient-Reported Experience Measures (PREMs) and Patient-Reported Outcome Measures (PROMs) can be linked. However, this linkage is not exhaustive across all domains. While the data allows for connections between medicine usage and some health outcomes, further development is required to achieve comprehensive linkage across all patient records and conditions. In particular, using non-structured data, such as clinical notes, could be improved through Natural Language Processing (NLP) for specific conditions. The implementation of additional linkages will need to be done on a case-by-case basis. In addition, it is possible to link socioeconomic information, e.g. for indicators like the FDEP (FDep or French Deprivation Index), a measure of neighbourhood deprivation.
11	Vital status	Vital status is retrieved from healthcare coverage details and covers date of death for all patients if they died inside the hospital.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	Fond G, Pauly V,Orleans V,Antonini F,Fabre C,Sanz M,Klay S,Jimeno MT,Leone M,Lancon C,Auquier P,Boyer L "Increased in-hospital mortality from COVID-19 in patients with schizophrenia." L'Encephale (2021): 32933762
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111141 Website: http://ap-hm.fr/

InGef Research Database (InGef RDB)

#	Section	Description
1	Database Identification and country	InGef Research Database (InGef RDB) Germany (DE)
2	Data partner information section	InGef – Institute for Applied Health Research Berlin GmbH InGef is a research service provider within the German statutory health insurance system.
3	Coverage and timespan	The Research Database contains approximately 10.5 million insured persons from 50 of the 94 statutory health insurances in Germany (as of Oct 2025). All individuals in the research database are included in the OMOP CDM. Due to data protection laws for personal data in Germany, only the last 10 years of data are available in the RDB and thus in the OMOP CDM.
4	Healthcare setting / type of data	In general, the data in the RDB reflects most of the health services paid for by statutory health insurances. In the OMOP CDM, the health services are reduced to primary and secondary care. This means that treatments by GPs and specialists (e.g., paediatricians), prescription medicines dispensed by pharmacies, hospital inpatient and outpatient care, as well as information about all insured individuals are included. The following data elements are presented in the OMOP CDM: demographic information, diagnoses, procedures, dispensing drugs and advanced therapy medicinal products, vaccinations, pregnancy data (via diagnoses and procedures) and contraception.
5	Data collection process	In Germany, data exchange between medical service providers and statutory health insurances is organized via central data collection centers. In addition to receiving and forwarding data, these centers also store data and provide access to third parties under strict regulations in data warehouses. The RDB contains selected, condensed, and anonymized information from one of these data warehouses. Data sovereignty remains with the individual health insurance companies.
6	General representativeness	The RDB covers about 11% of the German population and is comparable to the German population in terms of the distribution of age and sex. Most health insurances that contribute to the RDB have nationwide coverage, meaning that the database covers all regions of Germany.

#	Section	Description
		Since almost all services covered by statutory health insurances are specified in national legislation, healthcare provision all over Germany is well represented in the RDB. Additionally, in Germany it is very common to stay with the same health insurance throughout life, which results in a good longitudinal coverage over the entire period of 10 years.
7	Data content /source coding	The coding in the research database complies with national classification and coding rules in Germany. Diagnoses are coded according to ICD-10-GM. Inpatient and outpatient surgeries or procedures are recorded as OPS codes (German classification of Operations and Procedures). The dispensing of drugs in pharmacies is recorded using the PZN (pharmaceutical registration number). For drugs that miss a PZN-to-RxNorm mapping, the ATC code is used instead. In some cases, dispensed drugs can be coded using OPS codes (e.g., in hospitals) or EBM codes (fee schedule for outpatient treatments).
8	Data Harmonization	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.
9	Quality control (database specific)	The data transmitted by healthcare providers complies with the standardized requirements and formats of the Association of Statutory Health Insurances (GKV-SV). Before being imported into the research database, the data elements are checked for data format, completeness, and plausibility. After each update of the research database, various counts are compared with the previous update to verify completeness. Due to the anonymity of the database, direct validation of the data (e.g., using medical records as the gold standard) is not possible.
10	Linkage	Due to the anonymization of the source data, linkage is not possible.
11	Vital status	The date of death is recorded as the last day of the quarter in which the death occurred (i.e., 30/31st of Mar/Jun/Sept/Dec) as reported to the health insurance (no linkage to death registry). The cause of death is not available.
12	Limitations	Ambulatory diagnoses and procedures are summarised in the source on a quarterly basis. Both are mapped to the observation table with the date set to the last day of the respective quarter (i.e. 30/31st of Mar/Jun/Sept/Dec) and the concept “History of event within 3 months” (observation_concept_id 1340222), with the actual diagnosis or procedure concept_id recorded in the field “value_as_concept_id”. There is no vocabulary for the German pharmaceutical product codes (PZN). A direct source-to-standard-mapping has been done manually by InGef but is incomplete. The drug exposure duration is unknown. Following OMOP conventions, the end date is always set to dispensing date + 29. Outpatient and inpatient procedures are recorded as OPS codes (German Procedure Classification), for which the vocabulary is incomplete.
13	Main references	Andersohn F, Walker J "Characteristics and external validity of the German Health Risk Institute (HRI) Database." Pharmacoepidemiology and drug safety (2016): PMID 26530279 | doi: 10.1002/pds.3895 Ludwig M, Enders D, Basedow F, Walker J, Jacob J: Sampling strategy, characteristics, and representativeness of the InGef research database. Public Health 2022. doi: 10.1016/j.puhe.2022.02.013
14	Link to HMA-EMA catalogue and database webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111207 Website: https://www.ingef.de/en/

IQVIA Disease Analyzer Germany (IQVIA DA Germany)

#	Section	Description
1	Data source identification and country	IQVIA DA Germany (IQVIA Disease Analyzer Germany) Germany
2	Data partner information section	IQVIA
3	Coverage and timespan	Data collection since: 1989 Extent: Nation-wide. GP and specialists in Germany using specific patient management software.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians). Diagnoses, medication, and procedures from an ambulatory setting. Medications are recorded as prescriptions of marketed products.
5	Data collection process	Outpatient electronic health records. By clinicians at healthcare contact.
6	General representativeness	No specific details on general representativeness given.
7	Data content /source coding	Prescription is on product code level (German PZN), ICD10, NFC, Local lab coding.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. There can be patients registered under different ID numbers, because there is no linkage between different GPs.
9	Quality control (data source specific)	Data is quality checked on plausibility.
10	Linkage	No.
11	Vital status	Death information is derived from medical events.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/104282 Website: https://www.iqvia.com/

Papageorgiou General Hospital (PGH)

#	Section	Description
1	Data source identification and country	PGH (Papageorgiou General Hospital) Central Macedonia, Greece
2	Data partner information section	Papageorgiou General Hospital Program Management Office
3	Coverage and timespan	Data collection since: 1999 Extent: Other. Patients from across the Macedonia region and from neighbouring countries as well.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). long-term/skilled nursing facility

#	Section	Description
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems. The data is gathered live (as the patient is examined) in the production EHR instance.
6	General representativeness	The database represents best the population of Northern Greece requiring hospital care. It captures data on hospitalizations, ICU admissions, medical procedures, and measurements. With a capacity of 745 beds, PGH supports over 200,000 hospitalization days annually. The OMOP CDM has around 1.41M patients.
7	Data content /source coding	Medications are coded using ATC codes. KEN («Κλειστά Ενοποιημένα Νοσήλεια») Coding System and Internal Hospital Coding System: This coding system is applied to procedural data, and it was enforced by the Greek Public Health Ministry. It has been mapped to ICD-10. Internal Hospital Coding System: This system is employed for coding medical devices. ICD-10 (International Classification of Diseases, Tenth Revision): This system is utilized for coding both 'Diagnoses' and 'Deaths' from the source. Internal LIS Coding System: This coding system, associated with the internal laboratory information system, is used for laboratory tests and examinations.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. In Greece, each patient is identified by a unique universal health insurance number.
9	Quality control (data source specific)	There is no formal quality assurance plan in place for the raw data collected during daily clinical practice, therefore, gaps or errors in the data should be anticipated.
10	Linkage	No known linkages.
11	Vital status	There are only in-hospital deaths available in the data.
12	Limitations	Drugs recorded at the active ingredient level. Drug dosages and routes are not captured at the moment. Drug exposure presents the drug ordered for the patient and not the administration (in the majority of the cases the patient is given the ordered drug, but this might not be true for 100% of the cases). Drug exposure dates also refer to the order date. Medical devices have not been mapped. Medical history is missing. Patient progress is not available after discharge. Vital measurements like body temperature, BMI, blood pressure are not captured and Observations in general are missing.
13	Main references	Papapostolou G, Chytas A,Rekkas A,Bigaki M,Zeimpekis D,Dermentzoglou L,Tortopidis G,Natsiavas P "Real-World Data in Greece: Mapping the Papageorgiou General Hospital Data to the OMOP Common Data Model." Studies in health technology and informatics (2024): 39176625
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/institution/1000000215 Website: https://www.papageorgiou-hospital.gr/?lang=en

Semmelweis University Clinical Data (SUCD)

#	Section	Description
1	Data source identification and country	SUCD (Semmelweis University Clinical Data) Budapest, Hungary
2	Data partner information section	Semmelweis University -
3	Coverage and timespan	Data collection since: 2010 Extent: Regional. The general catchment area of SU is the central region of the country, Budapest city and Pest county, although patients can be referred from anywhere in Hungary. The total population of Budapest and Pest county is approximately 4,200,000 people. The total population of Hungary is around 9,500,000.

#	Section	Description
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and claims data, and other (specify). diagnostic data (laboratory tests, radiology, pathology)
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries. Data is extracted directly from the source database. From there, the data entry in the system is heavily controlled and validated on the user interface before being made available for further research.
6	General representativeness	SU captures information on patients who are covered by the public health insurance system. This covers all Hungarian citizens, and therefore the database should mirror the source population well. Although, besides Semmelweis University Clinics, there are multiple hospitals in the region, and data on visits in other hospitals in not represented in the database. Therefore, the patient population is not directly representative of the general population.
7	Data content /source coding	Regarding SU's source data, procedures and diagnoses are coded in SNOMED, measurements are coded in LOINC, and drugs are stored in RxNorm and ATC.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a unique identifier (SSN).
9	Quality control (data source specific)	The clinical database is the source database and therefore it has to be treated as a trusted database. Data entry in the systems is heavily controlled by validation on the user interface, and there are large number of rules that controls the data on the insurer's side that has to be corrected in the system by the users to be able to close the encounters. OMOP mapping is done in the framework by EHDEN recognized partners under quality check by the EHDEN society.
10	Linkage	No known linkages.
11	Vital status	Source for vital status unknown.
12	Limitations	Medication prescribed in secondary care is fully present in our database, but medication given in the hospital is rarely documented. General limitations for the data type applicable. General practitioner data is not present in our database; therefore, this part of the patient journey is not represented.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1000000184 Website: https://www.semmelweis.hu

Integrated Primary Care Information (IPCI)

#	Section	Description
1	Data source identification and country	IPCI (Integrated Primary Care Information) The Netherlands
2	Data partner information section	Erasmus University Medical Center Department of Medical Informatics
3	Coverage and timespan	Data collection since: 2006 Extent: Nation-wide. IPCI is a Dutch database that contains patient records from 2006 onwards. However, it mainly covers the central part of the country, including the most densely populated area (the 'Randstad') and non-urban areas. IPCI contains information on all patients registered with GPs responsible for non-emergency care and referrals. A patient is registered at birth or at first encounter with the GP.

#	Section	Description
4	Healthcare setting / type of data	Primary care – General Practitioner. Data is collected from primary care EHR. This includes demographic information, complaints and symptoms, diagnoses, laboratory test results, lifestyle factors (in limited amount), and correspondence with secondary care, such as referral and discharge letters.
5	Data collection process	Outpatient electronic health records. Data is entered into the EHR system by the GPs, during or after the visit. The patient dossiers are collected by Erasmus MC data managers and combined in one harmonized database. Several checks are done on this database to ensure correct data processing. Persons can have dossiers at multiple GPs.
6	General representativeness	More than 99% of the Dutch population has health insurance, and almost all citizens are registered with a general practitioner. Over 12 months, around 78% of the population has at least one contact with their GP. IPCI included around 350 GP practices out of around 5000 in the country (~ 7%). The demographic composition of the IPCI population mirrors that of the general Dutch population in terms of age and sex.
7	Data content /source coding	Dutch GPs use mainly Dutch standard codes, like ICPC-1 and Diagnostische Bepalingen maintained by NHG. And for therapy the G-Standard is used, maintained by ZIndex.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients can be registered under different IDs, but since a patient can only be registered at one GP at a time, the observations periods will not overlap.
9	Quality control (data source specific)	Prior to each data release, extensive quality control steps are performed, e.g., comparison of patient characteristics between practices, and checks to identify abnormal temporal data patterns in practices. For each practice, around 200 quality indicators are obtained. Of these indicators, a quarter refer to population characteristics, e.g. number of birth and mortalities relative to practice size, temporal consistency. The other indicators are based on medical data, e.g. distribution of measurement values, frequencies of diagnoses and procedures relative to age, completeness of data. The indicators are combined in a couple of quality scores for each practice. For these scores, cut-off values for acceptable quality have been defined. Practices with a score below a cut-off are excluded for research. This approach has shown to be very important, for example to check if data from practices that just joined the database are at an acceptable level of quality. The details of the approach, like the cut-off values for acceptance, are based on years of experience. In addition, trends are compared with the previous database release. Extensive quality control steps are performed before each data release. These include comparing patient characteristics between practices and checks to identify abnormal temporal data patterns in practices. Additional checks include over 200 indicators related to population characteristics (e.g., reliability of birth and mortality rates) and medical data (e.g., availability of durations of prescriptions and completeness of laboratory results). Records of low quality are excluded from the database.
10	Linkage	Linkage requires additional approval steps and needs to be assessed on a case-by-case basis. IPCI is not routinely linked with other databases.
11	Vital status	Vital status (death date and cause) is collected based on GP records.
12	Limitations	The main limitation comes with the fact that IPCI is limited to GP records, and although it contains information on referrals and discharge letters, it may not fully capture specific hospital information. IPCI does not include coded/detailed data about medications/procedures/test results from the hospital or other care-providers.
13	Main references	de Ridder MAJ, de Wilde M,de Ben C,Leyba AR,Mosseveld BMT,Verhamme KMC,van der Lei J,Rijnbeek PR "Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands." International journal of epidemiology (2022): 35182143
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/42618 Website: http://www.ipci.nl

Norwegian Linked Health Registry data (NLHR)

#	Section	Description
1	Data source identification and country	NLHR (Norwegian Linked Health Registry data) Norway
2	Data partner information section	University of Oslo Faculty of Mathematics and Natural Science – Department of Pharmacy
3	Coverage and timespan	Data coverage: 2008-2023 (primary care data), 2018-2023 (secondary care data and drug dispensation data) Extent: Nation-wide. 5.5M active population. 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.
4	Healthcare setting / type of data	Primary care – gps, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following registries are included: 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).
5	Data collection process	Registries. 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.
6	General representativeness	The NLHR data covers the full Norwegian population.
7	Data content /source coding	NPR: ICD-10 for diagnosis, ATC and some special codes for drug use, Norwegian codes for clinical procedures (surgery (NCSP), medicine (NCMP) and diagnostic imaging, image-guided intervention, and nuclear medicine (NCRP)). KUHR: ICD-10 and ICPC-2 and ICPC-2B for diagnosis/procedure. NorPD: ATC. SYSVAK and MSIS: national classifications. MBRN: custom classifications by questionnaires (incl. check box variables in Maternity health care card)
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. 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.
9	Quality control (data source specific)	In-house data quality checks of rates of common conditions, drug exposures, and outcomes. We compare obtained rates with official national statistics (e.g., birth statistics, yearly rates of drug dispensing, and diagnosis by age and gender). We also review missing data and outliers and inform registry holders of any unusual patterns.
10	Linkage	The NLHR is, by definition, a linkage of datasets. Helsedata.no is one central portal to apply for 11 national health registries, including all the registries that have been mapped to the OMOP CDM.
11	Vital status	The national death registry is linked.
12	Limitations	Diagnostic codes in the secondary care are limited to the comprehensive list requested from the registries. The list of codes will be revised annually with each data delivery. (ii) ICD-10 codes vary in granularity, ranging from 2-character to full-length codes. The granularity of data will be revised annually with each data delivery (iii) Drug dispensations to outpatients are included; however, over-the-counter (OTC) medications are not captured in the data. Only a few selected expensive drugs given in hospital, are included..
13	Main references	Trinh et al. Harmonizing Norwegian registries onto OMOP common data model: Mapping challenges and opportunities for pregnancy and COVID-19 research. Int J Med Inform 2024: 191: 105602.

#	Section	Description
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1000000409 Website: https://www.mn.uio.no/farmasi/english/research/groups/pharma-safe/

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

#	Section	Description
1	Data source identification and country	EMDB-ULSEDV (Egas Moniz Health Alliance database – Unidade Local de Saúde de Entre Douro e Vouga) Distrito de Aveiro, Portugal
2	Data partner information section	Clinical Academic Center Egas Moniz Health Alliance
3	Coverage and timespan	Data collection since: 1999 Extent: Regional. ULSEDV includes 37 primary care centres assisted by four hospitals (Hospital de São Sebastião, Hospital São João da Madeira, Hospital Franscisco Zagalo and Hospital São Miguel). It fully serves approximately 330,000 patients of the municipalities of Santa Maria da Feira, Arouca, São João da Madeira, Oliveira de Azeméis, Vale de Cambra and Ovar.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). The integrated database captures information about demographics, visit occurrences, diagnoses, medications, and procedures.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. The data is extracted directly from the source EHR system.
6	General representativeness	The healthcare facilities collecting the data are part of the national health service and treat patients of all socioeconomic levels. Therefore, the patient sample should be representative of the population.
7	Data content /source coding	Source data terminologies include ATC, RxNorm, ICD-9, ICD-10 codes, and ICPC-2 codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. The same patients cannot be registered under a different id in the OMOP CDM data.
9	Quality control (data source specific)	We have a two-step approach for data quality: A set of unit tests are implemented as data quality scenarios. These scenarios are tested at the ETL level and ran every time the ETL is executed. Those tests ensure that known issues in data quality are addressed and check automatically on every run. In addition, data quality is assessed through the Observational Health Data Sciences and Informatics (OHDSI) DataQualityDashboard project and reported as well. There is room for implementation of specific data quality checks that may be relevant at the study level, if required.
10	Linkage	No known linkages.
11	Vital status	Vital status captured through two mechanisms: 1) recording of deaths in hospital admissions (secondary and tertiary care); and 2) verification against the national patient registry, which is triggered by follow-up appointments in primary care.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111133 Website: https://www.emha.pt/

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

#	Section	Description
1	Data source identification and country	EMDB-ULSGE (Egas Moniz Health Alliance database - Unidade Local de Saúde de Gaia e Espinho) Distrito de Aveiro, Portugal
2	Data partner information section	Clinical Academic Center Egas Moniz Health Alliance
3	Coverage and timespan	Data collection since: 2004 Extent: Regional. 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). It serves the population of the municipalities of Gaia and Espinho, with a total population of 350,000 patients (100% coverage) and indirect coverage of 1,2 million inhabitants considering the high differentiation of clinical care.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). The integrated database captures information about demographics, visit occurrences, diagnoses, medications, and procedures.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. The data is extracted directly from the source EHR system.
6	General representativeness	The healthcare facilities collecting the data are part of the national health service and treat patients of all socioeconomic levels. Therefore, the patient sample should be representative of the population.
7	Data content /source coding	Source data terminologies include ATC, RxNorm, ICD-9, ICD-10 codes, and ICPC-2 codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. The same patients cannot be registered under a different id in the OMOP CDM data.
9	Quality control (data source specific)	We have a two-step approach for data quality: A set of unit tests are implemented as data quality scenarios. These scenarios are tested at the ETL level and ran every time the ETL is executed. Those tests ensure that known issues in data quality are addressed and check automatically on every run. In addition, data quality is assessed through the OHDSI DataQualityDashboard project and reported as well. There is room for implementation of specific data quality checks that may be relevant at the study level, if required.
10	Linkage	No known linkages.
11	Vital status	Vital status captured through two mechanisms: 1) recording of deaths in hospital admissions (secondary and tertiary care); and 2) verification against the national patient registry, which is triggered by follow-up appointments in primary care.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111133 Website: https://www.emha.pt/

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

#	Section	Description
1	Data source identification and country	ULSM-RT (Unidade Local de Saúde de Matosinhos Realtime Database) Distrito do Porto, Portugal
2	Data partner information section	Unidade Local de Saúde de Matosinhos Department of Research, Clinical Epidemiology and Public Health
3	Coverage and timespan	Data collection since: 1998 Extent: Regional. Complete primary and secondary public healthcare coverage for the region Matosinhos, which also includes the general regional referral patients to secondary care (318.000 patients (+600.000 patients with occasional contact with ULSM)).
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. All care settings covered are from EHR data from primary and secondary care. For procedures performed outside, claims data exists because ULSM is the payer of such procedures.
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems.
6	General representativeness	Includes all patients using public sector (100% coverage).
7	Data content /source coding	Medicines prescribed for outpatient pharmacy are coded natively in ATC and were translated to RxNorm at the ingredient level. Medications prescribed from the hospital pharmacy are not coded natively. Source data terminologies used are ATC, ICD-10-CM, ICD-9-CM.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. All patients are assigned a unique, non-repetitive number for healthcare that is the same across institutions both public and private in Portugal. There are no duplicate numbers.
9	Quality control (data source specific)	Source data used represents raw EHR data as it was input by healthcare professionals for care purposes. Thus, no source of error related to data cleaning or transformations is expected to exist in source data. The national identification number is used as primary key, persons without a national id are not included. All quality control is done on ETL with unit test and OMOP Quality Assurance scripts.
10	Linkage	Linked with National Death Certificate System (Sistema de Informação dos Certificados de Óbito (SICO).
11	Vital status	Date and cause of death information available (EHR and SICO).
12	Limitations	Dispensing information only available for internal prescriptions. Genetic data and patient-generated data not available.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111171 Website: https://www.ulsm.min-saude.pt

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

#	Section	Description
1	Data source identification and country	BIFAP (Base de Datos para la Investigación Farmacoepidemiológica en el Ámbito Público (Pharmacoepidemiological Research Database for Public Health Systems)) Spain

#	Section	Description
2	Data partner information section	AEMPS Pharmacoepidemiology and Pharmacovigilance Division - Medicines for human use Department
3	Coverage and timespan	Data collection since: 2001 Extent: Regional. Spanish National Health Service (SNS) from 9 of the 17 regions in Spain. The population currently included represents 36% of the total Spanish population.
4	Healthcare setting / type of data	Primary care – General Practitioner, and community pharmacists, and primary care specialists (e.g. paediatricians), and hospital inpatient care. 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.
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries. Data in BIFAP is collected from Primary Care and Hospital EHR.
6	General representativeness	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.
7	Data content /source coding	The BIFAP source data is coded in SNOMED, ICD, ICPC-2 (diagnoses), AEMPS (drugs), and local lab codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Pseudonymized ID numbers are generated at regional level. The Personal Identification Code for the Autonomous Community (CIPA) is used to perform the pseudonymisation procedure. Therefore, upon changing practice or de- and re-registration within the same region, (Autonomous Community) the patient in BIFAP is correctly identified as the same person with the same ID number. However, the same patient would obtain different ID numbers if the patient moves to a different region and is registered in a primary care practice in the new region. The percentage of people who are de-registered due to moving to other region in relation to the BIFAP population is, for example, 5% in Madrid and 4% Castilla y Leon. This situation would have a very limited impact on the data analysis due to the following: - The proportion is low (less than 5%) in relation to the overall population in BIFAP. - In BIFAP, only stable residents are included. This means that patients living in another region for a foreseen short time period and are provisionally assigned to a primary care practice are not included in BIFAP. - Medical events of those patients who have more than one ID do not overlap in time, since dates of events correspond to different periods. This means that counts of these events are never duplicated. - A number of study designs allows the same patient to be part of different cohorts or to be selected both as case and as control, provided that their person-time experience correspond to a different period of time. In all these cases, the impact in study analysis of duplicated IDs would be negligible.
9	Quality control (data source specific)	Patients who meet any of the following disability criteria are discarded: - Non-owners of the individual health card - Date of birth before 01/01/1801 - Active patients over 115 years of age - Patients without clinical records (only contains administrative information) - Patients marked as "fictitious" in the clinical history - Badly coded sex

#	Section	Description
		- Inactive without termination date - Start date = End date - Clinical records prior to date of birth
10	Linkage	The following data are also linked at individual patient level and available. For a subset of the BIFAP population (regions and/or periods of time): • Information on dispensation of medicines at hospital pharmacies from outpatients and inpatients. • Registration of Causes of Death by the National Institute for Statistics. From the start of the COVID-19 pandemic: • Vaccines COVID-19 Administration Registry linked to patients included in BIFAP. • Diagnosis Tests of COVID-19 linked to patients included in BIFAP, for some regions.
11	Vital status	Source for vital status unknown.
12	Limitations	Primary care is available from 2001 but is considered complete since 2005. Hospital discharge has different coverage periods per region Spain, with most starting between 2014–2016. This means that for different regions and different time periods there is a different coverage of healthcare events. In the release of July 2025, the laboratory results are not covered. These will be added again at the next release, expected at the end of 2025.
13	Main references	Maciá-Martínez MA, Gil M,Huerta C,Martín-Merino E,Álvarez A,Bryant V,Montero D "Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP): A data resource for pharmacoepidemiology in Spain." Pharmacoepidemiology and drug safety (2020): 32337840
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/21501 Website: http://www.bifap.org/index_EN.html

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

#	Section	Description
1	Data source identification and country	PRISIB (Platafoma de Recerca en Informació Sanitària de les Illes Balears) Balearic Islands, Spain
2	Data partner information section	IdISBa Health Data Research Platform of the Balearic Islands
3	Coverage and timespan	Data collection since: 2008 Extent: Regional. The geographic area of catchment for the PRISIB database includes all of the Balearic Islands (Mallorca, Menorca, Ibiza, Formentera, etc.). This is estimated to encompass the whole population of the archipelago including a significant number of citizens from other countries established in the region. The approximate population is around 1245000 inhabitants.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The PRISIB database includes hospital and primary care data comprising visits, measurements, diagnoses, procedures, laboratory results as well as medication prescriptions. This covers 60 primary care centres and 7 hospitals. Younger generations are registered at birth and enter a youth health protocol that registers many visits with their paediatrician. Home care and palliative care, when performed by the primary care team, is also recorded.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Other. The regional public health system data is centralized by the health care administration, including all hospitals and primary care centres as well as the electronic system for outpatients' drug

#	Section	Description
		prescriptions. All interactions with the public health care system are recorded, including dispensings, vaccination campaigns, ambulance service, and GP phone consultation.
6	General representativeness	The source of data should be representative of the whole population. Patients with private insurance might use less public services but are represented as it is required for drug reimbursement.
7	Data content /source coding	The following terminology systems are used: ATC, ICD9, ICD10, ICPC-2, National drug agency catalogue of products and the Spanish radiology association catalogue. Numeric ISO codes for countries. National Institute of Statistics codification for census districts.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a unique id across healthcare practices.
9	Quality control (data source specific)	When a dataset is generated from this source, a validation script is run to check that what was extracted conforms to the criteria stated in the data model for each particular project and generates a descriptive and data quality report.
10	Linkage	In the source data identifying information can be requested in order to link new sources of data. All patients are linked to the census district their home is in allowing for geographic linkage.
11	Vital status	The database only captures deaths occurring in the geographic area as recorded in the regional death certificate registry.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	Ruiz-Pérez M, Moragues A,Seguí-Pons JM,Muncunill J,Pou Goyanes A,Colom Fernández A "Geographical Distribution and Social Justice of the COVID-19 Pandemic: The Case of Palma (Balearic Islands)." GeoHealth (2023): 36819934
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111142 Website: https://www.idisba.es/en/Support-Services/Scientific-Technical-Platforms/Research-in-Health-Information

The Information System for Research on Primary Care (SIDIAP)

#	Section	Description
1	Data source identification and country	SIDIAP (The Information System for the Development of Research in Primary Care) Catalunya, Spain
2	Data partner information section	IDIAPJGol
3	Coverage and timespan	Data collection since: 2006 Extent: Regional. SIDIAP is a database of primary care electronic health records of the population of Catalonia, North-East Spain. It contains pseudo-anonymised records of more than 8 million people , of which 6.1 million are active as of 2022, representing around 76% of the Catalan population.
4	Healthcare setting / type of data	Primary care – General Practitioner, and hospital discharge data. SIDIAP captured data includes routine visits, socio-demographics, diagnoses, laboratory tests, drugs (prescribed and dispensed), referrals, sick leaves, and lifestyle information.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. Data is entered by primary care physicians upon healthcare contact, supplemented with hospital discharge records. The Institut Catala de la Salut (Catalan Health Institute) is the data controller.
6	General representativeness	It was previously shown that the captured SIDIAP population is highly representative of the entire Catalan region in terms of geographic, age, and sex distributions.

#	Section	Description
7	Data content /source coding	SIDIAP data covers all services that occur at the Primary Care Centres, as well as support services, such as sexual and reproductive health or home end-of-life care. Drugs are coded in ATC-WHO terminology in the source data. Health outcomes are captured in ICD-10CM codes. The SIDIAP contains all laboratory tests and results performed in primary health centres. Demographics, geographical, as well as socio-economic factors are recorded for each patient.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. A patient has a unique id, also upon changing practice or re-registration.
9	Quality control (data source specific)	Internal and external validation processes are carried out to determine the data quality of the SIDIAP information at each data update. These include stratifying the data by geographical regions and year in order to identify differences in data collection that need to be harmonized (e.g. recording of specific information under different codes). The measurement units of variables measuring one characteristic are also homogenized (e.g. transformation of the data from every laboratory that measures haemoglobin to grams per decilitre). Visual inspection of all data included in the database by week is also conducted, allowing one to see temporal patterns in the registry of a certain variable. With this information, the SIDIAP team can issue recommendations to researchers about the most common variable(s) where certain information is recorded (e.g., there are several variables with information concerning the women's menopausal status and with these visual inspection tools the SIDIAP team can inform the researchers about which related variables have the largest number of records and could be more helpful to capture menopause). Data availability (longitudinally and reliability), plausibility (range checks and unusual values), and consistency are inspected through visualisation tools. In addition, before accessing the data for a requested project, research teams have access to a quality-control report. This document contains counts, years, percentiles, maximums and minimums, incidences, and prevalence of the data requested for the project, allowing detection of inconsistencies in the data extraction prior to data delivery. External validation processes of the SIDIAP database mainly include assessing the data recorded in SIDIAP through linkage to external gold standard data sources, by analysing free text, or by sending questionnaires to health professionals.
10	Linkage	SIDIAP is linked to a hospital discharge database, pharmacy dispensation, and primary care laboratories. It can also be linked to other registries in Catalonia on a project by project basis.
11	Vital status	Mortality is fully captured in SIDIAP. The cause of death is not available but can be linked to the Spanish death registry on a project by project basis.
12	Limitations	The SIDIAP data is not representative of individuals not using public primary care, and conditions that are usually followed by specialist care might not be properly captured. In addition, there is limited information on lifestyle variables. Patients are followed until Death or when transferring to another primary health care centre that does not contribute to SIDIAP.
13	Main references	Recalde M, Rodríguez C,Burn E,Far M,García D,Carrere-Molina J,Benítez M,Moleras A,Pistillo A,Bolíbar B,Aragón M,Duarte-Salles T "Data Resource Profile: The Information System for Research in Primary Care (SIDIAP)." International journal of epidemiology (2022): 35415748
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/50190 Website: https://www.sidiap.org/index.php/en

Valencia Health System Integrated Dataset (VID)

#	Section	Description
1	Data source identification and country	VID (Valencia Health System Integrated Dataset) Comunitat Valenciana, Spain
2	Data partner information section	FISABIO Health Services Research & Pharmacoepidemiology Unit
3	Coverage and timespan	Data collection since: 2009 Extent: Regional. The VID covers the general population of the Valencia region, comprising 10.7% of the Spanish population. The total population is estimated to be around 5,300,000.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). Both primary and secondary care settings are covered, where visits, diagnoses, medications, measurements, and procedures are recorded. The population information system collects sociodemographics, health coverage, and mortality data. The Electronic prescription and dispensing system captures all information related to medication (active ingredient, strength, duration, indication, etc.). Emergency department and hospital admissions are registered, providing information on dates, diagnoses, and procedures. Measurements are captured additionally from the vaccine information system and the Microbiological surveillance network. Mortality is also captured.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and Other. Data extraction is performed by clinical IT personnel. Data is released by the health authorities on a project basis and can only be used for such purposes.
6	General representativeness	The population captured by the VID should represent the Valencia region well, as the VID contains data of the general population covered by the universal public health care system. About 97% of the population in this region is covered by public care.
7	Data content /source coding	Prescribed and dispensed medications are coded with the ATC system. The indications of each prescription, as well as procedures are coded using ICD9CM and ICD10ES.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a unique id between practices.
9	Quality control (data source specific)	The data is reviewed carefully with the IT personnel who perform the extraction of data and then by a senior researcher with expertise in RWD management in the HSRP unit. Several quality check scripts are run against the received data. Finally, a senior researcher with RWD and clinical expertise assesses the completeness, consistency, and quality of the data extraction. If any inconsistency or error is detected, the dataset is requested and extracted again.
10	Linkage	VID also contains hospital discharge records, emergency care discharge records, birth registry, congenital anomaly registry, perinatal mortality registry, cancer registry, pharmacy prescription and dispensing records, vaccine records, and microbiology records. . Mother- and father-child linkage is also available. Most databases are updated daily, but certain registries, such as the congenital anomaly registry and perinatal mortality registries, are updated yearly.
11	Vital status	Mortality dates and causes of death are available in the mortality registry and the perinatal mortality registry.
12	Limitations	A subgroup of women (born before 1953) is not mapped into OMOP CDM yet. Note that another DARWIN Data Partner, BIFAP, also covers the Valencia region and patient information will overlap with VID. Biological sex is not captured, only has the legal gender. The last year's information for gender is used and can change upon data refresh.

#	Section	Description
13	Main references	García-Sempere A, Orrico-Sánchez A,Muñoz-Quiles C,Hurtado I,Peiró S,Sanfélix-Gimeno G,Diez-Domingo J "Data Resource Profile: The Valencia Health System Integrated Database (VID)." International journal of epidemiology (2020): 31977043
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111174 Website: https://www.san.gva.es/ca/web/salut-publica

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

#	Section	Description
1	Data source identification and country	HI-SPEED (Health Impact - Swedish Population Evidence Enabling Data-linkage) Sweden
2	Data partner information section	Pharmacoepidemiology and Analysis Department (FeA), SMPA-GU, Läkemedelsverket, Box 26, 751 03 Uppsala, Sweden School of Public Health and Community Medicine, Institute of Medicine, Box 469, 405 30 Gothenburg, Sweden
3	Coverage and timespan	Data collection since: 2015 Extent: Nation-wide. The catchment area includes the whole of Sweden, covering the full population of approximately 11.7 million.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Primary care (GPs) is available only for the 2 largest regions (~40% of national population) The following data elements are collected: Socio-demographics, dispensed drug prescriptions, cause of death, diagnoses and procedures from secondary (specialist) care and inpatient visits or clinical events, as well as from primary care visits (40%pop only).
5	Data collection process	Registries. The data is acquired from the relevant Swedish national and regional registries, only once all legislative, GDPR and ethical approvals have been granted. Therefore, only relevant data is passed on, which will then be entered and processed by the study team. The data are updated several times annually.
6	General representativeness	The coverage includes all patients of all sociodemographic characteristics. Therefore, it should mirror the source population to a very good extent.
7	Data content /source coding	Medicines are coded with ATC and NPLID (National Product ID), ICD10-SE is used for diagnoses, and the Swedish procedure coding system (KVA) is used for clinical procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a uniquely id across datasets.
9	Quality control (data source specific)	The source data are obtained from the relevant Swedish National and Regional Registers. The registers perform some regular quality controls on their data. After receiving the data, we perform additional checks and cleaning. We also run regular quality checks on the data we manage.
10	Linkage	Data on specialist care is acquired from the National Patient Register; mortality information is provided by the Cause-Of-Death Registry. Drug data is provided by the National Prescribed Drug Register. Data are linked very accurately using the national personal ID number and pseudonymized before delivery to HI-SPEED. All data are updated 2-4 times per year.
11	Vital status	Data on date of death and underlying + contributing causes-of-death are extracted from the Cause-of-Death registry (i.e. based on death certificates).

#	Section	Description
12	Limitations	General limitations for the data type applicable. This is a research project where all studies require ethics approval. Data collection since: 2015 for most data, except prescribed drug register (from 2018), and some COVID-related data (tests, vaccination) from 2020 Primary care is only available for a subset.
13	Main references	Nyberg F, Franzén S,Lindh M,Vanfleteren L,Hammar N,Wettermark B,Sundström J,Santosa A,Björck S,Gisslén M "Swedish Covid-19 Investigation for Future Insights - A Population Epidemiology Approach Using Register Linkage (SCIFI-PEARL)." Clinical epidemiology (2021): 34354377
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/node/4463/ Website: https://www.gu.se/en/research/scifi-pearl

Clinical Practice Research Datalink GOLD (CPRD GOLD)

#	Section	Description
1	Data source identification and country	CPRD GOLD (Clinical Practice Research Datalink GOLD) The United Kingdom
2	Data partner information section	University of Oxford NDORMS
3	Coverage and timespan	Data collection since: 1987 Extent: Nation-wide. CPRD GOLD consists of patients in contributing practices using Vision software. Historically this covered the whole of the UK, but the number of contributing practices in the England is dropping. In January 2025 only 3 practices from England were a part of CPRD GOLD, while historical patient data were from the whole of the UK, and will continue to be so. In the future, no practices from England will be present, only practices from Scotland, Wales, and Northern Ireland.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. CPRD GOLD data include patient demographics, biological measurements, clinical symptoms and diagnoses, referrals to specialist/hospital and their outcome, laboratory tests/results, and prescribed medications.
5	Data collection process	Outpatient electronic health records. Data are entered by clinicians into the EHR. Data is processed by CPRD that provides data releases for research.
6	General representativeness	In the last 10 years, the CPRD GOLD regional distribution of currently contributing general practitioner (GP) practices has significantly shifted, resulting in many new practices joining from Scotland, Wales, and Northern Ireland, and fewer participating from England. These changes have affected the CPRD GOLD population size, regional coverage, and eligibility for data linkages. CPRD GOLD January 2024 contains >21.3 million historical and current patients (12.9 in England, 3.1 in Wales, 4.7 in Scotland, 0.7 in Northern Ireland). Of these, nearly 3 million are currently registered in a GP practice and represent ~4.3% of the estimated current UK population (0.1% in England, 32.3% in Wales, 28.6% in Scotland, 16.2% in Northern Ireland). Patients currently registered in CPRD GOLD January 2024 are broadly representative of the UK population with respect to age and sex. Reference: https://doi.org/10.1093/ije/dyaf077
7	Data content /source coding	Gemscript, Read, dm+d
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified

#	Section	Description
		upon onboarding into the DARWIN EU® data network. In GOLD, a patient can be registered under different ID numbers upon changing practice or re-registration. Researchers are not able to identify these patients, as the data are anonymised. However, GOLD covers less than 5% of the current UK GP practices and it is unlikely that an individual who does change GP practice ends up in another GP practice which uses the Vision software and accepts the CPRD data collection agreement. The very small number of duplicated IDs will have different observation periods and should not have an impact on the data analyses.
9	Quality control (data source specific)	CPRD GOLD only includes practices whose data quality is assessed to be up-to-standard (UTS). Each practice is associated to an UTS date set when the data quality standards become satisfactory, and CPRD recommend using only longitudinal data starting from this UTS date. Every time CPRD collect the EHR from a practice, checks are run for the data quality standards, and if they are not adequate, the EHR is not accepted. When the data quality becomes acceptable again, CPRD updates the practice UTS date. CPRD also checks data quality standards at the patient level, and associates each patient with a flag, reporting if its data are acceptable for clinical research. Only patients with acceptable data quality are included in the population to be mapped to CDM.
10	Linkage	CPRD GOLD can be linked to several sources, however our Oxford OMOP CDM is only linked to the CPRD GOLD Ethnicity Record and to the CPRD Townsend Deprivation Index at the Practice Level.
11	Vital status	The date of death in CPRD GOLD has been validated against the Population registry (ONS) mortality data. Reference: https://doi.org/10.1002/pds.4747
12	Limitations	The main limitation is due to the fact that CPRD GOLD is limited to GP records, and although it contains information on referrals and discharge letters, it may not fully capture specific hospital information. Events from hospital and specialist care are not covered.
13	Main references	Sanchez-Santos MT, Axson EL,Dedman D,Delmestri A "Data Resource Profile Update: CPRD GOLD." International journal of epidemiology (2025): 40499193
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111113 Website: https://www.cprd.com/data/primary-care-data/cprd-gold

ANNEX II. Fitness for use assessment

Data source justification for inclusion and key characteristics

IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium)

IQVIA LPD Belgium will be included in this study because it is a nation-wide primary care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in IQVIA LPD Belgium is sufficient, as data availability in IQVIA LPD Belgium starts in 2005, and the date of the most recent data extraction is 07/05/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in IQVIA LPD Belgium is 870 days (IQR 77–2,200), taken from the Portal.

There are no specific limitations present in IQVIA LPD Belgium.

Lastly, IQVIA LPD Belgium does not need approval for DARWIN EU® studies, which makes the execution of this study feasible within the current study timelines.

Croatian National Public Health Information System (NAJS)

NAJS will be included in this study because it is a nation-wide primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in NAJS is sufficient, as data availability in NAJS starts in 1998, and the date of the most recent data extraction is 14/08/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in NAJS is 3,840 days (IQR 3,310–3,930), taken from the Portal.

With regard to specific limitations of NAJS, a previous study within the network has shown that cancer-related records may include both unconfirmed diagnoses and follow-up visits after remission, potentially leading to an overestimation of cancer cases within the data source. However, follow-up visits after remission should not affect the definition of disease occurrence for this study, as the analysis aims to describe the history of a given condition.

Lastly, NAJS has umbrella approval for DARWIN EU® studies, which makes the execution of this study feasible within the current study timelines.

Estonian Biobank (EBB)

EBB will be included in this study because it is a nation-wide biobank data source covering primary care, secondary care, and hospital inpatient care that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in EBB is sufficient, as data availability in EBB starts in 2004, and the date of the most recent data extraction is 27/02/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in EBB is 7,300 days (IQR 7,300–7,300), taken from the Portal.

No specific limitations were identified for EBB. However, it is important to note for the interpretation of results that female participants are over-represented in EBB and that older individuals tend to participate less frequently; however, all age groups are represented.

Lastly, EBB has blanket approval, which makes the execution of this study feasible within the current study timelines.

Assistance publique Hôpitaux de Marseille (APHM)

APHM will be included in this study because it is a hospital data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in APHM is sufficient, as data availability in APHM starts in 2014, and the date of the most recent data extraction is 11/01/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in APHM is 129 days (IQR 0–1,500), taken from the Portal.

There are no specific limitations present in APHM.

Lastly, APHM can obtain IRB approval within one month, which makes the execution of this study feasible within the current study timelines.

InGef Research Database (InGef RDB)

InGef RDB will be included in this study because it is a nation-wide primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in InGef RDB is sufficient, as data availability in InGef RDB starts in 2016, and the date of the most recent data extraction is 02/01/2026 (as of 01/2026), which aligns with the study period. The median follow-up of the first observation period in InGef RDB is 3,380 days (IQR 1,280–3,560), taken from the Portal.

With regard to specific limitations of InGef RDB, it should be noted that outpatient diagnoses are recorded using the last day of the corresponding calendar quarter rather than the exact diagnosis date, resulting in an inaccurate index date.

Lastly, InGef RDB has umbrella approval for DARWIN EU® studies, which makes the execution of this study feasible within the current study timelines.

IQVIA Disease Analyzer Germany (IQVIA DA Germany)

IQVIA DA Germany will be included in this study because it is a nation-wide primary care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in IQVIA DA Germany is sufficient, as data availability in IQVIA DA Germany starts in 1989, and the date of the most recent data extraction is 17/10/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period IQVIA DA Germany is 121 days (IQR 0–1,570), taken from the Portal.

There are no specific limitations present in IQVIA DA Germany.

Lastly, IQVIA DA Germany does not need approval for DARWIN EU® studies, which makes the execution of this study feasible within the current study timelines.

Papageorgiou General Hospital (PGH)

PGH will be included in this study because it is a hospital data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in PGH is sufficient, as data availability in PGH starts in 1999, and the date of the most recent data extraction is 03/05/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in PGH is 56 days (IQR 0–2,310), taken from the Portal.

With regard to specific limitations of PGH, medical history information is missing, as expected for a hospital-based data source.

Lastly, PGH can obtain IRB approval within one month, which makes the execution of this study feasible within the current study timelines.

Semmelweis University Clinical Data (SUCD)

SUCD will be included in this study because it is a hospital data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in SUCD is sufficient, as data availability in SUCD starts in 2010, and the date of the most recent data extraction is 28/07/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in SUCD is 243 days (IQR 0–2.130), taken from the Portal.

There are no specific limitations present in SUCD.

Lastly, SUCD can obtain IRB approval within one month, which makes the execution of this study feasible within the current study timelines.

Integrated Primary Care Information (IPCI)

IPCI will be included in this study because it is a nation-wide primary care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in IPCI is sufficient, as data availability in IPCI starts in 2006, and the date of the most recent data extraction is 22/10/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in IPCI is 1,790 days (IQR 791–3,160), taken from the Portal.

There are no specific limitations present in IPCI.

Lastly, IPCI has blanket approval, which makes the execution of this study feasible within the current study timelines.

Norwegian Linked Health Registry data (NLHR)

NLHR will be included in this study because it is a nation-wide primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in NLHR is sufficient, as data availability in NLHR starts in 2008, and the date of the most recent data extraction is 28/01/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in NLHR is 5,820 days (IQR 4,200–5,820), taken from the Portal.

With regard to specific limitations of NLHR, hospital data are available only from 2018. As a result, for clinical conditions predominantly diagnosed in hospital settings, the history of disease described in the 2023 period prevalence estimates will be limited to diagnoses recorded from 2018 onwards among individuals under observation in 2023.

Lastly, NLHR can obtain IRB approval within one month, which makes the execution of this study feasible within the current study timelines.

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

EMDB-ULSEDV will be included in this study because it is a hospital data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in EMDB-ULSEDV is sufficient, as data availability in EMDB-ULSEDV starts in 1999, and the date of the most recent data extraction is 20/10/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in EMDB-ULSEDV is 3,680 days (IQR 499–6,630), taken from the Portal.

There are no specific limitations present in EMDB-ULSEDV.

Lastly, EMDB-ULSEDV can obtain IRB approval within two months, which makes the execution of this study feasible within the current study timelines.

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

EMDB-ULSGE will be included in this study because it is a hospital data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability in EMDB-ULSGE starts in 2004, and the date of the most recent data extraction is 01/11/2023 (as of 12/2025). As this extraction does not cover the entire study period, an updated release of the OMOP CDM will be required prior to study execution. The median follow-up of the first observation period in EMDB-ULSGE is 1,590 days (IQR 58–4,770), taken from the Portal.

No other specific limitations of EMDB-ULSGE were identified.

Lastly, EMDB-ULSGE can obtain IRB approval within two months, which makes the execution of this study feasible within the current study timelines.

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

ULSM-RT will be included in this study because it is a hospital data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in ULSM-RT is sufficient, as data availability in ULSM-RT starts in 1998, and the date of the most recent data extraction is 31/08/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in ULSM-RT is 1,550 days (IQR 8–5,310), taken from the Portal.

There are no specific limitations present in ULSM-RT.

Lastly, ULSM-RT can obtain IRB approval within two months, which makes the execution of this study feasible within the current study timelines.

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

BIFAP will be included in this study because it is a regional primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in BIFAP is sufficient, as data availability in BIFAP starts in 2001, and the date of the most recent data extraction is 31/10/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in BIFAP is 2,550 days (IQR 2,050–5,430), taken from the Portal.

No specific limitations were identified for BIFAP. However, it is important to note for the interpretation of results that it includes an influx of approximately six million patients from the Valencian region in 2019, for whom the observation period begins in 2019, resulting in shorter available historical follow-up for these individuals.

Lastly, BIFAP can obtain IRB approval within one month, which makes the execution of this study feasible within the current study timelines.

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

PRISIB will be included in this study because it is a regional primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Data availability in PRISIB starts in 2008, and the date of the most recent data extraction is 01/12/2023 (as of 12/2025). As this extraction does not cover the entire study period, an updated release of the OMOP CDM will be required prior to study execution. The median follow-up of the first observation period in PRISIB is 3,620 days (IQR 1,590–4,700), taken from the Portal.

No other specific limitations of PRISIB were identified.

Lastly, PRISIB can obtain IRB approval within two months, which makes the execution of this study feasible within the current study timelines.

The Information System for Research on Primary Care (SIDIAP)

SIDIAP will be included in this study because it is a regional primary care data source linked to hospital discharge records that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in SIDIAP is sufficient, as data availability in SIDIAP starts in 2006, and the date of the most recent data extraction is 31/12/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in SIDIAP is 5,760 days (IQR 2,170–6,940), taken from the Portal.

There are no specific limitations present in SIDIAP.

Lastly, ethics approval for SIDIAP is estimated to take 60 days from ethics submission of protocol, which makes the execution of this study feasible within the current study timelines.

Valencia Health System Integrated Dataset (VID)

VID will be included in this study because it is a regional primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in VID is sufficient, as data availability in VID starts in 2009, and the date of the most recent data extraction is 01/10/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in VID is 5,840 days (IQR 4,460–5,840), taken from the Portal.

With regard to specific limitations of VID, a subgroup of women born before 1953 has not yet been mapped to the OMOP CDM, and this needs to be considered when interpreting the results.

Lastly, VID can obtain IRB approval within two weeks, which makes the execution of this study feasible within the current study timelines.

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

HI-SPEED will be included in this study because it is a regional primary, secondary, and hospital inpatient care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in HI-SPEED is sufficient, as data availability in HI-SPEED starts in 2015, and the date of the most recent data extraction is 03/10/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in HI-SPEED is 3,530 days (IQR 3,420–3,530), taken from the Portal.

There are no specific limitations present in HI-SPEED. However, it is important to note for the interpretation of results that primary care data are available only for the two largest regions, covering approximately 40%

of the national population. As a result, conditions primarily captured in primary care may not be representative of the national population but instead reflect patterns from these two regions.

Lastly, HI-SPEED can obtain IRB approval within one month, which makes the execution of this study feasible within the current study timelines.

Clinical Practice Research Datalink GOLD (CPRD GOLD)

CPRD GOLD will be included in this study because it is a nation-wide primary care data source that provides relevant information for the characterisation of the DARWIN EU® network.

Moreover, data availability and follow-up in CPRD GOLD is sufficient, as data availability in CPRD GOLD starts in 1987, and the date of the most recent data extraction is 15/08/2025 (as of 12/2025), which aligns with the study period. The median follow-up of the first observation period in CPRD GOLD is 2,150 days (IQR 728–4,940), taken from the Portal.

There are no specific limitations present in CPRD GOLD. However, it is important to note for the interpretation of results that the current patient population is predominantly drawn from Scotland, Wales, and Northern Ireland, while historically CPRD GOLD has also included data from England.

Lastly, CPRD GOLD can obtain IRB approval within two months, which makes the execution of this study feasible within the current study timelines.

ANNEX III. Operational and reporting considerations

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 European Union (EU) member states will process personal data from individuals that 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® Digital Research Environment (DRE). These output files do not contain any data that allow identification of subjects included in the study. The Digital Research Environment (DRE) 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

Data source 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 package allows the user to define a search strategy and will use this to query the vocabulary tables of the OMOP common data model so as to find potentially relevant codes. In addition, the *PhenotypeR* (https://github.com/OHDSI/phenotypeR) R package will be run 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 DARWIN EU® R packages: *IncidencePrevalence* to estimate Incidence and Prevalence, and *CohortCharacteristics* to characterise the cohort by indication. 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® Coordination Centre (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.

In addition, we plan to prepare a manuscript and an abstract for submission to a scientific conference.

ANNEX IV. Preliminary list of condition definitions

Table S1. Preliminary list of condition definitions.

Phenotype	Concept name	Concept ID (including descendants)	Exclude concept ID	Vocabulary
Cancers				
Colorectal and anus cancer	Malignant colorectal neoplasm	37168850	NA	SNOMED
Colorectal and anus cancer	Malignant neoplasm of anorectum	40481902	NA	SNOMED
Breast cancer	Malignant tumor of breast	4112853	NA	SNOMED
Prostate cancer	Malignant neoplasm of prostate	4163261	NA	SNOMED
Lung cancer	Malignant neoplasm of lung	443388	NA	SNOMED
Diseases of the circulatory system				
Atrial fibrillation	Atrial fibrillation	313217	NA	SNOMED
Cardiac arrhythmia	Cardiac arrhythmia	44784217	NA	SNOMED
Coronary heart disease not otherwise specified	Coronary arteriosclerosis	317576	NA	SNOMED
Heart failure	Heart failure	316139	NA	SNOMED
Hypertension	Hypertensive disorder	316866	NA	SNOMED
Ischaemic stroke	Ischemic stroke	4310996	NA	SNOMED
Haemorrhagic stroke	Haemorrhagic stroke	35609033	NA	SNOMED
Myocardial infarction	Myocardial infarction	4329847	NA	SNOMED
Peripheral arterial disease	Peripheral arterial disease	3654996	NA	SNOMED
Stable angina	Stable angina	4119942	NA	SNOMED
Transient ischaemic attack	Transient cerebral ischemia	373503	NA	SNOMED
Unstable angina	Unstable angina co-occurrent and due to coronary arteriosclerosis	36712982	NA	SNOMED
Unstable angina	Unstable angina co-occurrent and due to arteriosclerosis of coronary artery bypass graft	35615053	NA	SNOMED
Unstable angina	Unstable angina due to arteriosclerosis of autologous arterial coronary artery bypass graft	608953	NA	SNOMED
Unstable angina	Unstable angina due to arteriosclerosis of autologous vein coronary artery bypass graft	37309713	NA	SNOMED
Diseases of the digestive system				
Cholecystitis	Cholecystitis	192956	NA	SNOMED
Chronic liver disease (excluding chronic viral hepatitis)	Chronic liver disease	4212540	Chronic viral hepatitis (4012113) including descendants	SNOMED

Phenotype	Concept name	Concept ID (including descendants)	Exclude concept ID	Vocabulary
Crohn's disease	Crohn's disease	201606	NA	SNOMED
Gastro-oesophageal reflux disease	Gastroesophageal reflux disease	318800	NA	SNOMED
Metabolic dysfunction-associated steatohepatitis	Metabolic dysfunction-associated steatohepatitis	40484532	NA	SNOMED
Ulcerative colitis	Ulcerative colitis	81893	NA	SNOMED
Diseases of the endocrine system				
Hypercholesterolaemia	Hypercholesterolemia	4029305	NA	SNOMED
Hyperlipidaemia	Hyperlipidemia	432867	NA	SNOMED
Hypothyroidism	Hypothyroidism	140673	NA	SNOMED
Obesity	Obesity	433736	NA	SNOMED
Type 1 diabetes mellitus	Type 1 diabetes mellitus	201254	NA	SNOMED
Type 2 diabetes mellitus	Type 2 diabetes mellitus	201826	NA	SNOMED
Diseases of the genitourinary system				
Acute kidney injury	Acute kidney injury	197320	NA	SNOMED
Chronic kidney disease	Chronic kidney disease	46271022	NA	SNOMED
Benign prostatic hyperplasia	Benign prostatic hyperplasia	198803	NA	SNOMED
Diseases of the respiratory system				
Asthma	Asthma	317009	NA	SNOMED
Chronic obstructive pulmonary disease	Chronic obstructive pulmonary disease	255573	NA	SNOMED
Haematological conditions				
Anaemia (limited to nutritional and metabolic anaemias, including iron, vitamin B12, and folate deficiency)	Nutritional anemia	4280354	NA	SNOMED
Anaemia (limited to nutritional and metabolic anaemias, including iron, vitamin B12, and folate deficiency)	Iron deficiency anemia	436659	NA	SNOMED
Anaemia (limited to nutritional and metabolic anaemias, including iron, vitamin B12, and folate deficiency)	Megaloblastic anemia due to vitamin B-12 deficiency	432588	NA	SNOMED
Anaemia (limited to nutritional and metabolic anaemias, including iron, vitamin B12, and folate deficiency)	Megaloblastic anemia due to folate deficiency	440977	NA	SNOMED
Infectious diseases				
Chronic viral hepatitis	Chronic viral hepatitis	4012113	NA	SNOMED
Human immunodeficiency virus	Human immunodeficiency virus infection	439727	NA	SNOMED
Mental health disorders				
Alcohol use-related disorders	Disorder caused by alcohol	36714559	NA	SNOMED
Anxiety disorders	Anxiety	441542	NA	SNOMED

Phenotype	Concept name	Concept ID (including descendants)	Exclude concept ID	Vocabulary
Bipolar affective disorder and mania	Bipolar disorder	436665	NA	SNOMED
Bipolar affective disorder and mania	Mania	4333677	NA	SNOMED
Depression	Depressive disorder	440383	NA	SNOMED
Schizophrenia, schizotypal, and delusional disorders	Delusional disorder	432590	NA	SNOMED
Schizophrenia, schizotypal, and delusional disorders	Schizophrenia	435783	NA	SNOMED
Schizophrenia, schizotypal, and delusional disorders	Schizotypal personality disorder	434010	NA	SNOMED
Musculoskeletal conditions				
Osteoporosis	Osteoporosis	80502	NA	SNOMED
Rheumatoid arthritis	Rheumatoid arthritis	80809	NA	SNOMED
Neurological conditions				
Alzheimer’s dementia	Alzheimer's disease	378419	NA	SNOMED
Dementia	Dementia	4182210	NA	SNOMED
Epilepsy	Epilepsy	380378	NA	SNOMED
Migraine	Migraine	318736	NA	SNOMED
Parkinson’s disease	Parkinson’s disease	381270	NA	SNOMED
Skin conditions				
Acne	Acne	141095	NA	SNOMED
Atopic dermatitis	Atopic dermatitis	133834	NA	SNOMED

NA = Not Applicable.

ANNEX V. ENCePP checklist for study protocols

European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Checklist for Study Protocols (Revision 4)

Study title: DARWIN EU® - Population demographics and disease frequency across the DARWIN EU® network

EU Post-Authorisation Studies (EU PAS) Register® number: EUPAS1000000962
Study reference number (if applicable): P4-C3-006, P4-C2-019, P4-C2-020

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]	☒	☐	☐	8.5
1.1.2 End of data collection[2]	☒	☐	☐	8.5
1.1.3 Progress report(s)	☐	☐	☒	
1.1.4 Interim report(s)	☐	☐	☒	
1.1.5 Registration in the EU PAS Register®	☒	☐	☐	
1.1.6 Final report of study results.	☒	☐	☐	5, 8.8.4

Comments:

There will be an Interim Preliminary Results Review (Shiny App) (Section 5).

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)	☒	☐	☐	6
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)	☒	☐	☐	8.3
2.1.4 Which hypothesis(-es) is (are) to be tested?	☐	☐	☒	
2.1.5 If applicable, that there is no *a priori* hypothesis?	☐	☐	☒	

Comments:

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

[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**
3.2 Does the protocol specify whether the study is based on primary, secondary, or combined data collection?	☒	☐	☐	8.4
3.3 Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.8.3
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?	☒	☐	☐	
4.2 Is the planned study population defined in terms of:				
4.2.1 Study time period	☒	☐	☐	8.3, 8.5
4.2.2 Age and sex	☒	☐	☐	8.3
4.2.3 Country of origin	☒	☐	☐	8.3, 8.4
4.2.4 Disease/indication	☒	☐	☐	8.3, 8.6.1
4.2.5 Duration of follow-up	☒	☐	☐	8.2, 8.3
4.3 Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	☒	☐	☐	8.3

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)	☐	☐	☒	
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)	☐	☐	☒	

Section 5: Exposure definition and measurement	**Yes**	**No**	**N/A**	**Section Number**
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?	☒	☐	☐	8.6.1
6.2 Does the protocol describe how the outcomes are defined and measured?	☒	☐	☐	8.6.1
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)	☐	☐	☒	
7.3 Does the protocol address information bias? (e.g. misclassification of exposure and outcomes, time-related bias)	☐	☐	☒	

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)	☐	☐	☒	

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)	☐	☐	☒	
9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	☒	☐	☐	11, 8.6.1
9.1.3 Covariates and other characteristics?	☒	☐	☐	11, 8.6.2
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)	☐	☐	☒	
9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	☒	☐	☐	11, 8.6.1
9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	☒	☐	☐	11, 8.6.2
9.3 Is a coding system described for:				
9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	☐	☐	☒	
9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	☒	☐	☐	11
9.3.3 Covariates and other characteristics?	☒	☐	☐	11
9.4 Is a linkage method between data sources described? (e.g. based on a unique identifier or other)	☐	☐	☒	

Comments:

Section 10: Analysis plan	**Yes**	**No**	**N/A**	**Section Number**
10.1 Are the statistical methods and the reason for their choice described?	☒	☐	☐	8.8.3
10.2 Is study size and/or statistical precision estimated?	☐	☐	☒	
10.3 Are descriptive analyses included?	☒	☐	☐	8.8.3
10.4 Are stratified analyses included?	☐	☐	☒	
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?	☒	☐	☐	8.8.3

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)	☒	☐	☐	11
11.2 Are methods of quality assurance described?	☒	☐	☐	11
11.3 Is there a system in place for independent review of study results?	☐	☐	☒	

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?	☒	☐	☐	9
12.1.2 Information bias?	☐	☐	☒	
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?	☒	☐	☐	11
13.2 Has any outcome of an ethical review procedure been addressed?	☐	☐	☒	
13.3 Have data protection requirements been described?	☒	☐	☐	8.8.2

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)?	☒	☐	☐	11
15.2 Are plans described for disseminating study results externally, including publication?	☒	☐	☐	11

Comments:

Name of the main author of the protocol: Anna Saura-Lazaro, Albert Prats-Uribe

Date: 27/01/2026

Signature: Anna Saura-Lazaro

ANNEX VI. 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, outlines, 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 outline 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 outline and study code, but with updated data and/or different data partners.

Study Outline

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.