



Study Protocol

P5-C3-006/P5-C2-004

DARWIN EU® - Identifying conditions of regulatory interest in Europe: a population-based cohort study

28/09/2026

Version 4.0

Authors: Alejandro Ballvé, Daniel Prieto-Alhambra, Elin Rowlands, Martí Català-Sabaté

Public

CONTENTS

LIST OF ABBREVIATIONS.....	5
1. TITLE	7
2. DESCRIPTION OF THE STUDY TEAM	7
3. ABSTRACT	9
4. AMENDMENTS AND UPDATES	11
5. MILESTONES	11
6. RATIONALE AND BACKGROUND	11
7. RESEARCH QUESTION AND OBJECTIVES.....	11
8. RESEARCH METHODS.....	12
8.1. Study design	12
8.2. Follow-up	13
Figure 1. Graphical description of study design related to Objective 1.	14
Figure 2. Included observation time for the denominator population for the estimation of period prevalence of selected clinical conditions (Objective 2).	15
Figure 3. Included observation time for the denominator population for the estimation of incidence rates of selected clinical conditions (Objective 3).	16
8.3. Study population with inclusion and exclusion criteria.....	16
8.4. Study setting and data sources	16
Table 1. Proposed data sources and their key characteristics.	17
8.5. Study period	18
8.6. Variables	18
8.6.1. Outcome	18
Table 2. List of outcomes and proposed flavours and subtypes.	19
8.6.2. Covariates	21
8.7. Study size	22
8.8. Analysis	22
8.8.1. Federated network analyses	22
8.8.2. Data privacy protection	23
8.8.3. Statistical model specification and assumptions of the analytical approach considered	23
Objectives 1 and 4	23
Objective 2.....	23
Objective 3.....	24
8.8.4. Output	24
MOCK Supplementary Table 1. Number of individuals under observation during the study period for each specific condition after applying each eligibility criteria (attrition table), by data source, for Objective 1.....	24
MOCK Figures 1a–1j. Population pyramid displaying the age distribution of the study population for each specific condition stratified by sex at the index date.	26
MOCK Figures 2a–2j. Bar chart displaying the distribution of subtypes for each main phenotype (where applicable).	27
MOCK Tables 1a–1x. Characteristics of the study populations with a first diagnosis of each specific condition during the study period.	28
MOCK Supplementary Tables 2a–2x. Top 10 differential features between cases and controls based on absolute standardised mean difference (ASMD) per phenotype (one table for each).....	37

MOCK Supplementary Tables 3a–3x. Top 10 most frequent concepts per phenotype (one table for each).	40
MOCK Figures 3a–3x. Period prevalence of each of the selected clinical conditions.	44
MOCK Supplementary Tables 4a–4x. Period prevalence of selected clinical conditions by data source.	45
MOCK Figures 4a–4x. Incidence of each of the selected clinical conditions (i.e. one figure for each of them) and colour-coded by data source.	47
MOCK Supplementary Tables 5a–5x. Incidence rates of selected clinical conditions by data source.	48
8.9. Evidence synthesis	50
9. STRENGTHS AND LIMITATIONS	50
Table 3. Limitations specific for each condition.	50
10. REFERENCES	53
11. ANNEXES	54
ANNEX I. Description of data sources	54
Croatian National Public Health Information System (NAJS)	54
Danish Data Health Registries (DK-DHR)	55
Finnish Care Register for Health Care (FinOMOP-THL)	56
InGef Research Database (InGef RDB)	58
Lithuanian Digital Health Platform (LDH)	60
Norwegian Linked Health Registry data (NLHR)	61
The Information System for Research in Primary Care (SIDIAP)	62
Valencia Health System Integrated Dataset (VID)	63
Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)	64
Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care (CPRD Aurum Linked)	66
ANNEX II. Fitness for use assessment	69
ANNEX III. Operational and reporting considerations	72
ANNEX IV. Suggested phenotyping plans	74
Spinal muscular atrophy (SMA)	74
Amyotrophic lateral sclerosis (ALS)	75
Multiple sclerosis (MS)	76
Prurigo nodularis (PN)	81
Bullous pemphigoid (BP)	81
Vitiligo	82
Lupus	83
Vasculitis	86
Psoriasis	89
Autoimmune hepatitis	90
Primary biliary cholangitis	91
Primary sclerosing cholangitis	92
Chronic viral hepatitis	92
Alcohol related liver disease (ALD)	93
Metabolic dysfunction-associated steatohepatitis	95
Heart failure (HF)	96
Cardio-renal-metabolic syndrome	97
ANNEX V. ENCePP checklist for study protocols	100
ANNEX VI. Glossary	106

Study title	DARWIN EU® - Identifying conditions of regulatory interest in Europe: a population-based cohort study
Protocol version	V4.0
Date	28/09/2026
EUPAS number	EUPAS1000001129
Active Substance	NA
Medicinal Product	NA
Research question and objectives	<u>Research question</u> The aim of this study is to develop standardised phenotypes for the identification and characterisation of conditions of regulatory interest using the DARWIN EU® Data Network, while assessing which data sources are most suitable to study each of these conditions. <u>Research objectives</u> The specific objectives of this study are: 1. To develop and characterise computable phenotypes to identify individuals affected by 17 conditions of regulatory interest. 2. To estimate the period prevalence of 17 conditions of regulatory interest in the period 2022–2025, overall, and by calendar year 3. To estimate the incidence rate of 17 conditions of regulatory interest in the period 2022–2025, overall, and by calendar year. 4. To estimate the cumulative incidence of clinically active multiple sclerosis and decompensated heart failure.
Countries of study	Denmark, Germany, Spain, Sweden, the United Kingdom Croatia, Finland, Lithuania, Norway, Spain
Authors	Alejandro Ballvé (a.ballve@darwin-eu.org) Daniel Prieto-Alhambra (d.prietoalhambra@darwin-eu.org) Elin Rowlands (e.rowlands@darwin-eu.org) Martí Català-Sabaté (m.catalasabate@darwin-eu.org)

LIST OF ABBREVIATIONS

Acronyms/terms	Description
ALD	Alcohol-related liver disease
ALP	Alkaline phosphatase
ALS	Amyotrophic lateral sclerosis
ALT	Alanine aminotransferase
ASMD	Absolute standardised mean difference
AST	Aspartate Aminotransferase
ATC	Anatomical Therapeutic Chemical
BP	Bullous pemphigoid
CC	Coordination Centre
CDM	Common Data Model
CKD	Chronic Kidney Disease
CKM	Cardiovascular-kidney-metabolic
CIPH	Croatian Institute of Public Health
CLE	Cutaneous lupus erythematosus
CPRD	Clinical Practice Research Datalink
CPRD Aurum Linked	Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DTZ	Data Transfer Zone
DK-DHR	Danish Data Health Registries
DKMA	Danish Medicines Agency
DRE	Digital Research Environment
EHR	Electronic Health Records
EMA	European Medicines Agency
ENCEPP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU	European Union
EUPAS	EU Post-Authorisation Studies Register
FinOMOP	Finnish Observational Medical Outcomes Partnership
FinOMOP-THL	Finnish Care Register for Health Care
FISABIO	Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana
FTD	Frontotemporal dementia
GGT	Gamma-glutamyl transferase
GP	General Practitioner
HES	Hospital Episode Statistics
HF	Heart failure
HI-SPEED	Health Impact - Swedish Population Evidence Enabling Data-linkage

Acronyms/terms	Description
ICD	International Classification of Diseases
ICD-10-GM	International Statistical Classification of Diseases and related health problems, 10 th revision, German Modification
IDIAPJGol	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina
IL	Interleukin
InGef	Institute for Applied Health Research Berlin GmbH
InGef RDB	InGef Research Database
IRB	Institutional Review Board
IQR	Interquartile Range
JAK	Janus kinase
LDH	Lithuanian Digital Health Platform
LOINC	Logical Observation Identifiers Names and Codes
MRI	Magnetic Resonance Imaging
MS	Multiple sclerosis
NAJS	Croatian National Public Health Information System
NLHR	Norwegian Linked Health Registry data
NSAIDs	Non-steroidal anti-inflammatory drugs
PD-1	Programmed Cell Death Protein 1
PD-L1	Programmed Death-Ligand 1
PN	Prurigo nodularis
PsA	Psoriatic arthritis
PEF	Preserved ejection fraction
REF	Reduced ejection fraction
SDA	State Data Agency
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
RxNorm	Medical prescription normalised
RWD	Real-World Data
SIDIAP	The Information System for Research in Primary Care
SLE	Systemic lupus erythematosus
SMA	Spinal muscular atrophy
SMPA-GU	Swedish Medical Products Agency - Gothenburg University
SNOMED	Systemised Nomenclature of Medicine
SOP	Standard operating procedure
TLR7	Toll-like receptor 7
UiO	University of Oslo
VID	Valencia Health System Integrated Dataset

1. TITLE

DARWIN EU® - Identifying conditions of regulatory interest in Europe: a population-based cohort study

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator	Alejandro Ballvé Daniel Prieto-Alhambra	University of Oxford
Data Scientist	Elin Rowlands Martí Català-Sabaté	University of Oxford
Clinical Domain Expert	Anna Saura-Lazaro Daniel Prieto-Alhambra Francesca d'Agnese Blanka Marton Hassan Shaffle	University of Oxford
Study Manager	Nathasa Yefimenko	Erasmus MC
Data source	Names	Data Partner Organisation*
Croatian National Public Health Information System (NAJS)	Ivan Pristaš Jakov Vuković Anamaria Jurčević Marko Čavlina Antea Jezidžić Karlo Pintarić	Croatian Institute of Public Health (CIPH)
Danish Data Health Registries (DK-DHR)	Elvira Bräuner Susanne Bruun	Danish Medicines Agency (DKMA)
Finnish Care Register for Health Care (FinOMOP-THL)	Toni Lehtonen Laura Salonen	FinOMOP
InGef Research Database (InGef RDB)	Raeleesha Norris Alexander Harms Annika Vivirito Viskovic, Vukasin	Institute for Applied Health Research Berlin GmbH (InGef)
Lithuanian Digital Health Platform (LDH)	Justas Trinkūnas Julius Juodakis	State Data Agency (SDA)
Norwegian Linked Health Registry data (NLHR)	Hedvig Marie Egeland Nordeng Saeed Hayati Peter Olteanu	University of Oslo (UiO)
The Information System for Research in Primary Care (SIDIAP)	Laura Granés González Agustina Giuliodori Picco Talita Duarte Salles Elena Roel Herranz	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol)

Valencia Health System Integrated Dataset (VID)	Gabriel Sanf��lix Gimeno Celia Robles Cabani��as Fran Llopis Cardona	Fundaci�� per al Foment de la Investigaci�� Sanitat��ria i Biom��dica de la Comunitat Valenciana (FISABIO)
Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)	Huiqi Li Fredrik Nyberg Rickard Ljung Ballin Marcel Talb��ck Mats	Swedish Medical Products Agency - Gothenburg University (SMPA-GU)
Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care (CPRD Aurum Linked)	Daniel Dedman Zara Cuccu Susan Hodgson Holly Bennett Anisha Khosla	Clinical Practice Research Datalink (CPRD)

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

3. ABSTRACT

Rationale and background

To use the DARWIN EU® Data Network for studying conditions of regulatory interest, we need to understand how to accurately identify them across different data sources and healthcare settings.

Research question and objectives

The aim of this study is to develop standardised phenotypes for the identification and characterisation of conditions of regulatory interest using the DARWIN EU® Data Network, while assessing which data sources are most suitable to study each of these conditions.

The specific objectives of this study are:

1. To develop and characterise computable phenotypes to identify individuals affected by 17 conditions of regulatory interest.
2. To estimate the period prevalence of 17 conditions of regulatory interest in the period 2022–2025, overall, and by calendar year
3. To estimate the incidence rate of 17 conditions of regulatory interest in the period 2022–2025, overall, and by calendar year
4. To estimate the cumulative incidence of clinically active multiple sclerosis and decompensated heart failure.

Methods

A cohort study using routinely collected health data from 10 data sources from 9 different countries across Europe. The study will include:

- Development, review, and characterisation of computable phenotypes and subtypes of regulatory interest. For some conditions, this characterisation will include the prescription of pre-defined drugs potentially associated with the conditions of interest as a proxy for the condition being drug induced.
- Disease occurrence analysis (incidence and prevalence), overall and stratified by calendar year, of all the phenotypes.

The four objectives will all contribute to the validation of the developed phenotypes.

The index date for Objective 1 will be the first recorded diagnosis of the target condition during the study period. Individuals will be followed up until the first of: one year after the index date, the end of the observation period, or the end of the study period.

Study population

Individuals under observation in the data sources at any time during the study period (1 January 2022–31 December 2025) who have ≥ 365 days of available history before the index date will be eligible. Those with a diagnosis of any of the conditions of interest during the study period will be included in Objective 1. Individuals with missing information on age or sex will be excluded. Additional exclusion criteria for Objective 1 and 3 will be individuals with a diagnosis of the same condition at any time before the index date ($-\infty$, -1), such that the diagnosis identified during the study period is incident (i.e., first diagnosis). For Objective 1, a population of sex- and age-matched individuals without the condition of interest and with the same index date will be defined for comparison of most frequent conditions.

Variables

Outcomes (and subtypes-):

Spinal muscular atrophy (type 1-4), amyotrophic lateral sclerosis (age at onset, associated/not with frontotemporal dementia), multiple sclerosis (with/without clinical activity), prurigo nodularis, bullous pemphigoid, vitiligo, lupus (systemic, cutaneous), vasculitis, psoriasis and psoriatic arthritis, autoimmune hepatitis, primary biliary cholangitis, primary sclerosing cholangitis, chronic viral hepatitis, alcohol related liver disease, metabolic dysfunction-associated steatohepatitis, heart failure (with preserved/reduced left ventricular ejection fraction, compensated/decompensated, right/left), cardio-renal-metabolic syndrome (cardiac-driven, renal-driven, metabolic-driven).

Relevant covariates:

- Age and sex.
- Medical history and medications, including the presence of certain pre-specified conditions in the year prior to the index date and initiations of new drugs known to be associated with some of the conditions (as a proxy for drug-induced phenotype)

Data sources

This study will be conducted using routinely collected data (nationwide and regional registries, primary care, and claims data sources) from 10 data sources within the DARWIN EU® Data Network, spanning 9 European countries: Denmark, Germany, Spain, Sweden, the United Kingdom, Croatia, Finland, Lithuania, Norway, and Spain.

Objectives 1 and 4

The resulting cohorts will be characterised at the index date and, where relevant, within pre-specified time windows before and after index date.

Objectives 2 and 3

Annual period prevalence and incidence rates of selected clinical conditions will be calculated for the total number of individuals in observation during the study period, overall and stratified by calendar year. Prevalence proportions and incidence rates will be reported with binomial and Poisson 95% confidence intervals, respectively.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study milestones and deliverables	Planned dates*
Final Study Protocol	10 September 2026
Clinical descriptions and Phenotype Plans Approval	10 September 2026
Code lists Peer Review and Approval	20 September 2026
Phenotypes Review and Approval	15 October 2026
Creation of Analytical code	15 October 2026
Execution of Analytical Code on the data	10 November 2026
Interim Study Report (Shiny App)	15 November 2026
Draft Study Report	30 November 2026 (to be confirmed by EMA)
Final Study Report	11 December 2026 (to be confirmed by EMA)

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

6. RATIONALE AND BACKGROUND

The DARWIN EU® Data Network comprises multiple Real-World Data sources that differ in data types, healthcare settings, coding practices, and population coverage.[1] Before conducting disease-specific or treatment-related studies, it is important to understand how to accurately identify the diseases of interest and which data sources are most suitable to study these conditions. Therefore, this study will serve as a critical first step in mapping the network's capacity to reliably capture conditions of regulatory interest. This knowledge can subsequently be used in other disease- or treatment-specific studies.

Furthermore, this study will support the development, documentation, and validation of standardised, fully reusable phenotypes and subtypes, which refer to subtypes of each disorder, for the selected clinical conditions within the DARWIN EU® Data Network. Testing and validation of these phenotypes and subtypes will be done through the characterisation of the individuals diagnosed with these conditions and via comparing estimated disease occurrence statistics (incidence/prevalence) with other published estimates. This process will involve characterising the population in terms of age, sex, and sub-phenotype distribution, as well as performing a large-scale characterisation at the time of first recorded diagnosis and estimating disease occurrence over a recent period. All this information will be contrasted with clinical and epidemiological descriptions, which are based on published literature and regulatory knowledge within the DARWIN EU® Coordination Centre, the local DARWIN EU® partners, and the European Medicines Agency (EMA). This will follow a standardised and reproducible phenotyping process, as previously described.[2]

7. RESEARCH QUESTION AND OBJECTIVES

Research question

The aim of this study is to develop standardised phenotypes for the identification and characterisation of conditions of regulatory interest using the DARWIN EU® Data Network, while assessing which data sources are most suitable to study each of these conditions.

Research objectives

The specific objectives of this study are:

1. To develop and characterise computable phenotypes to identify individuals affected by 17 conditions of regulatory interest.
2. To estimate the period prevalence of 17 conditions of regulatory interest in the period 2022–2025, overall, and by calendar year.
3. To estimate the incidence rate of 17 conditions of regulatory interest in the period 2022–2025, overall, and by calendar year.
4. To estimate the cumulative incidence of clinically active multiple sclerosis and decompensated heart failure.

8. RESEARCH METHODS

8.1. Study design

A cohort study will be conducted using routinely collected health data from 10 data sources from 9 countries across Europe. The study design will comprise:

- **Objectives 1 and 4:** Development, review, and characterisation of computable phenotypes and subtypes of regulatory interest in line with the DARWIN EU® Phenotyping Standard Operating Procedure (SOP) and will include the following steps:
 - Drafting and approval of clinical descriptions and phenotyping plans for each of the proposed phenotypes and subtypes. For some conditions, this characterisation will include the prescription of pre-defined drugs potentially associated with the conditions of interest during the 30 and 90 days prior to the index date. For some conditions, the characterisation will include the description of pre-defined comorbidities of interest. Notably, the appropriate definition of some subtypes (i.e. multiple sclerosis (MS), heart failure (HF), and amyotrophic lateral sclerosis (ALS)) will need to incorporate data beyond the index date. In other words, prospective longitudinal data will be needed for some subtypes.
 - Generation, peer review, and approval of code lists for each of the proposed phenotypes and, where applicable, of subtypes, with their different potential variations (e.g. narrow versus broad, primary versus overall, etc.). Review and approval of the resulting phenotypes and subtypes and comparison against pre-specified clinical descriptions, including:
 - Descriptive epidemiology (age at onset, sex distribution, prevalence, incidence).
 - Pre-index characteristics (signs/symptoms, observations, medications), overall and compared to age-sex matched specific disease-free controls.
 - On and post-index characteristics (complications, medications, differential diagnoses), overall and compared to age-sex matched controls.
- **Objectives 2 and 3:** Descriptive epidemiology analysis, including:
 - Estimation of the period prevalence proportion of each phenotype, overall, and stratified by calendar year
 - Estimation of the background incidence rate of each phenotype, overall, and stratified by calendar year

The four objectives will contribute to the validation of the developed phenotypes. To ensure full reproducibility and traceability for the entire phenotype generation and validation pipeline, the R package *CodeListGenerator*, developed by DARWIN EU®, will be used for the generation and validation of code lists, and *PhenotypeR* will be used for a full phenotype review and validation, including benchmarking against

disease-free controls.[3, 4] Additionally, existing analytical pipelines will be used to generate results in line with the DARWIN EU® Catalogue of Standard Analytics, including *CohortConstructor* (for cohort curation), *CohortCharacteristics* and *CohortSurvival* (for characterisation, Objectives 1 and 4), and *IncidencePrevalence* (for Objectives 2 and 3).[5–8] The latter three packages have also been developed by DARWIN EU®.

8.2. Follow-up

Objectives 1 and 4: Characterisation of pre-, on, and post-index features

During the review of computable phenotypes, a characterisation of the individuals will be conducted to compare these with existing knowledge, as detailed in the pre-specified Clinical Descriptions (an example for multiple sclerosis (MS) can be found in [Annex IV](#)). This characterisation will use the date of the first diagnosis of the target condition/phenotype during the study period as the index date, as depicted in [Figure 1](#).

In general terms, Objective 1 characterisation will draw on the data present on the index date, as well as in the 365 days before and the 365 days after. This will allow for the identification of indicated treatments (medicines and non-medicinal treatments), as well as conditions for the identification of differential diagnoses (disqualifiers) and complications.

Additionally, the presence of certain pre-specified conditions of interest will be assessed in the window (-365, 0) for metabolic dysfunction-associated steatohepatitis, heart failure (HF), and cardio-renal-metabolic syndrome. For other conditions (i.e. prurigo nodularis (PN), bullous pemphigoid (BP), vitiligo, lupus, vasculitis, and psoriasis) the window for assessing pre-defined medications of interest (i.e. potentially associated with the outcome) will be (-90,0) and (-30,0). Finally, the window for the characterisation of the phenotypes of amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), and HF will include post-diagnosis follow up (-365, 730) to ensure that the necessary information for the subtypes is captured. The end of follow-up will be defined as the earliest of one year from the index date (two years for ALS with FTD, MS with clinical activity, and decompensated HF), the end of the study period, or the end of the individual observation period.

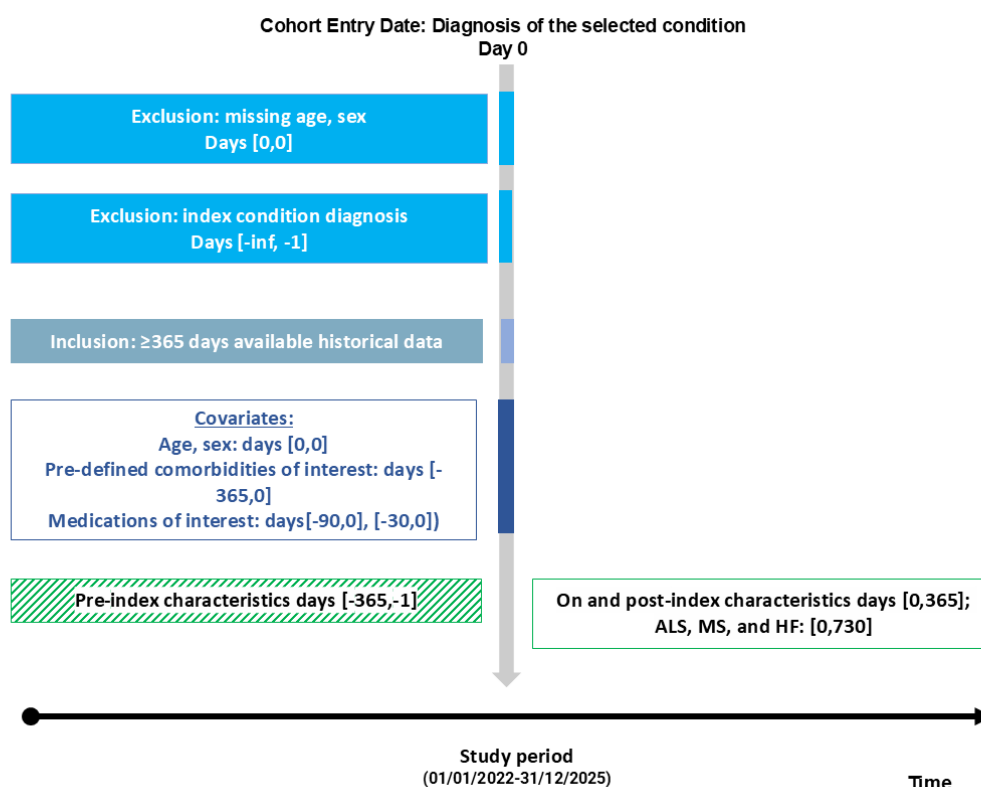


Figure 1. Graphical description of study design related to Objective 1.

ALS=Amyotrophic lateral sclerosis, HF=Heart failure, MS=Multiple sclerosis.

Objective 2 and 3: Descriptive epidemiology of selected clinical conditions

The index date for Objectives 2 and 3 will be the beginning of the study period (or of each calendar year of the study period for analyses stratified by calendar year).

Follow-up for the overall period prevalence estimation will begin on 1 January 2022 (or later, if an individual's eligible observation period starts after this date) and will end on 31 December 2025 (or earlier, if the observation ends before that date). Estimates stratified by calendar year will use annual periods, e.g. from 1 January 2023 (or entry date, if after 1 January 2023) to 31 December 2023 (or exit date, if before 31 December 2023) for the period prevalence in 2023.

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. Individuals will not be required to be present for the entire follow-up period to contribute to the prevalence estimates. See [Figure 2](#) for an illustration of follow-up periods for prevalence estimation. In this example, the observation period of person ID 1 starts before the beginning of the study period and ends after the end of the study period. Therefore, this individual will contribute the entire 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 only contribute to the denominator cohort 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.

Given that most, but not all, of the conditions of interest are chronic and not curable, an any-time prior lookback window will be used for each calendar year to estimate prevalence. However, some of the conditions (i.e. chronic viral hepatitis, metabolic dysfunction-associated hepatitis, alcohol related liver disease, PN, and BP) may be curable and/or self-limited. For these conditions, a 1-year lookback period will be applied instead.

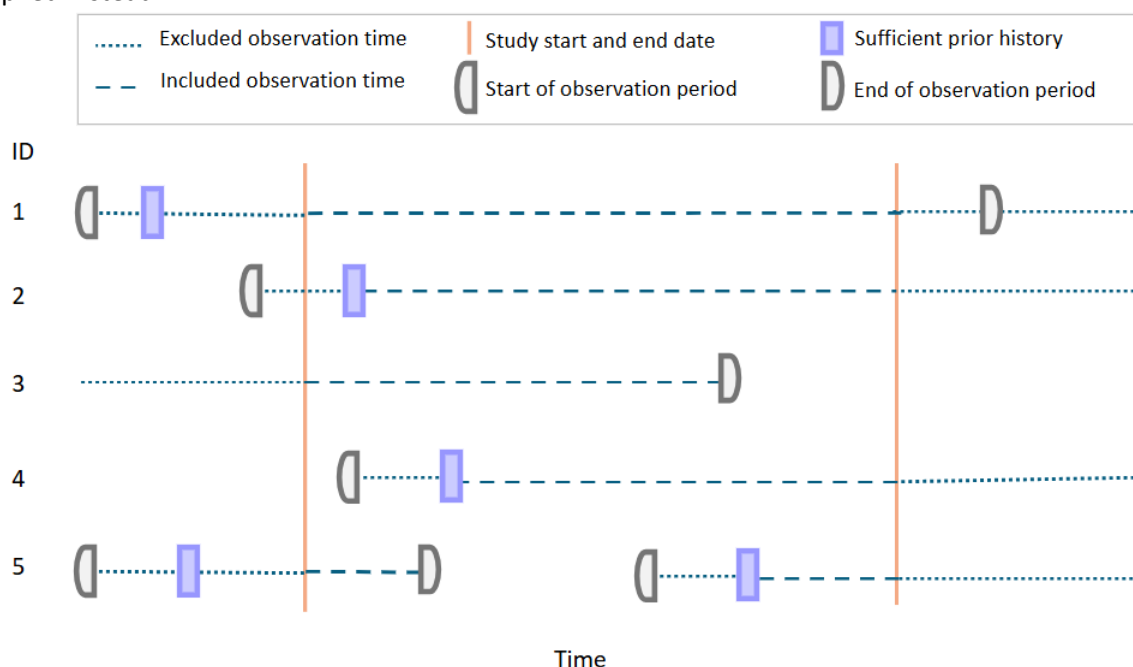


Figure 2. Included observation time for the denominator population for the estimation of period prevalence of selected clinical conditions (Objective 2).

In the analysis of incidence rates, individuals will begin contributing person-time on the latest of the study start date (1 January 2022) or date at which they have sufficient prior data availability (365 days). Individuals will be followed up until the earliest of first diagnosis of the condition of interest, death, transfer out of the data source, end of the observation period in the data source, or end of the study period (31 December 2025). Given the nature of the conditions of interest, individuals who had a previous record of the condition will not enter the cohort and, thus, will not contribute to the denominator. Similarly, individuals who develop the condition will not re-enter the cohort. See Figure 3 for an illustration of the follow-up periods used for incidence estimation. In this example, Person IDs 1 and 3 start contributing person-time in January 2022, whereas Person ID 2 starts contributing person-time later because they do not have sufficient prior observation at the study start date. The follow-up period for Person IDs 1 and 2 end before the study end date because they are diagnosed with a condition of interest or die, respectively. In contrast, the follow-up period for Person ID 3 ends before the end of the study period because the observation period in the data source ends before the end of the study period.

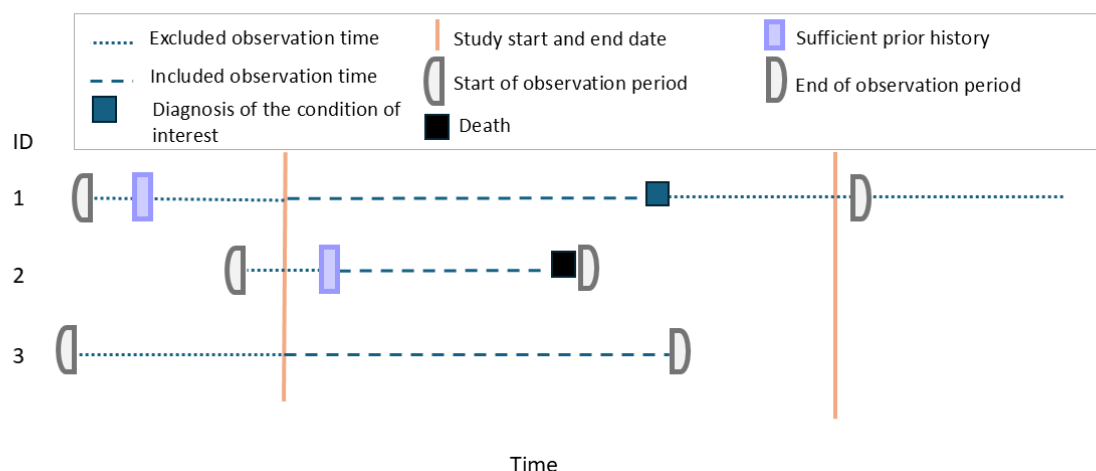


Figure 3. Included observation time for the denominator population for the estimation of incidence rates of selected clinical conditions (Objective 3).

In the estimation of incidence, any restrictions (e.g. age-based restrictions) added to the numerator will also be applied to the denominator.

8.3. Study population with inclusion and exclusion criteria

All objectives

Inclusion criteria

- Under observation at any time during the study period (1 January 2022 to 31 December 2025).
- At least 365 days of prior history for all individuals aged 1 year or older at the index date. No prior history is required for individuals under the age of 1.

Exclusion criteria

- Missing information on age or sex.

Additional eligibility criteria for Objective 1 and 3

Exclusion criteria

- Diagnosis of the index condition at any time before the index date (-inf, -1).

Additionally, for Objective 1, a population of sex- and age-matched individuals without the condition of interest and with the same index date will be defined.

8.4. Study setting and data sources

This study will be conducted using routinely collected data from 10 data sources within the DARWIN EU® Data Network, spanning 9 European countries. The proposed list of data sources and their characteristics are listed in [Table 1](#).

Table 1. Proposed data sources and their key characteristics.

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals	Calendar period covered by each data source	Contributing to
Denmark	DK-DHR	Primary, secondary, and hospital inpatient care	Nationwide registry	6.0M	January 1995–October 2025	Objective 1,2, 3
Germany	InGef RDB	Primary, secondary, and hospital inpatient care	Claims	7.4M	January 2016–September 2025	Objective 1,2, 3
Spain	SIDIAP	Primary care (linked to hospital discharged records)	EHR	6.1M	January 2006–December 2024	Objective 1,2, 3
Sweden	HI-SPEED	Primary, secondary, and hospital inpatient care	Nationwide registry	10.6M	January 2015–August 2025	Objective 1,2, 3
United Kingdom	CPRD Aurum Linked	Primary care linked to hospital discharged records.	EHR	17.1M	April 1997–March 2025	Objective 1,2, 3
Croatia	NAJS	Primary, secondary, and hospital inpatient care	Nationwide Registry	4.3M	February 2014–December 2025	Objective 1,2, 3
Finland	FinOMOP-THL	Primary care (GPs and specialists), secondary care, hospital inpatient care	Nationwide Registry	6.2M	January 2015–December 2025	Objective 1,2, 3
Lithuania	LDH	Primary care (GPs and specialists), secondary care, hospital inpatient care	Claims	2.7M	January 2015–June 2026	Objective 1,2, 3

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals	Calendar period covered by each data source	Contributing to
Norway	NLHR	Primary, secondary, and hospital inpatient care	Nationwide registry	5.6M	2008–December 2023	Objective 1, 2, 3
Spain	VID	Primary, secondary, and hospital inpatient care	EHR	5.6M	January 2009–December 2024	Objective 1, 2, 3

EHR=Electronic Health Record, NAJS=Croatian National Public Health Information System, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NLHR=Norwegian Linked Health Registry data, SIDIAP=The Information System for Research in Primary Care, VID=Valencia Health System Integrated Dataset, HES=Hospital Episode Statistics, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care

Data sources selection

These data sources fulfil the criteria required in terms of data quality, completeness, and representativeness for characterisation and descriptive epidemiology of the selected clinical conditions in different European countries, covering Northern, Central, Western, and Southern European regions. ([Annex I](#)). Furthermore, the data sources have been chosen based on preliminary feasibility counts of the selected conditions. Indeed, the selected data sources include data with outcomes for all or most phenotypes of interest.

8.5. Study period

The study period for patient identification will be defined as from 1 January 2022 to 31 December 2025. This study period was selected to ensure recent and accurate data capture across the entire interval for the estimation of period prevalence across all participating data sources.

However, it should be noted that in several data sources the availability of the accurate data ends before 12/2025 (see [Table 1](#)). Therefore, more recent data might be available at the time of the analysis, in which case the analyses will be executed on the most recent and quality checked data source available at each data partner site.

8.6. Variables

8.6.1. Outcome

All outcomes will be defined using condition and/or observation codes. The development and validation of these algorithms based on pre-specified clinical descriptions for each selected clinical condition will be undertaken as part of the study execution. However, a preliminary approach is provided in [Annex IV](#) in the form of Phenotyping plans. These preliminary concepts will be refined during study execution, in accordance with the standard DARWIN EU® phenotyping processes.[2] These processes will include the review of code lists by clinical experts and the review of phenotype performance across participating data sources using the *PhenotypeR* R package.[4]

First, code lists and phenotypes will be developed, reviewed, and approved during study execution. Code lists will be presented in an electronic format (Shiny app). All decisions made regarding the inclusion and/or exclusion of codes/concepts will be aligned with the definitions of the phenotypes and subtypes previously

developed in the clinical descriptions and operationalised in the phenotyping plans. All these decisions will be accurately reported in the Cohort diagnostics review forms. A short summary of such decisions will be provided along with the final code lists selected for each condition and sub-phenotype as part of the study output, as specified in the corresponding section.

When appropriate, more than one variant or flavour of each phenotype will be developed (e.g. broad and narrow definitions to include more sensitive/inclusive vs specific code lists respectively). For instance, it could be that more than one flavour is needed to estimate incidence and prevalence, as some coded diagnoses cannot be used for estimating incidence (e.g. concepts including “old” or “history of”). The characteristics of each variant will be described, together with a discussion of their respective strengths and limitations and recommendations regarding the contexts in which each variant may be most appropriate. Notably, different variants may be preferable for different use cases. Of note, most of the conditions of interest are chronic and not curable.

In addition to the phenotype overall and its variants, subtypes will be developed as shown in [Table 2](#). However, this might change during the phenotyping process. Further details are provided in [Annex IV](#).

Table 2. List of outcomes and proposed flavours and subtypes.

Condition	Subtypes and particularities
Spinal Muscular Atrophy (SMA)	Subtypes proposed are: • Type I (Werdnig-Hoffmann) • Type II • Type III (Kugelberg-Welander) • Type IV (adult-onset)
Amyotrophic lateral sclerosis (ALS)	Two different, non-exclusive categories of subtypes are proposed: - Age at onset: ○ Adult-onset ALS (index date at age ≥ 25 years old) ○ Juvenile-onset ALS (< 25 years old) - Association with frontotemporal dementia (FTD): ALS-FTD Description of the subtypes will require a longer (i.e. 2-year) follow-up for the assessment of associations and/or complications that are part of the phenotype definition
Multiple sclerosis	The main subtypes proposed are MS with clinical activity (i.e. relapsing MS) and without clinical activity (by exclusion). subtypes Characterisation of this sub-phenotype will require post-diagnosis information follow-up
Prurigo nodularis	The proportion of individuals with a potentially drug-induced condition will be operationalised as the percentage of prescription of pre-defined relevant drugs (i.e. drugs which have been strongly associated with the condition of interest) in the prior 30 and 90 days to the index date.
Bullous pemphigoid	The proportion of individuals with a potentially drug-induced condition will be operationalised as the percentage of prescription of pre-defined relevant drugs (i.e. drugs which have been strongly associated with the condition of interest) in the prior 30 and 90 days to the index date.
Vitiligo	The proportion of individuals with a potentially drug-induced condition will be operationalised as the percentage of prescription of pre-defined relevant drugs (i.e. drugs which have been strongly associated with the condition of interest) in the prior 30 and 90 days to the index date.
Lupus	The two main subtypes proposed are: - Systemic lupus erythematosus (SLE) - Cutaneous lupus erythematosus (CLE)

Condition	Subtypes and particularities
	The proportion of individuals with a potentially drug-induced condition will be operationalised as the percentage of prescription of pre-defined relevant drugs (i.e. drugs which have been strongly associated with the condition of interest) in the prior 30 and 90 days to the index date.
Vasculitis	Two subtypes will be developed: - Broad vasculitis: A more sensitive one which includes primary, secondary, and unspecified vasculitis. This flavour would be composed by the non-specific concepts of “vasculitis” OR any of all the following: Giant cell arteritis, Takayasu arteritis, Kawasaki disease, Polyarteritis nodosa, Microscopic polyangiitis, Granulomatosis with polyangiitis, Eosinophilic granulomatosis with polyangiitis, IgA vasculitis, Cutaneous leukocytoclastic angiitis, Primary central nervous system vasculitis (“cerebral vasculitis”) - Narrow vasculitis flavour that would only include vasculitis which are most likely primary. This narrow variant will apply exclusion criteria The proportion of individuals with a potentially drug-induced condition will be operationalised as the percentage of prescription of pre-defined relevant drugs (i.e. drugs which have been strongly associated with the condition of interest) in the prior 30 and 90 days to the index date.
Psoriasis and psoriatic arthritis	The two subtypes developed will be: - Psoriasis: Formed by unspecific terms, such as “psoriasis” and/or of the following terms: plaque psoriasis, guttate psoriasis, pustular psoriasis, erythrodermic psoriasis, and inverse psoriasis. - Psoriatic arthritis, including only terms referring to psoriasis-related arthropathy. The proportion of individuals with a potentially drug-induced condition will be operationalised as the percentage of prescription of pre-defined relevant drugs (i.e. drugs which have been strongly associated with the condition of interest) in the prior 30 and 90 days to the index date.
Autoimmune hepatitis	Two subtypes will be ascertained: with and without hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (Aspartate aminotransferase (AST)/Alanine aminotransferase (ALT) elevations), and cholestasis (alkaline phosphatase (ALP)/gamma-glutamyl transferase (GGT)) and bilirubin will be reported as part of the characterisation (Objective 1).
Primary biliary cholangitis	Two subtypes will be ascertained: with and without hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).
Primary sclerosing cholangitis	Two subtypes will be ascertained: with and without hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).
Chronic viral hepatitis	Two subtypes will be ascertained: with and without hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).
Alcohol related liver disease	Two subtypes will be ascertained: with and without hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).
Metabolic dysfunction-associated steatohepatitis	Two subtypes will be ascertained: with and without hepatic cirrhosis. Additionally, the proportion of pre-defined comorbidities of interest (i.e. obesity, diabetes, dyslipidaemia) before index date will be described.

Condition	Subtypes and particularities
	Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).
Heart failure	Three different, non-exclusive categories of subtypes are proposed based on: The ejection fraction: - Heart failure with reduced ejection fraction - Heart failure with preserved ejection fraction Compensated/decompensated: Decompensated HF, operationalised as at least one exacerbation event during the follow-up, defined as an inpatient admission in which HF is recorded as the primary diagnosis. Compensated (by exclusion). Location - Right side - Left side Additionally, the proportion of pre-defined comorbidities of interest (i.e. ischaemic heart disease, arterial hypertension, cardiomyopathies, valvular heart disease) before index date will be described.
Cardio-renal-metabolic syndrome	The three subtypes proposed are: - Cardiac-driven - Renal-driven - Metabolic-driven. Additionally, the proportion of pre-defined comorbidities of interest (i.e. heart failure, CKD, dyslipidaemia, diabetes mellitus, obesity, hypertension, and hepatic comorbidities) before index date will be described.

ALP=Alkaline phosphatase, ALS=Amyotrophic lateral sclerosis, ALT=Alanine aminotransferase, AST=Aspartate Aminotransferase, BP=Bullous pemphigoid, CLE=Cutaneous lupus erythematosus, GGT=Gamma-glutamyl transferase, HF=Heart failure, MS=Multiple sclerosis, PN=Prurigo nodularis, SLE=Systemic lupus erythematosus, SMA=Spinal muscular atrophy

8.6.2. Covariates

For Objective 1, the following covariates will be used to characterise the study population at the index date [0, 0], defined as the first ever diagnosis of the target condition within the study period. However, for selected conditions, an additional assessment window of 90 to 1 days before the index date will be included, as specified below:

- Sex
- Age
- Medical history
- Medications
- For prurigo nodularis, bullous pemphigoid, vitiligo, lupus, vasculitis, and psoriasis, the proportion of individuals who initiated treatment with a drug of interest, as defined below, within the 30- and 90-day periods preceding the index date will be reported. Where applicable, this assessment will also be

performed for relevant subtypes (e.g. drug-induced SLE and drug-induced CLE). Drugs of interest include all medications that have been described as associated with the corresponding drug-induced condition, as well as other drugs within the same therapeutic class, where appropriate and consistent with criteria established by the clinical domain experts. Examples of drugs of interest include:

- Drug-induced pruritus, potentially leading to PN: antihypertensive drugs, opioids, biologics (e.g. EGFR inhibitors, IL-6/IL-17/IL-23 inhibitors, janus kinase (JAK) inhibitors, checkpoint inhibitors), antibiotics, oral antifungal agents, antimalarials, chemotherapeutics, antidepressants, antiseizure, antipsychotics.[9]
- Drug-induced BP: antibiotics, amiodarone, captopril, furosemide, non-steroidal anti-inflammatory drugs (NSAIDs), lithium, phenytoin, DPP-4 inhibitors, anti-programmed cell death protein 1 (PD-1), anti-programmed death-ligand 1 (PD-L1), sulfasalazine/mesalazine.[10]
- Drug-induced vitiligo: anti-programmed cell death protein 1 (PD-1), anti-programmed death-ligand 1 (PD-L1), Toll-like receptor 7 (TLR7) agonist (imiquimod), Tumour Necrosis Factor-alpha (TNF - α) inhibitors (adalimumab, infliximab, etanercept), anti-CD52 (alemtuzumab), anti-CTLA-4 (ipilimumab), anti-interleukin (IL)-12/23 (ustekinumab).[11]
- Drug-induced lupus: procainamide, hydralazine, quinidine, sulphonamides, penicillin, amino salicylic acid, hydrochlorothiazide, cimetidine, phenylbutazone, and mesantoin.[12]
- Drug-induced vasculitis: antibiotics, anti-thyroid drugs, anti-TNF- α inhibitors, clozapine, allopurinol, D-penicillamine, hydralazine, phenytoin, sulfasalazine, propylthiouracil, methimazole, carbimazole, benzyl thiouracil.[13]
- Drug-induced psoriasis: beta-blockers, lithium, TLR7 agonist (imiquimod), antimalarial terbinafine, anti-PD-1, anti-PD-L1, sorafenib, bupropion.[14]
- For metabolic dysfunction-associated steatohepatitis, HF, and cardio-renal-metabolic syndrome, the presence in the year prior to the index date of certain, previously defined comorbidities will be specifically identified. These comorbidities of interest will differ across phenotypes. They will be described as a proportion of individuals with the condition of interest and, where relevant, of individuals within the sub-phenotype of interest. For example, as part of the HF characterisation, the proportion of individuals with HF (overall, HF with preserved ejection fraction, or HF with reduced ejection fraction) who have a history of hypertension, ischaemic heart disease, valvular heart disease, or cardiomyopathy prior to the index date will be reported.

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. All available already collected data will be used to identify individuals with the phenotypes of interest and estimate the period prevalence and incidence rate 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 secure online repository of DTZ (Data Transfer Zone). All results will be locked and timestamped for reproducibility and transparency. The study results of 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 and reported as "<5".

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

Objectives 1 and 4

Once the phenotypes (and each of their variants/cohorts) have been developed, the resulting cohorts will be characterised at the index date and within the characterisation windows defined in [Section 8.2](#). This characterisation will be based on OMOP CDM mapped data using the *CohortCharacteristics* R package, developed by DARWIN EU®.[5] It will include individual demographics, medical history (comorbidities and/or differential diagnoses), and comedications. Age at index date will be summarised using the median and interquartile range (IQR), and sex will be summarised as proportions. Medical history and general medication use will be assessed during the following time windows: 365 to 31 days before the index date, 30 to 1 day before the index date, on the index date, 1 to 30 days after the index date, and 31 to 365 days after the index date. Comorbidities and comedications will be presented as the 10 most frequent concepts overall and the concepts with the largest absolute standardised mean differences (ASMD) compared with age- and sex-matched individuals without the condition of interest ("controls"). This comparison will help identify the key features that differentiate individuals with the phenotypes from matched general population controls. In general terms, the distribution of subtypes will be documented and summarised as proportions. However, the subtypes of MS with/without clinical activity and HF compensated/decompensated will be measured using cumulative incidence, as implemented in the *CohortSurvival* R package (Objective 4).[8]

Objective 2

Prevalence of the selected clinical conditions will be estimated as period prevalence proportion, defined as the total number of individuals in observation during the study period with the clinical condition of interest, divided by the population under observation during the same period. For chronic and non-curable conditions, an individual will be considered to have the condition of interest if they have a recorded diagnosis either during or at any moment before the study period. Disease onset will be defined as the date of the individual's first recorded diagnosis of the condition, including diagnoses recorded before the study period. In other words, and to minimise underestimation of prevalence, all available historical data will therefore be used as the lookback period for these conditions. For conditions that may resolve or be cured a 1-year lookback period will be applied instead, as stated in [Section 8.2](#). Binomial 95% confidence intervals will be calculated for all prevalence proportions. All analyses will be estimated using the *IncidencePrevalence* R package, developed by DARWIN EU®.[2]

Objective 3

Incidence of the selected clinical conditions will be estimated as incidence rates, defined as the total number of individuals in observation during the study period with an incident diagnosis of the clinical condition of interest, divided by the person-years (expressed as per 100,000) under observation for the same period. As stated in [Section 8.2](#), individuals with a prior diagnosis of the condition will be excluded. Poisson 95% confidence intervals will be calculated for all incidence rates. All analyses will be estimated using the *IncidencePrevalence* R package, developed by DARWIN EU®.[2]

As a sensitivity analysis and acknowledging that most SMA cases are diagnosed during childhood, additional incidence analyses will be conducted restricting the denominator population to paediatric individuals, defined as those aged <18 years.

8.8.4. Output

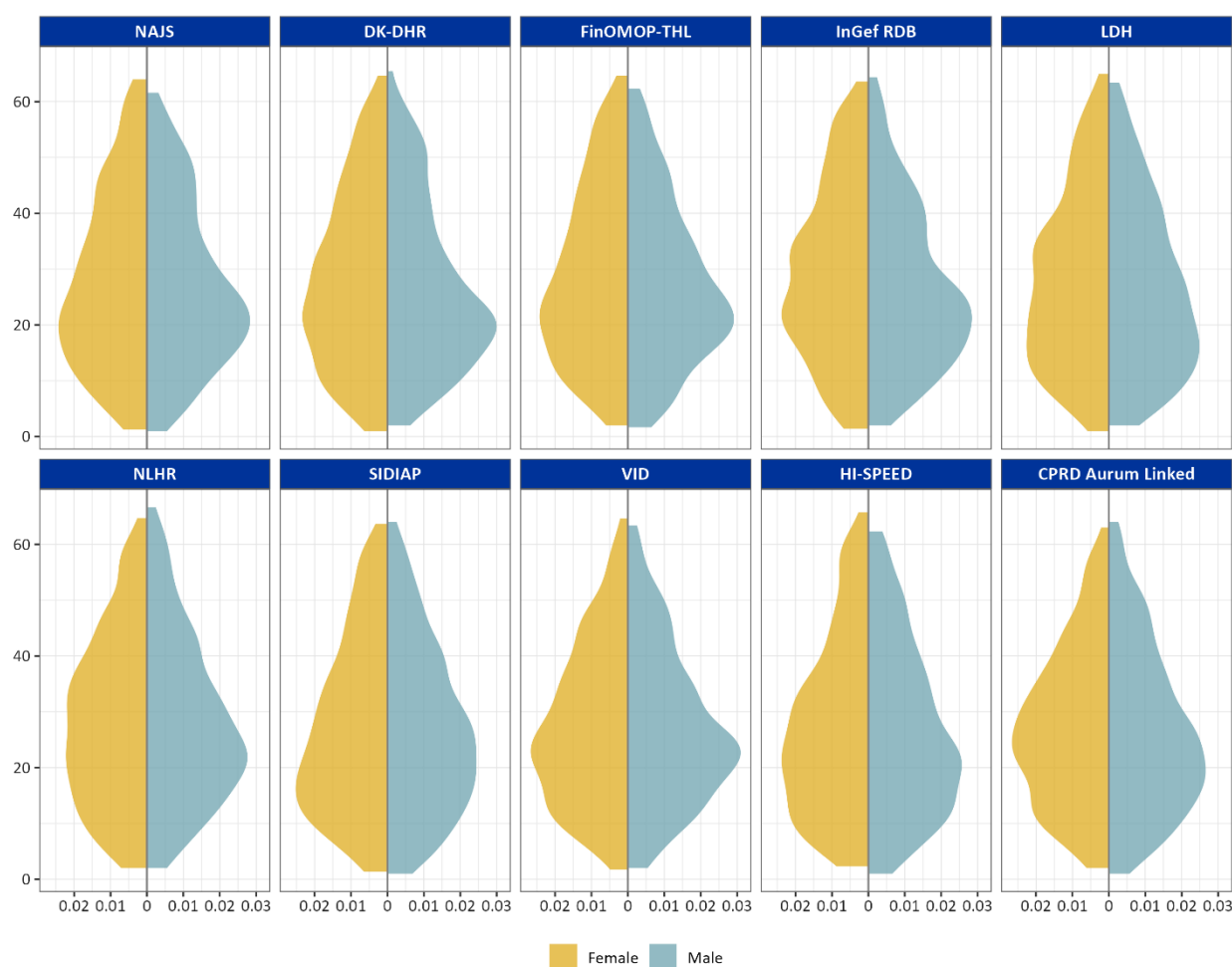
Below, the expected output to be provided in the report is summarised. Of note, figures will not be reported if there are less than 5 individuals in a given stratum. All tables and figures have been generated using synthetic data and do not represent real results. For brevity, Supplementary Tables 2–4 only provide an example of a condition (spinal muscular atrophy (SMA)).

- A PDF report, including an executive summary and the following tables, figures, and materials.
- An interactive dashboard (Shiny app) will be generated by incorporating all the results (tables and figures).

MOCK Supplementary Table 1. Number of individuals under observation during the study period for each specific condition after applying each eligibility criteria (attrition table), by data source, for Objective 1.

Condition 1		Number of subjects	
	Data source 1	Data source X	
Initial population (entire data source)	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	
First diagnosis of the condition during the study period (index diagnosis)	5,100	507	
Diagnosis of the index condition at any time before the index diagnosis (-inf, -1)	4,880	390	
Condition 2		Number of subjects	
	Data source 1	Data source X	
Initial population (entire data source)	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	27,912	
First diagnosis of the condition during the study period (index diagnosis)	5,100	507	
Diagnosis of the index condition at any time before the index diagnosis (-inf, -1)	4,880	390	

Condition 1		Number of subjects	



MOCK Figures 1a–1j. Population pyramid displaying the age distribution of the study population for each specific condition stratified by sex at the index date.

NAJS=Croatian National Public Health Information System, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NLHR=Norwegian Linked Health Registry data, SIDIAP=The Information System for Research in Primary Care, VID=Valencia Health System Integrated Dataset, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care

One figure will be generated for each condition, with one panel for each data source. Consideration will also be given to adding, for each data source, a pyramid showing the distribution of the underlying population.



MOCK Figures 2a–2j. Bar chart displaying the distribution of subtypes for each main phenotype (where applicable).

SMA=spinal muscular atrophy, NAJS=Croatian National Public Health Information System, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NLHR=Norwegian Linked Health Registry data, SIDIAP=The Information System for Research in Primary Care, VID=Valencia Health System Integrated Dataset, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care.

One figure will be generated for each condition, with one panel for each data source.

MOCK Tables 1a–1x. Characteristics of the study populations with a first diagnosis of each specific condition during the study period.

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
Spinal muscular atrophy												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 -Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	Type III	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Type IV	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Type I	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Type II	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Amyotrophic lateral sclerosis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 - Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	Adult-onset	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	With FTD	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	ALS only	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Juvenile-onset	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Multiple sclerosis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 - Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	With clinical activity	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Without clinical activity	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Prurigo nodularis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 - Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Drug-induced		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Bullous pemphigoid												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 - Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Drug-induced		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Vitiligo												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Drug-induced		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Lupus												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	SLE	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	CLE	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Drug-induced		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Vasculitis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Primary vasculitis (narrow)		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Drug-induced		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Psoriasis and psoriatic arthritis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	Psoriasis	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	PsA	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Drug-induced		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Autoimmune hepatitis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hepatic cirrhosis		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AST*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALP*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
GGT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Bilirubin*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Primary biliary cholangitis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hepatic cirrhosis		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AST*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALP*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
GGT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Bilirubin*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Primary sclerosing cholangitis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hepatic cirrhosis		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AST*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALP*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
GGT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Bilirubin*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Chronic viral hepatitis												

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hepatic cirrhosis		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AST*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALP*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
GGT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Bilirubin*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Alcohol related liver disease												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hepatic cirrhosis		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AST*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALP*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
GGT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Bilirubin*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Metabolic dysfunction-associated steatohepatitis												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hepatic cirrhosis		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AST*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
ALP*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
GGT*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Bilirubin*		Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Obesity		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Diabetes mellitus		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Dyslipidaemia		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Heart failure												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	With REF	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Compensated	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Decompensated	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Right-sided	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	With PEF	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Left-sided	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Ischaemic heart disease		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Arterial hypertension		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Cardiomyopathy		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Valvular heart disease		N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Cardio-renal-metabolic syndrome												
Number subjects	–	N	0	0	0	0	0	0	0	0	0	0
Age	–	Median [Q25 – Q75]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
Sex	Female	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Male	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Sub-phenotype	Cardiac driven	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Renal driven	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Data source												
Variable name	Variable level	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Metabolic driven	N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

ALS=Amyotrophic lateral sclerosis, Alkaline phosphatase=ALP, Alanine aminotransferase=ALT, Aspartate aminotransferase=AST, CLE=Cutaneous lupus erythematosus, CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, FTD=Frontotemporal dementia, Gamma-glutamyl transferase=GGT, HI-SPEED=Health Impact – Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, PEF=Preserved ejection fraction, PsA=Psoriatic arthritis, NAJS=Croatian National Public Health Information System, REF=Reduced ejection fraction, SLE=Systemic lupus erythematosus, SIDIAP=The Information System for Research in Primary Care, NLHR=Norwegian Linked Health Registry data, VID=Valencia Health System Integrated Dataset, –=data not available yet

*For ALP, ALT, AST, GGT, and bilirubin, the percentage of missing values in each data source will be specified in the footnote.

MOCK Supplementary Tables 2a–2x. Top 10 differential features between cases and controls based on absolute standardised mean difference (ASMD) per phenotype (one table for each).

		Data sources									
Cohort name	Concept Name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
-365 to -31											
Phenotype	Concept 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
-30 to -1											
Phenotype	Concept 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

		Data sources									
Cohort name	Concept Name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Concept 8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0 to 0											
Phenotype	Concept 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1 to 30											
Phenotype	Concept 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Data sources											
Cohort name	Concept Name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Concept 7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31 to 365											
Phenotype	Concept 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Concept 10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NAJS=Croatian National Public Health Information System, SIDIAP=The Information System for Research in Primary Care, NLHR=Norwegian Linked Health Registry data, VID=Valencia Health System Integrated Dataset

MOCK Supplementary Tables 3a–3x. Top 10 most frequent concepts per phenotype (one table for each).

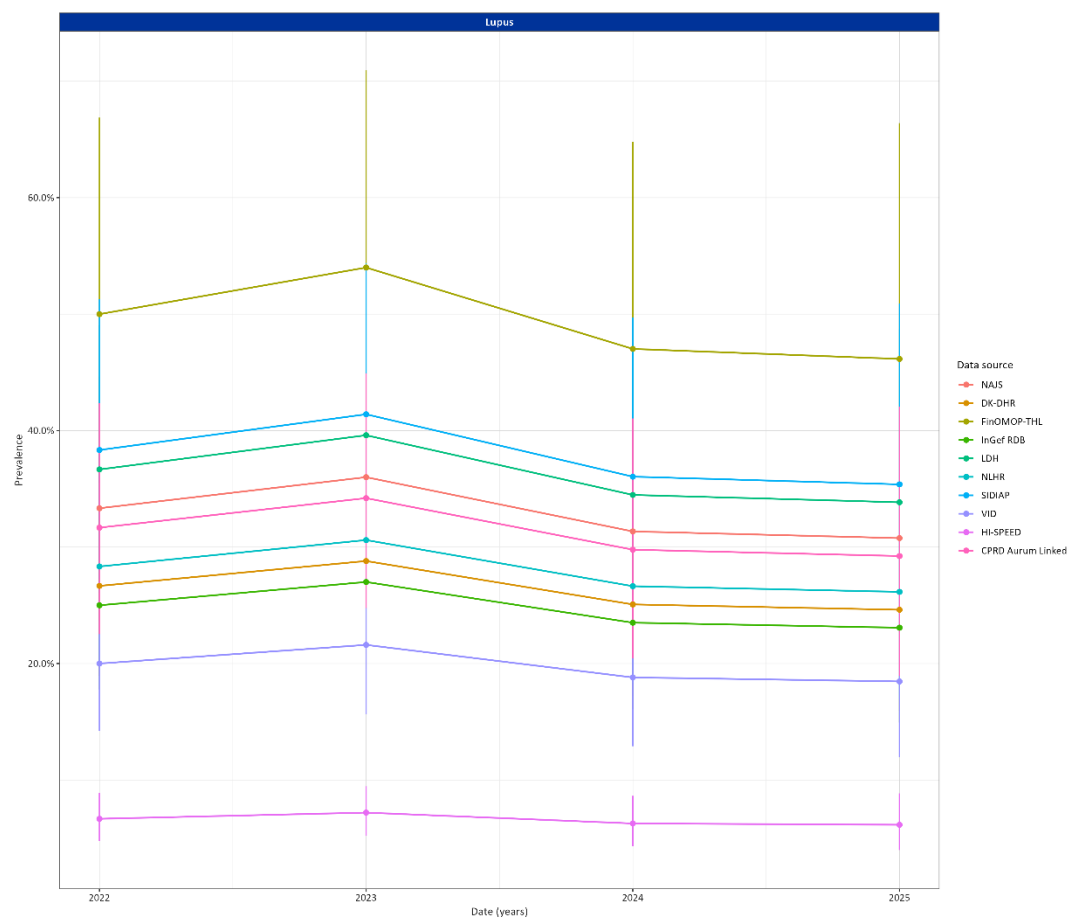
		Data sources									
Cohort name	Variable name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
-365 to -31											
Phenotype	Concept 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
-30 to -1											
Phenotype	Concept 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Data sources											
Cohort name	Variable name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Concept 8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0 to 0											
Phenotype	Concept 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1 to 30											
Phenotype	Concept 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

		Data sources									
Cohort name	Variable name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
	Concept 6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
31 to 365											
Phenotype	Concept 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Concept 10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NAJS=Croatian National Public Health Information System, SIDIAP=The Information System for Research in Primary Care, NLHR=Norwegian Linked Health Registry data, VID=Valencia Health System Integrated Dataset.

- Supplementary material (provided in [Annex IV](#)): A cohort/phenotyping logic and plan, including all the types of variables/parameters to be included for each phenotype and sub-phenotype.
- Supplementary material: Summary of the key decisions taken in the selection of code lists for each phenotype.
- Supplementary material: A link to a shiny app with all the standard codes for each of the proposed phenotypes.



MOCK Figures 3a–3x. Period prevalence of each of the selected clinical conditions.

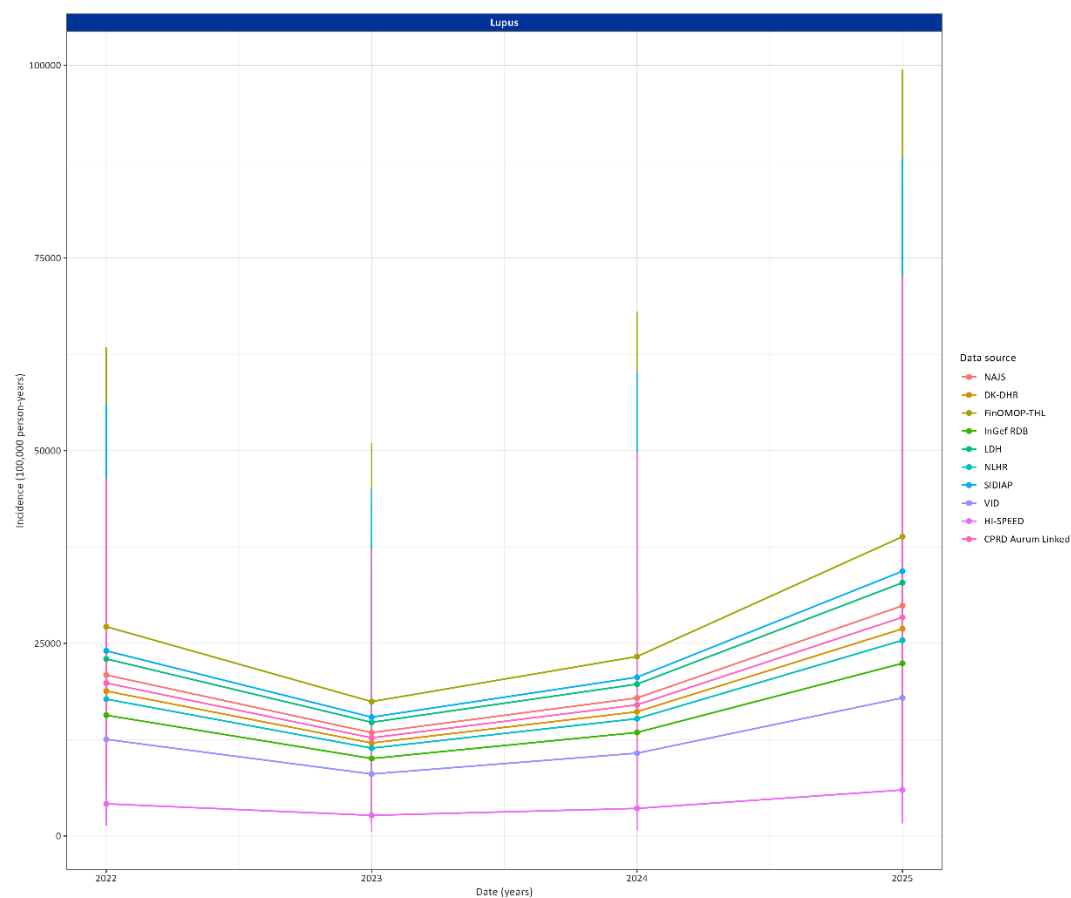
CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NAJS=Croatian National Public Health Information System, NLHR=Norwegian Linked Health Registry data, SIDIAP=The Information System for Research in Primary Care, VID=Valencia Health System Integrated Dataset.

MOCK Supplementary Tables 4a–4x. Period prevalence of selected clinical conditions by data source.

Data source											
Prevalence (year)	Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
Spinal muscular atrophy											
2022	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	prevalence	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2023	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	prevalence	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2024	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	prevalence	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2025	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	prevalence	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	prevalence_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00



CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NAJS=Croatian National Public Health Information System, NLHR=Norwegian Linked Health Registry data, SIDIAP=The Information System for Research in Primary Care, VID=Valencia Health System Integrated Dataset.



MOCK Figures 4a–4x. Incidence of each of the selected clinical conditions (i.e. one figure for each of them) and colour-coded by data source.

CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NAJS=Croatian National Public Health Information System, NLHR=Norwegian Linked Health Registry data, SIDIAP=The Information System for Research in Primary Care, VID=Valencia Health System Integrated Dataset.

MOCK Supplementary Tables 5a–5x. Incidence rates of selected clinical conditions by data source.

Incidence (year)		Data source									
		Estimate name	NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED
Spinal muscular atrophy											
2022	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	person_days	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	person_years	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2023	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	person_days	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	person_years	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2024	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	person_days	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	person_years	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Incidence (year)	Estimate name	Data source									
		NAJS	DK-DHR	FinOMOP-THL	InGef RDB	LDH	NLHR	SIDIAP	VID	HI-SPEED	CPRD Aurum Linked
2025	incidence_100000_pys_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	denominator_count	0	0	0	0	0	0	0	0	0	0
	numerator_count	0	0	0	0	0	0	0	0	0	0
	person_days	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	person_years	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_lower	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	incidence_100000_pys_95CI_upper	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, DK-DHR=Danish Data Health Registries, FinOMOP-THL=Finnish Care Register for Health Care, HI-SPEED=Health Impact - Swedish Population Evidence Enabling Data-linkage, InGef RDB=InGef Research Database, LDH=Lithuanian Digital Health Platform, NAJS=Croatian National Public Health Information System, SIDIAP=The Information System for Research in Primary Care, NLHR=Norwegian Linked Health Registry data, VID=Valencia Health System Integrated Dataset.

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® Data Network.

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 (i.e. primary care, secondary care, hospital inpatient care) and type of data (i.e. population/national registries, claims, electronic health records (EHR)).

9. STRENGTHS AND LIMITATIONS

The study will be informed by routinely collected healthcare data; therefore, data quality issues must be considered. EHR and other routinely collected data have certain inherent limitations because they were collected for clinical and/or administrative 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. Regarding this study in particular, some of the included data sources do not have specific codes for the conditions of interest. Moreover, the documentation of comorbidities, necessary for characterisation, may vary across data sources. As disease prevalence and incidence will be estimated using recorded clinical history, undiagnosed or unrecorded conditions will not be captured. Moreover, certain groups of individuals may be less likely to seek clinical attention and/or have their health-related events recorded, potentially leading to selection bias. Furthermore, misclassification of diagnoses at source or data mapping errors may occur, leading to information bias. 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.[2] Finally, the results estimated from this study will only reflect the populations from the included data sources. The limitations specific for each condition are provided in [Table 3.](#)

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® Data Network, supporting consistent study population descriptions 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 sources across Europe will cover a wide range of healthcare settings and data types, including EHR, claims data, and registries, thus supporting broad generalisability within the network.

Table 3. Limitations specific for each condition.

Condition	Subtypes and comments
Drug-induced conditions (i.e. PN, BP, vitiligo, lupus, vasculitis, psoriasis)	It is acknowledged that caution is warranted when using prior exposure to a medication associated with the development of a given condition as a proxy for a drug-induced disease. Many of these medications are commonly prescribed in the general population (e.g. NSAIDs), and for several of them, the association between drug exposure and disease onset is rare. Moreover, from a clinical perspective, it is often difficult to establish with certainty that exposure to a particular medication has caused the condition. This diagnostic uncertainty likely contributes to the fact that drug-induced conditions are infrequently coded in routine clinical practice. Instead, a description of the proportion of individuals who have a drug highly associated with the condition of interest prescribed in the 30 and 90 days prior to the index date (i.e. the diagnosis of the condition) will be provided as part of Objective 1, as mentioned above.

Condition	Subtypes and comments
SMA	SMA type 0 will not be included because it is foetal and very rare, so it could be difficult to capture using EHR. Although the different SMA subtypes are associated with substantial differences in prognosis, subtype-specific diagnostic codes are not routinely used. Nevertheless, the age at diagnosis is highly suggestive of the underlying subtype, given the distinct ages of onset and life expectancies associated with each phenotype. This operational definition has important limitations, as there is likely to be considerable overlap between types II and III. Although this limitation could potentially be mitigated by incorporating survival into the validation process, doing so would neither ensure accurate differentiation between these two phenotypes nor justify the additional complexity it would introduce to the study. Therefore, the potential overlap between types II and III will be acknowledged as a study limitation.
ALS	Some data sources, such as FinOMOP-THL and InGef RDB, do not have specific codes for ALS, but for “motor neuron disease” instead. This broader diagnostic category cannot be used interchangeably with ALS, so this condition will not be studied in these data sources. Although the adult vs juvenile onset subtypes will be operationalised as age at diagnosis, in this condition there is often a significant gap between clinical onset and diagnosis. There is no available data to establish other potential subtypes, such as genetic forms or familial versus sporadic. The proportion of the ALS-FTD sub-phenotype will only be calculated among individuals with the phenotype who have at least two years of follow-up, so they will not reflect the true proportion of these subtypes among those with the phenotype.
MS	There are no counts for diagnoses specifying the type of MS (relapsing-relapsing, primary progressive, secondary progressive etc), except for CPRD Aurum Linked, which has a specific diagnosis for the progressive forms. Distinguishing whether MS has a progressive component is not feasible without a specific diagnosis of a progressive form, which is only available in CPRD Aurum Linked. While an increase in the Expanded Disability Status Scale score in the absence of relapses would be suggestive of disease progression, these data are not available in the data sources used in this study. The use of alternative proxies (e.g. clinical signs and symptoms or wheelchair prescriptions) could lead to misclassification, either because these data are incompletely recorded or because it is difficult to distinguish disability resulting from an unrecorded relapse from that attributable to progressive disease. Furthermore, radiologic activity, which is a critical aspect of MS, is potentially not well recorded in RWD. The proportion of the “MS with clinical activity” sub-phenotype will only be calculated among individuals with the phenotype who have at least two years of follow-up, so they will not reflect the true proportion of these subtypes among those with the phenotype.
Bullous pemphigoid	There are no specific diagnoses for classic BP vs non-bullous pemphigoid.
Vasculitis	Despite having developed a broad and a narrow flavour to address the issue of primary/secondary vasculitis, the conditions and/or drugs potentially leading to a “secondary” cause might not be well recorded. Further the causal role of these conditions/drugs is often uncertain. As another limitation, some single-organ vasculitis, and, particularly, isolated/idiopathic aortitis and cutaneous arteritis (formerly known as cutaneous polyarteritis nodosa) cannot be studied with the available concepts/codes.
Psoriasis and psoriatic arthritis	There are a lot of counts for the concept “psoriasis” with no subtype specified. Erythrodermic psoriasis is only available in CPRD Aurum Linked.
Autoimmune hepatitis	NA
Primary biliary cholangitis	NA

Condition	Subtypes and comments
Primary sclerosing cholangitis	NA
Chronic viral hepatitis	NA
Alcohol related liver disease	NA
Metabolic dysfunction-associated steatohepatitis	Given that probably there is a lot of overlap between the diagnoses of obesity, diabetes, and dyslipidaemia, subtypes based on the metabolic phenotype could be difficult to operationalise. Therefore, and as mentioned, comorbidities of interest will be reported as part of Objective 1 instead. In InGef RDB (using ICD-10), the source concept is "Other specified inflammatory liver diseases, including NASH", which is mapped to the standard (SNOMED) concept "Inflammatory disease of liver".
Heart failure	Subtypes based on the underlying aetiology might be difficult to operationalise. For example, it may be difficult to ascertain whether arterial hypertension is the primary cause of heart failure, as it is a highly prevalent diagnosis among the age groups affected by heart failure. Therefore, and as mentioned, comorbidities of interest will be reported as part of Objective 1 instead. The proportion of the "HF decompensated" sub-phenotype will only be calculated among individuals with the phenotype who have at least two years of follow-up, so they will not reflect the true proportion of these subtypes among those with the phenotype.
Cardio-renal-metabolic syndrome	Cardiorenal metabolic syndrome as a single diagnostic entity/concept is not available in some data sources. In some, the former concept "cardio-renal-syndrome" is used instead. Others, such as FinOMOP-THL, SIDIAP, VID, and InGef RDB (using ICD-10), do not have specific concepts/codes for cardio-renal-metabolic syndrome. Given that the most recent descriptions of the syndrome also include hepatic involvement, comorbid hepatic conditions will be described as part of the characterisation.

ALS=Amyotrophic lateral sclerosis, BP=Bullous pemphigoid, CPRD Aurum Linked=Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care, EHR=Electronic Health Record, FinOMOP-THL=Finnish Care Register for Health Care, InGef RDB=InGef Research Database, MS=Multiple sclerosis, NA=Not applicable, PN=Prurigo nodularis, RWD=Real world data, SIDIAP=The Information System for Research in Primary Care, SMA=Spinal muscular atrophy, VID=Valencia Health System Integrated Dataset

10. REFERENCES

- [1] Data Analysis and Real World Interrogation Network (DARWIN EU) | European Medicines Agency (EMA), <https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/real-world-evidence/data-analysis-real-world-interrogation-network-darwin-eu> (2021, accessed 13 June 2026).
- [2] Dernie F, Corby G, Robinson A, 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* 2024; 33: e70042.
- [3] Burn E, Alcalde-Herraiz M, Català M, et al. CodelistGenerator: Identify Relevant Clinical Codes and Evaluate Their Use. 2023; 4.0.2.
- [4] Zhao P. Phenotype: A Tool for Phenotypic Data Processing. 2020; 0.1.0.
- [5] Du M, Prats-Urbe A, Mercadé-Besora N, et al. CohortCharacteristics: an R package for population characterisation in observational studies using the OMOP common data model. *Eur J Epidemiol*. Epub ahead of print 3 April 2026. DOI: 10.1007/s10654-025-01352-4.
- [6] Raventós B, Català M, Du M, et al. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. *Pharmacoepidemiol Drug Saf* 2024; 33: e5717.
- [7] Build and Manipulate Study Cohorts Using a Common Data Model, <https://ohdsi.github.io/CohortConstructor/> (accessed 15 June 2026).
- [8] López-Güell K, Burn E, Català M, Li X, Newby D, Mercade-Besora N. CohortSurvival: Estimate Survival from Common Data Model Cohorts., <https://darwin-eu.github.io/CohortSurvival/>. (2026).
- [9] Drug-induced pruritus. *DermNet®*, <https://dermnetnz.org/topics/drug-induced-pruritus> (2023, accessed 10 July 2026).
- [10] Ramdial PK, Naidoo DK. Drug-induced cutaneous pathology. *J Clin Pathol* 2009; 62: 493–504.
- [11] Anthony N, Bourneau-martin D, Ghamrawi S, et al. Drug-induced vitiligo: a case/non-case study in Vigibase®, the WHO pharmacovigilance database. *Fundam Clin Pharmacol* 2020; 34: 736–742.
- [12] Rubin RL. Drug-induced lupus. In: *Systemic Lupus Erythematosus*. Elsevier, pp. 817–830.
- [13] Radić M, Martinović Kaliterna D, Radić J. Drug-induced vasculitis: a clinical and pathological review. *Neth J Med* 2012; 70: 12–17.
- [14] Balak DM, Hajdarbegovic E. Drug-induced psoriasis: clinical perspectives. *Psoriasis Auckl NZ* 2017; 7: 87–94.

- [15] Ndumele CE, Rangaswami J, Chow SL, et al. Cardiovascular-Kidney-Metabolic Health: A Presidential Advisory From the American Heart Association. *Circulation* 2023; 148: 1606–1635.

11. ANNEXES

ANNEX I. Description of data sources

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 (CIPH) 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. 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 electronic health records, and Inpatient hospital billing systems, and Registries, 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 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 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.

#	Section	Description
		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 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/

Danish Data Health Registries (DK-DHR)

#	Section	Description
1	Data source identification and country	DK-DHR (Danish Data Health Registries) Denmark
2	Data partner information section	Danish Medicines Agency (DKMA) Data & Data Analysis (DAD)
3	Coverage and timespan	Data collection since: 1995 Extent: Nation-wide. The data is representative of the entire Danish population.
4	Healthcare setting / type of data	Community pharmacists, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following data elements are collected: diagnosis (including rare diseases and pregnancy data), hospital admissions, discharge and ICU data, Cause of death, Drug prescriptions, dispensing, vaccination and contraception, Procedures (surgical and non-surgical hospital), and Sociodemographic information (sex and age only).
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and Other. All causes of deaths, all retrieved drug prescriptions, all records of vaccinations, all hospital inpatient and outpatients contacts including disease diagnoses and hospital surgical and non-surgical procedures, histologically confirmed incident cancers, laboratory test results for the entire Danish population from 1/1/1995 onwards.
6	General representativeness	The data is representative of the entire Danish population. Healthcare is free in Denmark, so we do not expect any bias in data collection based on socio-economic status.

#	Section	Description
7	Data content / source coding	Diagnoses and causes of death are collected using the ICD-10 vocabulary. ATC and RxNorm are used for Drugs. SNOMED codes are used for Procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have unique identifiers used to link datasets.
9	Quality control (data source specific)	The data we have received relating to nationwide Danish Health Data registries offer an opportunity for large-scale, population-based studies with several advantages 1) Their large size improves the precision of estimates and enables the study of rare exposures and outcomes with long-term latency, 2) Inclusion of nearly all individuals in the target population ensures that the data reflect routine clinical care and all clinical segments of the source population, 3) Data are collected independently of each research study, thus minimising certain types of bias, e.g., non-response, and the influence from attention to the research question on the diagnostic process. Before the source data is sent to us, the Danish Health Data Authority runs and does comprehensive checks of the registry table data validity of the variables, breaks in data, changes in variable coding, missingness, etc. We perform checks of missingness/completeness in relation to requested variables. In essence, we are receiving a dump of a mirror of the data that is controlled by the SDS. The documentation performed by SDS is available online, in Danish primarily https://www.esundhed.dk/Dokumentation (all variables), but also in English https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/national_health_registers
10	Linkage	There is no linkage in this data source.
11	Vital status	The Cause of Death registry (DAR) is used, the cause of death is collected using ICD-10 codes.
12	Limitations	- Lifestyle factors (such as smoking status and weight) are only complete for pregnant women. - Stillborn children will not have any records in our CDM. This means that e.g. birth weight and birth length of stillborns is not recorded. - There is no information on key socio-economic status (SES) factors, such as occupation, education, and income. - Patient contacts in primary care (visits to the GP) are not included. Consequently, the incidence of chronic diseases like Type 2 Diabetes (T2D) and asthma must be determined using drug prescriptions as a proxy.
13	Main references	Schmidt M, Schmidt SAJ, Adelborg K, Sundbøll J, Laugesen K, Ehrenstein V, Sørensen HT "The Danish health care system and epidemiological research: from health care contacts to database records." Clinical epidemiology (2019): 31372058
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111217 Website: https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/healthdatadenmark

Finnish Care Register for Health Care (FinOMOP-THL)

#	Section	Description
1	Data source identification and country	FinOMOP-THL (Finnish Care Register for Health Care) Finland
2	Data partner information section	FinOMOP (FinOMOP) The Department of Data and Analytics
3	Coverage and timespan	Data collection since: 1998 Extent: Nation-wide. The current CDM population comprises all persons having been alive and residing in Finland since the beginning of 2011.
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). THL maintains health registers that cover both public and private, primary, and specialised

#	Section	Description
		inpatient, urgent and outpatient health care encounters in Finland, starting from 2011. The entire public sector and private inpatient encounters have been included since 2011, while private outpatient encounters, including occupational care, are included since 2020. Since 1998, the Care Register for Health Care (TerveysHilmo) has covered both public outpatient and inpatient specialized care and private inpatient care. Since 2009, the Finnish National Vaccination Register has covered all vaccinations from the public sector and from a large part of private vaccination providers, with the data coverage from both sections being very good to complete from 2020 onwards. Since 2011, the Register of Primary Health Care Visits (AvoHilmo) has covered public primary care. Since 2020, the register has also covered private outpatient care and occupational care. In addition, the CDM also contains positive COVID-19 test results from the Finnish National Infectious Diseases Register. The register itself covers all laboratory confirmations for around 70 specific microbes from 1995 onwards, but only COVID-19 has currently been mapped to the CDM. The CDM also includes prescription records from multiple different sources. Both Care Register for Health Care and Register of Primary Health Care Visits contain very basic prescription data recorded during health care encounters that include just the ATC-code and trade name of medication. More comprehensive prescription data from Kanta Prescription Centre, maintained by Social Insurance Institution of Finland (Kela), has been integrated into the CDM since its 2024 release. The Kanta Prescription records are based electronic prescriptions, which were adopted by most public health care providers in 2010 and by most private providers by 2017.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries. Data is entered by clinicians upon healthcare contact and processed by THL (Kela in the case of Kanta Prescription Centre).
6	General representativeness	The THL data has national coverage and is therefore well representative of the Finnish population. Using the complete population as a basis for the person table also serves to facilitate calculations on a population level, e.g. incidence rates.
7	Data content / source coding	The following coding systems have been OMOP-mapped, typically to a good level of completeness: ICD10fi Finnish Extension, ICPC-2, ATC, Toimenpideluokitus (procedure classification adapted from the Nordic Classification of Surgical Procedures (NCSP)), Terveystieteiden tutkimuskeskusten erikoissalat (Hilmo specific provider speciality), Rokotustapa (AR/YDIN National classification for vaccine administration), Tupakointistatus (AR/YDIN National classification for smoking status). Vaccinations are identified on product level based on batch number, trade name, vaccine title, and ATC-code. This is mapped on brand and type in the OMOP CDM.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Each patient in THL has a unique identifier.
9	Quality control (data source specific)	The source data collection undergoes a structural and semantic validation before entry into the source database. Additionally, some coded variables undergo quality assessment against the respective code systems post entry into the database. The source registers are also assessed for completeness and coverage, with the aim of improving future collection in the areas where data is lacking.
10	Linkage	THL is already a linkage of multiple Finnish registries (see above).
11	Vital status	The National Population registry data forms the basis for forming the patient population. This ensures an up-to-date location (municipality of residence) of patients, as well as complete death occurrences (although not the cause of death).
12	Limitations	All drug records in the CDM are currently based on prescriptions. Kanta Prescription Centre also includes information on drug dispensings, but these have not currently been converted into OMOP CDM. Depending on the type of the medication, a single prescription can be for up to two years dosage of the drug. This means a patient can have up to two-year breaks between observations while actively using the drug. Observation of a prescription also does not mean that the patient necessarily bought or used the medication. The CDM does not currently have meaningful end dates or days of supply for drug exposure. This

#	Section	Description
		information is not available for Care Register for Health Care or Register of Primary Health Care Visits, and for Kanta Prescription Centre it is only available in unstructured, free-form format that has not been converted into OMOP CDM. The source system for prescription drug data is only available to the DP until Dec-2023, therefore drug_exposure records from this source are only available in that time range. Some vital signs, like height and weight measurements, are recorded even if the person has not been actually measured, just that the measurement is deemed relevant to the condition or it's treatment. This may lead to duplicate measurements as the value from previous visit might be taken. Not all private health care records are covered for the CDM's entire follow-up time from 2011 onwards. For Register of Primary Health Care Visits and Finnish National Vaccination Register, records from private health care have been available from 2020 onwards. For Kanta Prescription Centre, the coverage of private health care records has been good from 2017 onwards. The inclusion of private health care mainly presents itself as an increase in the number of observations, meaning that it has to be accounted for when interpreting any time series data from the CDM.
13	Main references	Häkkinen, Pirjo; Mölläri, Kaisa; Saukkonen, Sanna-Mari; Väyrynen, Riikka; Mielikäinen, Lasse; Järvelin, Jutta "Hilmo - Sosiaali- ja terveydenhuollon hoitoilmoitus 2020 : Määrittelyt ja ohjeistus : Voimassa 1.1.2020 alkaen" Terveyden ja hyvinvoinnin laitos (2019):
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111187; https://catalogues.ema.europa.eu/data-source/1111191 Website: https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions; https://www.kanta.fi/en/research-and-knowledge-management

InGef Research Database (InGef RDB)

#	Section	Description
1	Data source identification and country	InGef RDB (InGef Research Database) Germany
2	Data partner information section	Institute for Applied Health Research Berlin GmbH (InGef)
3	Coverage and timespan	Data collection since: 2014 Extent: Nation-wide. 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	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, and claims data. InGef is a research service provider within the German statutory health insurance system. In general, the data in the RDB reflects most of the health services paid for by the 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	Insurance/administrative claims. 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

#	Section	Description
		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. 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 Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. 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 (data source 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. Pharmacy & Hospital data having a general lag of 2 calendar months, while outpatient data has a lag in delivery of up to 9 calendar months before it can be considered complete. Data deliveries occur separately per health insurance, and the information about which health insurance a person is insured with is removed during the anonymisation process. This means that, per patient, we cannot explicitly determine whether the lack of data means that their insurance ended/they were not diagnosed or treated/they did not die in that time period, or whether we don't see any records because their insurance company has not yet delivered its most recent data for that sector. This means that conclusions drawn using data for the last few quarters in the release could be inconsistent. Drugs mostly contain pharmacy-dispensed drugs. We generally don't have records of drugs administered by a healthcare professional as part of a

#	Section	Description
		treatment (inpatient as well as outpatient. An exception are drugs that hospitals bill separately using OPS procedures, for example expensive treatments.
13	Main references	Andersohn F, Walker J "Characteristics and external validity of the German Health Risk Institute (HRI) Database." Pharmacoepidemiology and drug safety (2016): 26530279
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111207 Website: https://www.ingef.de/en/

Lithuanian Digital Health Platform (LDH)

#	Section	Description
1	Data source identification and country	LDH (Lithuanian Digital Health Platform) Lithuania
2	Data partner information section	State Data Agency (SDA) State Data Governance Group
3	Coverage and timespan	Data collection since: 2015 Extent: Nation-wide. The catchment area is the whole country of Lithuania.
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. The following data elements are collected in the source: diagnoses, medications and vaccines, procedures, lab values, hospital admissions (including ICU), pregnancy data, cause of death, biomarker data and devices. For the initial OMOP mapping, only indication and prescription data was included.
5	Data collection process	Outpatient electronic health records. The data is entered by clinicians at healthcare contact (EHR) but can be fully used for analyses by LDH.
6	General representativeness	Near-complete, but varies by dataset (e.g. almost all hospitalizations, reimbursed prescriptions, but not all elective procedures)
7	Data content / source coding	The following source coding is used: ICD10 (AM) for Diagnoses, Cause of Death and Indication for use, ATC codes for medications, ACHI and national vocabularies for Procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. All Lithuanian citizens have unique IDs. Except: new-born infants (0 days to 28 days), not identified persons and foreigners. It should have very small impact on the data analyses.
9	Quality control (data source specific)	Various checks are performed at the time of data creation (hospital or pharmacy EHR systems) and submission to the national EHR system. These may include validation of form fields against input structures, data types, classifiers, and removal of unsigned documents. At the State Data Agency, we additionally remove records marked as retracted due to error, a small number of duplicated death records, clearly erroneous dates, and apply some other minor corrections.
10	Linkage	Yes, family linkage can be created on ad hoc basis. If needed, data can be linked to student and school registries, civil registry, tax registries, social welfare data and to medical notes, discharge letters and diagnostical reports. Linkage to Death Registry is possible.
11	Vital status	Death and Cause of Death will be captures with a linkage to the Death Registry.
12	Limitations	There is no data about medication used in hospital during inpatient visit. Medicine usage is captured, but outcomes only if part of indication. Laboratory tests are currently not available. In the future only laboratory test from some providers can be made available.

#	Section	Description
13	Main references	Juodakis J, Zubavičiūtė S, Dapkutė A, Greičienė S, Masiliūnas R, Narkevičienė I, Trinkūnas J, Žemaitytė J, Literskė G "Data Resource Profile: The Lithuanian health data reuse pathway." International journal of epidemiology (2025): 40618226
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1000000683 Website: https://duomenys.stat.gov.lt/health-data/

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 (UiO) Faculty of Mathematics and Natural Science – Department of Pharmacy
3	Coverage and timespan	Data collection since: 2008 Extent: Nation-wide. 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 – General Practitioner, 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 and the OMOP standard vocabularies. 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.

#	Section	Description
12	Limitations	(i) 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. (iv) The NLHR comprises the following core datasets: • Primary care data: covers the period from 2008 to 2023 • Hospital (secondary care) data, and drug prescription and vaccination data: cover the period from 2018 to 2023 The observation period is defined as follows: • Observation_period_start_date: the later of 2008 or the individual's birth date • Observation_period_end_date: the earlier of the individual's death date or the end of 2023 Immigration and emigration data are also considered when determining the observation period. While most individuals have a single continuous observation period, a small number have two non-overlapping periods. These differing coverage periods across datasets can affect study outcomes. For example, when analyzing hospital-diagnosed conditions, the prevalence appears low from 2008 to 2018, with a sharp increase in incidence in 2018.
13	Main references	Trinh NT, Houghtaling J, Bernal FL, Hayati S, Maglanoc LA, Lupattelli A, Halvorsen L, Nordeng HM "Harmonizing Norwegian registries onto OMOP common data model: Mapping challenges and opportunities for pregnancy and COVID-19 research." International journal of medical informatics (2024): 39153282
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/

The Information System for Research in Primary Care (SIDIAP)

#	Section	Description
1	Data source identification and country	SIDIAP (The Information System for Research in Primary Care) Catalunya, Spain
2	Data partner information section	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (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 inpatient care. 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 Català 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.
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 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the

#	Section	Description
		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, Bolívar 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	Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (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 active population of the dataset is estimated to be around 5.6 million.

#	Section	Description
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 sociodemographic, 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. 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 and the OMOP standard vocabularies. 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. Biological sex is recorded upon registration, but it can be modified upon request by the individual.
13	Main references	García-Sempere A, Orrico-Sánchez A, Muñoz-Quiles C, Hurtado I, Peiró S, Sanfélix-Gimeno G, Díez-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

#	Section	Description
2	Data partner information section	Swedish Medical Products Agency - Gothenburg University (SMPA-GU) Pharmacoepidemiology and Analysis Department (FeA) - School of Public Health and Community Medicine, Institute of Medicine
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 and the OMOP standard vocabularies. 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 Patient Drug Register. And similarly for other registers. 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 death and underlying + contributing causes-of-death are extracted from the Cause-of-Death registry (i.e. based on death certificates).
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 Aurum with Hospital Episode Statistics Admitted Patient Care (CPRD Aurum Linked)

#	Section	Description
1	Data source identification and country	CPRD Aurum Linked (Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care) England, the United Kingdom
2	Data partner information section	Clinical Practice Research Datalink (CPRD) Safety and Surveillance
3	Coverage and timespan	Data collection since: 1995 Extent: Nationwide. The CPRD Aurum database includes just over 17M currently registered patients, or around 25.5% of the UK population. All currently contributing practices (n= 1603) are in England, but coverage varies between English regions. The Aurum OMOP CDM database covers the period from 01/01/1995 to the most recent data collection (December 2025). For the Aurum + linked HES APC database, the coverage period is truncated to ensure that all patients are under observation in both the primary care and linked HES data, therefore the maximum allowable coverage period is currently 01/04/1997 to 31/03/2025 https://www.cprd.com/doi/cprd-primary-care-and-linked-data-omop-common-data-model-march-2026
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. CPRD Aurum includes data captured in UK primary care. In the UK, almost all individuals are registered with a single general practitioner (GP), and around 90% of all healthcare contacts with the UK National Health Service (NHS) occur in primary care (also known as general practice). GPs have a central role as gatekeepers for access to secondary and tertiary services, and other community based services and intervention programmes. Widespread adoption of dedicated computer software in UK primary care began in the late 1980s and was almost universal by 1995. GPs and general practice staff use these software to manage patients' electronic health records which include registration and demographic information, symptoms, diagnoses, clinical investigations and laboratory tests, prescriptions for medicines and appliances, a wide range of behavioural and social factors relevant to health and wellbeing, and administrative information. Some information is captured on contacts with other health services, including referral details and discharge summaries and other communications from those services. CPRD Aurum does not include any associated free text notes, scanned documents or images from the patient record. Hospital Episode Statistics Admitted Patient Care (HES APC) data captures information on inpatient care delivered in NHS hospitals and independent treatment centres funded by the NHS. Key variables include age; sex; ethnicity, episode start/end date; a single primary diagnosis (ICD-10), with up to 19 other comorbidity diagnoses; up to 24 procedures and other interventions (with date), coded using the UK Office of Population Censuses and Surveys Classification of Interventions and Procedures classification (OPCS-4). Medical treatments (drugs and devices) and lab values are not recorded in HES APC.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. Patient EHR data are entered by practice staff using the EMIS Web GP practice management software. Data may also be captured automatically via electronic links from pathology and diagnostic laboratories. EMIS Web is a cloud based system, and practice data are uploaded to the vendor data centre. Pseudonymised data is extracted from the vendor system and downloaded to CPRD via a dedicated cloud based interface. CPRD undertake additional checks to ensure data conform to specification, and comply with established data quality metrics, and apply additional pseudonymisation steps. HES APC data are derived from hospital activity data which are abstracted from clinical records and discharge summaries by trained clinical coders according to centrally issued coding rules and standards. The data is collected centrally for monitoring, planning, and payment purposes. For secondary uses including research, additional cleaning and curation is undertaken to generate the HES APC data. This includes linking episodes for the same individual, and addition of a Master Person Service identifier (MPS ID) which allows linkage to primary care and other health datasets for research.

#	Section	Description
6	General representativeness	CPRD Aurum provides a purposive sample of English general practices using EMIS Web GP software. EMIS Web has an uneven regional coverage. Participation in the CPRD data collection is voluntary. While not a random sample, CPRD Aurum is broadly representative of the overall population in terms of age and gender (https://pubmed.ncbi.nlm.nih.gov/30859197/), with minor differences in terms of area based deprivation and urban-rural distribution. (https://pubmed.ncbi.nlm.nih.gov/35902219/)
7	Data content / source coding	Data is captured using READ, EMIS2, SNOMED, DM+D, ICD10 and OPSC4 codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients can move and receive a different ID number upon changing practice, but a patient can only be registered at one practice at a time. Upon registering at a new practice, a patient's records will be transferred with them, and they will be deregistered at their previous practice. CPRD cannot track patients across practices as the data is de-identified and a patient key is generated for each patient at a practice level, not a country level. The patient key will remain the same for as long as the patient is registered at the practice. The main impact of patients changing practice is to reduce the average follow-up time available for each analytical unit (i.e. patient-within-practice). A secondary issue is that records for the same individual split across two or more practices are not strictly independent. The impact on inferential statistics is likely to be very limited, and analytical adjustment is typically not required (or applied).
9	Quality control (data source specific)	CPRD follows the same three-staged processing pipeline for all the data that it curates: curate individual data collections; integrate data into master database; and extract study specific datasets. For each part of the processing pipeline, CPRD applies checks to ensure that the quality of the data is maintained. At the data curation stage, conducted at both a collection level and at a record level, the data is verified to ensure that the data conforms to an agreed specification. Collection level verification ensures that the collection contains only expected data files, that the collection files are from a known source, and that the collections' files are in the correct sequence. The record level verifications performed at Level 1 are structural in nature and are applied to every data element that CPRD receives without considering the context of the data item against the entirety of the master database. These checks include that the data item is of the correct type, that fields that should have values do indeed contain a value, and that, where specified, the format and length of the data is accurate. Any violations of these checks result in the collection being rejected, the data not being integrated into the master database, and further inspection or recollection of the data taking place. At the data integration stage, the actual content of the data is validated, considering any existing content already collected. These validations are contextual in nature and consider the values of multiple data items within the master database. As CPRD provide real world data no data is modified before being made available for research. The data is either rejected (as per the data curation checks) or integrated. Once integrated, data is not corrected, but patient, practice, and database level data quality metrics and markers are put in place, based on internal consistency of the data, to aid researchers in deciding if the data is suitable for research. There are some exceptions to this no modification principle. In the case that a patient is validated as being opted out, then their data is removed from future versions of the database to ensure their wishes are respected. Further validation and cleaning of study specific datasets may be undertaken by researchers or as part of good data management, and is dependent upon the research question and the researchers performing the analysis.
10	Linkage	Linkage of CPRD Aurum to HES APC is undertaken by NHS England. Identifiers for patients in CPRD Aurum (NHS number, gender, full date of birth, residential postcode) flow directly from the vendor data centre to NHS England, and are linked to HES APC via the MPS ID. A bridging file of patient index numbers for each dataset is returned to CPRD to allow linkage of the clinical data. Linkage is undertaken at approximately 6 month intervals, but this may be impacted by organisation factors external to CPRD. Linkage to additional datasets is expected over the next 12-24 months, including linked death registrations (for accurate cause of death), hospital outpatient activity, and cancer registrations.

#	Section	Description
11	Vital status	The great majority of deaths are eventually captured in patients' primary care record. Published comparisons found that over 98% of national death registrations (the gold standard) were also captured in the CPRD GOLD primary care database https://pubmed.ncbi.nlm.nih.gov/30908785/ , and results are similar for CPRD Aurum (unpublished). Estimated date of death in the primary care record shows good agreement with registered date of death (within 30 days in 98.8% of occurrences after 2013), but cause of death is captured inconsistently.
12	Limitations	CPRD Aurum uses data collected for the primary purpose of managing and delivering safe and effective patient care in general practice. There are requirements and incentives for practice staff to record high quality information, but there is heterogeneity between practices and over time. UK primary care databases have been extensively validated and found to be of sufficient quality to support observational research. CPRD Aurum captures information on prescriptions issued by practice staff, from which drug exposures can only be inferred. There is no information about drug exposures in secondary care via the HES APC.
13	Main references	Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P "Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum." International journal of epidemiology (2019): 30859197
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/30008 Website: https://www.cprd.com/data/primary-care-data/cprd-aurum

ANNEX II. Fitness for use assessment

The included data sources provide coverage of Northern, Central, Western, and Southern European regions.

The included data sources have relatively large and stable denominator data while having linkages to hospital and specialised care data, which is considered key in achieving a representative numerator for the studied conditions.

Croatian National Public Health Information System (NAJS)

NAJS will be included in this study because it is a nation-wide registry with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in NAJS starts in 2014, and the date of the most recent data release is 26/02/2026 (as of June 2026), which aligns with the study period. The median follow-up of the first observation period in NAJS is 4020 days (IQR 3400–4110).

No other specific limitations were identified.

Lastly, NAJS has umbrella approval and therefore does not need to obtain approval on a per study basis, which makes the execution of this study feasible within the current study timelines.

Danish Data Health Registries (DK-DHR)

DK-DHR will be included in this study because it is a nation-wide registry with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in DK-DHR starts in 1995, and the date of the most recent data release is 18/01/2025 (as of June 2026). As the current data extraction does not cover the entire study period, the study will be conducted using the latest available data. The median follow-up of the first observation period in DK-DHR is 7200 days (IQR 2610–10900).

No other specific limitations were identified.

Lastly, DK-DHR has umbrella approval and therefore does not need to obtain approval on a per study basis, which makes the execution of this study feasible within the current study timelines.

Finnish Care Register for Health Care (FinOMOP-THL)

FinOMOP-THL will be included in this study because it is a nation-wide registry with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in FinOMOP-THL starts in 2015, and the date of the most recent data release is 14/04/2026 (as of June 2026), which aligns with the study period. The median follow-up of the first observation period in FinOMOP-THL is 4020 days (IQR 4020–4020).

FinOMOP-THL does not have a specific source code for ALS, but one for the broader and more unspecific “motor neuron disease” concept instead. Similarly, it does not have a specific source code for cardio-renal-metabolic syndrome. Therefore, these two conditions will not be studied in this data source. No other specific limitations present in FinOMOP-THL.

Lastly, FinOMOP-THL is expected to obtain IRB approval within 1–2 months, 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 claims-based data source with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in InGef RDB starts in 2016, and the date of the most recent data release is 01/11/2025 (as of June 2026). As the current data extraction does not cover the entire study period, the study will be conducted using the latest available data. The median follow-up of the first observation period in InGef RDB is 3380 days (IQR 1280–3560).

InGef RDB does not have a specific source code (ICD-10-GM) for ALS, but one for the broader and more unspecific “motor neuron disease, including amyotrophic lateral sclerosis” concept instead. It does not have a specific source code for cardio-renal-metabolic syndrome either. Therefore, these two conditions will not be studied in this data source.

No other specific limitations were identified.

Lastly, InGef RDB does not need Institutional Review Board (IRB) approval, which makes the execution of this study feasible within the current study timelines.

Lithuanian Digital Health Platform (LDH)

LDH will be included in this study because it is a claims-based data source with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in LDH starts in 2015, and the date of the most recent data release is 26/02/2026 (as of June 2026), which aligns with the study period. The median follow-up of the first observation period in LDH is 4020 days (IQR 4020 –4020).

No other specific limitations were identified.

Lastly, LDH is expected to obtain IRB approval within 2 months 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 registry with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in NLHR starts in 2008, and the date of the most recent data release is 29/10/2024 (as of June 2026). As the current data extraction does not cover the entire study period, the study will be conducted using the latest available data. The median follow-up of the first observation period in NLHR is 5820 days (IQR 4200–5820).

No other specific limitations were identified.

Lastly, NLHR is expected to obtain IRB approval within 2–4 weeks, which makes the execution of this study feasible within the current study timelines.

The Information System for Research in Primary Care (SIDIAP)

SIDIAP will be included in this study because it is an electronic health records source with primary care data linked to hospital discharged records that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in SIDIAP starts in 2006, and the date of the most recent data release is 31/12/2024 (as of June 2026). As the current data extraction does not cover the entire study period, the

study will be conducted using the latest available data. The median follow-up of the first observation period in SIDIAP is 5760 days (IQR 2170–6940). SIDIAP does not have a specific source code for cardio-renal-metabolic syndrome.

No other specific limitations were identified.

Lastly, SIDIAP is expected to obtain IRB approval within 2–3 months, 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 an electronic health records source with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in VID starts in 2009, and the date of the most recent data release is 01/10/2025 (as of June 2026). As the current data extraction does not cover the entire study period, the study will be conducted using the latest available data. The median follow-up of the first observation period in VID is 5840 days (IQR 4460–5840).

No other specific limitations were identified.

Lastly, VID is expected to obtain IRB approval within 1–3 months, which makes the execution of this study feasible within the current study timeline.

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

HI-SPEED will be included in this study because it is a nation-wide registry care data source with primary, secondary, and hospital inpatient care data that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in HI-SPEED starts in 2015, and the date of the most recent data release is 31/08/2025 (as of June 2026). As the current data extraction does not cover the entire study period, the study will be conducted using the latest available data. The median follow-up of the first observation period in HI-SPEED is 3890 days (IQR 3330–3890).

No other specific limitations were identified.

Lastly, HI-SPEED will need to submit an ethical amendment application, which usually takes one month. This makes the execution of this study feasible within the current study timelines.

Clinical Practice Research Datalink Aurum with Hospital Episode Statistics Admitted Patient Care (CPRD Aurum Linked)

CPRD Aurum Linked will be included in this study because it is an electronic health records source with primary care data linked to Hospital Episode Statistics admitted patient care that provides relevant information on the conditions of interest in the general population.

Data availability and follow-up in CPRD Aurum Linked starts in 1995, and the date of the most recent data release is 01/09/2025 (as of June 2026). Further, the linked data source coverage only spans until the end of March 2025. As the current data extraction does not cover the entire study period, the study will be conducted using the latest available data. The median follow-up of the first observation period in CPRD Aurum Linked is 1830 days (IQR 684–4530).

No other specific limitations were identified.

Lastly, CPRD Aurum Linked is expected to obtain IRB approval in 2 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 standardised 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) 679/2016 in the various member states.

QUALITY CONTROL

Data source quality control

When defining cohorts for indications, a systematic search of possible codes for inclusion will be identified using the *CodelistGenerator* R package.[3] 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. The *CohortConstructor* R package will be used to generate the relevant cohorts.[7] In addition, the *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.[4] 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.[5, 6] Additionally, *CohortSurvival* will be used to estimate cumulative incidence.[8] 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, the authors plan to prepare a manuscript and an abstract for submission to a scientific conference.

ANNEX IV. Suggested phenotyping plans

For all phenotyping plans, the domains looked for will be “conditions” and “observations”.

Spinal muscular atrophy (SMA)

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Mercuri, Eugenio, et al. "Spinal muscular atrophy." *Nature Reviews Disease Primers* 8.1 (2022): 52.
- Pierzchlewicz, K., & Kotulska, K. (2025). Spinal muscular atrophy clinical aspects—review article. *Next Research*, 100769.

Short description

SMA is a neurodegenerative disease inherited in an autosomal recessive manner which most often affects the anterior horn cells of the spinal cord, leading to the loss of the lower motor neurons, severe muscle weakness, and often early death. In this context, SMA is the genetic disease with the highest mortality rate in infancy and early childhood.

Most cases are caused by mutations in the SMN1 gene located on chromosome 5q (5q-SMA). There are also rare forms of SMA caused by mutations in other genes (non-5q SMA), including distal SMA/hereditary distal motor neuropathy, which differ according to whether the primary pathology affects the anterior horn cells or the motor nerve itself, and other less common forms. When SMN1 is mutated and nonfunctional, the amount of functional SMN protein depends almost entirely on SMN2. SMN2 is a partially functional gene that arose from an evolutionary mutation of SMN1 in primates and is present in nearly all individuals. It only produces a fully functional SMN protein in about 10% of transcripts. The greater the number of SMN2 copies, the milder the disease phenotype when SMN1 is defective. Based on clinical severity, SMA is classified into four types (0–4), with Type I being the most severe and Type IV the mildest. These SMA subtypes can be divided in those with paediatric onset (types 1 to 3) and two less common types, one with antenatal onset (type 0) and one with adult onset and milder phenotype (type 4). They also differ in the survival without treatment. In SMA type 0, the onset is foetal, and the survival with treatment based solely on supportive care is days-weeks. In SMA type 1, the typical age at presentation is 3 months, and the survival with treatment based solely on supportive care is months. In SMA type 2, the onset is typically at 12 months (7–18 months), and the survival with treatment based solely on supportive care is (median) 20 years. In SMA type 3, the onset is typically at 3 years (1.5–10 years), and the survival with treatment based solely on supportive care is normal. In SMA type 4, the typical age at onset is >18 years, and the survival with treatment based solely on supportive care is also normal. This classification is based on the age of onset and maximum motor function achieved in untreated patients and does not capture the change in disease phenotype due to treatment with the new therapies.

Names of variants (flavours) of cohorts and resulting proposed logic for the cohorts

A priori, only one flavour will be developed.

4 main subtypes can be identified: type I, type II, type III, and type IV. Type 0 will not be included because it is foetal and very rare.

Since the counts for these subtypes are very low, it is suggested to operationalise them as age at onset, which should be similar to age at diagnosis, given the high impact that these diseases have. It is noted that there might be, however, substantial overlap between types II and III. While these two types might differ in survival, this would require a significant extension of the follow-up and would still not be able to differentiate with certainty the two subtypes.

Proposed Search strategy

- Spinal muscular atrophy OR
- Werdnig-Hoffmann OR
- Kugelberg-Welander OR
- //SMA//

Amyotrophic lateral sclerosis (ALS)

Information source

- Dynamed (Oxford – https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny/)
- Quinn, C., & Elman, L. (2020). Amyotrophic lateral sclerosis and other motor neuron diseases. *CONTINUUM: Lifelong Learning in Neurology*, 26(5), 1323-1347.
- Souza PVS, Serrano PL, Farias IB, Machado RIL, Badia BML, Oliveira HB, Barbosa AS, Pereira CA, Moreira VF, Chieia MAT, Barbosa AR, Braga VL, Pinto WBVR, Oliveira ASB. Clinical and Genetic Aspects of Juvenile Amyotrophic Lateral Sclerosis: A Promising Era Emerges. *Genes (Basel)*. 2024 Feb 28;15(3):311. doi: 10.3390/genes15030311. PMID: 38540369; PMCID: PMC10969870.
- Antonioni A, Raho EM, Lopriore P, Pace AP, Latino RR, Assogna M, Mancuso M, Gragnaniello D, Granieri E, Pugliatti M, Di Lorenzo F, Koch G. Frontotemporal Dementia, Where Do We Stand? A Narrative Review. *Int J Mol Sci*. 2023 Jul 21;24(14):11732. doi: 10.3390/ijms241411732. PMID: 37511491; PMCID: PMC10380352.

Short description

ALS is a disease within the motoneuron disease spectrum that remains a devastating, progressive, and terminal condition with relatively limited treatment options. The presentation is heterogeneous, and prognosis is variable. The diagnosis of ALS is still clinical. Clinicians must also be familiar with mimics of ALS and other less-common motor neuron diseases to ensure accurate diagnosis. Importantly, some of the disorders within the spectrum will evolve into ALS with time. Emerging disease-modifying therapies provide modest benefit in slowing progression and prolonging survival, and the advent of gene-based therapies offers hope.

Approximately 90% of patients with ALS have no known family history and are thus classified as having sporadic or acquired disease. Familial ALS related to mendelian, monogenic inheritance occurs in at least 10% of patients with ALS, and most of these cases are inherited in an autosomal dominant manner. The C9orf72 expansion is the most common monogenic cause of disease, accounting for 40% of familial ALS with a penetrance of up to 100% by age 80. Patients with the C9orf72 expansion can present with ALS, frontotemporal dementia (FTD), or a combination of both. More rarely, C9orf72 expansions can be associated with features of other neurodegenerative diseases, including parkinsonism, tremor, myoclonus, corticobasal syndrome, Huntington disease-like syndrome, Creutzfeldt-Jakob disease, and Alzheimer's disease. The SOD1 variation is the second most common cause of familial ALS.

Juvenile Amyotrophic Lateral Sclerosis is a genetically heterogeneous neurodegenerative disorder, which is frequently misdiagnosed due to low clinical suspicion and little knowledge about disease characteristics. While different age thresholds are used, the most common one is before the age of 25.

Names of variants (flavours) of cohorts and resultant Proposed Logic for Cohort/s

A priori, only one main broad flavour will be developed.

The two main subtypes will be:

- One regarding the age at onset:
 - o Adult-onset ALS
 - o Juvenile-onset ALS (<25 years old)
- One regarding whether it is associated with FTD:
 - o ALS only
 - o ALS-FTD

There is no available data to establish other potential subtypes, such as genetic forms or familial versus sporadic.

Although there is a specific code for juvenile-onset ALS, the counts for this specific diagnosis are low. Therefore, this sub-phenotype will be operationalised based on age at first diagnosis. It should be acknowledged as a limitation that there is often a gap between clinical onset and diagnosis in most cases.

Given that the counts for the concepts “ALS with dementia” and “ALS with frontotemporal dementia” are extremely low, these subtypes will be operationalised as those individuals who have the two diagnoses at any time within a two-year pre- and post- index date window (given that both are chronic, incurable conditions). The index date will be defined as first ALS diagnosis.

Proposed Search strategy

- ALS concept
Amyotrophic lateral sclerosis OR
Lou Gehrig's disease OR
\\ALS\\
- FTD concept
Frontotemporal dementia OR
Frontotemporal lobar degeneration OR
Frontotemporal degeneration OR
Frontal lobe dementia OR
\\FTD\\

Multiple sclerosis (MS)

Information source

- Dynamed (Oxford – https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Cocco, E. (2026). Continuum. Lifelong learning in Neurology. Multiple sclerosis and related disorders

- Leray, Emmanuelle, et al. "Epidemiology of multiple sclerosis." *Revue neurologique* 172.1 (2016): 3–13.

Introduction

MS is a chronic immune-mediated disease of the central nervous system characterised by inflammation, demyelination, and neurodegeneration. Its aetiology remains elusive, and a combination of genetic, environmental, and epigenetic factors has been associated with MS. Its heterogeneous clinical presentation and course require a comprehensive approach to diagnosis that integrates clinical, imaging, and cerebrospinal fluid assessments.

Typically, there are two distinct phases in the disease:

- The relapsing-remitting (RR) phase is characterised by clinical events affecting different regions of the CNS that are separated in time (by at least one month).
- The progressive phase is defined by at least one year of continuous clinical worsening, with or without ongoing disease activity. This phase is most strongly associated with neurodegeneration and long-term disability and represents the stage of the disease in which therapeutic interventions are least effective.

When the progressive phase starts after a RR phase, it is called secondary progressive multiple sclerosis. Approximately 80% of RR patients eventually develop a progressive phase, and around 50% of patients with relapsing-remitting MS progress to secondary progressive MS within 10–15 years. Progressive disease, however, can appear without a prior RR phase and is then called primary progressive multiple sclerosis. Primary progressive MS can arise from a radiologically isolated syndrome or from a single attack (single-attack progressive MS).

Regardless of the disease phase, disease activity should be always specified. Although the traditional distinction between primary and secondary progressive MS remains in use, current classifications increasingly group these entities under the broader category of progressive MS, placing greater emphasis on whether the disease is active or inactive and whether there is ongoing progression. Notably, progression may occur independently of inflammatory activity. Indeed, the MS Phenotype Group recommended in 2013 that the diagnosis of relapsing-remitting, primary progressive, and secondary progressive MS includes a current time-defined assessment of the patient's disease activity and progression. Therefore, all cases of MS should be classified according to:

- Disease course: Relapsing-remitting MS versus progressive MS (either secondary (progressive accumulation of disability after initial relapsing course) or primary (progressive accumulation of disability from onset) progressive MS).
- Disease activity: Active or inactive, based on either clinical evidence (relapses) or radiological evidence (new T2 lesions and/or gadolinium-enhancing lesions on magnetic resonance imaging (MRI)).
- For progressive forms: With progression or without progression.

Clinical presentation/Symptoms

Most patients with MS initially present with a clinically isolated syndrome, which is a monophasic episode characterized by symptoms and objective findings that reflect a monofocal or multifocal inflammatory demyelinating event in the CNS. These syndromes occur acutely or subacutely and last at least 24 hours, with or without recovery, in the absence of fever or infection. The terms attack, relapse, or exacerbation are synonymous, but a clinically isolated syndrome refers to the first episode suggestive of MS. The most frequent presentations of clinically isolated syndrome in MS are optic neuritis, brainstem or cerebellar syndromes, and myelitis.

- Optic neuritis: In MS, patients with optic neuritis usually present with a combination of mild-to-moderate unilateral visual loss, a central scotoma, mild periorbital or retro-orbital pain worsening on eye movements, impairment in colour vision or low-contrast visual acuity, a relative afferent pupillary defect, or a normal-appearing or mildly swollen optic disc.
- Brainstem or cerebellar syndromes: The most frequent symptoms of MS involvement in this region include diplopia, oscillopsia, unilateral facial numbness, and gait instability. On examination, typical findings include internuclear ophthalmoplegia; sixth nerve palsy; gaze-evoked nystagmus; and ocular, limb, or gait ataxia.
- Myelitis: In relapsing MS, patients with acute myelitis usually present with mild-to-moderate asymmetric sensory or motor symptoms and deficits, with or without sphincter dysfunction.

Background/Epidemiology

Although most people with MS are between 20 and 40 years old at symptom onset, MS can develop later in life. When MS appears in adults older than 50 years, it is referred to as late-onset MS.

Multiple sclerosis (MS) is the most common demyelinating disease seen in high-income countries, and has a heterogeneous prevalence worldwide: it is highest in North America (140/100,000 population) and Europe (108/100,000), and lowest in East Asia (2.2/100,000 population) and sub-Saharan Africa (2.1/100,000). Global age-standardized incidence of MS in 2019 was 0.7 cases per 100,000 persons, with the highest incidences occurring in regions with a high socio-demographic index.

The typical ages of onset for MS are a mean age of 25-29 years for relapsing-remitting MS and a mean age of 39-41 years primary progressive MS. Over the past 30 years, the MS gender ratio has changed due to an increase in the incidence of MS in women.

Assessment/Diagnosis

Comprehensive knowledge of the clinical, MRI, and cerebrospinal fluid findings typical of MS is critical for the application of the 2024 McDonald criteria. The criteria were not developed to differentiate MS from other conditions, but rather to identify MS or a high likelihood of the disease in patients with typical presentations once other diagnoses have been reasonably excluded. That is, MS remains a diagnosis of exclusion, and there must be no better explanation other than MS prior to its diagnosis.

The accurate interpretation of MRI findings is essential for the diagnosis of MS, particularly as the 2024 McDonald criteria now permit diagnosis in people with nonspecific clinical presentations or incidental imaging findings alone. Specific anatomic locations in the central nervous system are typical for MS lesions: periventricular, juxtacortical, spinal cord, optic nerve.

The diagnosis requires dissemination in space (≥ 2 locations affected) and in time of the lesions, but these concepts have changed in the 2024 McDonald criteria (with respect to the prior version (2017) of the criteria). The optic nerve now serves as a fifth location (in addition to periventricular, cortical or juxtacortical, infratentorial, spinal cord), and Optical coherence tomography and visual evoked potential can now be used (in addition to MRI) for assessing its involvement. In the 2024 revisions, an alternative criterion for dissemination in space can be used exclusively in patients with progressive onset: in these patients, at least two spinal cord lesions are considered enough to demonstrate dissemination in space. Dissemination in time is fulfilled when a patient experiences a new attack, has new T2 lesions, or has simultaneous contrast-enhancing and non-enhancing lesions. Cerebrospinal fluid-restricted IgG Oligoclonal bands or the central vein sign select-6 criteria can substitute for dissemination in time.

Disqualifiers/Differential Diagnoses *

It is widely acknowledged that many diseases can mimic the clinical and MRI appearance of MS. Like prior iterations, the 2024 McDonald criteria do not distinguish MS from other disorders and continue to rely on

clinical expertise for the consideration of a better explanation than MS as a key element of its application to patient presentations.

There are several clinical red flags that suggest diagnoses other than MS, such as transient neurologic symptoms lasting less than 24 hours, severe or bilateral optic neuritis, severe unilateral optic neuritis with poor recovery, severe transverse myelopathy with hyperacute onset and bilateral sensorimotor symptoms, multiple cranial neuropathies, isolated headache, and isolated fatigue.

There are also MRI red flags, such as meningeal enhancement, increasing lesions, macrobleeds, infarcts, cavities, and symmetric lesions.

Strengtheners/pathognomonic signs*

Since MS is a diagnostic of exclusion, there are no pathognomonic symptoms nor signs.

Therapeutic plan/Treatment

There are different approaches to the anti-inflammatory, disease-modifying treatment: early effective with or without subsequent de-escalation or an escalation approach.

Among those with low relative efficacy: glatiramer acetate and interferon; among those with medium relative efficacy: dimethyl fumarate, teriflunomide, fingolimod, cladribine; among those with high relative efficacy: natalizumab, antiCD20 (ocrelizumab, rituximab), antiCD52 (alemtuzumab).

Complications and prognosis

Several complications may occur, including cognitive impairment and psychiatric complications, seizures, and a higher risk of cardiovascular events.

Further, MS with or without disease-modifying therapy is associated with an increased risk of both outpatient and serious infection.

Patients with MS are estimated to live a mean 5 years less than healthy controls. However, this may not take in account greater benefits of recent advances in treatments for MS.

Stressful life events associated with increased risk of exacerbations of MS. Factors associated with a poorer prognosis in terms of disease course and disability accumulation in patients with MS include incomplete recovery after the first attack, a shorter interval between the first 2 attacks, early bowel and bladder symptoms, smoking, and Black American ethnicity. 59% of patients with primary progressive MS require unilateral walking assistance within 10 years

Names of variants (flavours) of cohorts and resultant Proposed Logic for Cohort/s

A priori, only one main broad flavour will be developed.

Distinguishing whether MS has a progressive component is not feasible without a specific diagnosis of a progressive form, which is only available in CPRD Aurum Linked.

Therefore, the two subtypes proposed are:

- MS with clinical activity (i.e. relapsing MS)
- MS without clinical activity

Clinical activity could be operationalised as follows in an individual with an MS diagnosis and at least one of the following:

- Another MS diagnosis recorded as the primary diagnosis during a hospital admission or emergency department visit (acknowledging that some of these events will inevitably represent pseudo-relapses).

- One of the following diagnoses: optic neuritis (or related terms), myelitis (or related terms), or brainstem syndrome (excluding vertigo, as the most common cause of vertigo in individuals with MS is not a relapse). Some symptoms and/or signs may be found across different clinical syndromes, such as paraesthesia (facial paraesthesia would indicate brainstem syndrome, whereas limb paraesthesia would suggest myelitis).
- An administration of glucocorticoids at high doses (usually 1000 mg, either as a single pulse or as two tablets of 500 mg).

Proposed Search strategy

- Multiple sclerosis

//MS// OR

Multiple sclerosis

- Optic neuritis

Optic neur* OR

//ON// OR

Inflammation of the optic OR

Papillitis

- Myelitis

Myelitis OR

Myelop* OR

Limb weakness OR

Paraesthesia OR

Urinary incontinence OR

Faecal incontinence

- Brainstem syndrome

Diplopia OR

Internuclear ophthalmoplegia OR

Nystagmus OR

Skew OR

Brainstem OR

Ataxia OR

Instability OR

Incoordination OR

Imbalance OR

Dysarthria OR

Dysphagia OR

Vertigo OR

Trigeminal neuralgia OR

Facial palsy OR

Paraesthesia

Prurigo nodularis (PN)

Information source

- Huang AH, Williams KA, Kwatra SG. Prurigo nodularis: Epidemiology and clinical features. J Am Acad Dermatol. 2020 Dec;83(6):1559-1565. doi: 10.1016/j.jaad.2020.04.183. Epub 2020 May 23. PMID: 32454098.
- Williams KA, Huang AH, Belzberg M, Kwatra SG. Prurigo nodularis: Pathogenesis and management. J Am Acad Dermatol. 2020 Dec;83(6):1567-1575. doi: 10.1016/j.jaad.2020.04.182. Epub 2020 May 24. PMID: 32461078.
- Kwatra SG, Pereira MP, Misery L, Mollanazar NK, Shah P, Wiggins S. Prurigo nodularis: disease burden, clinical features and approach to management. Br J Dermatol. 2025 Sep 18;193(4):642-652. doi: 10.1093/bjd/ljaf213. PMID: 40665436.

Short description

PN is a chronic skin condition characterised by severely pruritic nodules that cause a profound negative impact on quality of life. The pathogenesis of PN is thought to be a cutaneous reaction pattern caused by vicious cycles of chronic itch followed by repeated scratching. The exact pathogenesis of PN remains unknown. However, previous studies show that significant interaction and dysregulation between immune cells and neuronal circuitry play important roles in the pathogenesis of PN.

Names of variants/flavours or cohorts

A priori, only one main broad flavour will be developed.

Since there are no specific codes to identify drug-induced PN, no specific drug-induced sub-phenotype will be developed. However, as part of the characterisation of PN, prescriptions of drugs that may be related to the development of PN in the 90 and 30 days prior to the diagnosis (index date) will be identified and described as a proportion. Such proportion could be interpreted as potentially drug-induced PN.

Proposed search strategy

Prurigo nodularis OR

Nodular prurigo OR

Hyde

Bullous pemphigoid (BP)

Information source

- Miyamoto D, Santi CG, Aoki V, Maruta CW. Bullous pemphigoid. An Bras Dermatol. 2019 Mar-Apr;94(2):133-146. doi: 10.1590/abd1806-4841.20199007. Epub 2019 May 9. PMID: 31090818; PMCID: PMC6486083.

Short description

BP presents with pruritic, tense, clear bullae (blisters) on the skin in older adults. There is currently no standardised classification of BP. However, it is possible to recognise distinct variants of the disease according to the age of onset, clinical presentation, and triggering factors:

- Classic BP
- No bullous BP
- Exfoliative erythroderma
- Infantile and childhood BP

Names of variants (flavours) of cohorts

A priori, only one main broad flavour will be developed.

Since there are no specific codes to identify drug-induced BP, no specific drug-induced sub-phenotype will be developed. However, as part of the characterisation of BP, prescriptions of drugs that may be related to the development of BP in the 90 and 30 days prior to the diagnosis (index date) will be identified and described as a proportion. Such proportions could be framed as a potentially drug-induced BP.

Proposed Search strategy

- Bullous#pemphigoid OR
- //BP// OR
- Pemphigoid bullous OR
- Lever* disease OR
- Senile pemphigoid

Vitiligo

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny/)
- Bergqvist C, Ezzedine K. Vitiligo: A Review. Dermatology. 2020;236(6):571-592. doi: 10.1159/000506103. Epub 2020 Mar 10. PMID: 32155629.
- Ezzedine K, Eleftheriadou V, Whitton M, van Geel N. Vitiligo. Lancet. 2015 Jul 4;386(9988):74-84. doi: 10.1016/S0140-6736(14)60763-7. Epub 2015 Jan 15. PMID: 25596811.

Short description

Vitiligo is an acquired chronic depigmenting disorder of the skin resulting from selective destruction of melanocytes which results in typical non-scaly, chalky-white macules. It is the most frequent cause of depigmentation worldwide. The autoimmune or autoinflammatory theory is the leading hypothesis for causation and is supported by strong evidence.

Vitiligo can be classified into two major forms: non-segmental vitiligo, also known as, simply, vitiligo, and segmental vitiligo. Non-segmental vitiligo, the commonest form of this unpredictable disease, is characterised by symmetrical and bilateral white patches. Different clinical subtypes have been described,

including generalised, acrofacial, and universalis, mixed types, all with a bilateral distribution. In terms of age of onset, two phenotypes of non-segmental vitiligo have been distinguished; the first is of early onset (before the age of 12 years) and it is often associated with halo naevus and a familial background of premature hair greying, whereas the second is of late onset and is most often associated with an acrofacial pattern.

Segmental vitiligo is less common (accounts approximately for 5–16% of vitiligo cases) than non-segmental vitiligo and usually has a unilateral distribution. Segmental vitiligo is further classified as uni-segmental, bi-segmental, or pluri-segmental.

Names of variants (flavours) of cohorts

Since there are no specific codes to identify segmental vs non-segmental nor drug-induced vitiligo, only one broad variant will be developed.

Since there are no specific codes to identify drug-induced vitiligo, no specific drug-induced sub-phenotype will be developed. However, as part of the characterisation of vitiligo, prescriptions of drugs that may be related to the development of vitiligo in the 90 days prior to the diagnosis (index date) will be identified and described as a proportion. Such sub-phenotype could be framed as a potentially drug-induced vitiligo.

Proposed Search strategy

- “Vitiligo”

Lupus

Information source

- Dynamed (Oxford – https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny/)
- Kiriakidou M, Ching CL. Systemic Lupus Erythematosus. Ann Intern Med. 2020 Jun 2;172(11):ITC81–ITC96. doi: 10.7326/AITC202006020. PMID: 32479157.
- Rubin RL. Drug-induced lupus. Expert Opin Drug Saf. 2015 Mar;14(3):361–78. doi: 10.1517/14740338.2015.995089. Epub 2015 Jan 2. PMID: 25554102.

Short description

Systemic lupus erythematosus (SLE)

SLE is a condition in which the immune system attacks healthy cells and tissues throughout the body. Production and defective elimination of antibodies, circulation, and tissue deposition of immune-complexes and complement and cytokine activation contribute to clinical manifestations that range from mild fatigue and joint pain to severe, life-threatening organ damage. The latter group involves cases of pre-existing SLE which remain or recur after withdrawal of the implicated medication, while bona fide drug-induced lupus usually occurs in the setting of a previously normal immune system and disappears after discontinuation of the medication.

A wide variety of drugs have been thought to exacerbate SLE, including antibiotics, antiseizures drugs, hormones, NSAIDs; particularly implicated are hydralazine, sulphonamides, penicillin, *para*-aminosalicylic acid, hydrochlorothiazide, cimetidine, phenylbutazone, and mesantoin.

Cutaneous lupus erythematosus (CLE)

CLE is an autoimmune disease affecting the skin, presenting as either an isolated condition or in the setting of SLE. They have been classified in lupus erythematosus specific and non-specific lesions. The former is further divided into acute (which includes a localised and a generalised form), subacute (annular form and papulosquamous form), chronic (discoid lupus erythematosus (localised and disseminated form), lupus erythematosus profundus, chilblain lupus erythematosus), and intermittent (lupus erythematosus tumidus). Chronic CLE accounts for the majority (reportedly about 80%) of CLE cases, with discoid lupus erythematosus being the most common subtype.

Drug-induced systemic lupus is more common than drug-induced cutaneous lupus, although both are possible.

Names of variants (flavours) of cohorts

A priori, three lupus flavours will be defined: overall lupus (including both SLE and CLE and unspecified forms of lupus), SLE, and CLE.

In addition, SLE and CLE will be treated as subtypes, allowing their prevalence to be described as a proportion of the overall lupus phenotype.

Since there are no specific codes to identify drug-induced lupus in most data sources, no specific drug-induced sub-phenotype will be developed. However, as part of the characterisation of lupus, prescriptions of drugs that may be related to the development of lupus in the 90 and 30 days prior to the diagnosis (index date) will be identified and described as a proportion. Such proportions will be described for each of the three flavours (lupus overall, CLE, and SLE), and they could be framed as a potentially drug-induced lupus (non-specified or either cutaneous or systemic).

Proposed Search strategy

Lupus overall

- Lupus OR
- systemic lupus erythematosus OR
- systemic lupus OR
- lupus erythematosus OR
- //SLE// OR
- disseminated lupus erythematosus OR
- generalized lupus erythematosus OR
- generalised lupus erythematosus OR
- lupus disseminatus OR
- cutaneous lupus erythematosus OR
- cutaneous lupus OR
- skin lupus OR
- discoid lupus OR
- lupus panniculitis OR
- lupus profundus OR
- hypertrophic lupus OR

- tumid lupus OR
- lupus erythematosus tumidus OR
- mucosal lupus OR
- //CLE// OR
- //DLE// OR

SLE:

- Lupus OR
- systemic lupus erythematosus OR
- systemic lupus OR
- lupus erythematosus OR
- //SLE// OR
- disseminated lupus erythematosus OR
- generalized lupus erythematosus OR
- generalised lupus erythematosus OR
- lupus disseminatus

Excluding: "cutaneous" OR skin

CLE:

- cutaneous lupus erythematosus OR
- cutaneous lupus OR
- skin lupus OR
- discoid lupus OR
- lupus panniculitis OR
- lupus profundus OR
- hypertrophic lupus OR
- tumid lupus OR
- lupus erythematosus tumidus OR
- mucosal lupus OR
- //CLE// OR
- //DLE//

Excluding: nephritis, neuropsychiatric, neurologic, psychiatric, carditis, vasculitis, pneumonitis, hematologic

Vasculitis

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Ecclestone T, Watts RA. Classification and epidemiology of vasculitis: Emerging concepts. *Best Pract Res Clin Rheumatol*. 2023 Mar;37(1):101845. doi: 10.1016/j.berh.2023.101845. Epub 2023 Jul 18. PMID: 37468418.
- Romero Gómez C, Hernández Negrín H, Ayala Gutiérrez MDM. Eosinophilic granulomatosis with polyangiitis. *Med Clin (Barc)*. 2023 Apr 6;160(7):310–317. English, Spanish. doi: 10.1016/j.medcli.2023.01.003. Epub 2023 Feb 9. PMID: 36774291.
- Parums DV. A Review of IgA Vasculitis (Henoch-Schönlein Purpura) Past, Present, and Future. *Med Sci Monit*. 2024 Jan 28;30:e943912. doi: 10.12659/MSM.943912. PMID: 38281080; PMCID: PMC10832303.
- Martins–Martinho J, Dourado E, Khmelinskii N, Espinosa P, Ponte C. Localized Forms of Vasculitis. *Curr Rheumatol Rep*. 2021 Jul 1;23(7):49. doi: 10.1007/s11926-021-01012-y. PMID: 34196889; PMCID: PMC8247627.
- Byram, K., Hajj-Ali, R.A. & Calabrese, L. CNS Vasculitis: an Approach to Differential Diagnosis and Management. *Curr Rheumatol Rep* **20**, 37 (2018). <https://doi.org/10.1007/s11926-018-0747-z>

Short description

Vasculitis can be broadly classified into systemic and organ-specific (or localised).

Systemic vasculitis

Systemic vasculitis are heterogeneous disabling diseases characterised by chronic inflammation of the blood vessels, potentially leading to tissue destruction and organ failure.

One classification and diagnostic criteria is that of the American College of Rheumatology/European Alliance of Associations for Rheumatology, which identifies the following:

- Large vessel:
 - Takayasu arteritis is a rare, chronic, often granulomatous vasculitis mainly affecting the aorta and its major branches, and the coronary and pulmonary arteries.
 - Giant cell arteritis is a chronic, systemic vasculitis involving large- and medium-sized arteries, often affecting the temporal artery, as well as the aorta and/or its major branches.
- Medium vessel vasculitis:
 - Kawasaki disease is an acute, self-limiting disease predominantly affecting medium-sized vessels, especially the coronary arteries, resulting in coronary artery aneurysms, as well as inflammation of the myocardium. Consequently, it is the leading cause of acquired childhood heart disease in high-income countries.
 - Polyarteritis nodosa is a systemic necrotising vasculitis that predominantly targets medium sized arteries.
- Small vessel vasculitis:
 - ANCA-associated vasculitis:
 - Microscopic polyangiitis is characterised by necrotising vasculitis of small-sized blood vessels without granulomatous inflammation or immune deposits,

commonly leading to pulmonary capillaritis and necrotising glomerulonephritis. The most affected organ is the kidney.

- Granulomatosis with polyangiitis (formerly Wegener granulomatosis) is characterised by necrotising vasculitis of small- to medium-sized blood vessels and necrotising granulomatous inflammation of the upper and lower respiratory systems.
- Eosinophilic granulomatosis with polyangiitis (formerly known as Churg-Strauss syndrome) is a rare, ANCA-associated necrotising small- to medium-vessel vasculitis characterised by eosinophil-rich granulomatous inflammation, asthma, and peripheral blood eosinophilia. It commonly affects the respiratory tract, lungs, skin, peripheral nerves, and heart.
- IgA vasculitis (formerly known as Henoch-Schoenlein Purpura) is an immune complex-mediated small-vessel vasculitis characterised by the deposition of IgA1-containing immune complexes in the walls of small vessels. It primarily affects the skin, joints, gastrointestinal tract, and kidneys and is the most common systemic vasculitis in children, although it also occurs in adults.

Localised or single organ vasculitis

Single organ vasculitis has been defined as vasculitis in arteries or veins of any size in a single organ which has no features that indicate whether it is a limited expression of a systemic vasculitis. Single organ vasculitis should be named according to the specific organ involved and pattern of inflammation observed. Their inflammatory process distribution may be multifocal/diffuse (e.g. skin, renal, peripheral nervous system) or unifocal (e.g. aorta, breast, gastrointestinal, and urogenital structures). A few rare cases of patients who are initially diagnosed as having single organ vasculitis may later evolve into systemic disease. Currently, it incorporates the following vasculitis:

- Cutaneous leukocytoclastic angiitis (formerly known as cutaneous small vessel vasculitis, hypersensitivity vasculitis, and cutaneous necrotising vasculitis) represents a vascular reaction pattern, mainly due to circulating immune complexes that may include idiopathic (or primary) and secondary presentations.
- Cutaneous arteritis (formerly known as cutaneous polyarteritis nodosa): Although it is still debated whether cutaneous arteritis should be considered a cutaneous form of polyarteritis nodosa, the 2012 Revised International Chapel Hill Consensus Conference clearly separates these entities.
- Primary central nervous system vasculitis is a rare and diffuse “single organ” vasculitis affecting the brain and spinal cord. It is often a severe form of vasculitis, with a wide range of possible neurological symptoms and signs that frequently requires aggressive immunosuppressive therapy.
- Isolated aortitis: Aortitis encompasses a group of non-infectious and infectious disorders leading to inflammation of the aorta. Many systemic diseases may involve the aorta, namely giant cell arteritis, Takayasu’s arteritis, other vasculitis (e.g. Behçet’s disease, Cogan’s syndrome, relapsing polychondritis), non-vasculitis rheumatic disorders (rheumatoid arthritis, ankylosing spondylitis, systemic lupus erythematosus), IgG4-related disease, and sarcoidosis. According to the 2012 Revised International Chapel Hill Consensus Conference nomenclature, “isolated aortitis” is the preferred term when aortitis is identified without other accompanying features suggestive of systemic vasculitis. Currently, it is still unclear whether isolated aortitis is a distinct disorder or, by contrast, a group of heterogeneous diseases with the same initial presentation.
- Others

Names of variants (flavours) of cohorts and resultant proposed Logic for Cohort/s

Two main flavours will be developed:

- A broad, more sensitive flavour which includes primary, secondary, and unspecified vasculitis. This flavour would comprise the non-specific concept of “vasculitis” OR any of all the following: Giant cell arteritis, Takayasu arteritis, Kawasaki disease, Polyarteritis nodosa, Microscopic polyangiitis, Granulomatosis with polyangiitis, Eosinophilic granulomatosis with polyangiitis, IgA vasculitis, Cutaneous leukocytoclastic angiitis, Primary central nervous system vasculitis (“cerebral vasculitis”), aortitis.
- A narrow flavour that would only include vasculitis which are most likely primary. This narrow flavour will apply exclusion criteria consisting of record of infection or drug of interest (i.e. which could potentially be associated with vasculitis according to existing literature) in the 30 days prior to the index date.

Since there are no specific codes to identify drug-induced vasculitis in most data sources, no specific drug-induced sub-phenotype will be developed. However, as part of the characterisation of vasculitis, prescriptions of drugs that may be related to the development of vasculitis in the 90 and 30 days prior to the diagnosis (index date) will be identified and described as a proportion for the vasculitis broad variant. Such proportion will be framed as the proportion of overall vasculitis that are potentially drug-induced.

Proposed Search strategy

Broad flavour:

- Vasculitis OR angiitis OR polyangiitis OR arteritis OR polyarteritis OR vessel inflammation
- Giant cell arteritis OR
- Takayasu OR
- Kawasaki OR
- Polyarteritis nodosa OR //PAN// OR necrotizing arteritis OR necrotizing vasculitis OR periarteritis nodosa OR
- Microscopic polyangiitis OR microscopic polyarteritis OR //MPA// OR
- Granulomatosis with polyangiitis OR //GPA// OR Wegener OR
- Eosinophilic granulomatous with polyangiitis OR Churg OR //EGPA// OR allergic granulomatosis OR allergic angiitis OR
- IgA vasculitis OR Ig A mediated vasculitis OR Henoch OR //HSP// OR
- Cutaneous leukocytoclastic OR hypersensitivity vasculitis OR
- Primary central nervous system vasculitis OR //PACNS// OR primary central nervous system angiitis OR angiitis of the central nervous system OR vasculitis of the central nervous system OR angiitis of the CNS OR vasculitis of the CNS OR cerebral vasculitis OR cerebral angiitis OR cerebral arteritis OR

Narrow flavour:

- Giant cell arteritis OR
- Takayasu OR
- Kawasaki OR

- Polyarteritis nodosa OR //PAN// OR necrotizing arteritis OR necrotizing vasculitis OR periarteritis nodosa OR
- Microscopic polyangiitis OR microscopic polyarteritis OR //MPA// OR
- Granulomatosis with polyangiitis OR //GPA// OR Wegener OR
- Eosinophilic granulomatous with polyangiitis OR Churg OR //EGPA// OR allergic granulomatosis OR allergic angiitis OR
- IgA vasculitis OR Ig A mediated vasculitis OR Henoch OR //HSP// OR
- Cutaneous leukocytoclastic OR hypersensitivity vasculitis OR
- Primary central nervous system vasculitis OR //PACNS// OR primary central nervous system angiitis OR angiitis of the central nervous system OR vasculitis of the central nervous system OR angiitis of the CNS OR vasculitis of the CNS OR cerebral vasculitis OR cerebral angiitis OR cerebral arteritis

Excluding: all those who have a record of any infection, oncologic disease and/or any of the following drugs in the prior 30 days: antibiotics (e.g. cephalexime), anti-thyroid drugs, anti-TNF- α inhibitors (adalimumab, etanercept, infliximab), clozapine, allopurinol, D-penicillamine, hydralazine, phenytoin, sulfasalazine, propylthiouracil, methimazole, carbimazole, benzyl thiouracil.[13]

Psoriasis

Short description

Psoriasis is an immune-mediated, genetic disease manifesting in the skin or joints or both. Psoriasis provides many challenges, including high prevalence, chronicity, disfigurement, disability, and associated comorbidity.

There are five types of psoriasis, according to the clinical presentation:

- Plaque psoriasis (psoriasis vulgaris): Chronic plaque psoriasis (psoriasis vulgaris) is the most common form of the disease and accounts for about 90% of cases.
- Guttate (droplet) or eruptive psoriasis: it is the typical form of onset of psoriasis in children, often after a streptococcal infection of the upper respiratory tract.
- Inverse or intertriginous or flexural psoriasis
- Pustular psoriasis (palmoplantar vs generalised)
- Erythrodermic psoriasis

Psoriatic arthritis (PsA) is a heterogeneous condition with musculoskeletal involvement, including arthritis, enthesitis, dactylitis, and axial involvement, as well as potential skin and nail disease. PsA accounts for around 20% of referrals to the early arthritis clinic and presents a significant diagnostic and management challenge. Early diagnosis is important to prevent long term functional disability and to ensure optimal management of arthritis and key comorbidities. In 1964, PsA was recognised as a separate disease by the American College of Rheumatology and is now included as a member of the spondyloarthropathy spectrum.

Names of variants (flavours) of cohorts and resultant proposed Logic for Cohort/s

The two main flavours proposed are:

- Psoriasis: Formed by unspecific terms, such as “psoriasis” and/or one of the following terms: plaque psoriasis, guttate psoriasis, pustular psoriasis, erythrodermic psoriasis, and inverse psoriasis.
- PsA, including only terms referring to psoriasis-related arthropathy.

No specific subtypes are proposed. Particularly, since there are no specific codes to identify drug-induced psoriasis nor PsA in most data sources, no specific drug-induced sub-phenotype will be developed. However, as part of the characterisation of psoriasis and PsA, prescriptions of drugs that may be related to the development of these conditions in the 90 and 30 days prior to the diagnosis (index date) will be identified and described as a proportion for both variants. Such proportion will be framed as the proportion of psoriasis or PsA, respectively, that are potentially drug-induced.

Proposed Search strategy

- Psoriasis concept:

Psoriasis OR

Psoriatic

Excluding: arthritis OR arthropathy OR spondyl OR enthesitis OR dactylitis

- PsA concept:

Psoriasis arthritits OR

psoriasis arthropathy OR

//PsA// OR

psoriatic spondyl OR

psoriatic enthesitis OR

psoriatic dactylitis

Excluding: skin OR plaque OR guttate OR pustular OR erythrodermic OR inverse OR nail OR scalp OR palmoplantar

Autoimmune hepatitis

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Muratori L, Lohse AW, Lenzi M. Diagnosis and management of autoimmune hepatitis. BMJ. 2023 Feb 6;380:e070201. doi: 10.1136/bmj-2022-070201. Erratum in: BMJ. 2023 Feb 10;380:p330. doi: 10.1136/bmj.p330. PMID: 36746473.

Short description

Autoimmune hepatitis is an inflammatory disease of the liver of unknown cause that may progress to liver cirrhosis and end stage liver failure if diagnosis is overlooked and treatment delayed. An abnormal immune response targeting liver autoantigens and inducing persistent and self-perpetuating liver inflammation is the pathogenic mechanism of the disease.

Names of variants (flavours) of cohorts and resultant Proposed Logic for Cohort/s

Only one variant is proposed: broad autoimmune hepatitis. No subtypes are proposed. However, prognosis/severity at baseline will be determined based on the presence or absence of hepatic cirrhosis.

Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1), acknowledging that biochemistry biomarkers often correlate poorly with the severity of chronic hepatic diseases.

Proposed Search strategy

- Autoimmune hepatitis OR
- Auto-immune hepatitis OR
- Chronic active autoimmune hepatitis OR
- Autoimmune chronic active hepatitis OR
- //AIH// OR
- Lupoid hepatitis OR
- Plasma cell hepatitis OR
- Exclusion terms: “viral”, “alcohol”, “metabolic”

Primary biliary cholangitis

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Lleo, A., Wang, G. Q., Gershwin, M. E., & Hirschfield, G. M. (2020). Primary biliary cholangitis. *The Lancet*, 396(10266), 1915-1926.
- Gazda, Jakub, Drazilova, Sylvia, Janicko, Martin, Jarcuska, Peter, The Epidemiology of Primary Biliary Cholangitis in European Countries: A Systematic Review and Meta-Analysis, *Canadian Journal of Gastroenterology and Hepatology*, 2021, 9151525, 11 pages, 2021. <https://doi.org/10.1155/2021/9151525>

Short description

Primary biliary cholangitis is an autoimmune liver disease that predominantly affects the small bile duct. It is characterised by a granulomatous lymphocytic cholangitis, with typical seroreactivity for antimitochondrial antibodies.

Names of variants (flavours) of cohorts and resultant proposed Logic for Cohort/s

Only one variant is proposed: primary biliary cholangitis.

No subtypes are proposed. However, prognosis/severity at baseline will be determined based on the presence or absence of hepatic cirrhosis.

Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).

Proposed Search strategy

- primary biliary cholangitis OR
- //PBC// OR
- primary biliary cirrhosis

Primary sclerosing cholangitis

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Karlsen, T. H., Folseraas, T., Thorburn, D., & Vesterhus, M. (2017). Primary sclerosing cholangitis-a comprehensive review. *Journal of hepatology*, 67(6), 1298-1323.

Short description

Primary sclerosing cholangitis is a rare disorder characterised by multi-focal bile duct strictures and progressive liver disease. The close association with inflammatory bowel disease is a hallmark of the condition as inflammatory bowel disease is present in most patients.

Names of variants (flavours) of cohorts and resultant proposed Logic for Cohort/s

Only one variant is proposed: primary sclerosing cholangitis. No subtypes are proposed. However, prognosis/severity at baseline will be determined based on the presence or absence of hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).

Proposed Search strategy

“primary sclerosing cholangitis” OR
 //PSC// OR
 “primary biliary cirrhosis”

Chronic viral hepatitis

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Wang LS, D'Souza LS, Jacobson IM. Hepatitis C-A clinical review. *J Med Virol*. 2016 Nov;88(11):1844-55. doi: 10.1002/jmv.24554. Epub 2016 Apr 29. PMID: 27097298.
- Guvenir M, Arikan A. Hepatitis B Virus: From Diagnosis to Treatment. *Pol J Microbiol*. 2020 Dec;69(4):391-399. doi: 10.33073/pjm-2020-044. Epub 2020 Dec 27. PMID: 33574867; PMCID: PMC7812357.

Short description

The main conditions leading to chronic viral hepatitis are chronic viral hepatitis C and chronic viral hepatitis B. Both are persistent infections that cause permanent hepatic inflammation which may remain asymptomatic for many years and can progress to fibrosis, hepatic cirrhosis, hepatocellular carcinoma, liver failure, and death, if untreated. Both can be cured by antiviral treatments. However, both chronic viral hepatitis B and C are still a huge health concern worldwide. Notably, individuals with chronic viral hepatitis B can be infected with hepatitis D, which might lead to severe hepatitis.

Names of variants (flavours) of cohorts and resultant proposed Logic for Cohort/s

Only one variant is proposed: chronic viral hepatitis, which will include chronic viral hepatitis B, chronic viral hepatitis C, and chronic viral hepatitis D. No subtypes are proposed. However, prognosis/severity at baseline will be determined based on the presence or absence of hepatic cirrhosis.

Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1).

Proposed Search strategy

- chronic viral hepatitis OR
- chronic hepatitis virus OR
- chronic viral liver OR
- chronic hepatitis B OR
- chronic //HBV// OR
- hepatitis B chronic OR
- //CHB// OR
- chronic hepatitis C OR
- chronic //HCV// OR
- hepatitis C chronic OR
- chronic hepatitis D OR
- chronic //HDV// OR
- hepatitis D chronic OR
- chronic delta hepatitis

Alcohol related liver disease (ALD)

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny/)
- Mackowiak B, Fu Y, Maccioni L, Gao B. Alcohol-associated liver disease. J Clin Invest. 2024 Feb 1;134(3):e176345. doi: 10.1172/JCI176345. PMID: 38299591; PMCID: PMC10836812.
- Hernández-Évole H, Jiménez-Esquivel N, Pose E, Bataller R. Alcohol-associated liver disease: Epidemiology and management. Ann Hepatol. 2024 Jan-Feb;29(1):101162. doi: 10.1016/j.aohep.2023.101162. Epub 2023 Oct 12. PMID: 37832648.
- Niu X, Zhu L, Xu Y, Zhang M, Hao Y, Ma L, Li Y, Xing H. Global prevalence, incidence, and outcomes of alcohol related liver diseases: a systematic review and meta-analysis. BMC Public Health. 2023 May 11;23(1):859. doi: 10.1186/s12889-023-15749-x. Erratum in: BMC Public Health. 2023 Jul 18;23(1):1380. doi: 10.1186/s12889-023-15986-0. PMID: 37170239; PMCID: PMC10173666.

Short description

ALD is one of most common chronic liver diseases worldwide, accounting for approximately 50% of cirrhosis in the USA. ALD has a complex pathogenesis that includes the interplay of both genetic and environmental factors, yet the precise mechanisms are largely unknown.

ALD includes an array of liver disorders, ranging from simple steatosis to more severe forms of pathological liver changes, including alcohol-associated steatohepatitis, cirrhosis, and hepatocellular carcinoma. In addition, patients with underlying ALD and excessive alcohol intake may develop acute alcohol-associated hepatitis, which may be life-threatening.

Names of variants (flavours) of cohorts and resultant Proposed Logic for Cohort/s

Only one variant is proposed: alcohol related liver disease. No subtypes are proposed. However, prognosis/severity at baseline will be determined based on the presence or absence of hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1), acknowledging that biochemistry biomarkers often correlate poorly with the severity of chronic hepatic diseases.

Proposed Search strategy

- Alcohol associated liver disease OR
- Alcohol related liver OR
- Alcohol-associated liver disease OR
- Alcohol-related liver OR
- Alcohol associated steat* OR
- Alcohol related steat* OR
- Alcohol-associated steat* OR
- Alcohol-related steat* OR
- Alcoholic steat* OR
- Alcohol associated fat* OR
- Alcohol related fat* OR
- Alcohol-associated fat* OR
- Alcohol-related fat* OR
- Alcoholic fat* OR
- Alcohol associated cirrh* OR
- Alcohol related cirrh* OR
- Alcohol-associated cirrh* OR
- Alcohol-related cirrh* OR
- Alcoholic cirrh* OR
- Alcohol associated hepatitis OR
- Alcohol related hepatitis OR
- Alcohol-associated hepatitis OR
- Alcohol-related hepatitis OR
- Alcoholic hepatitis OR

Metabolic dysfunction-associated steatohepatitis

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Fouad Y, Alбораie M, Shiha G. Epidemiology and diagnosis of metabolic dysfunction-associated fatty liver disease. Hepatol Int. 2024 Oct;18(Suppl 2):827-833. doi: 10.1007/s12072-024-10704-3. Epub 2024 Jul 5. PMID: 38967907; PMCID: PMC11450050.

Short description

Metabolic dysfunction-associated steatohepatitis/steatotic liver disease (MASLD) (previously known as non-fatty liver disease) or metabolic dysfunction linked to fatty liver disease (MAFLD) is the most common chronic liver illness worldwide.

Names of variants (flavours) of cohorts and resultant Proposed Logic for Cohort/s

Only one variant is proposed: metabolic dysfunction-associated steatohepatitis (MASLD). No subtypes are proposed. However, comorbidities of interest (i.e. obesity, diabetes, dyslipidaemia) before the index date will be reported in a non-exclusive manner.

Prognosis/severity at baseline will be determined based on the presence or absence of hepatic cirrhosis. Where available, biochemistry biomarkers (those of hepatocellular injury (AST/ALT), and cholestasis (ALP/GGT) and bilirubin will be reported as part of the characterisation (Objective 1), acknowledging that biochemistry biomarkers often correlate poorly with the severity of chronic hepatic diseases.

Proposed Search strategy

- MASLD concept

Metabolic dysfunction-associated steat* OR

Metabolic dysfunction associated steat* OR

non-alcoholic fat* OR

Non alcoholic fat* OR

Non-alcoholic liver OR

Non alcoholic liver OR

//MASLD// OR

//MAFLD// OR

//NAFLD// OR

//MASH// OR

//NASH// OR

- Comorbidities concepts have already developed for other studies of DARWIN EU® Data Network (e.g. P4-C3-006, P4-C2-019, and P4-C2-020 <https://catalogues.ema.europa.eu/node/4866>) and include obesity, dyslipidaemia, diabetes mellitus, and hypertension.

Heart failure (HF)

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Chen J, Aronowitz P. Congestive Heart Failure. Med Clin North Am. 2022 May;106(3):447-458. doi: 10.1016/j.mcna.2021.12.002. Epub 2022 Apr 4. PMID: 35491065.
- Mosterd A, Hoes AW. Clinical epidemiology of heart failure. Heart. 2007 Sep;93(9):1137-46. doi: 10.1136/hrt.2003.025270. PMID: 17699180; PMCID: PMC1955040.
- Campbell P, Rutten FH, Lee MM, Hawkins NM, Petrie MC. Heart failure with preserved ejection fraction: everything the clinician needs to know. Lancet. 2024 Mar 16;403(10431):1083-1092. doi: 10.1016/S0140-6736(23)02756-3. Epub 2024 Feb 14. Erratum in: Lancet. 2024 Mar 16;403(10431):1026. doi: 10.1016/S0140-6736(24)00494-X. PMID: 38367642.
- 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines

Short description

HF is a syndrome consisting in symptoms and signs caused by cardiac dysfunction, resulting in reduced longevity. HF is characterised by symptoms and signs, so asymptomatic left ventricular systolic dysfunction is not equivalent to HF.

HF is generally a chronic condition in which bouts of worsening symptoms and signs can occur that may require hospitalisation or more frequent doctor visits (decompensation of chronic HF). Alternatively, HF may present acutely, with occurrence of severe symptoms and signs within 24 hours. Acute HF may present in several forms: acute pulmonary oedema secondary to cardiac dysfunction, cardiogenic shock, and/or acute worsening (decompensation) of chronic HF.

HF can result from impairment in the ability of the heart to pump enough blood into the circulation during systole. Alternatively, HF can also occur in patients with normal left ventricular systolic function in whom higher filling pressures are needed to obtain a normal end-diastolic volume of the left ventricle, so called heart failure with preserved left ventricular ejection fraction or “diastolic” HF. The definition of HF with preserved left ventricular ejection fraction offered by the European Society of Cardiology is more complex: “Those with symptoms and signs of HF, with evidence of structural and/or functional cardiac abnormalities and/or raised natriuretic peptides, and with a left ventricular ejection fraction of $\geq 50\%$ ”. This nuanced and lengthy definition makes it more difficult to diagnose HF with preserved left ventricular ejection fraction than HF with reduced ejection fraction ($\leq 40\%$) or mildly reduced ejection fraction (41–49%).

Names of variants (flavours) of cohorts and resultant proposed logic for the cohorts

Three different, non-exclusive categories of subtypes could be developed based on:

- The ejection fraction: This category will only be applied if the ejection fraction is specified in the diagnostic concept and/or in the conducted measurements or procedures (echocardiography).
 - o Heart failure with (mildly, moderately, or severely) reduced ejection fraction
 - o Heart failure with preserved ejection fraction
- Whether HF is:
 - o Decompensated: Chronic HF decompensation could be operationalised as at least one exacerbation event during the follow-up, defined as an inpatient admission with HF recorded as the primary diagnosis, and/or an “acute HF” diagnosis during the follow-up of an individual with a prior “chronic HF” diagnosis.

- Compensated: Absence of the above criteria for the “decompensated” sub-phenotype.
- Location: This category will only be applied if it is specified in the diagnostic concept, acknowledging that this specification might be missing in many cases.
 - Right side
 - Left side

Comorbidities of interest (i.e. ischaemic heart disease, arterial hypertension, cardiomyopathies, valvular heart disease) in the year prior to the index date will be reported as proportions in a non-exclusive manner.

Proposed Search strategy

- Heart failure concept:
 - “heart failure” OR
 - “cardiac failure” OR
 - “ventricular failure” OR
 - “biventricular failure” OR
 - “myocardial failure”
- The comorbidities concepts will be needed. These have previously been developed for other DARWIN EU® studies (such as, P4-C3-006, P4-C2-019, and P4-C2-020 <https://catalogues.ema.europa.eu/node/4866>).

Cardio-renal-metabolic syndrome

Information source

- Dynamed (Oxford - https://dpa-pde-oxford.shinyapps.io/Phenotyping_ESBCO_Shiny)
- Young JB, Eknayan G. Cardiorenal Syndrome: An Evolutionary Appraisal. *Circ Heart Fail.* 2024 Jun;17(6):e011510. doi: 10.1161/CIRCHEARTFAILURE.123.011510. Epub 2024 May 17. PMID: 38757274.
- Gunnarsson S, Vito O, Unwin RJ. Cardiovascular-kidney-metabolic syndrome: prevalence, risks, disease trajectories, and early-stage management. *Am J Physiol Cell Physiol.* 2026 Jan 1;330(1):C1-C8. doi: 10.1152/ajpcell.00499.2025. Epub 2025 Nov 21. PMID: 41269265.
- Zannad F, Rossignol P. Cardiorenal Syndrome Revisited. *Circulation.* 2018 Aug 28;138(9):929-944. doi: 10.1161/CIRCULATIONAHA.117.028814. PMID: 30354446.

Short description

Building on the foundation of metabolic syndrome in 2023, the American Heart Association introduced the Cardiovascular-kidney-metabolic (CKM) syndrome, which has been defined as “A systemic disorder characterised by pathophysiological interactions among metabolic risk factors, chronic kidney disease, and the cardiovascular system leading to multiorgan dysfunction and a high rate of adverse cardiovascular outcomes”. [15] CKM syndrome includes both individuals at risk for cardiovascular disease due to the presence of metabolic risk factors, chronic kidney disease, or both, as well as individuals with existing cardiovascular disease which is potentially related to or complicates metabolic risk factors or CKD.

Although this model addresses the interplay between the cardiovascular, renal, and metabolic systems, it does not fully capture the role of the liver. MASLD, and, particularly, its progression to cirrhosis, is

bidirectionally associated with metabolic dysregulation, atherosclerosis, HF, and CKD. Therefore, claims have been made to expand the CKM framework to a cardiovascular-renal-hepatic-metabolic framework.

Names of variants (flavours) of cohorts and Resultant Proposed Logic for Cohort/s

Two flavours could be developed:

- Narrow flavour focused on cardio-renal-metabolic syndrome.
- Broader flavour encompassing:
 - Cardiorenal syndrome (specified as such), CKM (specified as such), or CRHM (specified as such), or
 - The presence of ≥ 2 of the following:
 - Metabolic syndrome (as defined by ≥ 2 of obesity, diabetes, dyslipidaemia, arterial hypertension)
 - Cardiovascular disease, defined as having had a major cardiovascular event (coronary artery disease, stroke, peripheral artery disease, heart failure).
 - Chronic kidney disease (any stage)

The three main subtypes will be:

- Cardiac driven
- Renal driven
- Metabolic driven.

These subtypes will be operationalised using the following hierarchical approach, according to the presence or absence of conditions such as HF (cardiac-driven) or CKD (renal-driven), or, in the absence of HF or CKD, diabetes, and dyslipidaemia (metabolic-driven). If none of these criteria are met, or if both HF and CKD are present, the individual will be classified as having an unspecified sub-phenotype.

Additionally, the presence of comorbidities of interest (heart failure, CKD, dyslipidaemia, diabetes mellitus, obesity, hypertension, and MASLD) will be described as proportions for the phenotype overall and each sub-phenotype (Objective 1).

Proposed Search strategy

- Cardio-renal-metabolic syndrome concept:

"Cardiorenalmetabolic " OR "cardiovascular-renal-metabolic" OR "cardiovascular-kidney-metabolic" OR "cardiovascular renal metabolic" OR "cardiovascular kidney metabolic" OR "cardio-renal-metabolic" OR "cardio renal metabolic " OR "cardiometabolic" OR "Cardiovascular-kidney-metabolic" OR "cardiovascular kidney metabolic" OR "heart-kidney"

OR

//CRMS// OR //CRM// OR //CKM//
- Metabolic syndrome concept:
 - "X syndrome" OR "syndrome X"
 - "Metabolic syndrome" OR "syndrome metabolic"
- Diabetes type 2 concept (already developed for P4-C3-006, P4-C2-019, and P4-C2-020 <https://catalogues.ema.europa.eu/node/4866>)

- Dyslipidaemia concept (already developed for P4-C3-006, P4-C2-019, and P4-C2-020 <https://catalogues.ema.europa.eu/node/4866>)
- Heart failure concept (already developed for P4-C3-006, P4-C2-019, and P4-C2-020 <https://catalogues.ema.europa.eu/node/4866>)
- Chronic kidney disease concept (already developed for P4-C3-006, P4-C2-019, and P4-C2-020 <https://catalogues.ema.europa.eu/node/4866>)

ANNEX V. ENCePP checklist for study protocols

Study title: DARWIN EU® - Identifying conditions of regulatory interest in Europe: a population-based cohort study

EU PAS Register® number: EUPAS1000001129

Study reference number (if applicable): P5-C3-006/P5-C2-004

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

Comments:

Section 2: Research question	Yes	No	N/A	Section Number
2.1 Does the formulation of the research question and objectives clearly explain:				
2.1.1 Why the study is conducted? (e.g. to address an important public health concern, a risk identified in the risk management plan, an emerging safety issue)	X			6
2.1.2 The objective(s) of the study?	X			7
2.1.3 The target population? (i.e. population or subgroup to whom the study results are intended to be generalised)	X			8.3.
2.1.4 Which hypothesis(-es) is (are) to be tested?			X	
2.1.5 If applicable, that there is no <i>a priori</i> hypothesis?			X	

Comments:

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

² Date from which the analytical dataset is completely available.

<u>Section 3: Study design</u>	Yes	No	N/A	Section Number
3.1 Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	X			8.1.
3.2 Does the protocol specify whether the study is based on primary, secondary or combined data collection?	X			8.4.
3.3 Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	X			8.1., 8.2.
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))			X	
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)			X	

Comments:

<u>Section 4: Source and study populations</u>	Yes	No	N/A	Section Number
4.1 Is the source population described?	X			8.3.
4.2 Is the planned study population defined in terms of:	X			
4.2.1 Study time period	X			8.3., 8.5.
4.2.2 Age and sex	X			8.3.
4.2.3 Country of origin	X			8.4.
4.2.4 Disease/indication	X			8.6.1
4.2.5 Duration of follow-up	X			8.2.
4.3 Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	X			8.3.

Comments:

<u>Section 5: Exposure definition and measurement</u>	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)			X	
5.2 Does the protocol address the validity of the exposure measurement? (e.g. precision, accuracy, use of validation sub-study)			X	
5.3 Is exposure categorised according to time windows?			X	

<u>Section 5: Exposure definition and measurement</u>	Yes	No	N/A	Section Number
5.4 Is intensity of exposure addressed? (e.g. dose, duration)			X	
5.5 Is exposure categorised based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?			X	
5.6 Is (are) (an) appropriate comparator(s) identified?			X	

Comments:

<u>Section 6: Outcome definition and measurement</u>	Yes	No	N/A	Section Number
6.1 Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated?	X			8.6.1.
6.2 Does the protocol describe how the outcomes are defined and measured?	X			8.6.1, 11
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)			X	
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)			X	

Comments:

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

Comments:

<u>Section 8: Effect measure modification</u>	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)			X	

Comments:

<u>Section 9: Data sources</u>	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)			X	
9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	X			8.6.1., 11
9.1.3 Covariates and other characteristics?	X			8.6.2., 11
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)			X	
9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	X			11
9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	X			11
9.3 Is a coding system described for:				
9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)			X	
9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	X			11
9.3.3 Covariates and other characteristics?	X			11
9.4 Is a linkage method between data sources described? (e.g. based on a unique identifier or other)			X	

Comments:

<u>Section 10: Analysis plan</u>	Yes	No	N/A	Section Number
10.1 Are the statistical methods and the reason for their choice described?	X			8.8
10.2 Is study size and/or statistical precision estimated?			X	
10.3 Are descriptive analyses included?	X			8.8.

<u>Section 10: Analysis plan</u>	Yes	No	N/A	Section Number
10.4 Are stratified analyses included?			X	
10.5 Does the plan describe methods for analytic control of confounding?			X	
10.6 Does the plan describe methods for analytic control of outcome misclassification?			X	
10.7 Does the plan describe methods for handling missing data?			X	
10.8 Are relevant sensitivity analyses described?			X	

Comments:

<u>Section 11: Data management and quality control</u>	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)	X			11
11.2 Are methods of quality assurance described?	X			11
11.3 Is there a system in place for independent review of study results?			X	

Comments:

<u>Section 12: Limitations</u>	Yes	No	N/A	Section Number
12.1 Does the protocol discuss the impact on the study results of:				
12.1.1 Selection bias?	X			9
12.1.2 Information bias?	X			9
12.1.3 Residual/unmeasured confounding? (e.g. anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).			X	
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:

<u>Section 13: Ethical/data protection issues</u>	Yes	No	N/A	Section Number
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	X			11

<u>Section 13: Ethical/data protection issues</u>	Yes	No	N/A	Section Number
13.2 Has any outcome of an ethical review procedure been addressed?			X	
13.3 Have data protection requirements been described?	X			11

Comments:

<u>Section 14: Amendments and deviations</u>	Yes	No	N/A	Section Number
14.1 Does the protocol include a section to document amendments and deviations?	X			

Comments:

<u>Section 15: Plans for communication of study results</u>	Yes	No	N/A	Section Number
15.1 Are plans described for communicating study results (e.g. to regulatory authorities)?	X			11
15.2 Are plans described for disseminating study results externally, including publication?	X			11

Comments:

Name of the main author of the outline:

Alejandro Ballvé, Daniel Prieto-Alhambra

Date: 27/08/2026

Signature:

Alejandro Ballvé Martín

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

Complex Studies (C3)

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

Coordination Centre (CC)

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

Data Access

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

Data Quality Framework

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

Data Source

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

DARWIN EU®

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

EMA (European Medicines Agency)

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

Evidence Generation

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

Federated Network

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

GDPR (General Data Protection Regulation)

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

Health Technology Assessment (HTA)

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

Metadata

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

Off-the-Shelf Studies (OTS)

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

OHDSI (Observational Health Data Sciences and Informatics)

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

Patient-Level Data

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

OMOP (Observational Medical Outcomes Partnership)

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

Real-World Data (RWD)

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

Real-World Evidence (RWE)

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

Regulatory Decision-Making

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

Routine Repeated Studies (RR)

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

Study Protocol

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

Very Complex Studies (C4)

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