Study Protocol

P4-C1-010

DARWIN EU® - Feasibility of studies on early (pre-symptomatic) stages of type 1 diabetes mellitus in the DARWIN EU® network

15/09/2025

Version 3.0

Authors: Julieta Politi, Nicholas Hunt, Katia Verhamme, Cesar Barboza, Maarten van Kessel, Ioanna Nika, Ross Williams, Ger Inberg, Natasha Yefimenko

Public

CONTENTS

Study title	DARWIN EU® - Feasibility of studies on early (pre-symptomatic) stages of type 1 diabetes mellitus in the DARWIN EU® network
Protocol version	V3.0
Date	15/09/2025
EUPAS number	EUPAS1000000756
Active substance	N/A
Medicinal product	N/A
Research question and objectives	Research questions What is the frequency and timing of autoantibody and glucose testing prior to clinical diagnosis of type 1 diabetes mellitus within the DARWIN EU® preselected network data sources? Objectives The aim of this study is to investigate the feasibility of conducting research on the early (pre-symptomatic) stages of type 1 diabetes mellitus within the DARWIN EU® network. It will focus on the frequency and timing of autoantibody and glucose testing before the disease becomes clinically apparent. The specific objectives for this study are: 1. To describe the characteristics of individuals newly diagnosed with type 1 diabetes mellitus in terms of demographics, prespecified comorbidities and medications, and diagnostic tests of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), prior to and at the time of type 1 diabetes mellitus diagnosis, and to assess selected characteristics at one-year post-diagnosis. 2. To estimate, for each diagnostic test of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), the median (IQR) time in days from 1) the earliest recorded test and 2) the earliest recorded abnormal result (where ascertainable) to the date of first-ever type 1 diabetes mellitus diagnosis. 3. To estimate the annual point prevalence of type 1 diabetes mellitus during 2015–2024, using population-based data sources.
Countries of study	Denmark, Finland, France, Hungary, Netherlands, Spain
Authors	Julieta Politi (j.politi@darwin-eu.org) Nicholas Hunt (n.hunt@darwin-eu.org) Katia Verhamme (k.verhamme@darwin-eu.org)

LIST OF ABBREVIATIONS

Acronyms/term	Description
ATC	Anatomical Therapeutic Chemical
CDM	Common Data Model
CC	Coordinating centre
CDW Bordeaux	Clinical Data Warehouse of Bordeaux University Hospital
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DKA	Diabetic Ketoacidosis
DK-DHR	Danish Data Health Registries
DQD	Data Quality Dashboard
DOI	Declaration of Interests
DQD	Data Quality Dashboard
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-TaUH Pirha	Tampere University Hospital patient cohort
GAD-65	Glutamic acid decarboxylase
GDPR	General Data Protection Regulation
H12O	Hospital Universitario 12 de Octubre
IA-2	Insulinoma-associated antigen-2
IAA	Insulin autoantibodies
ICA	Islet Cell autoantibodies
ICD	International Classification of Diseases
IP	Inpatient
IPCI	Integrated Primary Care Information
IQR	Interquartile range
IRB	Institutional Review Board
MODY	Maturity-Onset Diabetes of the Young
OGLD	Oral Glucose-lowering Drug
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OP	Outpatient
RxNorm	Medical prescription normalized
SNOMED	Systematized Nomenclature of Medicine
WHO	World Health Organisation
ZnT8	Anti-Zinc transporter 8

1. TITLE

DARWIN EU® - Feasibility of studies on early (pre-symptomatic) stages of type 1 diabetes mellitus in the DARWIN EU® network

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator	Julieta Politi Nicholas Hunt Katia Verhamme	Erasmus MC
Data Scientist	Cesar Barboza Maarten van Kessel Ioanna Nika Ross Williams Ger Inberg	Erasmus MC
Study Project Manager	Natasha Yefimenko	Erasmus MC
Data Partner*	**Names**	**Organisation**
DK-DHR	Elvira Bräuner Susanne Bruun	Danish Medicines Agency (DKMA)
FinOMOP-TaUH Pirha	Hakkarainen Leena Kati Kristiansson Tiina Wahlfors	Pirkanmaa Welfare Services County, Tampere University Hospital
CDW Bordeaux	Guillaume Verdy Romain Griffier	Clinical Data Warehouse of Bordeaux University Hospital – Direction Generale
SUCD	Bagyura Zsolt István Ágota Mészáros Kiss Loretta Zsuzsa Héja Tibor	Semmelweis University
IPCI	Mees Mosseveld Katia Verhamme	Erasmus MC
H12O	Juan Luis Cruz Bermudez Noelia Garcia Barrio Paula Rubio Mayo	Fundación Investigación Biomédica Hospital 12 de Octubre

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

3. ABSTRACT

Title

DARWIN EU® - Feasibility of studies on early (pre-symptomatic) stages of type 1 diabetes mellitus in DARWIN EU® network

Rationale and background

Identifying type 1 diabetes mellitus at an early, presymptomatic stage offers clinical advantages. These benefits include a decreased risk of diabetic ketoacidosis (DKA) at the onset of the disease and a notable reduction in clinical symptoms. Additionally, products such as Tzield (teplizumab) are being developed to target early stages of type 1 diabetes mellitus, aiming to delay disease progression. There is also increasing attention in clinical practice to early screening (via specific antibodies), which helps in identifying candidates for disease-modifying therapies and provides early access to diabetes-related education and disease management.

Research question and objectives

Research questions

What is the frequency and timing of autoantibody and glucose testing prior to clinical diagnosis of type 1 diabetes mellitus within the DARWIN EU® preselected network data sources?

Objectives

The aim of this study is to investigate the feasibility of conducting research on the early (pre-symptomatic) stages of type 1 diabetes mellitus within the DARWIN EU® network. It will focus on the frequency and timing of autoantibody and glucose testing before the disease becomes clinically apparent.

The specific objectives for this study are:

1. To describe the characteristics of individuals newly diagnosed with type 1 diabetes mellitus in terms of demographics, prespecified comorbidities and medications, and diagnostic tests of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), prior to and at the time of type 1 diabetes mellitus diagnosis, and to assess selected characteristics at one-year post-diagnosis.
2. To estimate, for each diagnostic test of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), the median (IQR) time in days from 1) the earliest recorded test and 2) the earliest recorded abnormal result (where ascertainable) to the date of first-ever type 1 diabetes mellitus diagnosis.
3. To estimate the annual point prevalence of type 1 diabetes mellitus during 2015–2024, using population-based data sources.

Methods

Study design

- Patient-level characterisation (retrospective cohort study to fulfil Objectives 1 and 2)
- Population-level descriptive epidemiology (descriptive cohort study to fulfil Objective 3)

Population

For objectives 1 and 2, the study population will include individuals with a first-ever recorded diagnosis of type 1 diabetes mellitus during the study period (see outcome below). A minimum of 365 days of prior observation time before the index date will be required (applied to non-hospital-based data sources and individuals aged 1 year or older).

The study population for objective 3 will include all individuals present in the database during the study period, 01/01/2015 to 31/12/2024, or to the end of available data, and with at least 365 days of database

history prior to the index date (applied to non-hospital-based data sources and individuals aged 1 year or older).

Variables

Exposure:

N/A.

Outcome (type 1 diabetes mellitus, main phenotype):

The main phenotype will require the presence of both SNOMED CT codes indicative of type 1 diabetes mellitus and initiation of insulin therapy (RxNorm codes), within 180 days of each other. The index date will be defined as the earlier of the two events.

For point prevalence estimations of type 1 diabetes mellitus, a person will be counted as a case if they have ever met the case definition at the time of estimation (reference date).

Relevant covariates:

The covariates of interest include age groups (0–9, 10–19, 20–29, 30–39, 40–49, ≥50 years), sex, diagnostic tests of interest, and prespecified medications and comorbidities. The diagnostic tests of interest include HbA1c, oral glucose tolerance test, fasting glucose, random glucose, insulin and/or C-peptide measurements, insulin antibody levels, islet cell antibodies, insulinoma-associated antigen-2 antibody measurement, anti-zinc transporter 8 antibodies, and anti-GAD antibodies. Prespecified medications to be assessed include the use of immune modulators (teplizumab), oral glucose-lowering medications (individually), and verapamil. Comorbidities of interest include diabetic ketoacidosis, thyroid disease (hypo- and hyper-thyroidism, individually), celiac disease, and other autoimmune diseases, among other conditions of interest.

Data sources

1. Denmark: Danish Data Health Registries (DK-DHR)
2. Finland: Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)
3. France: Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)
4. Hungary: Semmelweis University Clinical Data (SUCD)
5. Netherlands: Integrated Primary Care Information (IPCI)
6. Spain: Hospital Universitario 12 de Octubre (H12O)

Study size

No sample size has been calculated, as this is an exploratory study that will not test a specific hypothesis.

Statistical analysis

Baseline characteristics (age, sex, predefined comorbidities, medications, and diagnostic tests of interest) will be summarised as n (%) for categorical variables and mean (SD), median (IQR), minimum, and maximum for continuous variables. Selected characteristics at one year post-diagnosis (type 2 diabetes mellitus and oral glucose-lowering drugs) will be assessed in individuals with at least 365 days of observation time following the index date. Analyses will be done using the *CohortCharacteristics* R package.

The point prevalence of type 1 diabetes mellitus will be estimated annually (as of January 1st of each year) in the general population in population-based data sources (IPCI and DK-DHR) and reported by calendar year. The statistical analyses will be performed based on OMOP CDM mapped data using the *IncidencePrevalence* R package.

Primary results will describe type 1 diabetes mellitus in the overall population. Results by age groups and sex (individually) will be presented in the **Annex**.

A minimum cell count of 5 will be used when reporting results, with any smaller count reported as “<5”.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study milestones and deliverables	Planned dates*
Final Study Protocol	To be confirmed by EMA
Creation of Analytical code	September 2025
Execution of Analytical Code on the data	October 2025
Draft Study Report	November 2025
Final Study Report	To be confirmed by EMA

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

6. RATIONALE AND BACKGROUND

Type 1 diabetes mellitus is a chronic autoimmune condition characterised by the destruction of insulin-producing pancreatic beta cells, leading to symptomatic hyperglycaemia and insulin dependence.(1) The disease develops gradually over time, progressing through defined stages, from the onset of islet autoimmunity (stage 1), to presymptomatic dysglycaemia due to declining β-cell function (stage 2), and eventually to clinically manifest diabetes (stage 3). The rate of progression between stages varies, ranging from a few months to several decades.(2) Identifying type 1 diabetes mellitus at an early, presymptomatic stage offers important clinical benefits, including a reduced risk of diabetic ketoacidosis (DKA) at the disease onset and less severe clinical symptoms.(3) Additionally, disease-modifying therapies, such as teplizumab (Tzield), have been developed to delay disease progression, highlighting the value of early detection. Screening programs based on pancreatic-islet autoantibodies are also being adopted in some settings, enabling earlier identification of at-risk individuals and facilitating timely access to diabetes-related education and disease management.(3)

Routinely collected electronic healthcare data represent a potentially valuable resource for studying the early stages of type 1 diabetes mellitus. However, a critical first step is the accurate identification of individuals with clinically manifest type 1 diabetes mellitus using an appropriate case definition. This provides a foundation for characterising the disease course and evaluating the availability and quality of relevant data—such as laboratory results, diagnoses, and treatments—that could support the identification of earlier disease stages. Understanding cohort characteristics and data completeness is also key to determining whether sufficient information exists to study disease stages and their progression.

The aim of this study is to evaluate the feasibility of conducting research on the early (pre-symptomatic) stages of type 1 diabetes within preselected DARWIN EU® network data sources. Specifically, this study will investigate the feasibility of conducting research on the early (pre-symptomatic) stages of type 1 diabetes mellitus within the DARWIN EU® network.

7. RESEARCH QUESTION AND OBJECTIVES

Research questions

What is the frequency and timing of autoantibody and glucose testing prior to clinical diagnosis of type 1 diabetes mellitus within the DARWIN EU® preselected network data sources?

Research objectives

The aim of this study is to investigate the feasibility of conducting future research on the early (pre-symptomatic) stages of type 1 diabetes mellitus within the DARWIN EU® network. In particular, we will focus on the frequency and timing of antibody and glucose testing before the disease becomes clinically apparent.

The specific objectives for this study are:

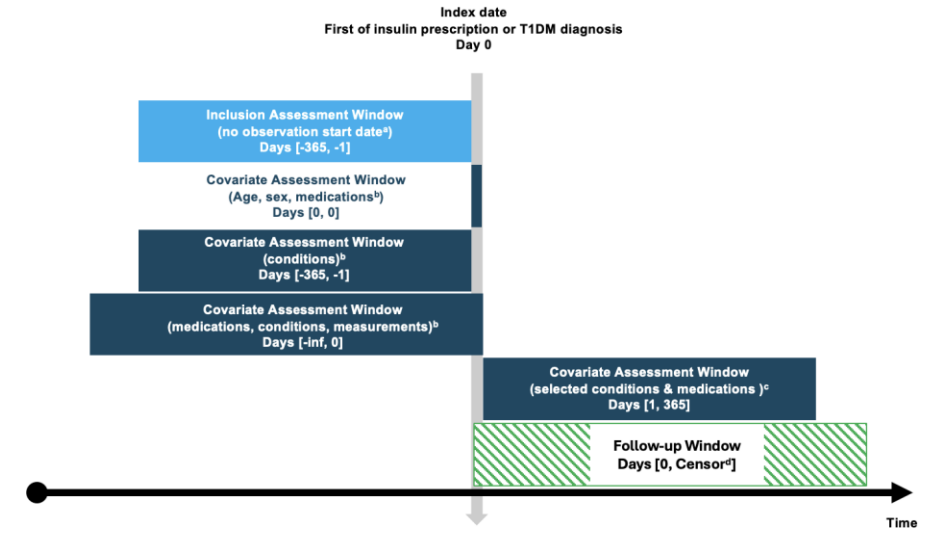
1. To describe the characteristics of individuals newly diagnosed with type 1 diabetes mellitus in terms of demographics, prespecified comorbidities and medications, and diagnostic tests of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), prior to and at the time of type 1 diabetes mellitus diagnosis, and to assess selected characteristics at one-year post-diagnosis.
2. To estimate, for each diagnostic test of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), the median (IQR) time in days from 1) the earliest recorded test and 2) the earliest recorded abnormal result (where ascertainable) to the date of first-ever type 1 diabetes mellitus diagnosis.
3. To estimate the annual point prevalence of type 1 diabetes mellitus during 2015–2024, using population-based data sources.

8. RESEARCH METHODS

8.1. Study design

A retrospective cohort study will be conducted using routinely collected health data from six data sources across six EU member states. The study will include

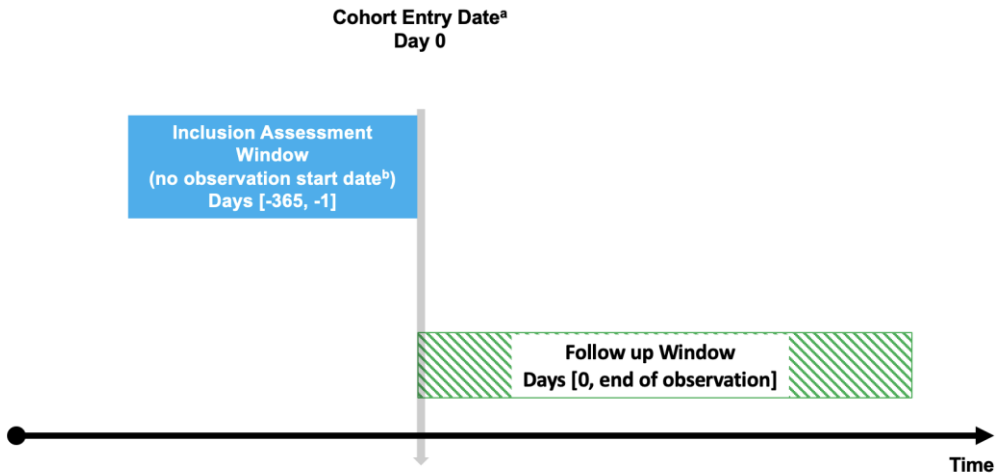
- Assessment of baseline characteristics of newly diagnosed type 1 diabetes mellitus in terms of demographics, prespecified comorbidities and medication, and available diabetes-related diagnostic testing (HbA1c, C-peptide, glucose measurements, antibodies) Error! Reference source not found.(Error! Reference source not found.).
- The study will also estimate the median time from the first record of each diabetes-related diagnostic test of interest to the first formal diagnosis of type 1 diabetes mellitus (Error! Reference source not found.). Time from the first abnormal test result (by individual test) to the first formal diagnosis of type 1 diabetes mellitus will also be estimated.
- Calculation of the annual (point) prevalence of type 1 diabetes mellitus (non-hospital databases) (**Figure**).



a. Applies to individuals older than one year and individuals not in hospital data sources
b. Covariates as defined in section 8.6.3
c. Selected conditions & medications: type 2 diabetes and oral-glucose-lowering drugs (section 8.6.3)
d. Death, disenrollment, end of data source availability, five years, or end of the study period (31/12/2024)

T1DM = type 1 diabetes mellitus

Figure 1. Study design for objectives 1 and 2, on characterising individuals at the time of type 1 diabetes mellitus diagnosis (stage 3).



a. The cohort entry date will be the date of inclusion in the denominator (presence during the study period, and 365 days of prior observation).
b. Applies to individuals older than one year and individuals not in hospital data sources

Figure 2. Study design for objective 3, on estimating the point prevalence of type 1 diabetes mellitus (stage 3).

8.2. Study setting and data sources

This study will be conducted using routinely collected data from 6 data sources, including primary care (n=1), hospital care (n=4), and registry-based data settings (n=1) within the DARWIN EU® network of data partners, and from 6 EU member states. All data were a priori mapped to the OMOP CDM.

Data sources

1. Denmark: Danish Data Health Registries (DK-DHR)
2. Finland: Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)
3. France: Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)
4. Hungary: Semmelweis University Clinical Data (SUCD)
5. Netherlands: Integrated Primary Care Information (IPCI)
6. Spain: Hospital Universitario 12 de Octubre (H12O)

Data Selection

These data sources fulfil the criteria required in terms of data quality, completeness, timeliness, and representativeness for a disease epidemiology study, while covering different regions of Europe.

When assessing the reliability of data sources, data partners are asked to describe their internal data quality process on the source data as part of the DARWIN EU® onboarding procedure. To further ensure data quality, we utilised the *Achilles* tool,(4) which systematically characterises the data and generates data characteristics such as age distribution, condition prevalence per year, and data density. Data density includes information on 1) monthly record counts by data domain, such as conditions, drug exposures, procedures (which offers insights into data collection patterns and the start date of each data source) and 2) measurement value distribution (i.e., min, max, quartiles for numeric values per measurement concept and per unit and counts for discrete measurement-value pairs). The latter can be compared against expectations for the data, based on predefined standards, historical trends, or known epidemiological patterns to identify potential anomalies or inconsistencies. Additionally, the data quality dashboard (DQD) provides more objective checks on plausibility of data completeness, consistency, and conformity across the data sources.

In terms of relevance, the DARWIN EU® portal, as well as information from the onboarding documents, was used to assess whether data sources have information on type 1 diabetes mellitus and the different diabetes-related measurements of interest.

The data sources were selected based on their available type 1 diabetes mellitus counts, availability of antibody testing, and the presence of diabetes-related measurements of relevance for the study (e.g., HbA1c, glucose, and C-peptide), in addition to their ability to support timely IRB approvals, thereby ensuring alignment with the timeline established by stakeholders for the conduct of this study.

Two data sources are representative of the target population, being a nationwide secondary care data or regional GP data, whereas the others only include hospitalised patients or patients with a secondary care encounter.

Data source justification and key characteristics

Danish Data Health Registries (DK-DHR)

DK-DHR will be included in this study because it is a nationwide registry that contains secondary care records, and it is representative of the general population.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus is approximately 120,300.

Moreover, data availability and follow-up in DK-DHR are sufficient, as data availability in DK-DHR began in 1995 and the most recent data extraction date is 2024, which aligns with the study period. The median follow-up of the first observation period is 7,920 days (IQR: 2,610–10,900).

There are some limitations present in DK-DHR, namely the absence of data on diabetes mellitus autoantibodies of interest.

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

Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)

FinOMOP-TaUH Pirha will be included in this study because it is a hospital data source that provides relevant information on individuals with type 1 diabetes mellitus who receive care in secondary care settings.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus is approximately 5,000. Additionally, approximately 1,500 person-counts for antibody measurements are expected (indicating the number of individuals who have had the test recorded).

Moreover, data availability and follow-up in FinOMOP-TaUH Pirha are sufficient, as data availability in FinOMOP-TaUH Pirha began in 2007, and the date of the most recent data extraction is 2025, which aligns with the study period and the median follow-up of the first observation period 4,230 days (IQR: 384–7,980).

Lastly, FinOMOP-TaUH Pirha IRB approval is estimated to take <1 week, which makes the execution of this study feasible within the current study timelines.

Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)

CDW Bordeaux will be included in this study because it is a hospital data source that provides relevant information on individuals with type 1 diabetes mellitus receiving care in secondary care settings.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus is approximately 14,700. Additionally, approximately 1,500 person-counts for antibody measurements are expected (indicating the number of individuals who have had the test recorded).

Moreover, data availability and follow-up in CDW Bordeaux are sufficient, as data availability in CDW Bordeaux began in 2005, and the date of the most recent data extraction is 2024, which aligns with the study period, and the median follow-up of the first observation period is 384 days (IQR: 60–2,450).

Lastly, IRB approval for CDW Bordeaux is estimated to take 1 month, which makes the execution of this study feasible within the current study timelines.

Semmelweis University Clinical Data (SUCD)

SUCD will be included in this study because it is a secondary care and hospital-based data source that provides relevant information on individuals with type 1 diabetes mellitus who receive care in secondary care settings.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus is approximately 11,500. Additionally, approximately 2,900 person-counts for antibody measurements are expected (indicating the number of individuals who have had the test recorded).

Moreover, data availability and follow-up in SUCD are sufficient, as data availability in SUCD began in 2011, and the date of the most recent data extraction is 2024, which aligns with the study period, and the median follow-up of the first observation period in SUCD is 266 days (IQR: 0–2,170).

Lastly, IRB approval for SUCD is estimated to take approximately 3 months, which makes the execution of this study feasible within the current study timelines.

Integrated Primary Care Information (IPCI)

IPCI will be included in this study because it is a primary care data source that provides relevant information on type 1 diabetes mellitus in the general population.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus is approximately 9,700. Additionally, approximately 1700 person-counts for antibody measurements are expected (indicating the number of individuals who have had the test recorded).

Moreover, data availability and follow-up in IPCI are sufficient, as data availability in IPCI began in 2006, and the date of the most recent data extraction is 2024, which aligns with the study period and the median follow-up of the first observation period in Data source 1 is 1730 days (IQR: 791–3070).

Lastly, IPCI has a blanket IRB approval for conducting DARWIN EU® studies, which makes the execution of this study feasible within the current study timelines.

Hospital Universitario 12 de Octubre (H12O)

H12O will be included in this study because it is a hospital data source that provides relevant information on individuals with type 1 diabetes mellitus who receive care in secondary care settings.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus is approximately 23,700. Additionally, approximately 6,000 person-counts for antibody measurements are expected (indicating the number of individuals who have had the test recorded).

Moreover, data availability and follow-up in H12O are sufficient, as data availability in H12O began in 2015 and the date of the most recent data extraction is 2024, which aligns with the study period, and the median follow-up of the first observation period is 529 days (IQR: 1–3,750).

Lastly, IRB approval for H12O is estimated to take 3 months, which makes the execution of this study feasible within the current study timelines.

More detailed information on the data sources planned for use in this study is provided in **ANNEX I**. Data sources description.

8.3. Study period

The study period will span from 01/01/2015 to 31/12/2024, or until the end of available data if earlier.

8.4. Follow-up

The follow-up will start on the latest of the following dates: i) study start date (01/01/2015) or ii) date at which individuals have at least 365 days of prior history recorded (applied to non-hospital-based data sources and individuals aged 1 year or older).

The end of follow-up will be defined as the earliest of loss to follow-up, death, or end of observation period (the latest available data), whichever occurs first.

8.5. Study population with inclusion and exclusion criteria

Objectives 1 and 2 (newly diagnosed type 1 diabetes mellitus cohort*):*

Inclusion criteria

First-ever recorded diagnosis of type 1 diabetes mellitus during the study period (as defined in **Section 8.6.2**).

Objective 3:

Inclusion criteria

- All individuals present in the period from 01/01/2015 to 31/12/2024 (or the first and latest available date)
- Minimum 365 days of available history before the index date (applied to non-hospital-based data sources and individuals aged 1 year or older).

8.6. Variables

8.6.1. Exposure

N/A, as no specific drugs of interest will be investigated.

8.6.2. Outcome

Type 1 diabetes mellitus phenotype (main definition)

Individuals will be classified as type 1 diabetes cases when they fulfil both requirements within a window of ≤180 days of each other (index date being the earliest of the 2 dates):

- Prescription of Insulin (at the ingredient level using RxNorm codes), AND
- Condition occurrence of type 1 diabetes mellitus* (based on SNOMED CT codes)

For Objectives 1 and 2, the first-ever (incident) diagnosis during the 2015–2024 period will be required. For Objective 3, all individuals who have ever met the type 1 diabetes mellitus case definition on or before the reference date and are under observation on that date will be included.

*Standard SNOMED concept and its descendants will be used. Stage-specific concepts are not available in the data sources used in this study. Consequently, it will be assumed that most individuals in this study will be identified at stage 3, corresponding to the initiation of insulin treatment.

Constructing an accurate phenotype for type 1 diabetes mellitus is complex. Prior work has shown that diagnostic codes alone may yield low sensitivity and suboptimal case identification, and that combining diagnosis codes with treatment records, such as prescriptions for insulin or insulin delivery devices, can substantially improve the positive predictive value of the case definition.(7) While this study does not aim to develop or validate a novel case definition, phenotypes will be further refined during the Cohort Diagnostics stage, which includes clinical experts' review of code lists and the observed counts across age groups, following their execution across the participating data sources. This process may incorporate or exclude disease codes, and/or add additional conditions to construct a fit-for-purpose definition as part of the phenotyping process.(8)

The preliminary concept sets used for identifying outcomes are described in **ANNEX I**. Data sources description.

8.6.3. Other covariates and other variables

The covariates to be used to fulfil Objectives 1 and 2 are as follows:

- Age/age groups defined 10-year age bands at the index date, namely:
    - 0–9; 10–19; 20–29; 30–39; 40–49, ≥50 years
- Sex
- Calendar year
- Medications (Assessment window at several windows: index date: [0, 0], and any time prior to index [-Inf, -1]):
    - Insulin (any)
    - Immune modulation
    - Oral glucose-lowering drugs by individual classes: Metformin, sulfonylureas, dipeptidyl peptidase-4 (DPP-4) inhibitors, sodium-glucose cotransporter 2 (SGLT-2)
    - Verapamil
- Comorbidities (Assessment window at index date: [0, 0], and any time prior [-Inf, -1]):
    - Ketoacidosis
    - Hypothyroidism

  - Hyperthyroidism
  - Celiac disease
  - Other autoimmune conditions (reported together: pernicious anaemia, Addison's disease, Autoimmune hepatitis)
  - Overweight and obesity
  - Hypertension
  - Type 2 Diabetes mellitus (T2 DM)
- Selected characteristics at one-year post-diagnosis (Assessment window [1, 365])
  - Type 2 diabetes mellitus
  - Oral glucose-lowering drugs by individual classes: Metformin, sulfonylureas, dipeptidyl peptidase-4 (DPP-4) inhibitors, sodium-glucose cotransporter 2 (SGLT-2)
- Type 1 diabetes mellitus-related diagnostic test measurements of interest (Assessment window any time prior to index date: [-Inf, -1]):
  - HbA1c measurements
  - Oral glucose tolerance test
  - Fasting glucose measurements
  - Random glucose measurements
  - Insulin and /or C-peptide measurement
  - Insulin autoantibodies (IAA)
  - Islet Cell autoantibodies (ICA)
  - Anti-IA-2 (insulinoma-associated antigen-2) antibodies
  - Anti-Zinc transporter 8 (ZnT8) antibodies
  - Anti-glutamic acid decarboxylase (GAD-65) antibodies
- The diagnostic test measurements will be classified as abnormal if they meet predefined clinical thresholds, specifically (9):
  - HbA1c ≥6.5% or ≥48 mmol/mol
  - Fasting glucose ≥126 mg/dL or ≥7.0 mmol/L
  - Random glucose ≥200 mg/dL or ≥11.1 mmol/L
  - Abnormal OGTT (2-hour plasma glucose ≥200 mg/dL or ≥11.1 mmol/L)
  - Low C-peptide <0.2 nmol/L or <0.6 ng/mL
  - Positive result for one or more islet autoantibodies (e.g., GAD-65, IA-2, ZnT8, IAA, ICA)

Thresholds for defining abnormality will be applied uniformly across data sources.

- Other measurements (Assessment window any time prior to index date: [-180, -1]):
  - BMI measurement

For Objective 2, both the first recorded occurrence of each measurement listed under Type 1 diabetes mellitus-related measurements and the first abnormal result (as defined by prespecified clinical

thresholds), when available, will be used to estimate the time from testing to the formal diagnosis of type 1 diabetes mellitus . Median time intervals will be calculated separately for each recorded test.

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

8.7. Study size

No sample size has been calculated, as this is a descriptive disease epidemiology study that will not test a specific hypothesis. Additionally, we will utilise previously collected data to estimate the prevalence of type 1 diabetes mellitus. Thus, the sample size is driven by the availability of data for patients with type 1 diabetes mellitus. Based on a preliminary feasibility assessment, the expected number of persons with type 1 diabetes mellitus in the data sources included in this study ranges from 5,000 (FinOMOP-TaUH Pirha) to 120,300 (DK-DHR).

8.8. Analysis

8.8.1. Federated network analyses

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

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

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

8.8.2. Patient privacy protection

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

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

R-packages

Tools such as *CohortDiagnostics* (5) and *DrugExposureDiagnostics* (6) will be used to provide additional insights into cohort characteristics, record counts, and index event misclassification. The *DrugExposureDiagnostics* package will be used to evaluate ingredient-specific attributes and patterns in drug exposure records. Upon finalisation of the study protocol and creation of the disease and drug cohorts of interest by the DARWIN EU® Coordination Centre, these packages will be executed in each data source by each data partner.

Patient characterisation (Objectives 1 and 2) will be done using the *CohortCharacteristics* R package, developed by DARWIN EU®.(11) This will include descriptive summary statistics as follows:

- Pre-specified patient-level characteristics on and before the index date (newly diagnosed type 1 diabetes mellitus), based on prespecified conditions and medications at their respective time window of interest.
- Pre-specified selected patient-level characteristics at one-year post index date (newly diagnosed type 1 diabetes mellitus), restricted to individuals with at least 365 days of observation time following index date.
- Among individuals newly diagnosed with type 1 diabetes mellitus, we will describe the proportion of individuals who have records of the relevant diagnostic test measurements of interest, as well as the proportion of individuals with abnormal measurement values (where possible), any time prior to the index date, by individual diagnostic test measurement of interest. Results will be reported separately for each test of interest.
- The time interval (in days) between type 1 diabetes-related diagnostic test measurements (see **Section 8.6.3**) and the formal diagnosis of type 1 diabetes mellitus will be summarised by individual measurement of interest. For each measurement type, both the earliest recorded test date and the earliest abnormal result (when available and based on predefined clinical thresholds) will be used. The distribution of time intervals will be reported using the median, interquartile range (IQR), mean, standard deviation, minimum, and maximum. In data sources where measurement values or units are unavailable, only the presence and date of the test will be used; the earliest abnormal test results will not be assessed in such cases.

The main results will describe type 1 diabetes mellitus in the overall population. Supplementary analysis will include results by age group and sex (except for the time interval estimation between testing and diagnosis, which will be provided only by age group).

The point prevalence of type 1 diabetes mellitus (Objective 3) will be calculated using the *IncidencePrevalence* R package, developed by DARWIN EU®.(10)

- Point prevalence calculations will use population-based, non-hospital databases (IPCI and DK-DHR).

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

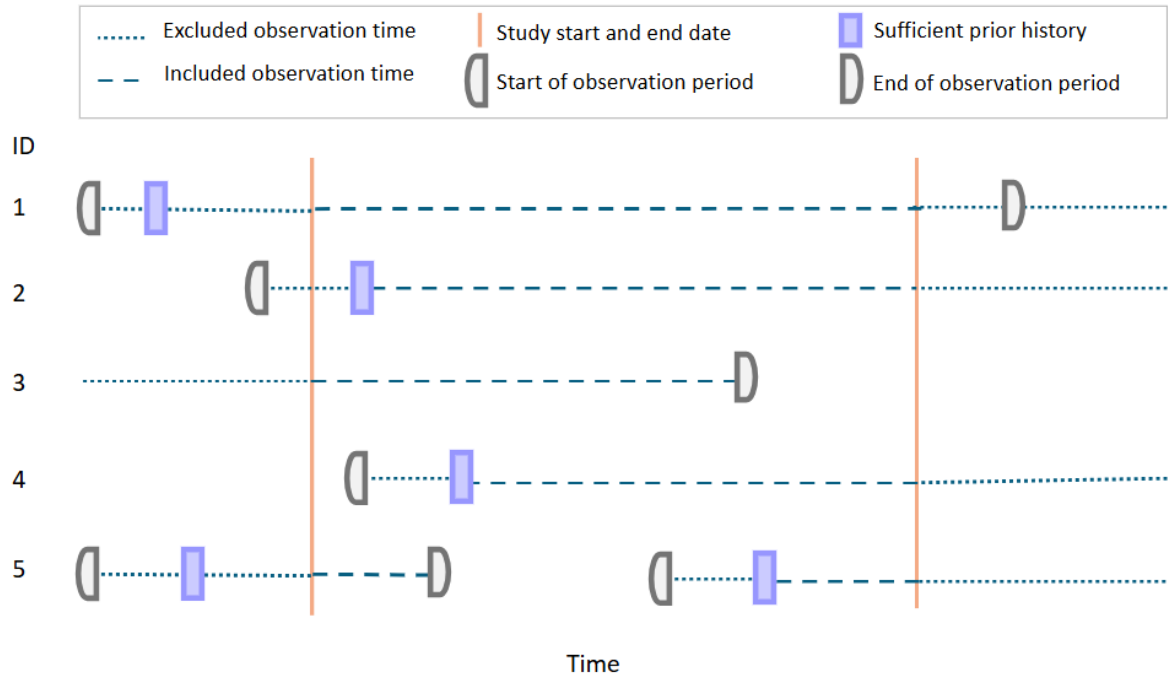


Figure 3. Included observation time for the denominator population.

For annual point prevalence, all individuals who have ever met the type 1 diabetes mellitus case definition on or before the date of interest and are under observation on that date and year will be included.

The point prevalence of type 1 diabetes mellitus will be estimated annually on January 1st (**Figure 4**), defined as the proportion of all individuals who have ever met the case definition on or before the reference date, and all individuals under observation on the reference date, as per the data source.

The main results will describe type 1 diabetes mellitus in the overall population. Supplementary analysis will include results by age groups and sex.

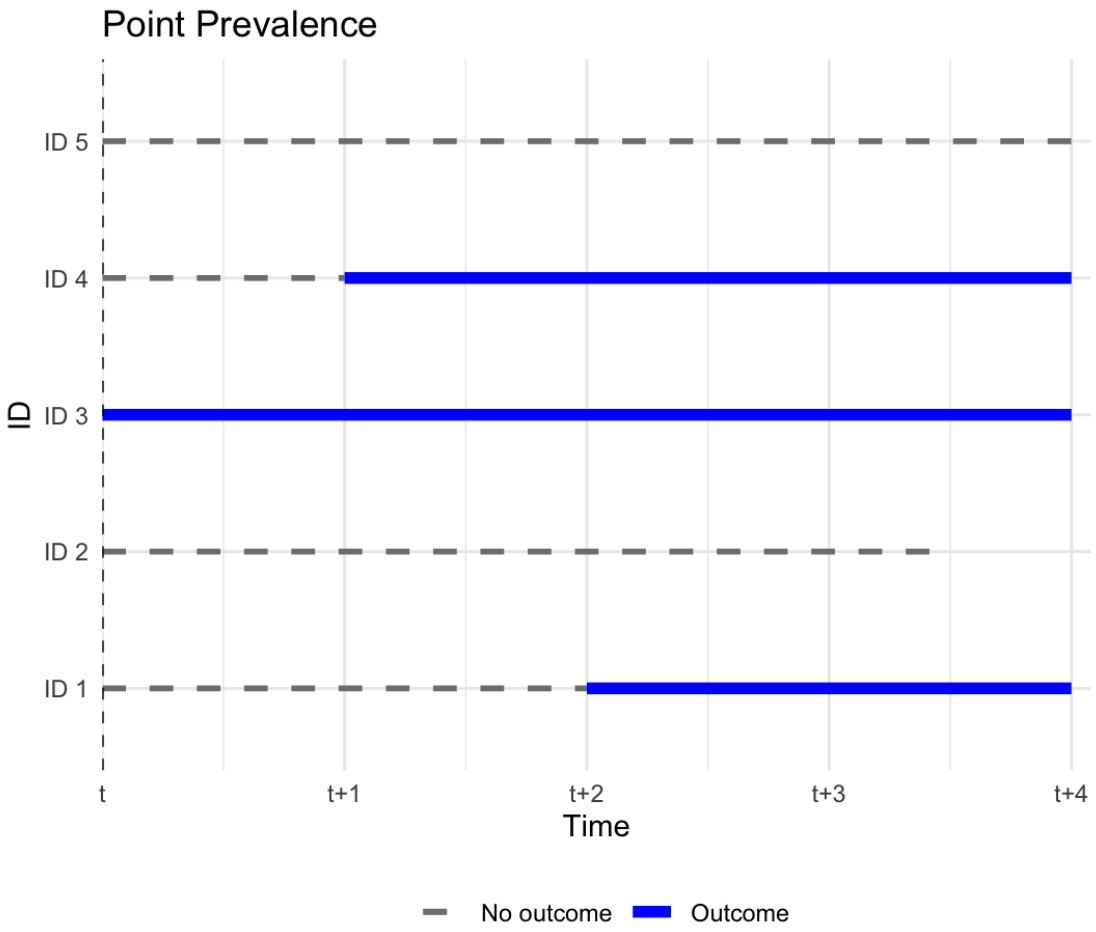


Figure 4. Illustration of Individual Follow-up Time for Point Prevalence Estimation.

Point prevalence is defined as the proportion of individuals who are in the outcome cohort at a specific time point, among those under observation at that time. For example, at time t+2, two of the five individuals are in the outcome cohort, with a point prevalence of 40% (IDs 3 and 4), while at t+3, three of the five individuals are in the outcome cohort (IDs 1, 3, and 4), with a point prevalence of 60%.

Methods to derive parameters of interest

Age

Age at index date will be calculated using January 1st of the year of birth as a proxy for the actual birthday. Date/month is either not present or cannot be made available for governance reasons. If available, the date is often set to the first of the month for the patient's privacy.

The following age groups will be used for stratification: 0–9; 10–19; 20–29; 30–39; 40–49; and ≥50 years.

Calendar time

Calendar time will be determined on the calendar year during which the index diagnosis was recorded.

8.8.4. Output

Output will include the following:

A PDF report including an executive summary, and the following tables and figures.

- Table 1. Overall study participants and number of newly diagnosed type 1 diabetes mellitus cases by data source (2015–2024).

- Table 2. Baseline and post-index characteristics at the time of type 1 diabetes mellitus diagnosis.
- Table 3. Proportion of individuals with diagnostic tests of interest prior to type 1 diabetes mellitus diagnosis, by individual test type.
- Table 4. Proportion of individuals with abnormal diagnostic tests of interest prior to type 1 diabetes mellitus diagnosis, by individual test type.
- Table 5. Timing of first diagnostic test measurements relative to type 1 diabetes mellitus diagnosis (index date), by individual test type.
- Table 6. Timing of first abnormal diagnostic test measurements relative to type 1 diabetes mellitus diagnosis (index date), by individual test type.
- Figure 1. Annual point prevalence of type 1 diabetes mellitus.
- Table 7. Underlying data reported in Figure 1 (annual point prevalence of type 1 diabetes mellitus).

Results in the main study report will be based type 1 diabetes mellitus (main definition) (see **Section 8.6.2.**) in the overall study population and reported per data source.

Age-group and sex stratifications (individually) will be included in the **Annex** for Tables 2–4. Tables 5–6 will be presented by age group only. Figure 1 and the underlying numerators/denominators (Table 7) will also be reported by age and by sex in the **Annex**.

An Interactive dashboard will be generated by incorporating all the results (tables and figures) included in the PDF report mentioned above.

8.9. Evidence synthesis

Results from analyses described in **Section 8.8. Analysis** will be presented separately for each data source. No meta-analysis of results will be conducted.

9. STRENGTHS AND LIMITATIONS

This study has several strengths. It will leverage routinely collected, electronic healthcare data from a wide range of healthcare systems, thereby enhancing the generalisability of its findings. Applying harmonised cohort definitions and phenotyping procedures will ensure consistency across data sources and facilitate reproducibility. Stratification by data source type will offer insights into how settings may affect patient characteristics and prevalence estimates. By focusing on annual point prevalence, the study will also provide an estimate of the burden of type 1 diabetes mellitus in the included populations.

Relevant limitations should be considered. This study will be informed by routinely collected healthcare data, and therefore, data quality issues must be considered. In particular, the recording of type 1 diabetes mellitus may vary across data sources. While relatively few false positives are expected, given the clinical specificity of the diagnosis, false negatives may be more likely, particularly in data sources with incomplete data capture across healthcare settings. In such cases, diagnostic updates made in specialist care may not be visible, leading to misclassification. Excluding individuals with a prior history of type 2 diabetes mellitus may also contribute to false negatives when individuals with true type 1 diabetes mellitus are initially misdiagnosed as having type 2. This is particularly relevant in older adolescents and adults, where type 1 diabetes mellitus can be difficult to distinguish from type 2 diabetes mellitus based solely on clinical presentation. The most reliable discriminator in this age group is the presence of autoantibodies; however, these results may not be consistently available in the data sources used for this study. For individuals without autoantibody results, classification will rely on diagnostic codes and treatment patterns, which may not fully distinguish between type 1 diabetes mellitus and type 2 diabetes mellitus. This limitation may lead to the inadvertent inclusion of individuals with type 2 diabetes mellitus in the type 1 diabetes mellitus cohort (false positive). Similarly, some individuals with insulin requirements at diagnosis may be

misclassified as type 1 diabetes mellitus (false positives), despite having distinct underlying conditions, such as monogenic diabetes mellitus (e.g., MODY) or youth-onset type 2 diabetes mellitus. To assess the potential impact of such misclassification, we will estimate the proportion of individuals with at least one recorded diagnosis of type 2 diabetes mellitus and exposure to oral glucose-lowering drugs (separately), during the year following the index date.

Hospital-based data sources, such as DK-DHR, FinOMOP-TaUH Pirha, H12O, CDW-Bordeaux, and SUCD, are likely to reflect a subset of individuals with more severe or complex presentations, compared to those managed exclusively in primary care. In sources not linked to primary care, relevant aspects of a patient's clinical history may be missing if they occurred outside the hospital setting, limiting the completeness of case identification and temporal reconstruction. These limitations also affect point prevalence estimation. In hospital-exclusive data sources, estimates based on hospital encounters may be biased due to an unstable denominator, defined by healthcare utilisation rather than a stable population, as well as underrepresenting the true prevalence of type 1 diabetes mellitus by excluding those managed entirely in primary care settings. In turn, hospital-based data sources without population-level denominators will not be used to estimate prevalence in this study. Last, the documentation of comorbidities, medications, and measurements, necessary for patient-level characterisation, may also vary across data sources.

Analyses will include mapped measurements that require records with a valid concept ID, a recorded value, and a unit. This data domain has not been systematically assessed for the quality or completeness of these measurement values within or across data sources. If values or units are missing, not properly recorded, or not correctly mapped to the OMOP CDM, the corresponding measurement will be excluded from the analysis. Some extreme or implausible values may be present and could impact the reliability of the results. Classification of "abnormal" glucose and other diabetes-related measurements will follow the prespecified thresholds, which may not reflect local reference standards, potentially leading to misclassification in some data sources. Body mass index (BMI) values will be based on those recorded in the CDM. For individuals <18 years of age, these values may be inaccurate, as they are not necessarily adjusted for age- and sex-specific growth curves.

Lastly, the results estimated in this study will only reflect the populations represented in the included data sources. Electronic health records have certain inherent limitations because they were primarily collected for clinical purposes rather than research use. Consequently, using 6 primary care, secondary care, hospital care, and registry-based data sources from Denmark, Finland, France, Hungary, the Netherlands, and Spain limits generalisability to those countries.

10. REFERENCES

1. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. The Lancet. 2014;383(9911):69-82.

2. Insel RA, Dunne JL, Atkinson MA, Chiang JL, Dabelea D, Gottlieb PA, et al. Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. Diabetes Care. 2015;38(10):1964-74.

3. Phillip M, Achenbach P, Addala A, Albanese-O'Neill A, Battelino T, Bell KJ, et al. Consensus guidance for monitoring individuals with islet autoantibody-positive pre-stage 3 type 1 diabetes. Diabetologia. 2024;67(9):1731-59.

4. DeFalco F, Ryan P, Schuemie M, Huser V, Knoll C, Londhe A, et al. Achilles: Achilles Data Source Characterization. R package version 1.7.2. 2023.

5. Gilbert J, Rao G, Schuemie M, Ryan P, Weaver J. CohortDiagnostics: Diagnostics for OHDSI Cohorts. R package version 3.3.0, https://github.com/OHDSI/CohortDiagnostics, https://ohdsi.github.io/CohortDiagnostics. 2024.

6. Inberg G, Burn E, Burkard T. DrugExposureDiagnostics: Diagnostics for OMOP Common Data Model Drug Records. R package version 1.0.9, https://github.com/darwin-eu/DrugExposureDiagnostics, https://darwin-eu.github.io/DrugExposureDiagnostics/. 2024.

7. Okui T, Nojiri C, Kimura S, Abe K, Maeno S, Minami M, et al. Performance evaluation of case definitions of type 1 diabetes for health insurance claims data in Japan. BMC Medical Informatics and Decision Making. 2021;21(1):52.

8. Dernie F, Corby G, Robinson A, Bezer J, Mercade-Besora N, Griffier R, et al. Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU). Pharmacoepidemiol Drug Saf. 2024;33(11):e70042.

9. Standards of medical care in diabetes--2011. Diabetes Care. 2011;34 Suppl 1(Suppl 1):S11-61.

10. Raventós B, Català M, Du M, Guo Y, Black A, Inberg G, et al. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. Pharmacoepidemiology and drug safety. 2024;33(1).

11. Catala M GY, Lopez-Guell K, Burn E, Mercade-Besora N, Alcalde M CohortCharacteristics: Summarise and Visualise Characteristics of Patients in the OMOP CDM. R package version 0.4.0 2024 [Available from: https://darwin-eu.github.io/CohortCharacteristics/.

11. ANNEXES

ANNEX I. Data sources description

DATA SOURCES DESCRIPTION

Danish Data Health Registries (DK-DHR)

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

Tampere University Hospital Cohort (FinOMOP - TaUH Priha)

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

Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)

The clinical data warehouse at Bordeaux University Hospital, in France, comprises electronic health records for over 2 million patients, with data collection beginning in 2005. The hospital complex comprises three main sites and has a total of 3,041 beds (as of 2021). The database currently holds information about the person (demographics), visits (inpatient and outpatient), conditions and procedures (billing codes), drugs (outpatient prescriptions and inpatient orders and administrations), measurements (laboratory tests and vital signs), and dates of death (in-hospital and out-of-hospital deaths).

Semmelweis University Clinical Data (SUCD)

Semmelweis University is the largest provider of health care services in Hungary. Most of the departments cater to the most serious cases and patients requiring complex treatment, thus making the university a national healthcare provider. The overwhelming majority of patient data originates from Hungary, primarily from the central region of the country, including Budapest and Pest County. The database contains approximately 2 million individual patients across all care settings of the University since 2011. The hospital information system (MedSolution) is an integrated IT system that provides functional support for inpatient and outpatient care processes, serving as a unified platform for various diagnostic areas. In some specific areas, it supports the registration of medications. It supports all kinds of hospital work processes from admission to discharge. The outpatient module serves as a platform for registering activities related to care episodes within outpatient specialist care. During care provision, data related to the patient's health state, diagnosis, documentation of requested examinations and medical consultations, prescribed medication, final reports, and performed interventions are recorded. The functions of the inpatient module support

care provision within inpatient settings. It documents the patient's health state at admission and during the hospital stay, along with the anamnesis, diagnosis, performed examinations and interventions, hospital final reports, and provided medication in certain areas of care provision, such as chemotherapy. Among other modules, the diagnostic module registers the requested laboratory and imaging examinations and records the laboratory results.

Integrated Primary Care Information (IPCI)

The database collects data in the Netherlands. It was started in 1992 by the Department of Medical Informatics of the Erasmus University Medical Center in Rotterdam, the Netherlands. The current database contains patient records from 2006 onwards, when the database's size began to increase significantly. IPCI is a nationwide Dutch database. However, it mainly covers the central part of the country, including the most densely populated and non-urban areas. The IPCI database contains data from records of general practitioners' (GPs) practices. It contains information on all patients registered with GPs responsible for non-emergency care and referrals. More than 99% of the Dutch population has health insurance, and almost all citizens are registered with a general practitioner. Over 12 months, around 78% of the population has at least one contact with their GP. IPCI included around 350 GP practices out of approximately 5,000 in the country (~7%). The demographic composition of the IPCI population mirrors that of the general Dutch population in terms of age and sex.

IPCI obtains data from computer-based patient records. Patient-level data includes demographic information, complaints and symptoms, diagnoses, laboratory test results, lifestyle factors, and correspondence with secondary care, such as referral and discharge letters. Dutch GPs use the International Classification of Primary Care (ICPC-1) coding for complaints, symptoms, and diagnoses, an international standard developed and updated by the World Organisation of Family Doctors (WONCA) International Classification Committee.

Extensive quality control steps are performed before each data release. These include comparing patient characteristics between practices and checking for abnormal temporal data patterns in practices. Additional checks include over 200 indicators related to population characteristics (e.g., reliability of birth and mortality rates) and medical data (e.g., availability of prescription durations, completeness of laboratory results, presence of hospital letters and prescriptions, proportion of patients with blood pressure measurements, etc.).

IPCI is not linked with other databases. Vital status (death date and cause) is collected based on GP records. The main limitation is that IPCI is limited to GP records, and although it contains information on referrals and discharge letters, it may not capture specific hospital information. The database profile was described by de Ritter et al., 2022. The database description is also available at ipci.nl.

Hospital Universitario 12 de Octubre (H12O)

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

ANNEX II. Operational and reporting considerations

DATA MANAGEMENT

Data management

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

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

Data storage and protection

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

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

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

QUALITY CONTROL

General data source quality control

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

Study specific quality control

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

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

The study code will be based on two R packages currently being developed to (1) estimate Incidence and Prevalence and (2) characterise drug utilisation using the OMOP common data model. These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via GitHub.

PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

A PDF report including an executive summary and the specified tables and/or figures will be submitted to EMA by the DARWIN EU® CC upon completion of the study.

An interactive dashboard incorporating all the results (tables and figures) will be provided alongside the PDF report. The full set of underlying aggregated data used in the dashboard will also be made available if requested.

ANNEX III: List of standalone documents

Table S1. Preliminary list of conditions definitions.

Phenotype	Concept name	Concept id (including descendants)	Exclude concept id	Vocabulary
Type 1 diabetes mellitus (concept-based)	Type 1 diabetes mellitus uncontrolled Type 1 diabetes mellitus Disorder due to type 1 diabetes mellitus	40484648 201254 435216	-	SNOMED CT
Hypertension	Hypertensive disorder	316866		
Ketoacidosis	Ketoacidosis	4209145	-	SNOMED CT
Overweight and obesity	Overweight Body mass index 25–29 - overweight Obesity Body mass index 30+ - obesity	437525 4060705 433736 4060985	-	SNOMED CT

Table S2. Preliminary list of medicine definitions.

Substance Name	Concept name	Class	ATC code	Ingredient Concept ID	Include descendants
Insulin	lente insulin, human INSULINS AND ANALOGUES	Ingredient ATC 3rd	- ATC 3rd (A10A)	1513849 21600713	Yes
Immunomodulators	Teplizumab	Ingredient		741995	Yes
Verapamil		Ingredient		1307863	Yes

ANNEX IV: ENCePP checklist for study protocols

Doc.Ref. EMA/540136/2009

ENCePP Checklist for Study Protocols (Revision 4)

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

Study title: DARWIN EU® - Feasibility of studies on early (pre-symptomatic) stages of type 1 diabetes mellitus in the DARWIN EU® network

EU PAS Register® number: EUPAS1000000756
Study reference number (if applicable): P4-C1-010

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

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)	☒	☐	☐	7, 8
2.1.2 The objective(s) of the study?	☒	☐	☐	
2.1.3 The target population? (i.e. population or subgroup to whom the study results are intended to be generalised)	☒	☐	☐	
2.1.4 Which hypothesis(-es) is (are) to be tested?	☐	☐	☒	
2.1.5 If applicable, that there is no *a priori* hypothesis?	☐	☐	☒	

Comments:

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

Section 3: Study design		**Yes**	**No**	**N/A**	**Section Number**
3.1	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	☒	☐	☐	8.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	8.2
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.1
3.4	Does the protocol specify measure(s) of association? (e.g. risk, odds ratio, excess risk, rate ratio, hazard ratio, risk/rate difference, number needed to harm (NNH))	☐	☐	☒	
3.5	Does the protocol describe the approach for the collection and reporting of adverse events/adverse reactions? (e.g. adverse events that will not be collected in case of primary data collection)	☐	☐	☒	

Comments:

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

Comments:

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.1	Does the protocol describe how the study exposure is defined and measured? (e.g. operational details for defining and categorising exposure, measurement of dose and duration of drug exposure)	☐	☐	☒	
5.2	Does the protocol address the validity of the exposure measurement? (e.g. precision, accuracy, use of validation sub-study)	☐	☐	☒	

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

Comments:

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

Comments:

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

Comments:

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

Comments:

Section 9: Data sources	**Yes**	**No**	**N/A**	**Section Number**
9.1 Does the protocol describe the data source(s) used in the study for the ascertainment of:				8.2
9.1.1 Exposure? (e.g. pharmacy dispensing, general practice prescribing, claims data, self-report, face-to-face interview)	☐	☐	☒	
9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	☒	☐	☐	8.5
9.1.3 Covariates and other characteristics?	☒	☐	☐	8.6
9.2 Does the protocol describe the information available from the data source(s) on:				
9.2.1 Exposure? (e.g. date of dispensing, drug quantity, dose, number of days of supply prescription, daily dosage, prescriber)	☐	☐	☒	
9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	☒	☐	☐	8.5
9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	☒	☐	☐	8.6
9.3 Is a coding system described for:				
9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	☐	☐	☒	
9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	☒	☐	☐	8.5, 11
9.3.3 Covariates and other characteristics?	☐	☒	☐	
9.4 Is a linkage method between data sources described? (e.g. based on a unique identifier or other)	☐	☐	☒	

Comments:

Section 10: Analysis plan	**Yes**	**No**	**N/A**	**Section Number**
10.1 Are the statistical methods and the reason for their choice described?	☒	☐	☐	8.8
10.2 Is study size and/or statistical precision estimated?	☒	☐	☐	8.7
10.3 Are descriptive analyses included?	☒	☐	☐	8.8

Section 10: Analysis plan	**Yes**	**No**	**N/A**	**Section Number**
10.4 Are stratified analyses included?	☒	☐	☐	8.8
10.5 Does the plan describe methods for analytic control of confounding?	☐	☐	☒	
10.6 Does the plan describe methods for analytic control of outcome misclassification?	☐	☐	☒	
10.7 Does the plan describe methods for handling missing data?	☒	☐	☐	11
10.8 Are relevant sensitivity analyses described?	☐	☐	☒	

Comments:

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

Comments:

Section 12: Limitations	**Yes**	**No**	**N/A**	**Section Number**
12.1 Does the protocol discuss the impact on the study results of:				
12.1.1 Selection bias?	☐	☒	☐	9
12.1.2 Information bias?	☒	☐	☐	
12.1.3 Residual/unmeasured confounding? (e.g. anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).	☐	☐	☒	
12.2 Does the protocol discuss study feasibility? (e.g. study size, anticipated exposure uptake, duration of follow-up in a cohort study, patient recruitment, precision of the estimates)	☒	☐	☐	8.7, 9

Comments:

Section 13: Ethical/data protection issues	**Yes**	**No**	**N/A**	**Section Number**
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	☒	☐	☐	8.2

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

Comments:

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

Comments:

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

Comments:

ANNEX V: Glossary

Aggregated Data

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

Benefit-Risk Assessment

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

Common Data Model (CDM)

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

Complex Studies (C3)

Studies requiring the development or customization 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, 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 database 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 harmonized 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 databases (9) in DARWIN EU® studies.

Off-the-Shelf Studies (OTS)

Studies for which a standard protocol per study/analysis type and standardized 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 utilization 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 individual patients, often de-identified, used for longitudinal or detailed analyses.

OMOP (Observational Medical Outcomes Partnership)

A common data model (CDM) that standardizes 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 patient health status or healthcare delivery that is collected from routine clinical practice rather than from randomized 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 authorize, 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 databases, 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.

Name of the main author of the protocol: Julieta Politi

Date: 21/07/2025

Signature: Politi