Study Report

P4-C1-010

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

17/03/2026

Version 4.0

Authors: Julieta Politi, Nicholas Hunt, Cesar Barboza, Maarten van Kessel, Natasha Yefimenko, Katia Verhamme

Public

CONTENTS

LIST OF ABBREVIATIONS **5**

1. TITLE **7**

2. DESCRIPTION OF THE STUDY TEAM **7**

3. ABSTRACT **8**

4. AMENDMENTS AND UPDATES **12**

5. MILESTONES **12**

6. RATIONALE AND BACKGROUND **12**

7. RESEARCH QUESTION AND OBJECTIVES **12**

8. RESEARCH METHOD **13**

8.1. Study design 13

Figure 1. Study design for objectives 1 and 2, on characterising individuals at the time of type 1 diabetes mellitus diagnosis (stage 3) 14

Figure 2. Study design for objective 3, on estimating the point prevalence of type 1 diabetes mellitus (stage 3). 14

8.2. Follow-up 15

8.3. Study population with inclusion and exclusion criteria 15

8.4. Study setting and data sources 15

8.5. Study period 17

8.6. Variables 17

8.6.1. Exposure 17

8.6.2. Outcome 18

8.7. Study size 18

8.7.1. Covariates, including confounders, effect modifiers, intercurrent events, and other variables 18

8.8. Data transformation 20

8.9. Statistical methods 20

8.9.1. Main summary measures and statistical methods 20

Figure 3. Included observation time for the denominator population. 21

Figure 4. Illustration of Individual Follow-up Time for Point Prevalence Estimation. 22

8.9.2. Missing values 22

8.9.3. Sensitivity analysis 23

8.10. Deviations from the protocol 23

Table 1: Deviations from the protocol 23

9. RESULTS **23**

9.1. Participants 23

Table 2. Attrition of the type 1 Diabetes Cohort definition across data sources (2015–2024). 24

9.2. Main results 24

9.2.1. Characterisation 24

Table 3. Baseline demographics of the Type 1 Diabetes study cohort by data source (age and sex), 2015-2024 24

Table 4. Baseline comorbidities and medications in the Type 1 Diabetes study cohort. 26

Table 5. Measurements including autoantibody testing in Type 1 diabetes mellitus. 29

Table 6. Index date and 1-year follow-up characterisation of Type 2 diabetes codes, Glucose-lowering Therapy, and Islet Autoantibody Testing in the Type 1 diabetes cohort. 33

9.2.2. Time from the first-ever and first abnormal measurement to index date 35

Figure 5. Time from First-Ever measurement of interest to Type 1 Diabetes Mellitus Diagnosis: First Recorded and First Abnormal Tests for a) Glycaemic measurements, and b) C-peptide and autoantibodies. 38
9.2.3. Annual point prevalence 39
Table 7. Annual point prevalence of Type 1 diabetes on 1 January (%, 95% CI) by data source (DK-DHR and IPCI), 2015–2024. 40
Figure 6. Temporal trends of annual point prevalence of Type 1 diabetes on 1 January by data source (DK-DHR and IPCI), 2015–2024. 41
Table 8. Sensitivity Analysis: Annual point prevalence of Type 1 diabetes on 1 January, excluding individuals with any prior history of Type 2 diabetes (%, 95% CI), by data source (Dk-DHR and IPCI), 2015–2024. 42
Figure 7. Temporal trends of annual point prevalence of type 1 diabetes on 1 January, Excluding Individuals with Any Prior History of Type 2 Diabetes, by Data Source (DK-DHR and IPCI), 2015–2024. 43

10. DISCUSSION 43
10.1. Key results 43
10.2. Strengths and limitations of the research methods 45
10.3. Interpretation 46
10.4. Generalisability 49

11. CONCLUSION 50

12. REFERENCES 51

13. ANNEXES 53
ANNEX I. Description of data sources 53
ANNEX II. Operational and reporting considerations 61
ANNEX III: List of conditions and medication definitions 62
Table S1. List of conditions definitions. 62
Table S2. List of medication definitions. 62
ANNEX IV: Supplementary results 66
Figure S1. Temporal trends of annual point prevalence of type 1 diabetes on 1 January by Data Source and (DK-DHR and IPCI) Age Group, 2015–2024 66
Figure S2. Temporal trends of annual point prevalence of type 1 diabetes on 1 January by Data Source (DK-DHR and IPCI) and Sex, 2015–2024. 66
Figure S3. Temporal trends of annual point prevalence of type 1 diabetes on 1 January, Excluding Individuals with Any Prior History of Type 2 Diabetes, by Data Source (DK-DHR and IPCI) and Age Group, 2015–2024. 67
Figure S4. Temporal trends of annual point prevalence of type 1 diabetes on 1 January, Excluding Individuals with Any Prior History of Type 2 Diabetes, by Data Source (DK-DHR and IPCI) and Sex, 2015–2024. 67
Table S3. Representation of autoantibody results across data sources. 68
ANNEX V: Glossary 69

Study title	DARWIN EU® - Feasibility of studies on early (pre-symptomatic) stages of type 1 diabetes mellitus in the DARWIN EU® network
Study report version	V4.0
Date	17/03/2026
EUPAS number	EUPAS1000000756
Active substance	NA
Medicinal product	NA
Research question and objectives	Research question 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 was to investigate the feasibility of conducting research on the early (pre-symptomatic) stages of type 1 diabetes mellitus within the DARWIN EU® network. It focused on the frequency and timing of autoantibody and glucose testing before the disease becomes clinically apparent. The specific objectives for this study were: 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, prior to and at the time of type 1 diabetes mellitus diagnosis, and to assess selected characteristics at one-year post-index date. 2. To estimate, for each diagnostic test of interest, 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, The Netherlands, Spain
Authors	Julieta Politi (j.politi@darwin-eu.org) Nicholas Hunt n.hunt@darwin-eu.org Cesar Barboza c.barboza@darwin-eu.org Maarten van Kessel m.vankessel@darwin-eu.org Natasha Yefimenko n.yefimenkonosova@darwin-eu.org Katia Verhamme (k.verhamme@darwin-eu.org)

LIST OF ABBREVIATIONS

Acronyms/term	Description
ATC	Anatomical Therapeutic Chemical
BMI	Body Mass Index
CC	Coordination centre
CDM	Common Data Model
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
DPP-4	Dipeptidyl peptidase-4 inhibitors
DQD	Data Quality Dashboard
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
GP	General Practitioner
H12O	Hospital Universitario 12 de Octubre
IA-2	Insulinoma-associated antigen-2
IAA	Insulin autoantibodies
ICA	Islet Cell autoantibodies
ICD	International Classification of Diseases
IPCI	Integrated Primary Care Information
IQR	Interquartile range
IRB	Institutional Review Board
MODY	Maturity-Onset Diabetes of the Young
NICE	National Institute for Health and Care Excellence
OGLD	Oral Glucose-lowering Drug
OGTT	Oral glucose tolerance test
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
RxNorm	Medical prescription normalised
SNOMED	Systematized Nomenclature of Medicine
SGLT-2	sodium-glucose cotransporter 2

Acronyms/term	Description
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, including a decreased risk of diabetic ketoacidosis (DKA) at disease onset and a notable reduction in clinical symptoms. In addition, 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 on early screening (using specific antibodies), which helps identify 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 was to investigate the feasibility of conducting research on the early (pre-symptomatic) stages of type 1 diabetes mellitus within the DARWIN EU® network. It focused on the frequency and timing of autoantibody and glucose testing before the disease becomes clinically apparent.

The specific objectives for this study were:

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, prior to and at the time of type 1 diabetes mellitus diagnosis, and to assess selected characteristics at one-year post-index date.
2. To estimate, for each diagnostic test of interest, the median (Interquartile range (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

For objectives 1 and 2, the study population included 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 was required (applied to non-hospital-based data sources and individuals aged 1 year or older).

The study population for objective 3 included all individuals present in the data source 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 data source 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 required the presence of both SNOMED CT codes indicative of first-ever type 1 diabetes mellitus and initiation of first-ever insulin therapy (RxNorm codes), within 180 days of each other. The index date was defined as the earlier of the two events.

Relevant covariates: The covariates of interest included 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 included HbA1c, oral glucose tolerance test, fasting and random glucose, insulin and/or C-peptide measurements, insulin antibodies (IAA), islet cell antibodies (ICA), insulinoma-associated antigen-2 antibody (IA-2A), anti-zinc transporter 8 antibodies (ZnT8), and anti-glutamic acid decarboxylase (GAD-65) antibodies. Prespecified medications included immune modulators (teplizumab), oral glucose-lowering medications (individually), and verapamil. Comorbidities of interest included diabetic ketoacidosis, thyroid disease (hypo- and hyper-thyroidism, assessed separately), coeliac disease, and other autoimmune diseases, among other conditions of interest.

Data sources: Two population-based data sources (Danish Data Health Registries (DK-DHR) and Integrated Primary Care Information (IPCI)), four hospitals (Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux), Semmelweis University Clinical Data (SUCD), Hospital Universitario 12 de Octubre (H12O), and Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha) were used in the study.

Statistical analysis

Baseline characteristics (age, sex, predefined comorbidities, medications, and diagnostic tests of interest, including antibodies) were summarised as n (%) for categorical variables and mean (SD), median (IQR), minimum, and maximum for continuous variables. Other selected characteristics were assessed at one-year post-index date (type 2 diabetes mellitus, oral glucose-lowering drugs, antibodies).

The point prevalence of type 1 diabetes mellitus was 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. A sensitivity analysis for point prevalence excluded individuals with any prior history of type 2 diabetes mellitus.

The main body of the report describes type 1 diabetes mellitus in the overall population. Results stratified by age group and by sex are presented separately in the Shiny app.

A minimum cell count of 5 was used for reporting results.

Results

Cohort identification and demographics

Across six data sources, type 1 diabetes varied by source and setting: hospital-based data sources included 615–3,651 individuals, while population-based data sources included between 695 individuals (IPCI; regional) and 10,219 (DK-DHR; nationwide). Median age ranged between 19–56 years across data sources, with most individuals in the mid-20s to early-30s, while identified individuals were older in one data source (CDW Bordeaux). Sex distributions showed male predominance (range 53–60% male) in most data sources.

Clinical characteristics

Presence of ketoacidosis at the index date ranged between 4% (DK-DHR) and 16% (FinOMOP-TaUH), and in all but DK-DHR, it was higher than at any time prior to the index date. Type 2 diabetes codes (any time prior to index date) ranged between 3% (FinOMOP-TaUH) and 28% (SUCD). Non-insulin glucose-lowering therapy (any time prior to index date) ranged between 1% (H12O) and 22% (DK-DHR) and was primarily driven by metformin. Insulin therapy at index date ranged from 16% (IPCI) to 85% (H12O) and was higher in individuals identified in hospital-based data sources.

Measurement records

Presence of "Any measurement" of interest record before the index date ranged between 33% to 76% and 9% to 40% at index date. Random glucose was the most consistently identified glycaemic test, followed by HbA1c and oral glucose tolerance test (OGTT). "Any autoantibody measurement" recorded before the index date ranged between 1% and 16% and was available in most sources (except IPCI, <5). Positive autoantibody results were generally low (<5% of individuals).

Changes in clinical characterisation during 1-year following index date (restricted to individuals with ≥365 days follow-up)

During the 1-year follow-up (1–365 days), diagnosis of type 2 diabetes (first-ever record) increased from 2–6% at index date to 4–12%. Non-insulin glucose-lowering drug therapy rose from 3–8% at index date to 10–24% during days 1-365, primarily driven by metformin. First-ever antibody testing was common at index date and during days 1–365 in all data sources, except IPCI; in some sources (e.g., FinOMOP-TaUH, H12O), >50% of first-ever antibody tests occurred during these 2 windows.

Time from earliest and earliest abnormal measurement to diagnosis

For random glucose and frequently for HbA1c and OGTT, the earliest recorded tests were generally years before diagnosis (medians ranging between 2–9 years in several data sources). Earliest abnormal results generally occurred closer to diagnosis (often days to a few months prior), with variability by data source. For autoantibodies tested prior to index date, GAD-65 had the highest counts.

Point Prevalence

Annual point prevalence (2015–2024) was estimated in the two population-based data sources. In the primary analysis in DK-DHR, the prevalence declined steadily (from 89.8 per 10,000 in 2015 to 81.8 per 10,000 in 2024), while in IPCI, it increased modestly between 2015–2022 (from 15.6 per 10,000 to 19.3 per 10,000 in 2022, then decreased slightly in 2023–2024 to 17.2). In the sensitivity analysis (excluding individuals with any prior type 2 diabetes history), point prevalence was lower in both sources across all years (with a range of 54.5–55.2 per 10,000 in DK-DHR and 14.0–17.1 per 10,000 in IPCI). Age-stratified analyses in the sensitivity analysis showed heterogeneous trends. In DK-DHR, prevalence increased in younger age groups and declined in ages 40–49 and ≥50 years; in IPCI, prevalence generally increased over time in age groups <40 years, while trends in older age groups were more variable.

Discussion

This multi-data source feasibility study assessed whether routinely collected healthcare data can support identification of early type 1 diabetes (prior to clinical diagnosis and insulin-dependence). Glycaemic testing was commonly available, and the earliest testing often occurred years prior to index date, while the earliest abnormal testing was closer to index date. Autoantibody testing prior to index date was recorded in ~1–16% of individuals. Autoantibody "positivity" (abnormal, defined as qualitative "positive" record or numeric value above the test/unit-specific threshold) was low in the included data sources, limiting their use in future studies in which confirmation or staging may be required without further standardisation.

The type 1 diabetes cohorts identified were broadly plausible in terms of demographic and clinical characteristics. Requiring a first-ever type 1 diabetes record and first-ever insulin exposure within 180 days of the first qualifying event reduced initially identified individuals, with variable attrition across data sources. Ketoacidosis at entry was higher in hospital-based data sources, consistent with the expected acute and more severe presentation pattern. Recording of type 2 diabetes codes and non-insulin glucose-lowering therapy, more frequent in older adults, suggested diabetes-type 2 overlap and classification uncertainty within the type 1 diabetes cohorts.

Point-prevalence estimates (DK-DHR and IPCI) differed markedly between the two population-based data sources. Sensitivity analyses excluding prior type 2 diabetes reduced absolute prevalence levels. Sensitivity

analyses estimates in DK-DHR were broadly consistent with published general-population prevalence for Denmark. In contrast, for IPCI, estimates remained lower than the available estimates for the Netherlands, which may reflect under-ascertainment related to regional coverage and incomplete capture of secondary-care information (e.g., fragmented type 1 diabetes and/or insulin initiation). Point-prevalence trends are not directly comparable to reported type 1 diabetes incidence trends, as incidence and prevalence reflect different processes; in age-stratified sensitivity analyses, prevalence increased in younger age groups (<40 years) in both sources, whereas older-age trends were less consistent.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study deliverable	Timelines (planned)	Timelines (actual)
Final Study Protocol	September 2025	September 2025
Creation of Analytical code	September 2025	September/October 2025
Execution of Analytical Code on the data	October 2025	November/December 2025
Draft Study Report	November 2025	January 2025
Final Study Report	February 2026	To be confirmed by EMA

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 was 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 investigated 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 was 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 focused

on the frequency and timing of antibody and glucose testing before the disease becomes clinically apparent.

The specific objectives for this study were:

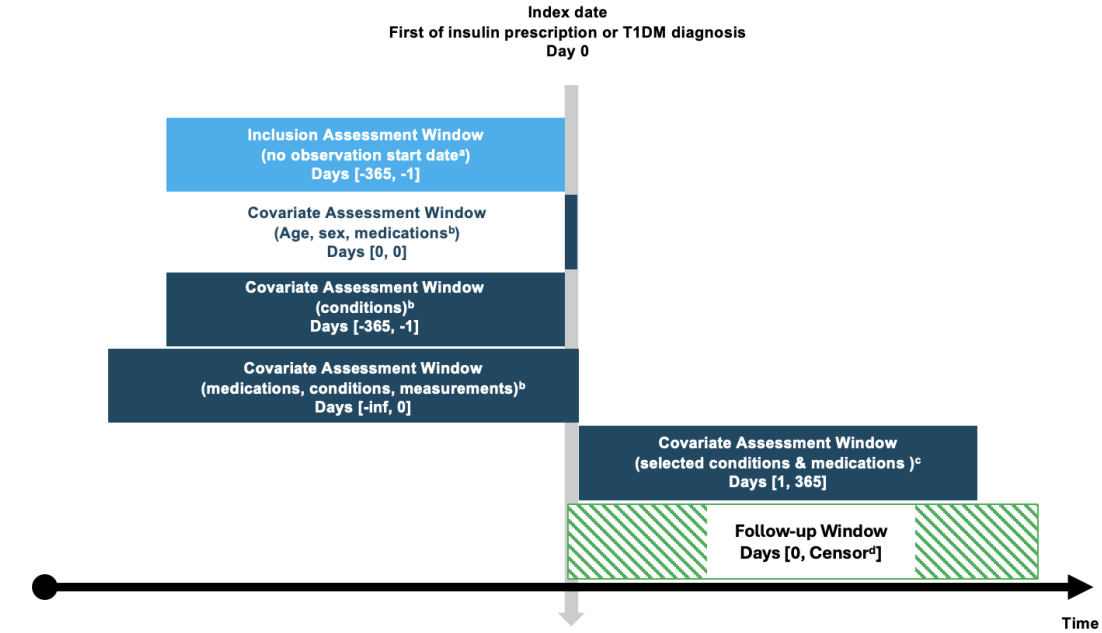
1. To describe the characteristics of individuals newly diagnosed with type 1 diabetes mellitus in terms of demographics, prespecified comorbidities and medications, and diagnostic tests of interest (HbA1C, C-peptide, glucose, and each autoantibody assay), prior to and at the time of type 1 diabetes mellitus diagnosis, and to assess selected characteristics at one-year post-index date.
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 METHOD

8.1. Study design

A retrospective cohort study was conducted using routinely collected health data from six data sources across six European Union (EU) member states. The study included:

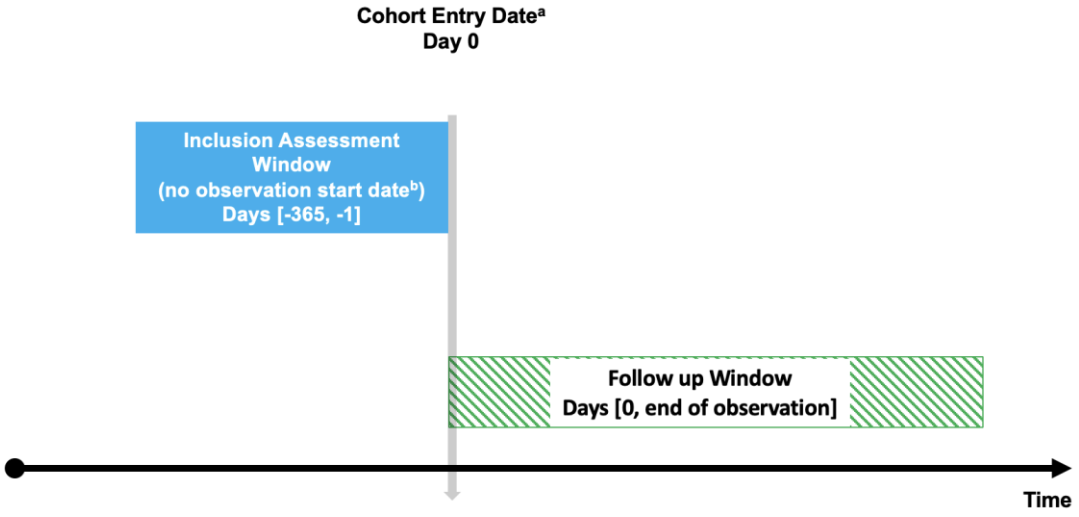
- Assessment of baseline characteristics of newly diagnosed type 1 diabetes mellitus in terms of demographics, prespecified comorbidities and medications, and diabetes-related diagnostic testing (HbA1c, C-peptide, glucose measurements, antibodies) (**Figure 1**).
- Estimation of the median time from the first record of each diabetes-related diagnostic test of interest to the first formal diagnosis of type 1 diabetes mellitus (**Figure 1**), and estimation of the median time from the first abnormal test result (by individual test) to the first formal diagnosis of type 1 diabetes mellitus.
- Calculation of the annual (point) prevalence of type 1 diabetes mellitus (in non-hospital data sources) (**Figure 2**).



a. Applies to individuals older than one year and individuals not in hospital data sources
b. Covariates
c. Selected conditions & medications: type 2 diabetes and oral-glucose-lowering drugs
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. Follow-up

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

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

8.3. 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 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).

8.4. Study setting and data sources

This study was 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, representing 6 EU member states. All data were a priori mapped to the Observational Medical Outcomes Partnership Common Data Model (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. The Netherlands: Integrated Primary Care Information (IPCI)
6. Spain: Hospital Universitario 12 de Octubre (H12O)

Data Selection

These data sources fulfilled the criteria for 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 metrics such as age distribution, condition prevalence by 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 the plausibility of data completeness, consistency, and conformity across data sources.

In terms of relevance, the DARWIN EU® portal and information from the onboarding documents were used to assess whether data sources include information on type 1 diabetes mellitus and relevant diabetes-related measurements.

The data sources were selected based on their type 1 diabetes mellitus counts, availability of antibody testing, and the presence of diabetes-related measurements of interest 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 were generally representative of the target population: nationwide secondary care data (DK-DHR) or regional GP data (IPCI), whereas the others included only hospitalised patients or patients with a secondary care encounter.

Data source justification and key characteristics

Danish Data Health Registries (DK-DHR)

DK-DHR was 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 was approximately 120,300.

Moreover, data availability and follow-up in DK-DHR were sufficient: data collection began in 1995, and the most recent data extraction date was 2024, which aligned with the study period. The median follow-up of the first observation period was 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.

Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)

FinOMOP-TaUH Pirha was 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 the secondary-care setting.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus was approximately 5,000. Additionally, approximately 1,500 person-counts for antibody measurements were expected (representing the number of individuals whose test results are recorded).

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

Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)

CDW Bordeaux was 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 the secondary-care setting.

Based on a preliminary feasibility assessment, the expected number of person-counts for type 1 diabetes mellitus was approximately 14,700. Additionally, approximately 1,500 person-counts for antibody measurements were expected (representing the number of individuals whose test results are recorded).

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

Semmelweis University Clinical Data (SUCD)

SUCD was included in this study because it is a secondary-care, 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 was approximately 11,500. Additionally, approximately 2,900 person-counts for antibody measurements were expected (representing the number of individuals whose test results are recorded).

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

Integrated Primary Care Information (IPCI)

IPCI was 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 was approximately 9,700. Additionally, approximately 1,700 person-counts for antibody measurements were expected (representing the number of individuals whose test results are recorded).

Moreover, data availability and follow-up in IPCI were sufficient: data collection began in 2006, and the most recent data extraction date was 2024, aligning with the study period. The median follow-up of the first observation period in Data source 1 was 1730 days (IQR: 791–3070).

Hospital Universitario 12 de Octubre (H12O)

H12O was 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 was approximately 23,700. Additionally, approximately 6,000 person-counts for antibody measurements were expected (representing the number of individuals whose test results are recorded).

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

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

8.5. Study period

The study period spanned from 01/01/2015 to 31/12/2024, or to the latest data availability, when a data source's observation period did not cover the entire study period.

8.6. Variables

8.6.1. Exposure

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

8.6.2. Outcome

Type 1 diabetes mellitus phenotype (main definition)

Individuals were classified as type 1 diabetes cases when they fulfilled 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 was required. For Objective 3, all individuals who met the type 1 diabetes mellitus case definition on or before the reference date and were under observation on that date were included.

*Standard SNOMED concept and descendants were used. Since stage-specific concepts were not available in the data sources used in this study, it is assumed that most individuals were identified at stage 3, corresponding to the initiation of insulin treatment.

The final concept sets used for identifying outcomes are presented in **Annex III**.

8.7. Study size

No sample size was calculated, as this was a descriptive disease epidemiology study that did not test a specific hypothesis. Additionally, previously collected data were used to estimate the prevalence of type 1 diabetes mellitus. Thus, the sample size was driven by the availability of data for patients with type 1 diabetes mellitus. Based on a preliminary feasibility assessment, the number of persons with a type 1 diabetes mellitus record (SNOMED CT condition recorded) in the selected data sources ranged from 5,000 (FinOMOP-TaUH Pirha) to 120,300 (DK-DHR).

8.7.1. Covariates, including confounders, effect modifiers, intercurrent events, and other variables

The covariates used for characterisation (Objectives 1 and 2) were the following:

- 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 index date : [0, 0], and any time prior to index date [-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 to index date [-Inf, -1]):
    - Ketoacidosis
    - Hypothyroidism
    - Hyperthyroidism
    - Coeliac disease

    - Other autoimmune conditions (including: pernicious anaemia, Addison's disease, and Autoimmune hepatitis; reported together)
    - Overweight and obesity
    - Hypertension
    - Type 2 Diabetes mellitus
- Selected characteristics at one-year post-index date (Assessment window [1, 365])
    - Type 2 diabetes mellitus (any record and first-ever record)
    - 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]. *For antibodies, also assessed for one year post index date [1,365]):
    - Any of the measurements below
    - 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 were classified as abnormal if they met predefined clinical thresholds, specifically [5]:
    - 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 oral glucose tolerance test (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 were applied uniformly across data sources.

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

For Objective 2, both the earliest recorded occurrence of each measurement listed under Type 1 diabetes mellitus-related measurements and the earliest abnormal result (as defined by prespecified clinical thresholds), when available, were used to estimate the time from testing to the formal diagnosis of type 1 diabetes mellitus. Median time intervals were calculated separately for each recorded test.

Measurement values for the first measurement occurring in the "any time prior" window [-Inf, -1] were obtained for all measurements described above.

The concept sets used for identifying covariates are provided in **Annex I**.

8.8. Data transformation

Analyses were conducted separately for each data source. Before study initiation, test runs of the analyses were performed on a subset of the data sources and quality control checks were performed. Once all the tests passed (see **Annex III**), the final study codes package was released in the version-controlled Study Repository for execution against all the participating data sources.

The data partners locally executed the analytics against the OMOP CDM in R Studio and reviewed and approved the, by default, aggregated results.

The study results of all data sources were checked, after which they were made available to the team, and the dissemination phase started. All results were locked and timestamped for reproducibility and transparency.

8.9. Statistical methods

8.9.1. Main summary measures and statistical methods

R-packages

Tools such as *CohortDiagnostics* [6] and *DrugExposureDiagnostics* [7] were used to provide additional insights into cohort characteristics, record counts, and index date event misclassification. The *DrugExposureDiagnostics* package was 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 were executed in each data source by each data partner.

Patient characterisation (Objectives 1 and 2) was done using the *CohortCharacteristics* R package, developed by DARWIN EU®.[8] This includes 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, 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, any time prior to the index date, by individual diagnostic test measurement of interest. Results are reported separately for each test of interest.
- The time interval (in days) between type 1 diabetes-related diagnostic test measurements (see **Section 8.7.1.**) and the formal diagnosis of type 1 diabetes mellitus was summarised by individual measurement of interest. For each measurement type, both the earliest recorded test date and the

earliest abnormal result (when available, based on predefined clinical thresholds) was used. The distribution of time intervals was reported using the median, interquartile range (IQR), mean, standard deviation, minimum, and maximum. In data sources where measurement values or units were unavailable, only the presence and date of the test was used; the earliest abnormal test results were not assessed in such cases.

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

The point prevalence of type 1 diabetes mellitus (Objective 3) was calculated using the *IncidencePrevalence* R package, developed by DARWIN EU®.[9]

Point prevalence calculations used population-based, non-hospital data sources (IPCI and DK-DHR).

The denominator included 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.

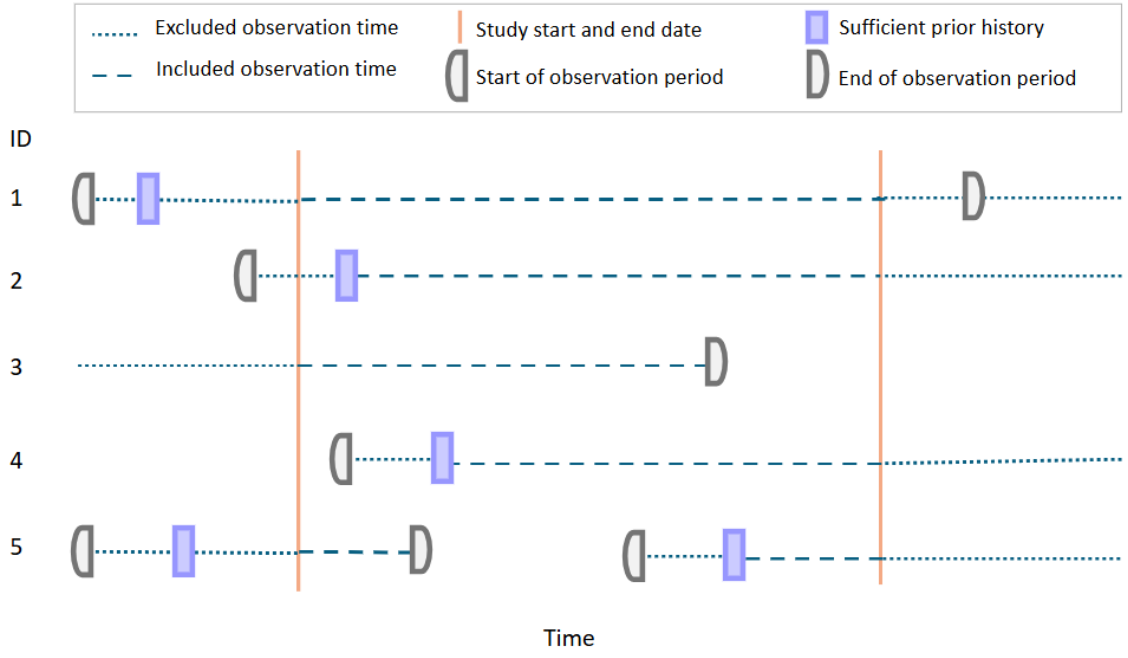


Figure 3. Included observation time for the denominator population.

The point prevalence of type 1 diabetes mellitus was 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 describe type 1 diabetes mellitus in the overall population. Results by age groups and sex are provided in the Shiny app.

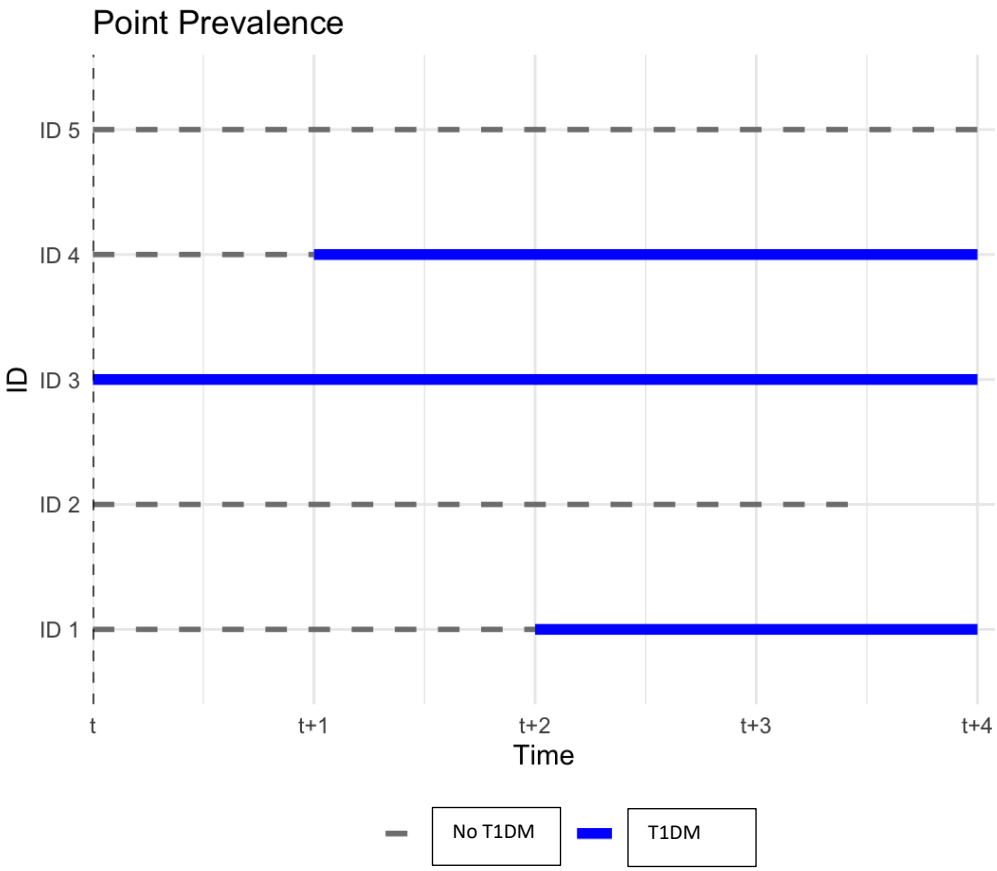


Figure 4. Illustration of Individual Follow-up Time for Point Prevalence Estimation.

Point prevalence is defined as the proportion of individuals who are in the type 1 diabetes mellitus (T1DM) 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%.

Period prevalence was used to obtain the total number of prevalent cases for the entire study period.

Methods to derive parameters of interest

Age

Age at index date was 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. When available, the date is often set to the first of the month for the patient's privacy.

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

Calendar time

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

8.9.2. Missing values

Characterisation and outcomes were based on information recorded in each data source. Conditions, medications, and measurements were identified only when documented; therefore, lack of a record was interpreted as absence of the condition, prescription, or measurement in the available data, acknowledging that this may reflect incomplete capture rather than true absence.

8.9.3. Sensitivity analysis

Point prevalence excluding individuals with a history of type 2 diabetes prior to index date

To assess the influence of potential diabetes subtype classification uncertainty, we conducted a sensitivity analysis applying an additional exclusion criterion to the type 1 diabetes case definition: individuals with any record of type 2 diabetes at any time prior to index date were excluded. Point prevalence was then re-estimated using the same calendar dates, population denominators, and analytic approach as in the primary analysis.

8.10. Deviations from the protocol

Table 1: Deviations from the protocol

Deviation number	Protocol version	Date	Section of study protocol	Deviation	Reason
1	V3	10/11/2025	8.6.3	First Antibodies measurement at index date and post diagnosis windows	To reflect tests that are done at diagnosis and in the year post-index date (A considerable proportion of testing was observed in CohortDiagnostics stage, so assessment in this window was added to best characterise the population).
2	V3	11/12/2025	8.6.3	First diabetes 2 and first OGLD index date and post diagnosis windows	This change was a further clarification to the protocol, as assessment in the post-index date window did not specify between first-ever record vs. any record.
3	V3	11/12/2025	8.8.3	Overall period prevalence	In order to provide an overall number of prevalent cases during the study period, calculation of overall period prevalence was added.
4	V3	11/12/2025	8.8.3	Point prevalence sensitivity analysis was added excluding any prior history of diabetes type 2 to the case definition	In order to quantify the impact of individuals with a prior diagnosis of type 2 diabetes in the prevalence estimates, a sensitivity analysis was done adding an exclusion criterion to the case definition, in which cases were required to satisfy a “no prior history of type 2 diabetes” criteria.

OGLD=Oral Glucose-lowering Drug.

9. RESULTS

The full set of results for this study are available through an interactive web application Shiny App, at EUPAS1000000756.

9.1. Participants

Cohort size varied across data sources and care settings. Hospital-based sources identified 615–3,651 first-ever type 1 diabetes cases. In the two population-based sources, first-ever type 1 diabetes case counts were 10,219 (DK-DHR; nationwide registry-based) and 695 (IPCI; primary care-based) (**Table 2**).

Table 2. Attrition of the type 1 Diabetes Cohort definition across data sources (2015–2024).

Reason	CDM name					
	DK-DHR	FinOMOP-TaUH	CDW Bordeaux	SUCD	IPCI	H12O
First-ever T1D diagnosis or first-ever insulin prescription (whichever occurs first)	107,396	47,683	36,087	17,172	18,060	42,528
Other required criterion met within 180 days of the initial qualifying event	11,035	1,789	3,651	2,751	804	615
Observation history requirement met* (≥365 days; applied to ages ≥1 years only)	10,219	-	-	-	695	-
Type 1 diabetes diagnosis	10,219	1,789	3,651	2,751	695	615

CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information. N=Number of subjects. T1D=Type 1 diabetes mellitus. Type 1 diabetes mellitus was defined as the occurrence of both: first-ever condition occurrence of type 1 diabetes mellitus (SNOMED CT), AND first-ever prescription of insulin at the ingredient level (RxNorm), occurring within 180 days of each other. The index date was defined as the earliest of the two qualifying events.
*In population-based data sources, a minimum of 365 days of observation prior to index date was required in individuals aged ≥ 1; infants aged 0 to <1 year were exempt and retained (DK-DHR: n=20; IPCI: n<5).

9.2. Main results

9.2.1. Characterisation

Demographics

Across data sources, the type 1 diabetes cohorts were generally young to middle-aged, with a median age ranging from 19 years (IPCI) to 56 years (CDW Bordeaux) (**Table 3**). Median age clustered in the mid-20s to early-30s in most data sources. CDW Bordeaux and SUCD had the highest proportion of individuals aged ≥ 50 (56.81% and 36.97% aged ≥50 years, respectively), compared with the other data sources.

Sex distributions showed a male predominance in most sources (52.7–60.1% male), whereas in SUCD, it was balanced.

Table 3. Baseline demographics of the Type 1 Diabetes study cohort by data source (age and sex), 2015-2024.

Variable name	Variable level	CDM name					
		DK-DHR	FinOMOP-TaUH	CDW Bordeaux	SUCD	IPCI	H12O
Type 1 diabetes							
Number subjects	N	10,219	1,789	3,651	2,751	695	615
Age	Median [Q25 - Q75]	28 [13 – 53]	25 [11 – 47]	56 [25 – 71]	36 [13 – 61]	19 [11 – 39]	30 [12 – 48]
	Mean (SD)	33.82 (23.43)	29.95 (21.6)	49.57 (26.5)	38.27 (24.9)	26.84 (21.0)	31.79 (22.6)
	Range	0 to 98	0 to 90	0 to 100	0 to 94	0 to 90	0 to 94
Age group	0 to 9	1,516 (14.8%)	369 (20.6%)	354 (9.7%)	442 (16.1%)	139 (20.0%)	117 (19.0%)

Variable name	Variable level	CDM name					
		DK-DHR	FinOMOP-TaUH	CDW Bordeaux	SUCD	IPCI	H12O
	10 to 19	2,348 (23.0%)	364 (20.4%)	420 (11.5%)	430 (15.6%)	210 (30.2%)	118 (19.2%)
	20 to 29	1,374 (13.5%)	272 (15.2%)	257 (7.0%)	261 (9.5%)	102 (14.7%)	70 (11.4%)
	30 to 39	1,065 (10.4%)	209 (11.7%)	260 (7.1%)	318 (11.6%)	74 (10.7%)	91 (14.8%)
	40 to 49	986 (9.7%)	184 (10.3%)	286 (7.8%)	283 (10.3%)	47 (6.8%)	81 (13.2%)
	≥ 50	2,930 (28.7%)	391 (21.9%)	2,074 (56.8%)	1,017 (37.0%)	123 (17.7%)	138 (22.4%)
Sex	Female	4,190 (41.0%)	812 (45.4%)	1,458 (39.9%)	1,379 (50.1%)	304 (43.7%)	291 (47.3%)
	Male	6,029 (59.0%)	977 (54.6%)	2,193 (60.1%)	1,372 (49.9%)	391 (56.3%)	324 (52.7%)

All values are N(%), unless otherwise specified. CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information. N=Number of subjects. T1D=Type 1 diabetes mellitus. Type 1 diabetes mellitus was defined as the occurrence of both: first-ever condition occurrence of type 1 diabetes mellitus (SNOMED CT), AND first-ever prescription of insulin at the ingredient level (RxNorm), occurring within 180 days of each other. The index date was defined as the earliest of the two qualifying events. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

Comorbidities and medications

Recorded comorbidities varied across sources (**Table 4**). Prior type 2 diabetes ranged between 2.9 (FinOMOP-TaUH)–27.9% (SUCD) (any time prior); and were >17% in DK-DHR (22.3%), CDW Bordeaux (17.5%), and SUCD (27.9%) (**Table 4**). Hypertension any time prior also varied, ranging from 2.42% (FinOMOP-TaUH) to 29.83% (SUCD). Ketoacidosis was more commonly recorded at index date than at any time prior to index date in several data sources (e.g., FinOMOP-TaUH 15.7%, CDW Bordeaux 12.3%, SUCD 7.6%, H12O 9.8% at index date). Overweight/obesity recorded any time prior to index date ranged from 0.6% (FinOMOP-TaUH) to 7.0% (DK-DHR), and the highest value recorded at index date was 13.1% in CDW Bordeaux. Autoimmune comorbidities were generally uncommon; coeliac disease was infrequent and was highest in H12O (3.3% prior; 3.1% at index date), while recorded thyroid disorders remained <7% across data sources and windows.

Medication capture at index date varied. Insulin exposure recorded on the index date ranged from 16.0% (IPCI) to 84.9% (H12O). Non-insulin glucose-lowering therapy was most commonly recorded in DK-DHR (e.g., metformin 21.49% any time prior) and was lower in other sources. Recorded use of DPP-4 inhibitors, SGLT2 inhibitors, and sulfonylureas was infrequent overall. Recorded teplizumab use was not observed in any data source.

Stratified results by age group and sex are provided in the Shiny app. Ketoacidosis at index date was highest in younger age groups and reached up to 38%. Type 2 diabetes codes and non-insulin glucose-lowering therapy increased with age.

Table 4. Baseline comorbidities and medications in the Type 1 Diabetes study cohort.

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=10,219		N=1,789		N=3,651		N=2,751		N=695		N=615	
	Assessment window											
	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]
Comorbidities												
Coeliac disease	47 (0.46%)	21 (0.21%)	10 (0.62%)	6 (0.34%)	<5	10 (0.27%)	11 (0.58%)	43 (1.56%)	5 (0.72%)	0 (0.00%)	18 (3.29%)	19 (3.09%)
Hypertension	1,549 (15.16%)	246 (2.41%)	39 (2.42%)	86 (4.81%)	410 (16.26%)	1,248 (34.18%)	565 (29.83%)	489 (17.78%)	34 (4.89%)	<5	28 (5.12%)	25 (4.07%)
Hyperthyroidism	180 (1.76%)	37 (0.36%)	5 (0.31%)	<5	8 (0.32%)	27 (0.74%)	19 (1.00%)	<5	5 (0.72%)	<5	0 (0.00%)	<5
Hypothyroidism	374 (3.66%)	73 (0.71%)	22 (1.36%)	30 (1.68%)	62 (2.46%)	238 (6.52%)	50 (2.64%)	64 (2.33%)	13 (1.87%)	<5	5 (0.91%)	<5
Ketoacidosis	455 (4.45%)	370 (3.62%)	5 (0.31%)	281 (15.71%)	19 (0.75%)	449 (12.30%)	35 (1.85%)	209 (7.60%)	0 (0.00%)	0 (0.00%)	11 (2.01%)	60 (9.76%)
Other autoimmune conditions*	44 (0.43%)	<5	<5	<5	10 (0.40%)	12 (0.33%)	11 (0.58%)	<5	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Overweight and obesity	719 (7.04%)	98 (0.96%)	10 (0.62%)	11 (0.61%)	161 (6.38%)	478 (13.09%)	53 (2.80%)	50 (1.82%)	19 (2.73%)	0 (0.00%)	17 (3.11%)	10 (1.63%)
Type 2 diabetes (any)	2,281 (22.32%)	715 (7.00%)	46 (2.85%)	72 (4.02%)	442 (17.53%)	346 (9.48%)	529 (27.93%)	327 (11.89%)	40 (5.76%)	52 (7.48%)	31 (5.67%)	17 (2.76%)
Medications												
Insulin	0 (0.00%)	2,271 (22.22%)	0 (0.00%)	1,340 (74.90%)	0 (0.00%)	1,966 (53.85%)	0 (0.00%)	977 (35.51%)	0 (0.00%)	111 (15.97%)	0 (0.00%)	522 (84.88%)
Glucose-lowering drugs	2,249 (22.01%)	216 (2.11%)	65 (4.03%)	65 (3.63%)	142 (5.63%)	268 (7.34%)	219 (11.56%)	169 (6.14%)	45 (6.47%)	51 (7.34%)	5 (0.91%)	23 (3.74%)

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=10,219		N=1,789		N=3,651		N=2,751		N=695		N=615	
	Assessment window											
	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]
(excluding insulin)												
Metformin	2,196 (21.49%)	210 (2.05%)	59 (3.66%)	49 (2.74%)	100 (3.97%)	154 (4.22%)	188 (9.93%)	116 (4.22%)	41 (5.90%)	45 (6.47%)	<5	18 (2.93%)
Dipeptidyl peptidase-4 inhibitors	681 (6.66%)	27 (0.26%)	18 (1.12%)	10 (0.56%)	47 (1.86%)	42 (1.15%)	63 (3.33%)	20 (0.73%)	7 (1.01%)	0 (0.00%)	<5	8 (1.30%)
Sodium-glucose cotransporter 2	464 (4.54%)	18 (0.18%)	10 (0.62%)	20 (1.12%)	<5	<5	43 (2.27%)	37 (1.34%)	5 (0.72%)	<5	0 (0.00%)	<5
Sulfonylureas	548 (5.36%)	6 (0.06%)	<5	0 (0.00%)	63 (2.50%)	54 (1.48%)	58 (3.06%)	18 (0.65%)	28 (4.03%)	11 (1.58%)	0 (0.00%)	0 (0.00%)
Teplizumab	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Verapamil	74 (0.72%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	10 (0.40%)	18 (0.49%)	<5	0 (0.00%)	<5	0 (0.00%)	<5	0 (0.00%)

All values are N (%), unless otherwise specified.
*Reported together: pernicious anaemia, Addison's disease, Autoimmune hepatitis.
CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information. N=Number of subjects. T1D=Type 1 diabetes mellitus. Type 1 diabetes mellitus was defined as the occurrence of both: first-ever condition occurrence of type 1 diabetes mellitus (SNOMED CT), AND first-ever prescription of insulin at the ingredient level (RxNorm), occurring within 180 days of each other. The index date was defined as the earliest of the two qualifying events. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

Measurements

Glycaemic measurements

The proportion of individuals with any recorded measurement varied across sources, ranging from 32.66% (IPCI) to 75.87% (H12O) in the any time prior window, and from 9.27% (SUCD) to 40.26% (CDW Bordeaux) at the index date (**Table 5**). Random glucose was the most consistently identified glycaemic test across data sources (prior to index date 20.72–74.77%; index date 8.91–39.55%). HbA1c was commonly recorded in DK-DHR and FinOMOP-TaUH (prior to index date 61.02% and 47.34%; at index date 27.97% and 25.10%, respectively) but was infrequently recorded in CDW Bordeaux (<1%). Fasting glucose recorded in FinOMOP-TaUH, IPCI, and DK-DHR (prior to index date 34.88%, 21.01%, and 11.45%).

BMI measurement

BMI measurement capture in the 180 days prior to index date exceeded 10% in CDW Bordeaux and H12O and was <7% in the remaining sources.

C-peptide and autoantibodies

C-peptide recording was heterogeneous, being highest in DK-DHR (prior to index date 21.56%; index date 12.44%), followed by CDW Bordeaux index date (8.93%); in all other data sources, frequency of C-peptide tests was ≤7% (**Table 4**). Recording of any autoantibody was available in most sources, except IPCI. Prior to index, the highest recording of any antibody measurement was most frequent in DK-DHR (prior to index date 16.24%), whereas at index date it was highest in FinOMOP-TaUH (28.06%); recording in all other sources was lower. GAD-65 largely mirrored overall autoantibody records; Recording of ICA was notable at the index date in FinOMOP-TaUH (26.66%), and IAA/ZnT8 were rarely recorded except in H12O. Across all sources, derived autoantibody positivity remained <5%. Stratified results by age group and sex are provided in the Shiny app.

Table 5. Measurements including autoantibody testing in Type 1 diabetes mellitus.

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=10,219		N=1,789		N=3,651		N=2,751		N=695		N=615	
	Assessment window											
	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]
Any Measurement (any of the measurements below)	6,883 (67.35%)	2,908 (28.46%)	1,123 (69.58%)	423 (23.64%)	975 (38.66%)	1,470 (40.26%)	825 (43.56%)	255 (9.27%)	227 (32.66%)	125 (17.99%)	415 (75.87%)	159 (25.85%)
Fasting glucose	1,170 (11.45%)	276 (2.70%)	563 (34.88%)	48 (2.68%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	146 (21.01%)	58 (8.35%)	0 (0.00%)	0 (0.00%)
Fasting glucose abnormal	691 (6.76%)	234 (2.29%)	323 (20.01%)	60 (3.35%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	85 (12.23%)	66 (9.50%)	0 (0.00%)	0 (0.00%)
Random glucose	6,603 (64.61%)	2,779 (27.19%)	835 (51.73%)	517 (28.90%)	872 (34.58%)	1,444 (39.55%)	716 (37.80%)	245 (8.91%)	144 (20.72%)	86 (12.37%)	409 (74.77%)	161 (26.18%)
Random glucose abnormal	4,682 (45.82%)	4,082 (39.95%)	342 (21.19%)	784 (43.82%)	118 (4.68%)	672 (18.41%)	369 (19.48%)	201 (7.31%)	37 (5.32%)	100 (14.39%)	261 (47.71%)	220 (35.77%)
Oral glucose tolerance test	1,081 (10.58%)	352 (3.44%)	551 (34.14%)	48 (2.68%)	633 (25.10%)	540 (14.79%)	38 (2.01%)	<5	144 (20.72%)	41 (5.90%)	386 (70.57%)	160 (26.02%)
Oral glucose tolerance test abnormal	481 (4.71%)	334 (3.27%)	190 (11.77%)	56 (3.13%)	46 (1.82%)	283 (7.75%)	9 (0.48%)	<5	58 (8.35%)	47 (6.76%)	243 (44.42%)	210 (34.15%)
HbA1c	6,236 (61.02%)	2,858 (27.97%)	764 (47.34%)	449 (25.10%)	23 (0.91%)	11 (0.30%)	311 (16.42%)	69 (2.51%)	81 (11.65%)	49 (7.05%)	197 (36.01%)	30 (4.88%)
HbA1c abnormal	4,933 (48.27%)	3,541 (34.65%)	493 (30.55%)	505 (28.23%)	15 (0.59%)	10 (0.27%)	247 (13.04%)	62 (2.25%)	61 (8.78%)	52 (7.48%)	164 (29.98%)	32 (5.20%)

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=10,219		N=1,789		N=3,651		N=2,751		N=695		N=615	
	Assessment window											
	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]
BMI measurement [-180, -1]	656 (6.42%)		0 (0.00%)		896 (35.53%)		0 (0.00%)		27 (3.88%)		70 (12.80%)	
C-peptide	2,203 (21.56%)	1,271 (12.44%)	123 (7.62%)	26 (1.45%)	8 (0.32%)	326 (8.93%)	130 (6.86%)	32 (1.16%)	<5	<5	40 (7.31%)	18 (2.93%)
C-peptide abnormal	0 (0.00%)	0 (0.00%)	29 (1.80%)	7 (0.39%)	0 (0.00%)	0 (0.00%)	21 (1.11%)	11 (0.40%)	0 (0.00%)	<5	17 (3.11%)	7 (1.14%)
Any autoantibody*	1,660 (16.24%)	1,999 (19.56%)	159 (9.85%)	502 (28.06%)	33 (1.31%)	57 (1.56%)	126 (6.65%)	<5	<5	0 (0.00%)	52 (9.51%)	0 (0.00%)
GAD-65	1,660 (16.24%)	1,999 (19.56%)	112 (6.94%)	36 (2.01%)	29 (1.15%)	40 (1.10%)	121 (6.39%)	<5	<5	0 (0.00%)	52 (9.51%)	0 (0.00%)
GAD-65 positive	0 (0.00%)	0 (0.00%)	71 (4.40%)	26 (1.45%)	0 (0.00%)	0 (0.00%)	73 (3.85%)	<5	0 (0.00%)	0 (0.00%)	25 (4.57%)	0 (0.00%)
Insulin autoantibodies (IAA)	0 (0.00%)	0 (0.00%)	9 (0.56%)	0 (0.00%)	5 (0.20%)	32 (0.88%)	40 (2.11%)	<5	0 (0.00%)	0 (0.00%)	50 (9.14%)	0 (0.00%)
Insulin autoantibodies (IAA) positive result	0 (0.00%)	0 (0.00%)	<5	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Islet cell autoantibodies (ICA)	0 (0.00%)	0 (0.00%)	68 (4.21%)	477 (26.66%)	16 (0.63%)	34 (0.93%)	109 (5.76%)	<5	0 (0.00%)	0 (0.00%)	52 (9.51%)	0 (0.00%)
Islet cell autoantibodies (ICA) positive result	0 (0.00%)	0 (0.00%)	20 (1.24%)	13 (0.73%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	9 (1.65%)	0 (0.00%)
ZnT8	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	12 (0.48%)	9 (0.25%)	20 (1.06%)	<5	0 (0.00%)	0 (0.00%)	42 (7.68%)	0 (0.00%)

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=10,219		N=1,789		N=3,651		N=2,751		N=695		N=615	
	Assessment window											
	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]	Any time prior [-inf, -1]	At index date [0, 0]
ZnT8 positive result	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.26%)	<5	0 (0.00%)	0 (0.00%)	<5	0 (0.00%)

All values are N (%), unless otherwise specified. *Any of the autoantibodies listed below. IA-2A not shown, as it was 0 or <5 counts in all data sources. Measurements described here represent "Earliest measurement" and "earliest abnormal measurement". Values were derived independently. "Earliest measurement" reflects the first-ever recorded measurement (any value), whereas "earliest abnormal (or positive result for antibodies)" reflects the first-ever abnormal result among individuals with ≥1 abnormal result (as defined by unit-specific thresholds/structured values). Therefore, counts/percentages for "earliest abnormal" are not a subset of "earliest measurement" within a given window and may be higher at index date when abnormality first occurs at index date after earlier (non-abnormal) testing. Absence of abnormal counts does not imply true absence of events (see limitations). CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information. N=Number of subjects. T1D=Type 1 diabetes mellitus. Type 1 diabetes mellitus was defined as the occurrence of both: first-ever condition occurrence of type 1 diabetes mellitus (SNOMED CT), AND first-ever prescription of insulin at the ingredient level (RxNorm), occurring within 180 days of each other. The index date was defined as the earliest of the two qualifying events. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

Index date and 1-year follow-up characterisation

Among individuals with ≥365 days of post-index date observation, recording of first-ever type 2 diabetes increased from 2.33–6.05% at index date to 4.06–12.17% during days 1–365 (highest in CDW Bordeaux: 12.17%) (**Table 6**).

Recording of glucose-lowering drugs excluding insulin also increased post-index date, from 3.42–8.05% at index date day to 10.33–23.61% during days 1–365 (highest in CDW Bordeaux: 23.61%, IPCI: 20.64%, DK-DHR: 19.70%), largely due to metformin, which increased from 2.44–6.88% at index date to 8.86–18.46% post-index date. Other classes were less common.

Occurrence of first-ever autoantibody testing varied by data source and generally increased during days 1–365. “Any autoantibody” testing was identified for 0–30.57% of individuals at index date and 2.85–62.22% of individuals during days 1–365, with particularly higher occurrence post-index date in H12O (62.22%), followed by DK-DHR (26.19%), FinOMOP-TaUH (25.77%), and SUCD (23.22%). GAD-65 resembled the “any autoantibody” patterns (e.g., DK-DHR at index date : 19.58% and post-index date 26.19%). Presence of other autoantibodies were data source-specific: testing for ICA was prominently recorded in FinOMOP-TaUH (29.09% at index date; 15.50% post-index date) and present post-index date in CDW Bordeaux and SUCD, while presence of IAA and ZnT8 tests were mainly observed post-index date in selected sources (notably H12O and SUCD). IA-2A was essentially absent across the data sources (only IPCI post-index date 2.18%).

Results by age group and sex are provided in the Shiny app.

Table 6. Index date and 1-year follow-up characterisation of Type 2 diabetes codes, Glucose-lowering Therapy, and Islet Autoantibody Testing in the Type 1 diabetes cohort.

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=8,942		N=1,626		N=2,448		N=2,011		N=596		N=487	
	Assessment window											
	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date
Type 1 Diabetes (with 365 days of observation post index date)												
Type 2 diabetes (any)	618 (6.91%)	2,143 (23.97%)	68 (4.18%)	134 (8.24%)	224 (9.15%)	589 (24.06%)	218 (10.84%)	433 (21.53%)	43 (7.21%)	72 (12.08%)	16 (3.29%)	62 (12.73%)
Type 2 diabetes (first-ever)	208 (2.33%)	730 (8.16%)	53 (3.26%)	66 (4.06%)	148 (6.05%)	298 (12.17%)	96 (4.77%)	174 (8.65%)	33 (5.54%)	32 (5.37%)	13 (2.67%)	38 (7.80%)
Glucose-lowering drugs (excluding insulin)	306 (3.42%)	1,762 (19.70%)	75 (4.61%)	168 (10.33%)	182 (7.43%)	578 (23.61%)	123 (6.12%)	350 (17.40%)	48 (8.05%)	123 (20.64%)	23 (4.72%)	53 (10.88%)
Glucose-lowering drugs (first-ever)	131 (1.46%)	674 (7.54%)	61 (3.75%)	109 (6.70%)	135 (5.51%)	369 (15.07%)	90 (4.48%)	233 (11.59%)	40 (6.71%)	70 (11.74%)	21 (4.31%)	40 (8.21%)
Metformin	218 (2.44%)	1,546 (17.29%)	54 (3.32%)	144 (8.86%)	101 (4.13%)	371 (15.16%)	81 (4.03%)	271 (13.48%)	41 (6.88%)	110 (18.46%)	18 (3.70%)	43 (8.83%)
Dipeptidyl peptidase-4 inhibitors	29 (0.32%)	263 (2.94%)	10 (0.62%)	23 (1.41%)	28 (1.14%)	125 (5.11%)	16 (0.80%)	52 (2.59%)	0 (0.00%)	5 (0.84%)	8 (1.64%)	5 (1.03%)
Sodium-glucose cotransporter 2	21 (0.23%)	314 (3.51%)	19 (1.17%)	43 (2.64%)	0 (0.00%)	35 (1.43%)	24 (1.19%)	78 (3.88%)	<5	<5	<5	9 (1.85%)
Sulfonylureas	9 (0.10%)	81 (0.91%)	0 (0.00%)	0 (0.00%)	43 (1.76%)	138 (5.64%)	12 (0.60%)	47 (2.34%)	10 (1.68%)	55 (9.23%)	0 (0.00%)	<5
Any autoantibody*	1,751 (19.58%)	2,342 (26.19%)	497 (30.57%)	419 (25.77%)	40 (1.63%)	371 (15.16%)	0 (0.00%)	467 (23.22%)	0 (0.00%)	17 (2.85%)	0 (0.00%)	303 (62.22%)

Variable level	Data source											
	DK-DHR		FinOMOP-TaUH		CDW Bordeaux		SUCD		IPCI		H12O	
	N=8,942		N=1,626		N=2,448		N=2,011		N=596		N=487	
	Assessment window											
	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date	Index date	1-year post index date
GAD-65	1,751 (19.58%)	2,342 (26.19%)	35 (2.15%)	403 (24.78%)	26 (1.06%)	281 (11.48%)	0 (0.00%)	453 (22.53%)	0 (0.00%)	17 (2.85%)	0 (0.00%)	301 (61.81%)
IA-2A	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (2.18%)	0 (0.00%)	0 (0.00%)
Insulin autoantibodies (IAA)	0 (0.00%)	0 (0.00%)	0 (0.00%)	<5	28 (1.14%)	207 (8.46%)	0 (0.00%)	239 (11.88%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	298 (61.19%)
Islet cell autoantibodies (ICA)	0 (0.00%)	0 (0.00%)	473 (29.09%)	252 (15.50%)	28 (1.14%)	245 (10.01%)	0 (0.00%)	336 (16.71%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	301 (61.81%)
ZnT8	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (0.20%)	56 (2.29%)	0 (0.00%)	127 (6.32%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	294 (60.37%)

All values are N (%), unless otherwise specified. *Any of the autoantibodies below. CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information. N=Number of subjects. T1D=Type 1 diabetes mellitus. Type 1 diabetes mellitus was defined as the occurrence of both: first-ever condition occurrence of type 1 diabetes mellitus (SNOMED CT), AND first-ever prescription of insulin at the ingredient level (RxNorm), occurring within 180 days of each other. The index date was defined as the earliest of the two qualifying events. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

9.2.2. Time from the first-ever and first abnormal measurement to index date

Glycaemic measurements

The median time from the earliest random glucose test to index date ranged from approximately −730 to −3,230 days (DK-DKR and FinOMOP-TaUH, respectively) (**Table 6**). In contrast, the earliest abnormal random glucose test generally occurred close to index date, with medians ranging from −4 to −134 days, except for FinOMOP-TaUH, which showed a longer interval (−1,736 days).

Median time from earliest HbA1c measurement to index date ranged from 210 days (CDW Bordeaux) to 1,936 days (FinOMOP-TaUH); earliest abnormal HbA1c measurement was generally recorded near index date in DK-DHR (median 22 days) and IPCI (48 days), though longer in the remaining data sources (range 1,840 days in FinOMOP-TaUH to 372 days in CDW Bordeaux).

Fasting glucose, where present, ranged from 970 days (IPCI) to 3669 days (FinOMOP-TaUH); the first abnormal fasting glucose tended to occur closer to index date (DK-DHR median 329 days; IPCI 166 days), but was recorded earlier in FinOMOP-TaUH (median 2,425 days).

Median time from first-ever OGTT to diagnosis ranged from 525 days (DK-DHR) to 3,304 days (FinOMOP-TaUH). First abnormal OGTT generally clustered near index date in DK-DHR, CDW Bordeaux, IPCI, and H12O (medians around 6 to 57 days), but was considerably longer in the remaining data sources (above 1000 days).

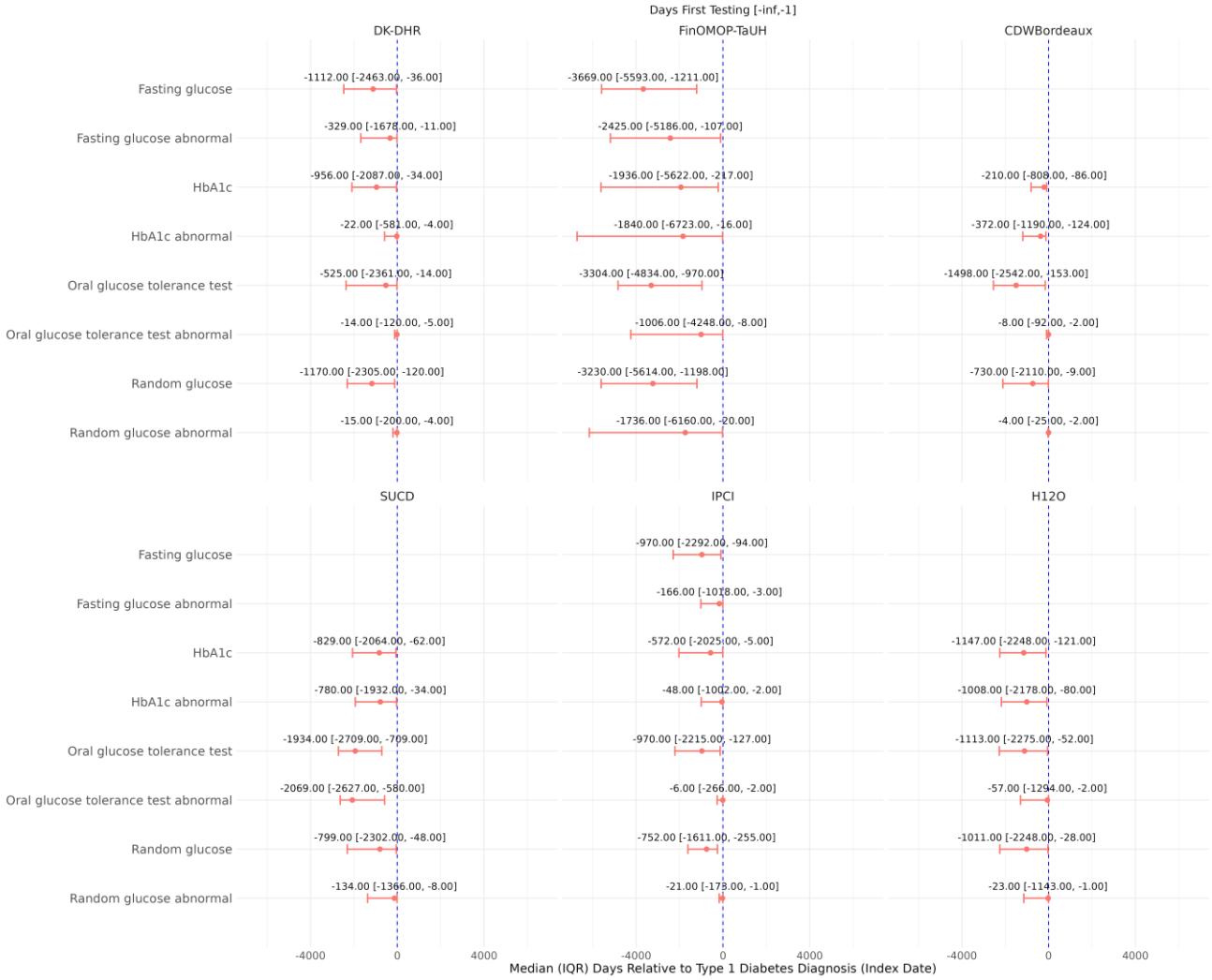
C-peptide and autoantibodies

In DK-DHR and CDW Bordeaux, the first C-peptide measurement occurred very near diagnosis (median 11 and 38 days, respectively), while in other sources, testing occurred at longer intervals prior to index date (range: 1,364 to 342 days, in FinOMOP-TaUH and SUCD, respectively). Abnormal C-peptide was rarely recorded (0% in several sources), but when present, timing varied by data source (range: 3,161 to 174 days, from first abnormal test to index date in FinOMOP-TaUH and SUCD, respectively).

For autoantibodies, GAD-65 first tests typically occurring close to index date in DK-DHR (median 10 days) but substantially earlier prior to index date in H12O (median 686 days). Recordings of other autoantibody tests (IAA, ICA, ZnT8) were sparse. Where present, medians for first tests ranged from near-index date (e.g., CDW Bordeaux IAA median 6 days) to multiple years prior to index date (e.g., H12O IAA median 714 days).

Results by age group are provided in the Shiny app.

a) *Glycaemic measurements*



b) C-peptide and autoantibodies

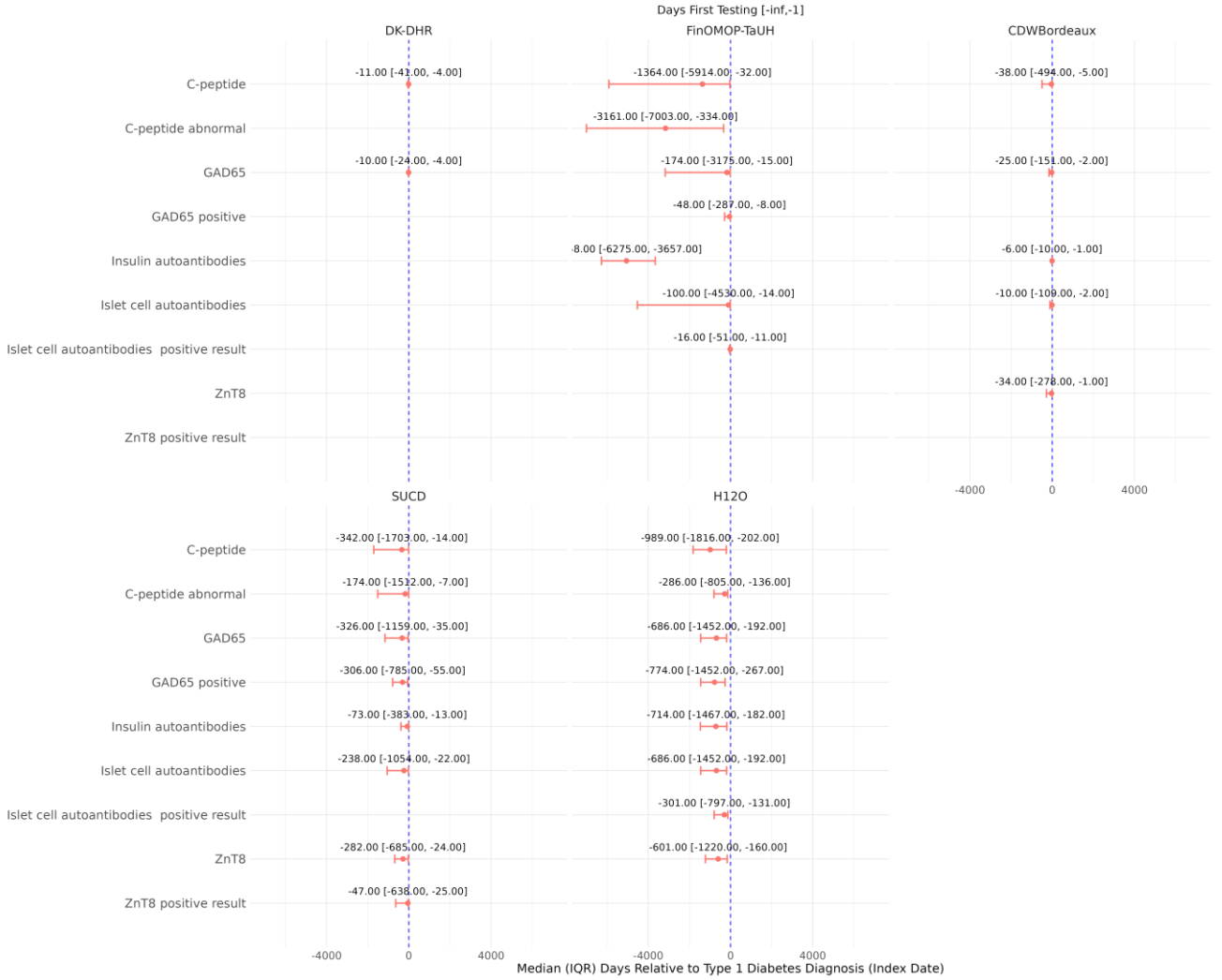


Figure 5. Time from First-Ever measurement of interest to Type 1 Diabetes Mellitus Diagnosis: First Recorded and First Abnormal Tests for a) Glycaemic measurements, and b) C-peptide and autoantibodies.

The x-axis shows the median and IQR of days from the first recorded test to the formal type 1 diabetes diagnosis (index date, Day 0). The dashed blue vertical line marks the index date. IA-2A results not shown, as they were 0 or <5 in all sources. CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information. N=Number of subjects. T1D=Type 1 diabetes mellitus. Type 1 diabetes mellitus was defined as the occurrence of both: first-ever condition occurrence of type 1 diabetes mellitus (SNOMED CT), AND first-ever prescription of insulin at the ingredient level (RxNorm), occurring within 180 days of each other. The index date was defined as the earliest of the two qualifying events. In population-based data sources, a minimum of 365 days of observation prior to index date was required. In panel B, the IPCI data source is omitted because all values were zero; and “positive” insulin autoantibodies are omitted for all data sources because values were zero across sources.

Earliest and earliest abnormal measurement values

Earliest and earliest abnormal measurement values for each measurement of interest were summarised by test and unit concept (available in the Shiny app). For glycaemic measures, values were generally within clinically plausible ranges, and medians in the abnormal strata were higher than in the corresponding non-abnormal strata, as expected, given the unit-specific value thresholds used to define abnormality.
For example, random glucose values were primarily recorded in mmol/L in DK-DHR, FinOMOP-TaUH, and SUCD (first-ever medians 5.20–8.00; first-ever abnormal medians 12.80–16.00, across the specified data sources), while capture in mg/dL was observed in CDW Bordeaux and H12O (first-ever medians 166.00 in CDW Bordeaux and 142.50 in H12O; first-ever abnormal medians 219.60 in FinOMOP-TaUH and 269.00 in H12O).

Across tests, unit capture and mapping varied across data sources. Examples include glucose records with missing or unmapped unit concepts, or “no matching concept” (available in the ShinyApp, e.g., random glucose values in CDW Bordeaux recorded under an unspecified unit category with medians 133.5 and 211, consistent with mg/dL-scale values). For C-peptide, measurement values were captured using multiple unit systems (e.g., pmol/L, nmol/L, ng/mL) with plausible medians in sources with available data. For GAD-65, values were recorded under heterogeneous unit concepts (e.g., IU/mL, kU/L) with wide distributions.

9.2.3. Annual point prevalence

Point prevalence estimates were only assessed in population-based data sources. During the study period, the number of individuals with type 1 diabetes totalled 68,332 in DK-DHR and 4,408 in IPCI.

Annual point prevalence differed markedly between data sources and showed distinct temporal patterns (**Figure 6, Table 7**).

In DK-DHR, point prevalence declined slightly over time, from 89.80 per 10,000 (95% CI 89.10–90.60) in 2015 to 81.80 per 10,000 (81.10–82.50) in 2024.

In IPCI, point prevalence was substantially lower and increased slightly between 2015 and 2022, rising from 15.60 per 10,000 (14.70–16.50) in 2015 to 19.30 per 10,000 (18.50–20.10) in 2022. It then decreased slightly in 2023 (16.70 per 10,000 [16.00–17.50]) and remained similar in 2024 (17.20 per 10,000 [16.50–18]).

Results by age group and sex are provided in the Shiny app and **Figures S1** and **S2**.

Table 7. Annual point prevalence of Type 1 diabetes on 1 January (%, 95% CI) by data source (DK-DHR and IPCI), 2015–2024.

Prevalence per 10,000						
Point Prevalence Date	Data source name					
	DK-DHR			IPCI		
	Estimate name					
Type 1 diabetes	Denominator (N)	Outcome: Type 1 diabetes (N)	Prevalence [95% CI]	Denominator (N)	Outcome: Type 1 diabetes (N)	Prevalence [95% CI]
01/01/2015	5,567,493	50,015	89.80 (89.10 – 90.60)	756,737	1,180	15.60 (14.70 – 16.50)
01/01/2016	5,601,185	50,230	89.70 (88.90 – 90.50)	922,444	1,580	17.10 (16.30 – 18.00)
01/01/2017	5,645,317	50,178	88.90 (88.10 – 89.70)	1,083,599	1,812	16.70 (16.00 – 17.50)
01/01/2018	5,681,835	49,765	87.60 (86.80 – 88.40)	1,114,128	1,939	17.40 (16.60 – 18.20)
01/01/2019	5,710,021	49,395	86.50 (85.70 – 87.30)	1,152,452	2,104	18.30 (17.50 – 19.10)
01/01/2020	5,735,841	49,130	85.70 (84.90 – 86.40)	1,174,115	2,235	19.00 (18.30 – 19.80)
01/01/2021	5,766,149	48,842	84.70 (84.00 – 85.50)	1,230,866	2,356	19.10 (18.40 – 19.90)
01/01/2022	5,792,793	48,612	83.90 (83.20 – 84.70)	1,262,247	2,434	19.30 (18.50 – 20.10)
01/01/2023	5,816,346	48,255	83.00 (82.20 – 83.70)	1,139,514	1,903	16.70 (16.00 – 17.50)
01/01/2024	5,866,791	47,992	81.80 (81.10 – 82.50)	1,185,002	2,043	17.20 (16.50 – 18.00)

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information; N=Number of subjects; CI=Confidence interval. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

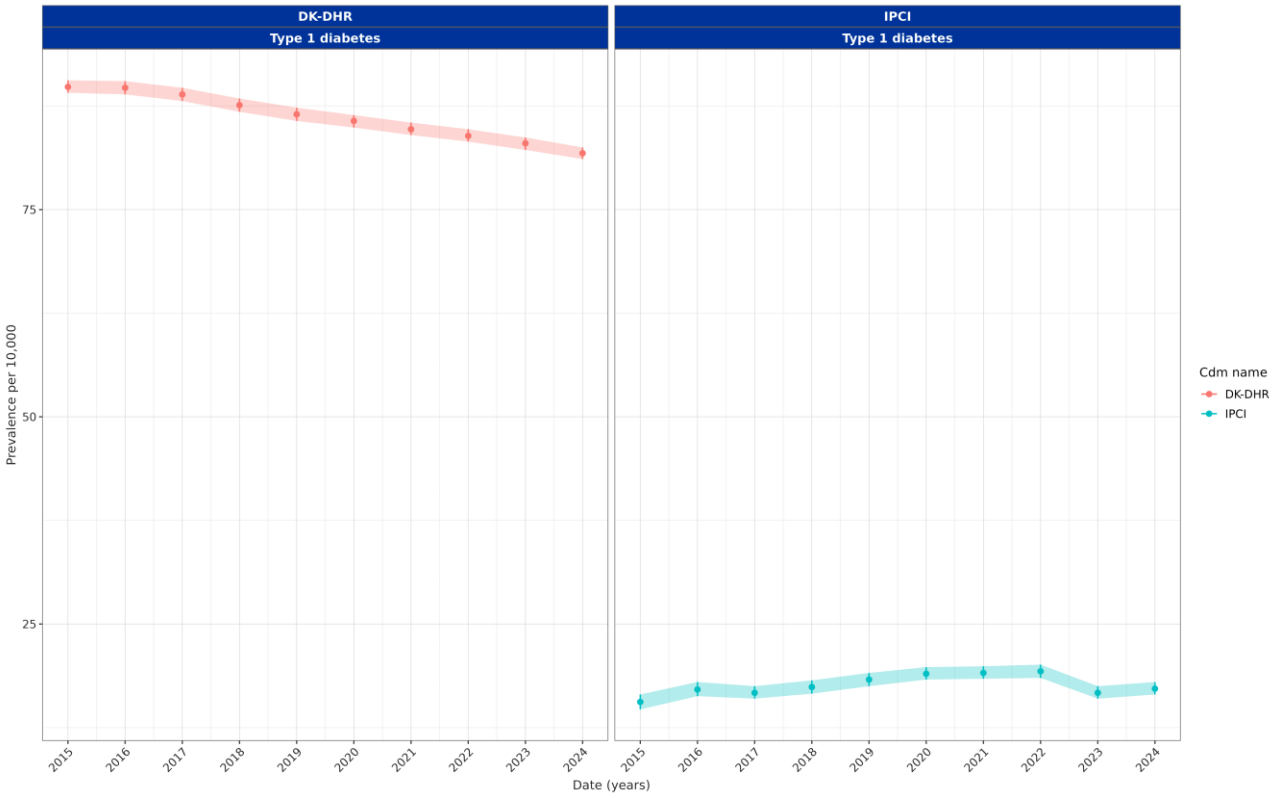


Figure 6. Temporal trends of annual point prevalence of Type 1 diabetes on 1 January by data source (DK-DHR and IPCI), 2015–2024.

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

Sensitivity analysis

When individuals with any prior type 2 diabetes in history were excluded, the number of individuals with type 1 diabetes decreased to 40,354 in DK-DHR and 3,876 in IPCI. Compared with the primary analysis, point prevalence estimates were lower in both DK-DHR and IPCI across all years after excluding individuals with any prior history of type 2 diabetes (**Figure 7, Table 8**). The overall temporal patterns were more stable in DK-DHR; they were generally similar for IPCI (slightly increasing to 2022, then declining).

Figures S3 and S4 display results by age group and sex. Tables are available in the ShinyApp.
Age-stratified analyses in the sensitivity analysis showed heterogeneous trends. In DK-DHR, prevalence increased in younger age groups and declined in ages 40–49 and ≥50 years; in IPCI, prevalence generally increased over time in age groups <40 years, while trends in older age groups were more variable.

Table 8. Sensitivity Analysis: Annual point prevalence of Type 1 diabetes on 1 January, excluding individuals with any prior history of Type 2 diabetes (%, 95% CI), by data source (Dk-DHR and IPCI), 2015–2024.

Prevalence per 10,000						
Point Prevalence Date	Data source name					
	DK-DHR			IPCI		
	Estimate name					
	Denominator (N)	Outcome (N)	Prevalence [95% CI]	Denominator (N)	Outcome (N)	Prevalence [95% CI]
Type 1 diabetes						
01/01/2015	5,567,493	30,731	55.20 (54.60 – 55.80)	756,737	1,061	14.00 (13.20 – 14.90)
01/01/2016	5,601,185	30,841	55.10 (54.50 – 55.70)	922,444	1,438	15.60 (14.80 – 16.40)
01/01/2017	5,645,317	30,944	54.80 (54.20 – 55.40)	1,083,599	1,647	15.20 (14.50 – 16.00)
01/01/2018	5,681,835	31,060	54.70 (54.10 – 55.30)	1,114,128	1,752	15.70 (15.00 – 16.50)
01/01/2019	5,710,021	31,154	54.60 (54.00 – 55.20)	1,152,452	1,890	16.40 (15.70 – 17.20)
01/01/2020	5,735,841	31,284	54.50 (53.90 – 55.10)	1,174,115	1,995	17.00 (16.30 – 17.80)
01/01/2021	5,766,149	31,423	54.50 (53.90 – 55.10)	1,230,866	2,097	17.00 (16.30 – 17.80)
01/01/2022	5,792,793	31,739	54.80 (54.20 – 55.40)	1,262,247	2,158	17.10 (16.40 – 17.80)
01/01/2023	5,816,346	31,897	54.80 (54.20 – 55.40)	1,139,514	1,678	14.70 (14.00 – 15.40)
01/01/2024	5,866,791	32,097	54.70 (54.10 – 55.30)	1,185,002	1,814	15.30 (14.60 – 16.00)

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

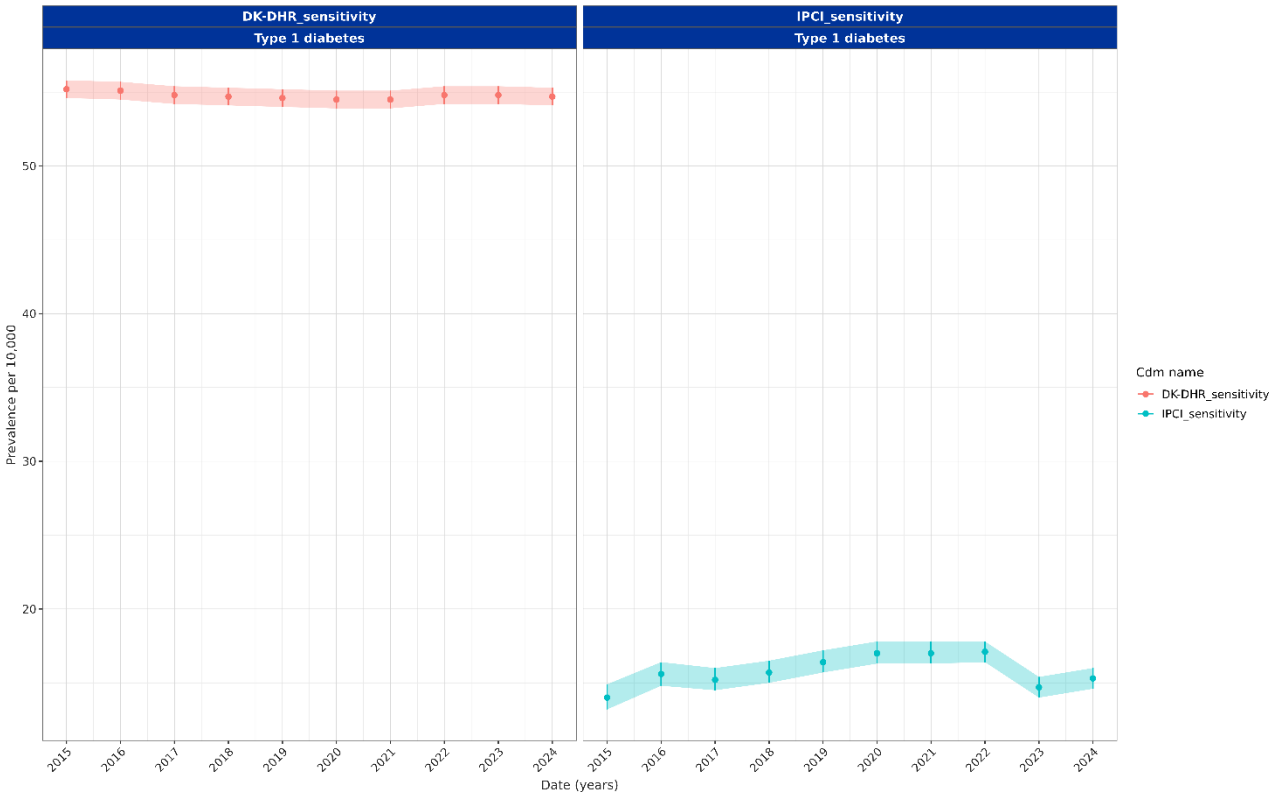


Figure 7. Temporal trends of annual point prevalence of type 1 diabetes on 1 January, Excluding Individuals with Any Prior History of Type 2 Diabetes, by Data Source (DK-DHR and IPCI), 2015–2024.

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

10. DISCUSSION

10.1. Key results

The total number of individuals diagnosed with first-ever type 1 diabetes ranged between 615 and 3,651 across hospital-based data sources and between 695 and 10,219 in population-based data sources. Given the study focus, we first highlight measurement availability below; detailed results are presented in **Section 9** (cohort characterisation first, then measurements).

Measurements

Glycaemic measurements

Regarding measurements, occurrence of “Any measurement” in individuals with type 1 diabetes ranged between 33% and 76% prior to index date and 9% and 40% at index date. Random glucose was the most consistently identified glycaemic test, followed by HbA1c and OGTT, which showed substantial variation across sources. Fasting glucose was source dependent, being available in DK-DHR, FinOMOP-TaUH, and IPCI.

C-peptide and autoantibody measurements

Autoantibody test presence was recorded in most sources (except IPCI, <5), mostly reflecting GAD-65. Availability of other antibodies was more limited and source-specific (IAA and ICA were available in FinOMOP-TaUH, CDW Bordeaux, SUCD, and H12O; ZnT8 in CDW Bordeaux, SUCD, and H12O; IA-2A was ≤5 or 0 in all sources). At any time prior to index date, proportion of individuals with “Any autoantibody

measurement" across data sources ranged between 1% and 16%. Positive results (abnormal) were uncommon (<5%) across data sources.

Time from the earliest and earliest abnormal measurement to index date

Across sources, the earliest recorded glycaemic test (random glucose and often HbA1c/OGTT), often occurred years before the first diagnosis of type 1 diabetes, with median lags ranging between 2–9 years in several data sources. The earliest abnormal results were generally recorded closer to the index date (often days to a few months prior), although timing varied by source.

Measurement values

For glycaemic measures, values were generally within clinically plausible ranges, and medians in the abnormal strata were higher than in the corresponding non-abnormal strata, as expected, given the unit-specific value thresholds used to define abnormality. Across tests, unit capture and mapping were heterogeneous in some sources.

Characterisation

Median age at index date ranged from 19 to 56 years. Sex was male-predominant (male range across data sources: 53–60%), except for SUCD, where it was equal.

Prior to index date, the prevalence of hypertension (range: 2–30%; FinOMOP-TaUH, SUCD) and overweight/obesity (range: 1–7%; FinOMOP-TaUH, DK-DHR) in the type 1 diabetes cohort showed wide variability across sources. Ketoacidosis was observed in all data sources except IPCI. At index date, proportions of individuals with ketoacidosis ranged between 4% (DK-DHR) and 16% (FinOMOP-TaUH), and in all but DK-DHR, proportion of individuals diagnosed with ketoacidosis at index date was higher than at any time prior to index date, consistent with acute presentation around cohort entry. Prior history of type 2 diabetes (presence of codes "any time prior to index date ") ranged between 2.9% (FinOMOP-TaUH) and 27.9% (SUCD), with values >17% in CDW Bordeaux, DK-DHR, and SUCD.

At the index date, 16%–85% of individuals were treated with insulin; this proportion was higher in hospital-based data sources (>25%). Non-insulin glucose-lowering therapy was recorded for 1%–22% of individuals, predominantly metformin.

Index date and 1-year follow-up characterisation

When describing characteristics post-index date (1–365 days) in those individuals with ≥365 days follow-up, presence of a diagnosis code for type 2 diabetes (first-ever record) increased from 2–6% at index date day to 4–12% during days 1–365. Treatment with non-insulin glucose-lowering drugs rose from in 3–8% of individuals at index date day to 10–24% post-index date, driven largely by metformin. First-ever autoantibody testing was frequent both at index date and during days 1–365 post-index date, in all data sources except IPCI. In some sources (e.g., FinOMOP-TaUH, H12O), ≥50% of first-ever testing occurred during these windows.

Type 1 diabetes prevalence

In the primary analysis, annual point prevalence (2015–2024) declined steadily from 89.80 per 10,000 (89.10–90.60) in 2015 to 81.80 per 10,000 (81.10–82.50) in 2024 in DK-DHR, whereas in IPCI, it increased from 15.60 per 10,000 (14.70–16.50) in 2015 to 19.30 per 10,000 (18.50–20.10) in 2022, then decreased in 2023–2024 (16.70 per 10,000 in 2023; 17.20 per 10,000 in 2024).

In the sensitivity analysis excluding individuals with any prior type 2 diabetes history, point prevalence was lower in both sources across all years. Temporal patterns changed slightly in DK-DHR (remaining stable), while remaining similar in IPCI. Age-stratified analyses in the sensitivity analysis showed heterogeneous trends. In DK-DHR, prevalence increased in younger age groups and declined in ages 40–49 and ≥50 years;

in IPCI, prevalence generally increased over time in age groups <40 years, while trends in older age groups were more variable.

The sensitivity analysis reduced absolute prevalence estimates, suggesting that the inclusion of individuals with prior type 2 diabetes contributed to higher point-prevalence estimates in the primary analysis.

10.2. Strengths and limitations of the research methods

Strengths

This study leveraged six routinely collected data sources spanning hospital/secondary care, primary care, and registry-linked data, including population-based sources (DK-DHR and IPCI), to identify and characterise individuals with type 1 diabetes diagnosis across care settings. Clinical characterisation used prespecified assessment windows (any time prior vs index date; and index date vs days 1–365 in individuals with ≥365 days follow-up), allowing clear separation of baseline history, diagnostic-period recording, and early follow-up. The report also provides detailed data on the availability and timing of glycaemic and autoantibody testing, including time from earliest and earliest abnormal tests to index date, informing the feasibility of future studies that depend on laboratory confirmation. Finally, point-prevalence estimates were complemented by a sensitivity analysis excluding individuals with any prior history of type 2 diabetes to evaluate the influence of potential phenotype overlapping on point-prevalence estimates.

Limitations

Findings should be interpreted considering data source-specific constraints.

Type 1 diabetes cohort identification

Cohort yield and attrition were influenced by the requirement for concordant recording of type 1 diabetes diagnosis and insulin exposure within 180 days and by restricting entry to first-ever events in all available history occurring within the study period. Both are sensitive to differences in outpatient prescribing capture, prescription vs dispensing representation, linkage across care settings, and historical observation length, which may lead to variable phenotype sensitivity across data sources.

Type 2 diabetes

The presence of type 2 diabetes codes and non-insulin glucose-lowering therapy (more frequent in older adults) suggests potential phenotype overlap or misclassification; however, the degree of misclassification could not be assessed.

Measurements

Measurement capture differed across data sources, which limited cross-data sources comparability and the interpretation of timing metrics; these metrics are also sensitive to left truncation and differences in available lookback. For all measurements, abnormal/positive results were derived using prespecified, measurement-specific threshold values stratified by unit and structured categorical interpretations (e.g., positive/negative). This approach performed well for glycaemic tests, where value/unit representation was generally consistent. In contrast, for C-peptide and autoantibodies, identification of positive results was low despite the ability to identify test occurrence in several sources. This pattern may reflect true low positivity in this cohort; however, it may also indicate under-ascertainment of positivity due to differences in result representation across data sources. Importantly, the absence of identified abnormal/positive results should be interpreted as “no positive detected in the available structured data” rather than confirmed negative testing. Contributing limitations included missing unit information for a subset of records (limiting the application of unit-dependent thresholds) and heterogeneity on how categorical result concepts (e.g., “positive”) were recorded in some implementations (e.g., use of other labels such as ‘reactive/detected’, or free-text entries that may not be mapped to standard “positive” concepts), which may have reduced the sensitivity of detecting positivity. Participating data sources confirmed this heterogeneity: some reported

limited or no availability of structured autoantibody results, and among those reporting availability, units and positivity definitions were not consistently specified (**Table S3**).

Prevalence estimates

Point-prevalence estimates are inherently source-specific and depend on population coverage and denominator definitions. For example, in primary care-based sources (IPCI), a diagnosis of type 1 diabetes is expected to be recorded, even when management occurs in specialist/secondary care; however, certain elements (e.g., some prescriptions, encounter details) may be incompletely represented and may influence prevalence and trends.

10.3. Interpretation

Overall, this multi-data source study demonstrates that identification and characterisation of type 1 diabetes in routinely collected health data are feasible but challenging and dependent on data source-specific records of diagnoses, medications, and measurements. Hospital data sources showed higher availability of specialised testing, while nationwide registries with primary and secondary care linkage (DK-DHR) may be better suited for longitudinal characterisation. Primary care data sources without hospital/specialist linkage are performing least well, as they are affected by care fragmentation and incomplete recording. Across sources, we observed high misclassification or initial misdiagnosis occurred between type 1 and 2 diabetes-type (especially in older adults) (classification uncertainty). This overlap should be considered when defining cohorts, particularly in studies extending beyond juvenile diabetes, and supports using longitudinal evidence beyond a single diagnosis code. In addition, heterogeneity in laboratory results, particularly regarding autoantibody positivity, may limit the feasibility of future analyses requiring antibody confirmation.

Type 1 diabetes phenotype

Phenotype performance varied across data sources. This was expected given the cohort definition required (i) concordant evidence of type 1 diabetes diagnosis and insulin exposure within 180 days and (ii) cohort entry based on the first-ever qualifying event in all available history, occurring within the study period. Both criteria are sensitive to how insulin exposure is captured (inpatient vs. outpatient coverage), how diagnoses are recorded (primary care vs. specialist vs. discharge coding), and the length of historical observation. In hospital-based data, type 1 diabetes is often documented around acute presentation and/or specialist care, where insulin is initiated and recorded in the same care setting, increasing the likelihood that insulin exposure is captured within the 180-day window. This is consistent with higher recording of insulin at the index date in hospital-based data sources (range: 36–86%) compared with primary care and registry sources (range: 16–22%). In primary care, both the diagnosis and insulin initiation may occur in hospital/secondary (specialist) care and only later be reflected in the primary care record; moreover, early insulin prescriptions may be incompletely captured when hospital prescribing is not included in the source. As a result, concordant recording of diagnosis and insulin may be delayed beyond the window. Consequently, variation in phenotype yield may reflect both differences in capture and timing, and between-data source comparisons should be interpreted cautiously.

Older ages observed in some hospital-based data sources may reflect that cohort entry captures the first observable type 1 diabetes-related complication/hospital contact rather than the true onset.

Overall, cohort characterisation was broadly consistent with expected clinical features of type 1 diabetes.[10-13]

Diabetes-type overlap

Recording of type 2 diabetes codes and non-insulin glucose-lowering therapy before and after cohort entry was observed in several sources and increased with age, particularly among individuals aged ≥50 years. This pattern is compatible with well-described challenges of distinguishing between type 1 and type 2 diabetes,

especially with later onset, using structured electronic healthcare data alone, as well as the potential for longitudinal overlap or reclassification of recorded diabetes type in routine care, particularly among adults.[13] On the other hand, the coexistence of both conditions, a state known as double diabetes, could affect up to one in four adults with type 1 diabetes.[14] Misclassification in adults may be substantial; epidemiological and simulation studies suggest that up to 40% of adults older than 30 years with type 1 diabetes may initially be diagnosed as type 2.[12] Consistent with this, data from UK BioBank suggest that in individuals with high genetic susceptibility for type 1 diabetes, diagnoses can take place across the first six decades of life, with ~40% of diagnoses occurring above age 30.[15] In addition, phenotype overlap may reflect heterogeneity in insulin-treated diabetes, including monogenic diabetes, such as Maturity-Onset Diabetes of the Young (MODY) or youth-onset type 2 diabetes misclassified as type 1 diabetes, and adult-onset autoimmune diabetes (LADA), initially recorded as type 2 diabetes. Non-insulin agents may also be used among insulin-treated individuals in some settings, which contributes to the overlap (e.g., adjunct metformin has been evaluated in adults with type 1 diabetes and may be used off-label).[16]

In the present study, no validation of the recorded diagnoses was undertaken and therefore, could not quantify true misclassification. However, the proportion of individuals in the type 1 diabetes cohort who also had recorded type 2 diabetes codes and/or non-insulin glucose-lowering therapy provides an empirical measure of diabetes-type overlap and classification uncertainty in these data. As mentioned previously, this overlap may reflect multiple clinical situations (dual diabetes, changes in diabetes type over time, and/or misclassification), which are not necessarily a data quality issue, and should be considered when interpreting data source-derived type 1 diabetes cohorts.

Based on these results, no single definition is optimal across all settings and purposes. For studies prioritising specificity, we recommend using a minimum of two-domain concordant phenotype (e.g., type 1 diagnosis in addition to insulin exposure within 180 days), consistent with multi-signal EHR phenotyping approaches, to which persistence over time could be added (e.g., repeat insulin records within 6–12 months post-index date).[17, 18] However, in hospital-based sources, particularly in adults, ‘first-ever’ qualifying events may reflect first captured complication/contact rather than true onset; disentangling incident onset from first recorded acute presentation is challenging in routine data and should be acknowledged. Use sensitivity analyses varying lookback requirements and confirmation rules should be explored.

For population prevalence, excluding prior type 2 diabetes was informative, especially in DK-DHR, where estimates were closer to external general-population figures (contextualised below under the *Point prevalence trends* subheading); additional exclusions (e.g., non-insulin glucose-lowering therapy, no type 2 diabetes in the year following index date) may increase specificity but could also exclude clinically plausible adult-onset autoimmune diabetes and reclassification pathways, so these should be treated as sensitivity specifications rather than defaults. In primary care sources, where specialist diagnoses and insulin initiation may be incompletely captured or returned to the GP record, relaxing concordance requirements (e.g., longer diagnosis–insulin windows, alternative index date rules, or persistence-based confirmation) may improve yield, though gains will be limited by availability of linkage across settings and capture.

Measurement availability and timing

Measurement records for glycaemic and autoantibody testing varied markedly across data sources and antibody testing, especially as it was not very well recorded.

In terms of specific tests, random glucose was the most frequently recorded measurement. HbA1c and OGTT were also available in all sources, but with substantial variation across sources; whereas fasting glucose was source dependent, being available in DK-DHR, FinOMOP-TaUH, and IPCI. Timing analyses showed that first-ever glycaemic testing often preceded recorded type 1 diabetes diagnosis by years, consistent with background routine testing in general healthcare, while first abnormal results tended to occur closer to diagnosis, consistent with intensified evaluation around clinical presentation. However,

these timing metrics are sensitive to historical data depth, left truncation, and local measurement capture practices; accordingly, they should be interpreted as reflecting recording processes as well as clinical pathways. These metrics reflect both clinical pathways and data capture processes, not clinical behaviour alone. For example, FinOMOP-TaUH, consistently showed earlier median timing for nearly all tests (>1900 days for all glycaemic measurements, and some antibodies). This pattern may reflect longer observable longitudinal history or a more complete laboratory capture in this source, as well as underlying clinical pathways, including source-country context (e.g., high background type 1 diabetes incidence and associated diagnostic awareness).[19] However, for glycaemic measurements, most tests and sources also showed multi-year prior to index date testing, compatible with longitudinal capture (e.g., HbA1c −956 in DK-DHR and −1147 in H12O; random glucose −1170 in DK-DHR and −1011 in H12O).

The proportion of individuals with any autoantibody testing prior to index date ranged between 1–16%. This proportion is difficult to benchmark against published studies because our cohort was designed to capture clinical type 1 diabetes diagnosis (stage 3, insulin dependence) in routine healthcare data rather than a screened high-risk surveillance cohort, in settings where population-wide pre-symptomatic autoantibody screening is not routinely implemented. Nevertheless, this indicates limited documented pre-index autoantibody testing and/or incomplete capture of testing in structured data. Although autoantibody test occurrence was captured in some sources, standardised identification of “positive” (abnormal) autoantibody records was limited using the prespecified positivity definition, suggesting current limitations in harmonisation of laboratory results (e.g., records in free text or use of less common units, and limitations in applying test-specific value-unit thresholds consistently across sources) (**Table S3**). This limits analyses requiring autoimmune confirmation or staging.[2] For autoantibody confirmation or staging (based on “positive”/abnormal results) to be used for future studies, further mapping or standardisation is likely required. With regards to the timing of the earliest autoantibody testing before diagnosis varied markedly: in DK-DHR, it occurred close to diagnosis (median ~10 days prior to index date), whereas in FinOMOP-TaUH, SUCD, and H12O, the first recorded test was often months to years prior to index date (median ~0.5–2 years). These longer prior to index date intervals may reflect differences in historical data length and laboratory capture and could be compatible with earlier risk-based evaluation in subsets of individuals; however, they should not be interpreted as evidence of systematic screening without additional context on local testing pathways and programmatic screening.

Routine population-wide autoantibody screening is not yet implemented as standard clinical practice in most settings, but risk-based approaches are increasingly endorsed. For example, the American Diabetes Association 2025 Standards of Care emphasise antibody-based screening for presymptomatic type 1 diabetes among individuals with a family history or known genetic risk.[20] However, only a minority of individuals who develop type 1 diabetes have an affected first-degree relative (~10%), so family-history–based screening alone would miss most future cases.[21] In parallel, Italy has established (by law) a nationwide paediatric program to identify type 1 diabetes in the general paediatric population.[22] Other population-based paediatric screening programs are currently being evaluated in Europe, including Germany’s primary-care-based Fr1da initiative.[23]

A substantial proportion of the earliest (first-ever) autoantibody testing was recorded at index date or in the 1–365 days post-index date in some sources. This timing is consistent with guideline-based practice in adults initially classified as having type 1 diabetes, where diabetes-specific autoantibodies can support classification. National Institute for Health and Care Excellence (NICE) guidelines note that false-negative rates are lowest close to diagnosis and may be further reduced by testing two different autoantibodies.[24]

Point prevalence trends

Point prevalence estimates differed by source, consistent with differences in underlying population coverage (e.g., national-registry-based vs regional-primary-care-based data sources) and differential

representation of specialist diabetes care. For context, published estimates in Denmark report general-population T1D prevalence of ~40–50 per 10,000, while in the Netherlands, ~40 per 10,000.[25, 26]

As discussed above, low prevalence in primary-care sources (IPCI) may reflect under-ascertainment because type 1 diabetes diagnosis and insulin initiation often occur in secondary care, and primary care records may receive these diagnoses/treatments incompletely or with delay. This care-pathway dependency could also reduce sensitivity for a 180-day diagnosis-insulin concordance requirement in primary care data sources. In addition, when there is diagnostic uncertainty, GPs may rely on specialist-led confirmation (including autoantibody testing) and treatment initiation, which may further delay complete capture in primary care records.[27]

In DK-DHR, point prevalence in 2024 was 81.8 per 10,000 in 2024, whereas in the sensitivity analysis, it decreased to 54.7 per 10,000 for the same year. In IPCI, point prevalence in 2024 was 17.2 per 10,000 in 2024, whereas in the sensitivity analysis it was 15.3 per 10,000 for the same year. The reduction of absolute prevalence in the sensitivity analysis, excluding individuals with any prior history of type 2 diabetes, indicates a marked inclusion of individuals with prior type 2 diabetes history, particularly among older age groups. For future analyses, a more restrictive definition, as applied in the sensitivity analysis, is recommended to increase the specificity of the diagnosis.

The decreasing prevalence observed in DK-DHR should not be interpreted as directly contradicting a globally increasing type 1 diabetes incidence, because incidence and prevalence reflect different processes (incident case accrual, survival/duration, exits from the population, and denominator composition).[26] In the sensitivity analysis, the overall prevalence in DK-DHR was broadly stable, suggesting that part of the apparent decline in the primary analysis may relate to diabetes subtype classification uncertainty in older adults. Age-stratified sensitivity analyses showed heterogeneous trends by age. In both data sources, prevalence generally increased over time in younger age groups (<40 years). In contrast, trends in older age groups differed: in DK-DHR, prevalence declined in ages 40–49 and ≥50 years, whereas in IPCI, trends in older age groups were more variable.

Collectively, the results represent a first approach to describing type 1 diabetes in routinely collected data and their feasibility for addressing future research questions focused on this population. The most notable identified constraint is the need to improve the harmonisation of measurements, especially concerning autoantibody positivity. Additionally, phenotype refinement can be considered, acknowledging potential reductions in cohort yield.

10.4. Generalisability

Generalisability of these findings should be interpreted at two levels (1) the underlying patient populations represented in each data source, and (2) the ability of each data source to capture the clinical information required for type 1 diabetes identification and characterisation. The included data sources span hospital-based, primary care-based, and registry-linked settings, and therefore represent different healthcare settings, coding practices, and care pathways. As a result, estimates are best interpreted as source-specific rather than as comparable across sources.

Findings on cohort yield, attrition, and the observed overlap with type 2 diabetes codes/non-insulin therapies are most applicable to studies using similar routinely collected data and similar phenotype requirements (concordant diagnosis and insulin exposure; first-ever in all available history). The 180-day concordance requirement may differentially affect sources depending on where and when diabetes diagnoses and insulin initiation are recorded. Consistent with this, insulin recorded on index date was more often identified in hospital-based sources (>25%) than in population/primary care-based sources (<25%). This likely reflects both where care is delivered and how events are recorded: in hospital-based data sources, diagnosis and prescribing are often captured within the same encounter and setting; in primary care-based sources, specialist diagnosis and medication dispensing are often recorded in separate systems

and may be reflected in the primary care record with delay, increasing date gaps between diagnosis and insulin. However, the true clinical sequence of events cannot be determined from these data.

Generalisability of laboratory-related findings are limited by variability in measurement records and by incomplete standardisation of measurements and autoantibody results across sources. In particular, conclusions about the feasibility of autoimmune confirmation or staging using autoantibody "positivity" (abnormal autoantibodies) are generalisable to other routinely collected datasets with similar laboratory result representation and mapping practices.

Finally, point-prevalence estimates should be interpreted as source-specific because they depend on population coverage, healthcare setting (e.g., primary care vs specialist/secondary care), and denominator definitions. Extrapolation to external populations should therefore be cautious and account for differences in data coverage and care pathways.

11. CONCLUSION

This multi-data source feasibility study assessed whether routinely collected healthcare data can support identification of early type 1 diabetes (prior to clinical diagnosis and insulin-dependence). Glycaemic measurement capture and timing were generally consistent with the expected diagnostic patterns, with the earliest recorded testing often preceding the index date, and the earliest abnormal results occurring closer to the index date. By contrast, although autoantibody test occurrence was identified in most sources, testing was infrequent and standardised identification of "positive" (abnormal) autoantibody results was very limited. Current use of antibodies for early-stage identification and staging is constrained by incomplete capture and laboratory result harmonisation.

More broadly, the phenotype requiring concordant recording of a type 1 diabetes diagnosis and insulin exposure identified cohorts with generally plausible demographic and clinical characteristics across data sources. Although heterogeneity was observed in cohort yield and data capture, this is consistent with differences in care setting and recording practices. The most actionable refinement supported by these results is to address diabetes-type overlap and classification uncertainty in older adults (e.g., excluding prior type 2 diabetes or other restrictions in older age strata), acknowledging the trade-off with cohort yield.

12. REFERENCES

1. Atkinson, M.A., G.S. Eisenbarth, and A.W. Michels, *Type 1 diabetes.* The Lancet, 2014. **383**(9911): p. 69-82.

2. Insel, R.A., 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): p. 1964-74.

3. Phillip, M., et al., *Consensus guidance for monitoring individuals with islet autoantibody-positive pre-stage 3 type 1 diabetes.* Diabetologia, 2024. **67**(9): p. 1731-1759.

4. DeFalco, F., et al., *Achilles: Achilles Data Source Characterization. R package version 1.7.2.* 2023.

5. *Standards of medical care in diabetes--2011.* Diabetes Care, 2011. **34 Suppl 1**(Suppl 1): p. S11-61.

6. Gilbert, J., et al., *CohortDiagnostics: Diagnostics for OHDSI Cohorts. R package version 3.3.0, https://github.com/OHDSI/CohortDiagnostics, https://ohdsi.github.io/CohortDiagnostics.* 2024.

7. Inberg, G., E. Burn, and T. Burkard, *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.

8. Catala M, G.Y., 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/.

9. Raventós, B., 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).

10. Fagherazzi, G., et al., *Nationwide Trends in Type 1 and Type 2 Diabetes in France (2010-2019): A Population-Based Study Using a Machine Learning Classification Algorithm.* Diabetes Ther, 2025. **16**(10): p. 1973-1991.

11. Rodríguez Escobedo, R., E. Delgado Álvarez, and E.L. Menéndez Torre, *Incidence of type 1 diabetes mellitus in Asturias (Spain) between 2011 and 2020.* Endocrinol Diabetes Nutr (Engl Ed), 2023. **70**(3): p. 189-195.

12. Harding, J.L., et al., *The Incidence of Adult-Onset Type 1 Diabetes: A Systematic Review From 32 Countries and Regions.* Diabetes Care, 2022. **45**(4): p. 994-1006.

13. Thomas, N.J., et al., *Identifying type 1 and 2 diabetes in research datasets where classification biomarkers are unavailable: assessing the accuracy of published approaches.* Journal of Clinical Epidemiology, 2023. **153**: p. 34-44.

14. Initiative, D.P.V., et al., *Prevalence and comorbidities of double diabetes.* Diabetes Research and Clinical Practice, 2016. **119**: p. 48-56.

15. Thomas, N.J., et al., *Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank.* Lancet Diabetes Endocrinol, 2018. **6**(2): p. 122-129.

16. Petrie, J.R., et al., *Cardiovascular and metabolic effects of metformin in patients with type 1 diabetes (REMOVAL): a double-blind, randomised, placebo-controlled trial.* Lancet Diabetes Endocrinol, 2017. **5**(8): p. 597-609.

17. Klompas, M., et al., *Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data.* Diabetes Care, 2013. **36**(4): p. 914-21.

18. Gajewska, K.A., et al., *Prevalence and incidence of type 1 diabetes in Ireland: a retrospective cross-sectional study using a national pharmacy claims data from 2016.* BMJ Open, 2020. **10**(4): p. e032916.

19. Parviainen, A., et al., *Decreased Incidence of Type 1 Diabetes in Young Finnish Children.* Diabetes Care, 2020. **43**(12): p. 2953-2958.

20. Committee, A.D.A.P.P., *2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025.* Diabetes Care, 2024. **48**(Supplement_1): p. S27-S49.

21. Sims, E.K., et al., *Screening for Type 1 Diabetes in the General Population: A Status Report and Perspective.* Diabetes, 2022. **71**(4): p. 610-623.

22. Cherubini, V., et al., *Follow-up and monitoring programme in children identified in early-stage type 1 diabetes during screening in the general population of Italy.* Diabetes Obes Metab, 2024. **26**(10): p. 4197-4202.

23. Winkler, C., et al., *Markedly reduced rate of diabetic ketoacidosis at onset of type 1 diabetes in relatives screened for islet autoantibodies.* Pediatric Diabetes, 2012. **13**(4): p. 308-313.

24. NICE, *Type 1 diabetes in adults: diagnosis and management (NG17)*. 2022.

25. Carstensen, B., P.F. Rønn, and M.E. Jørgensen, *Prevalence, incidence and mortality of type 1 and type 2 diabetes in Denmark 1996-2016.* BMJ Open Diabetes Res Care, 2020. **8**(1).

26. Ogle, G.D., et al., *Global type 1 diabetes prevalence, incidence, and mortality estimates 2025: Results from the International diabetes Federation Atlas, 11th Edition, and the T1D Index Version 3.0.* Diabetes Research and Clinical Practice, 2025. **225**: p. 112277.

27. Lal, R.A., et al., *Primary Care Providers in California and Florida Report Low Confidence in Providing Type 1 Diabetes Care.* Clinical Diabetes, 2020. **38**(2): p. 159-165.

13. ANNEXES

ANNEX I. Description of data sources

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 Analytics Centre (DAC)
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.
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 v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have 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	DK-DHR has the following limitations, which may be relevant confounders for certain complex Darwin EU® studies:

#	Section	Description
		- We lack information on key socio-economic status (SES) factors, such as occupation, education, and income. These variables may be important for analysis in some studies. - We only have complete data on lifestyle factors (such as smoking status and weight) for pregnant women. - We have no information on patient contacts in primary care (visits to the GP). Consequently, the incidence of chronic diseases like Type 2 Diabetes (T2D) and asthma must be determined using drug prescriptions as a proxy. Stillborn children will not have any records in our CDM. This means that e.g. birth length of stillborns is not recorded.
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 data source 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

Tampere University Hospital patient cohort (FinOMOP-TaUH)

#	Section	Description
1	Data source identification and country	FinOMOP-TaUH Pirha (Tampere University Hospital patient cohort) Pirkanmaa, Finland
2	Data partner information section	Pirkanmaa Welfare Services County, Tampere University Hospital Department of Research, Development and Education
3	Coverage and timespan	Data collection since: 2007 Extent: Regional. TaUH Research Data source includes all specialities/all patient groups treated in the Tampere University Hospital,
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Secondary, and tertiary care given in the region, including given clinical and pathology diagnoses, diagnostic and therapeutic procedures, laboratory findings, radiology and pathology reports, medication given in the hospital and electronic prescriptions, and continuous medical records (free text), including discharge letters since 2007.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. The data is entered by the clinician at point of contact and processed in a data warehouse for secondary use.
6	General representativeness	The data covers patients visiting the hospital and therefore will not reflect the general population.
7	Data content /source coding	ICD10fi: Finnish modification of ICD10; SNOMED2-TMP: a local SNOMED2-based vocabulary of organ-diagnosis pairs in pathology; NCSPfi: Finnish national version of NCSP/Nomesco vocabulary for procedures; LABfi: Finnish national vocabulary (Kuntaliitto) for laboratory measurements; LABfi_TMP: Local additions to the national vocabulary for laboratory measurements; UNIT_FIN: measurement units; MICROBEfi_TMP: Microbe names in measurement data; ATC: all drugs except anticancer drugs; active ingredient (in Finnish): for anticancer drugs; VNR: Nordic medicine product identifiers (https://pharmaca.fi/en/health/pharma/vnr/); Hilmo_eala: Finnish national medical specialty vocabulary.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. In some cases, patients can be re-registered with a new ID. Patients are identified by their social

#	Section	Description
		security number, and each patient gets a patient ID in the EHR, associated with the social security number. However, there are cases when the social security number has changed, and the associated patient ID also has changed. Approximately 4% of all patient IDs in the data source are such duplicates. Majority of such duplicates are newborns who have first been assigned a temporary social security number, which has later been changed to a permanent social security number. An unidentified person can have been given a temporary social security number, and sometimes a person can change their social security number, for example, because of threat (stalking) or gender reassignment. It is technically possible to link all the instances of a person in the data source however this is not currently done.
9	Quality control (data source specific)	The source data source tables are checked for completeness on a general level: whether there are null values in the fields, what the date ranges are for each table, are there gaps in the data, and how many unique patients can be found in each data table. Before mapping to the OMOP CDM, we have identified the data tables and the fields in the tables that are required to build the OMOP CDM. The White Rabbit and Rabbit-in-a-Hat tools were used in this process.
10	Linkage	Data from many nation-wide registries can be combined, which is subject to obtaining special data permits.
11	Vital status	The source of vital status (date of death) is the Digital and Population Data Services Agency.
12	Limitations	No data source-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111135 Website: https://www.pirha.fi/en/web/english

Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)

#	Section	Description
1	Data source identification and country	CDW Bordeaux (Clinical Data Warehouse of Bordeaux University Hospital) Nouvelle-Aquitaine, France
2	Data partner information section	CHU DE BORDEAUX - DIRECTION GENERALE Gironde / Nouvelle-Aquitaine
3	Coverage and timespan	Data collection since: 2005 Extent: Regional. It covers the population of Bordeaux metropolitan area, and possibly beyond, as the health care centre for referrals and expertise for the Nouvelle Aquitaine region. The data source contains data from 2005 onwards.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and claims data. The data source currently captures information about patient demographics, visit details, conditions, procedures, drugs, measurements, and mortality.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Biobank. The integrated data is extracted from the hospital production information system via a real-time research data source. The data is then processed and quality controlled by a team dedicated to maintaining the data source. Internal evaluations were carried out to ensure consistency between the research data source and the patient bedside software.
6	General representativeness	This is the 6th largest metropolitan area in France, and CHUBX is the largest hospital in the region. More than 75% of the patients administered to Bordeaux university hospital reside in

#	Section	Description
		the Gironde departments, with almost 50% coming directly from the Bordeaux metropolitan area. The hospital also captures additional cases from Nouvelle-Aquitaine region.
7	Data content /source coding	Diagnosis source data is coded in ICD-10 terminology. Procedures are coded in CCAM (French terminology). Laboratory measurements are coded in local terminology and partially mapped to LOINC. Drugs are coded through a local terminology and then mapped to UCD (French terminology), as well as ATC codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. We use the hospital's unique identifier to generate the patient identifier in OMOP. If two identities are merged at the hospital, the merge is taken into account in the CDW. An automatic (hourly) detection of suspected duplicated identities has been implemented at the hospital since 2020, with merging of duplicated identities by a specialized team. Identities since 2015 were processed retrospectively. Thus, the rate of identity duplication in the data source is low, especially since 2015.
9	Quality control (data source specific)	- The integrated data comes from the hospital production information system through a real-time replicated data source. Consistency evaluations between the replicated data source and the production system are performed by the technical team in charge of maintaining the replicated data source. - In the same way, consistency checks are performed between the replicated data source and the data integrated into the i2b2 CDW. In addition, dashboards enable monitoring the data integrated into the i2B2 CDW, in particular by controlling the amount of data available over time, and its evolution, according to the various data sources. - An internal evaluation was carried out to ensure the consistency between the data integrated into i2b2 and the data available in the software used at the patient's bedside. In addition, many use cases were performed on the i2b2 CDW, with return to the patient chart and comparison of the data integrated into i2b2 and the data available in the care file.
10	Linkage	Death certificates (without the cause of death).
11	Vital status	The data source is linked to the French death registry.
12	Limitations	CDW Bordeaux is limited to events captured in the hospital setting and thus does not include patient events not treated by the hospital (e.g., rare cancers). Patient events that are not included in CDW Bordeaux are rare disease treatments or specialist events that occur outside of CHUBX.
13	Main references	Cossin S, Diouf S,Griffier R,Le Barrois d'Orgeval P,Diallo G,Jouhet V "Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning." JAMIA open (2021): 33709061
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111112 Website: https://www.chu-bordeaux.fr/

Semmelweis University Clinical Data (SUCD)

#	Section	Description
1	Data source identification and country	SUCD (Semmelweis University Clinical Data) Budapest, Hungary
2	Data partner information section	Semmelweis University -
3	Coverage and timespan	Data collection since: 2010 Extent: Regional.

#	Section	Description
		The general catchment area of SU is the central region of the country, Budapest city and Pest county, although patients can be referred from anywhere in Hungary. The total population of Budapest and Pest county is approximately 4,200,000 people. The total population of Hungary is around 9,500,000.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and claims data, and other (specify). diagnostic data (laboratory tests, radiology, pathology)
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries. Data is extracted directly from the source data source. From there, the data entry in the system is heavily controlled and validated on the user interface before being made available for further research.
6	General representativeness	SU captures information on patients who are covered by the public health insurance system. This covers all Hungarian citizens, and therefore the data source should mirror the source population well. Although, besides Semmelweis University Clinics, there are multiple hospitals in the region, and data on visits in other hospitals in not represented in the data source. Therefore the patient population is not directly representative of the general population.
7	Data content /source coding	Regarding SU's source data, procedures and diagnoses are coded in SNOMED, measurements are coded in LOINC, and drugs are stored in RxNorm and ATC.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a unique identifier (SSN).
9	Quality control (data source specific)	The clinical data source is the source data source and therefore it has to be treated as a trusted data source. Data entry in the systems is heavily controlled by validation on the user interface, and there are large number of rules that controls the data on the insurer's side that has to be corrected in the system by the users to be able to close the encounters. OMOP mapping is done in the framework by EHDEN recognized partners under quality check by the EHDEN society.
10	Linkage	No known linkages.
11	Vital status	Source for vital status unknown.
12	Limitations	Medication prescribed in secondary care is fully present in our data source, but medication given in the hospital is rarely documented. General limitations for the data type applicable. General practitioner data is not present in our data source; therefore, this part of the patient journey is not represented.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1000000184 Website: https://www.semmelweis.hu

Integrated Primary Care Information (IPCI)

#	Section	Description
1	Data source identification and country	IPCI (Integrated Primary Care Information) The Netherlands
2	Data partner information section	Erasmus University Medical Center Department of Medical Informatics
3	Coverage and timespan	Data collection since: 2006 Extent: Nation-wide.

#	Section	Description
		IPCI is a Dutch data source that contains patient records from 2006 onwards. However, it mainly covers the central part of the country, including the most densely populated area (the 'Randstad') and non-urban areas. IPCI contains information on all patients registered with GPs responsible for non-emergency care and referrals. A patient is registered at birth or at first encounter with the GP.
4	Healthcare setting / type of data	Primary care – General Practioner. Data is collected from primary care EHR. This includes demographic information, complaints and symptoms, diagnoses, laboratory test results, lifestyle factors (in limited amount), and correspondence with secondary care, such as referral and discharge letters.
5	Data collection process	Outpatient electronic health records. Data is entered into the EHR system by the GPs, during or after the visit. The patientdossiers are collected by Erasmus MC data managers and combined in one harmonized data source. Several checks are done on this data source to ensure correct data processing. Persons can have dossiers at multiple GPs.
6	General representativeness	More than 99% of the Dutch population has health insurance, and almost all citizens are registered with a general practitioner. Over 12 months, around 78% of the population has at least one contact with their GP. IPCI included around 350 GP practices out of around 5000 in the country (~ 7%). The demographic composition of the IPCI population mirrors that of the general Dutch population in terms of age and sex.
7	Data content /source coding	Dutch GPs use mainly Dutch standard codes, like ICPC-1 and Diagnostische Bepalingen maintained by NHG. And for therapy the G-Standard is used, maintained by ZIndex date.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients can be registered under different IDs, but since a patient can only be registered at one GP at a time, the observations periods will not overlap.
9	Quality control (data source specific)	Prior to each data release, extensive quality control steps are performed, e.g., comparison of patient characteristics between practices, and checks to identify abnormal temporal data patterns in practices. For each practice, around 200 quality indicators are obtained. Of these indicators, a quarter refer to population characteristics, e.g., number of birth and mortalities relative to practice size, temporal consistency. The other indicators are based on medical data, e.g., distribution of measurement values, frequencies of diagnoses and procedures relative to age, completeness of data. The indicators are combined in a couple of quality scores for each practice. For these scores, cut-off values for acceptable quality have been defined. Practices with a score below a cut-off are excluded for research. This approach has shown to be very important, for example to check if data from practices that just joined the data source are at an acceptable level of quality. The details of the approach, like the cut-off values for acceptance, are based on years of experience. In addition, trends are compared with the previous data source release. Extensive quality control steps are performed before each data release. These include comparing patient characteristics between practices and checks to identify abnormal temporal data patterns in practices. Additional checks include over 200 indicators related to population characteristics (e.g., reliability of birth and mortality rates) and medical data (e.g., availability of durations of prescriptions and completeness of laboratory results). Records of low quality are excluded from the data source.
10	Linkage	Linkage requires additional approval steps and needs to be assessed on a case-by-case basis. IPCI is not routinely linked with other data sources.
11	Vital status	Vital status (death date and cause) is collected based on GP records.
12	Limitations	The main limitation comes with the fact that IPCI is limited to GP records, and although it contains information on referrals and discharge letters, it may not fully capture specific hospital information. IPCI does not include coded/detailed data about medications/procedures/test results from the hospital or other care-providers.

#	Section	Description
13	Main references	de Ridder MAJ, de Wilde M,de Ben C,Leyba AR,Mosseveld BMT,Verhamme KMC,van der Lei J,Rijnbeek PR "Data Resource Profile: The Integrated Primary Care Information (IPCI) data source, The Netherlands." International journal of epidemiology (2022): 35182143
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/42618 Website: http://www.ipci.nl

Hospital Universitario 12 de Octubre (H12O)

#	Section	Description
1	Data source identification and country	H12O (Hospital Universitario 12 de Octubre) Comunidad de Madrid, Spain
2	Data partner information section	Fundación Investigación Biomédica Hospital 12 de Octubre Instituto de Investigación (i+12)
3	Coverage and timespan	Data collection since: 2015 Extent: Regional. H12O is a national reference for the treatment of certain pathologies covering patients in the southern area of Madrid region.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). The H12O data source, INFOBANCO, contains information from the different health domains (laboratory, prescriptions, treatments, administrative, diagnoses, etc.). In addition, information is also obtained from other data sources, such as the pathological anatomy system, which provides information about sample analysis, and the cost system, containing information on the cost associated with a contact with the hospital. Work for the inclusion of further data is ongoing, among others, radiological information, or PROMs.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries, and Other. Data is entered by clinicians and processed in the regional 'INFOBANCO' platform. This platform allows combining data from multiple heterogeneous sources, and provides mechanisms for data governance, adequacy, interrogation, visualization and analysis for real-world evidence generation and decision support.
6	General representativeness	H12O only covers patients visiting the hospital due to a medical condition and is therefore not necessarily representative of the general population.
7	Data content /source coding	Diagnosis: ICD-10-CM, IDC-9, SNOMED CT, ORPHA; Procedures: ICD-10-PCS, ICD-9, SNOMED CT; Medication: ATC and SNOMED CT; Laboratory tests: LOINC; Histopathology: ICD-O-3.1 and SNOMED CT; Clinical observations: SNOMED CT
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients cannot be registered with different IDs. Each patient has a unique identifier in the hospital's EHR (NHC), a regional unique identifier (CIPA), and a national unique identifier (CIP-SNS).
9	Quality control (data source specific)	The EHR contains certain rules for recording information, e.g., it does not allow closing reports that do not have a coded diagnosis. In addition, it allows the use of terminologies, such as SNOMED or LOINC, for recording information, which helps to reduce unstructured data.
10	Linkage	The hospital data can be linked with pharmacy and the Madrid public health service.
11	Vital status	The hospital is connected to the CIBELES system of the Madrid public health service, which allows the unique identification of users and contains the date of death of the patient. In this

#	Section	Description
		way, the hospital has the information of the death, regardless of whether it occurred in the hospital or not.
12	Limitations	No data source-specific limitations documented. General limitations for the data type applicable.
13	Main references	Miguel Pedrera-Jiménez, Noelia Garcia Barrio, Antonio J Díaz Holgado, Pablo Serrano-Balazote, et al. "INFOBANCO: Standardized platform for management, semantic interoperability, and transparent reuse of EHRs" Conference: EIT Health German-Spanish Symposium on Health DataAt: Mannheim, Germany (2022):
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111145 Website: https://cpisanidadcm.org/infobanco

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 enabled the use of standardised analytics and using DARWIN EU® tools across the network since the structure of the data and the terminology system was 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 was written in R and used standardized analytics wherever possible. Each data partner executed the study code against their data source containing individual data and then returned the results (csv files) which only contained aggregated data. The results from each of the contributing data sites were then combined in tables and figures for the study report.

Data storage and protection

For this study, personal data from individuals in various EU member states were processed, using information collected from 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 were 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 were adhered to. In agreement with these regulations, rather than combining person level data and performing only a central analysis, local analyses were run, which generate non-identifiable aggregate summary results.

QUALITY CONTROL

Data source quality control

Define the quality control and packages that are specific to your study and remove any sections that are not relevant.

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

When defining cohorts for indications, a systematic search of possible codes for inclusion will be identified using the *CodelistGenerator* R package (https://github.com/darwin-eu/CodelistGenerator). This package allows the user to define a search strategy and will use this to query the vocabulary tables of the OMOP common data model so as to find potentially relevant codes. In addition, the *CohortDiagnostics* (https://github.com/OHDSI/CohortDiagnostics) and *DrugExposureDiagnostics* (https://cran.r-project.org/web/packages/DrugExposureDiagnostics/index date.html) R packages 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 *DrugExposureDiagnostics* package evaluates ingredient-specific attributes and patterns in drug exposure records.

The study code will be based on DARWIN EU® R packages: *IncidencePrevalence* to estimate Incidence and Prevalence, *DrugUtilisation* to characterise the drug use, and *CohortCharacteristics* to characterise the cohort by indication. These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via GitHub.

ANNEX III: List of conditions and medication definitions

Table S1. 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 42709887	-	SNOMED CT
Ketoacidosis	Ketoacidosis	4209145 42535540 4226238	-	SNOMED CT
Overweight and obesity	Overweight Body mass index date 25–29 - overweight Obesity Body mass index date 30+ - obesity	437525 4060705 433736 4060985	-	SNOMED CT
Coeliac disease	Coeliac disease	194992	-	SNOMED CT
Hypothyroidism	Hypothyroidism Hashimoto thyroiditis	140673 135215	-	SNOMED CT
Hyperthyroidism	Thyrotoxicosis Hyperthyroidism	138387 4142479	-	SNOMED CT
Other autoimmune disorders	Pernicious anemia Autoimmune hepatitis Addison's disease	432295 200762 443394	-	SNOMED CT

Table S2. List of medication definitions.

Substance Name	Concept name	Class	ATC code	Ingredient Concept ID	Include descendants
Insulin	lente insulin, human	Ingredient		46221581 44506754 40170911 35198096 1531601 1567198 35602717 1516976 1502905	Yes

Substance Name	Concept name	Class	ATC code	Ingredient Concept ID	Include descendants
				1544838 1588986 1550023 1513876 19090244 19090229 19090247 19090249 19090180 19013926 19091621 19090187 19013951 1590165 1596977 1586346 19090204 1513843 1513849 1562586 19090226 19090221 1586369	
Immunomodulators	Teplizumab	Ingredient		741995	Yes
Verapamil		Ingredient		1307863	Yes
Glucose lowering drugs		Ingredient		1503297 1580747 19059796 19122137 1516766 1597756 44785829 45774751 1559684 40239216	Yes

Substance Name	Concept name	Class	ATC code	Ingredient Concept ID	Include descendants
				793143	
				40170911	
				1525215	
				45774435	
				1529331	
				1560171	
				1547504	
				43013884	
				43526465	
				40166035	
				1583722	
				1502855	
				793293	
				44506754	
				1510202	
				19097821	
				1000979	
				1502826	
				1594973	
				19033909	
				44816332	
				43009089	
				1502809	
				19001409	
				40798673	
				40798860	
				19035533	
				19033498	
				1530014	
				19001441	
				43009020	
				779705	
				1301517	
				43009070	
				43009032	

Substance Name	Concept name	Class	ATC code	Ingredient Concept ID	Include descendants
				43009055 1517998 43009051 35198021 36854701 43009094 36863042 1515249	

ANNEX IV: Supplementary results

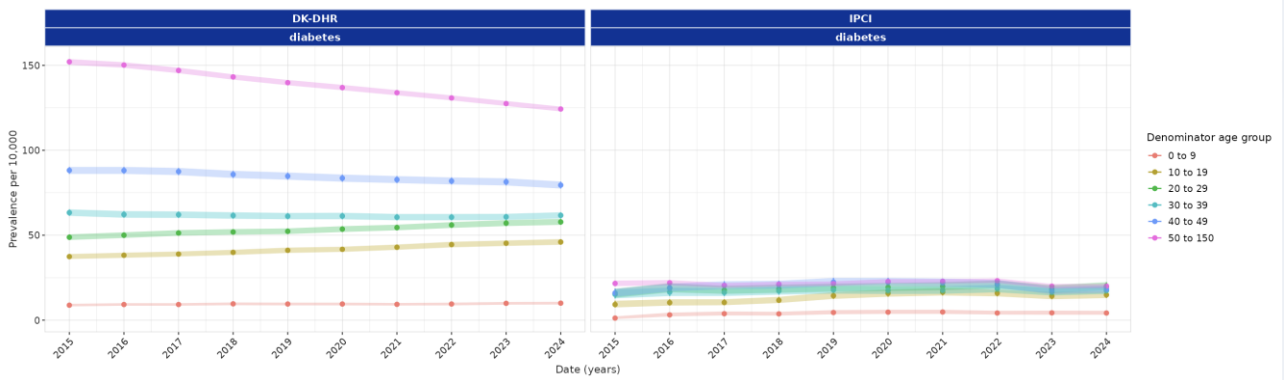


Figure S1. Temporal trends of annual point prevalence of type 1 diabetes on 1 January by Data Source and (DK-DHR and IPCI) Age Group, 2015–2024.

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

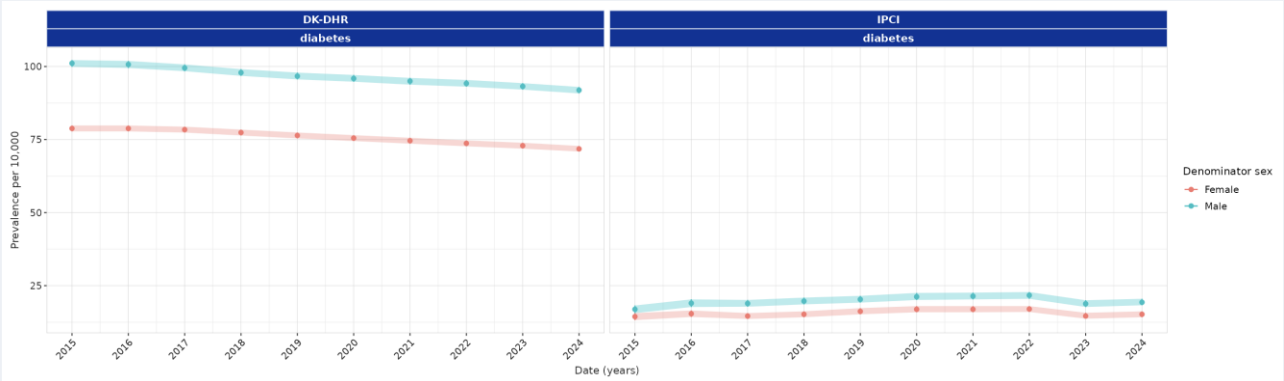


Figure S2. Temporal trends of annual point prevalence of type 1 diabetes on 1 January by Data Source (DK-DHR and IPCI) and Sex, 2015–2024.

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

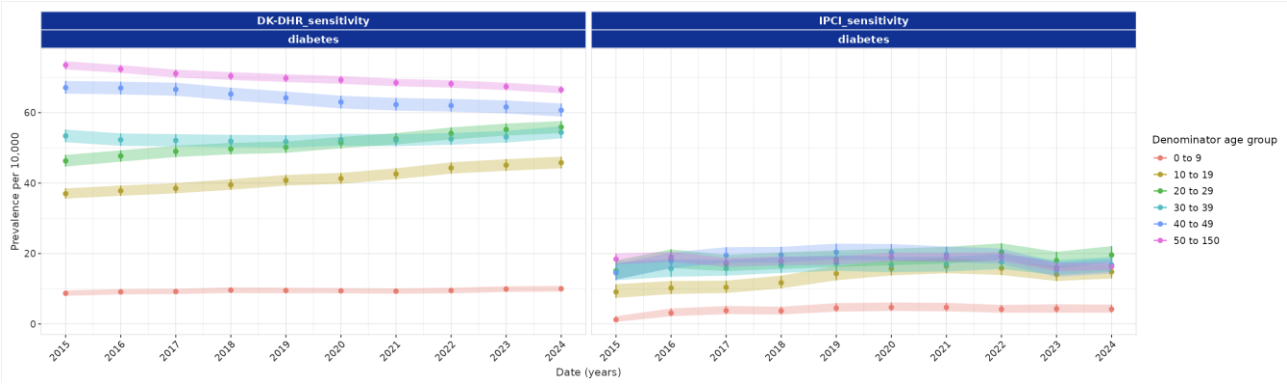


Figure S3. Temporal trends of annual point prevalence of type 1 diabetes on 1 January, Excluding Individuals with Any Prior History of Type 2 Diabetes, by Data Source (DK-DHR and IPCI) and Age Group, 2015–2024.

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

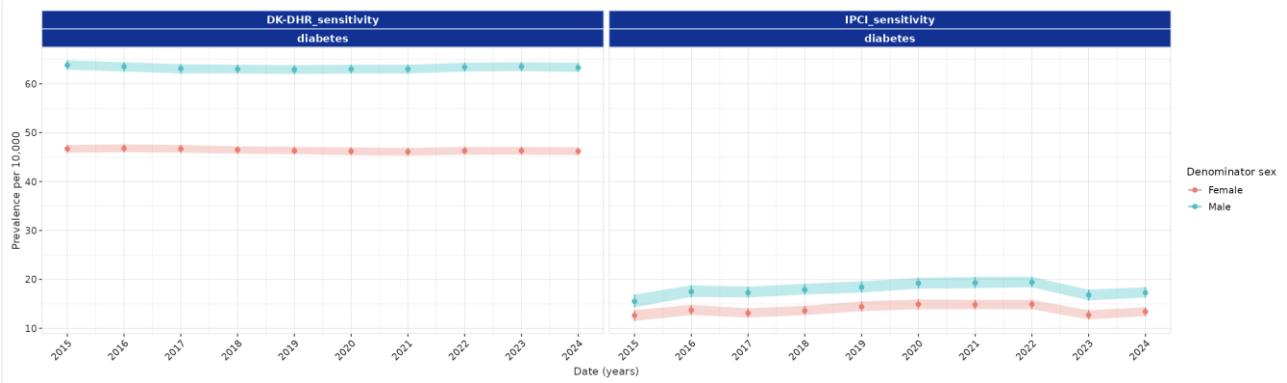


Figure S4. Temporal trends of annual point prevalence of type 1 diabetes on 1 January, Excluding Individuals with Any Prior History of Type 2 Diabetes, by Data Source (DK-DHR and IPCI) and Sex, 2015–2024.

DK-DHR=Danish Data Health Registries; IPCI=Integrated Primary Care Information. N=Number of subjects. In population-based data sources, a minimum of 365 days of observation prior to index date was required.

Table S3. Representation of autoantibody results across data sources.

Antibody	FinOMOP-TaUH	CDW Bordeaux	SUCD	H12O
GAD-65	No binary; **IU/mL**; cutoff **5**	Units **UI/mL**; cutoff **not provided**	Units **IU/mL**; cutoff **not provided**	No binary; **UI/mL**; cutoff **5.00**
IA-2A	No binary; **U/mL**; cutoff **15**; result may be "<15" vs numeric	Units **UI/mL**; cutoff **not provided**	"–" (not specified)	No binary; **U/mL**; cutoff **10.00**
ICA	No binary; **titer** (pre-2015-06-07: U/mL); cutoff **10** (earlier cutoff unknown)	Unit/cutoff **not provided**	**Binary pos/neg=yes** (unit/cut off n/a)	No binary; unit/cutoff **not specified**
ZnT8	**Not available**	Units **UA/mL**; cutoff **not provided**	Units **RU/mL**; cutoff **not provided**	No binary; **U/mL**; cutoff **0.00–15.00** *(as reported)*
IAA	No binary; **kU/L or U/mL**; cutoff **0.4 kU/L**	Units **UA**; cutoff **not provided**	Units **uLU/mL**; cutoff **not provided**	No binary; **U/mL**; cutoff **18.00**

The table summarises data sources survey responses on result encoding/positivity definitions; availability and mapping in the analytic dataset may differ (e.g., test occurrence present but positivity fields not consistently mapped). DK-DHR and IPCI indicated that results were not available. Binary=yes/no. Cutoff=threshold value considered positive at source. IU=International units; mL=mililliters; UI=Universal units; U=Units; kU=Kilo unit; UA=Unit actuation; RU=Relative units. IAA=Insulin autoantibodies; ICA=Islet Cell autoantibodies; IA-2A=Anti-IA-2 (insulinoma-associated antigen-2) antibodies; ZnT8=Anti-Zinc transporter 8 antibodies; GAD-65=Anti-glutamic acid decarboxylase antibodies. CDW Bordeaux=Clinical Data Warehouse of Bordeaux University Hospital; FinOMOP-TaUH=Tampere University Hospital patient cohort; SUCD=Semmelweis University Clinical Data; DK-DHR=Danish Data Health Registries; H12O=Hospital Universitario 12 de Octubre; IPCI=Integrated Primary Care Information.

ANNEX V: Glossary

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

Aggregated Data

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

Benefit-Risk Assessment

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

Common Data Model (CDM)

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

Complex Studies (C3)

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

Coordination Centre (CC)

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

Data Access

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

Data Quality Framework

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

Data Source

A 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 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 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 , 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.