Study Report

P2-C1-013 DARWIN EU® - Comparing Direct and Indirect Methods to Estimate Prevalence of Chronic Diseases using Real-World Data

24/06/2024

Version 3.0

Contents

Study Title	DARWIN EU® - Comparing direct and indirect methods to estimate prevalence of chronic diseases using real-world data
Study Report Version identifier	3.0
Dates Study Report updates	v1.0: 26/4/2024 v2.0: 28/5/2024 v3.0: 24/6/2024
EU PAS register number	EUPAS1000000088
Active substance	NA
Medicinal product	NA
Research question and objectives	In the context of chronic diseases with relatively low prevalence, how do direct and indirect RWD-based estimates of prevalence compare with each other? The specific objectives of this study are to: 1) Estimate the disease prevalence (direct estimate based on the proportion of individuals with the condition). 2) Estimate the disease incidence rate. 3) Estimate duration of disease using Kaplan-Meier survival curves. Of particular interest is the estimate of median survival as a summary measure of disease duration. 4) Produce an indirect estimation of prevalence as the product of incidence and median survival. for the following diseases: • Cystic fibrosis • Haemophilia (A and/or B) • Pulmonary arterial hypertension • Pancreatic cancer • Sickle cell disease Results will be provided overall and where possible stratified by age group: paediatrics (0-17 years old) and adults (18 years old and above).
Country(-ies) of study	United Kingdom, the Netherlands, Spain
Author	Maria de Ridder (m.deridder@darwin-eu.org) Katia Verhamme (k.verhamme@darwin-eu.org)

1. DESCRIPTION OF STUDY TEAM

Study team Role	Names	Organisation
Study Project Manager/Principal Investigator	Maria de Ridder	Erasmus MC
Data Scientist	Maarten van Kessel Ross Williams Cesar Barbosa	Erasmus MC Erasmus MC Erasmus MC
Epidemiologist	Katia Verhamme	Erasmus MC
Clinical Domain Expert	Katia Verhamme	Erasmus MC
Statistician	Maria de Ridder	Erasmus MC
Local Study Coordinator/Data Analyst	Antonella Delmestri Mees Mosseveld Talita Duarte Salles	University of Oxford – CPRD data Erasmus MC – IPCI data IDIAP – SIDIAP data

2. DATA SOURCES

Country	Name of Database	Health Care setting	Type of Data	Total number of subjects	Calendar period covered
UK	CPRD GOLD	Primary care	EHR	17,267,137	09/09/1987 to 23/06/2023
the Netherlands	IPCI	Primary care	EHR	2,817,331	01/01/2006 to 30/06/2023
Spain	SIDIAP	Primary care	EHR	8,553,325	01/01/2006 to 30/06/2023

3. ABSTRACT (STAND-ALONE SUMMARY OF THE STUDY REPORT)

Title

DARWIN EU® - Comparing direct and indirect methods to estimate prevalence of chronic diseases using real-world data.

Rationale and background

Prevalence of a disease or condition is defined as the proportion of individuals in a population affected by a condition at a given point in time. Quantifying disease prevalence is important from a public health perspective, e.g. to understand the impact of diseases on the population, or to plan and allocate health care resources. Measuring disease prevalence is also important from a medicine regulatory viewpoint, as regulatory agencies grant incentives for the development of new therapies for rare diseases, i.e. diseases with low prevalence.

Disease prevalence depends on the rate of incidence of the disease in the population as well as on the average duration of the disease. Under the assumption that both the incidence of the disease and its average duration are stable over time, a well-known mathematical relationship between prevalence (P), incidence (I) and average duration (D) is:

$$\frac{P}{1-P} = I \cdot D$$

When P is low, $(1-P) \approx 1$ and the equation reduces to the following expression for the indirect estimated prevalence:

$$P = I \cdot D$$

In this study direct and indirect estimated prevalences were compared using real-world data sources.

Research question and objective

The objective of this study was to compare direct and indirect estimations of prevalence of five rare, chronic diseases: Cystic fibrosis, Haemophilia, Pulmonary arterial hypertension (PAH), Pancreatic cancer and Sickle cell disease. This was done considering all patients with one of the diseases as well as separately for patients with paediatric diagnosis (age 0-17 years) and for patients with adult diagnosis (age 18 and older).

Research Methods

Study design

Retrospective cohort design to estimate disease point prevalence and incidence.
Retrospective cohort design to estimate median survival as a proxy for disease duration.
Data from three databases with routinely collected electronic healthcare records of general practices were used.

Population

All individuals present in one of the databases during the study period 01/01/2010 to 31/12/2022 were included to estimate incidence and prevalence.

For each disease, all patients with a diagnosis were included to estimate median disease duration.

Variables

- Presence of a diagnosis of
  - Cystic fibrosis
  - Haemophilia
  - Pulmonary arterial hypertension
  - Pancreatic cancer
  - Sickle cell disease
- Age at first diagnosis.
- Time from first diagnosis to death.

Data sources

1. Clinical Practice Research Datalink GOLD (CPRD GOLD), United Kingdom
2. Integrated Primary Care Information Project (IPCI), The Netherlands
3. Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària (SIDIAP), Spain

Sample size

No sample size has been calculated. All patients with the disease were included.

Data analyses

Population-level disease epidemiology:

For each disease of interest, the complete point prevalence at the middle of the study period, i.e. 01/01/2016, was calculated. For each patient, the first diagnosis of the diseases of interest was considered. The disease was considered to stay present during patient's lifetime. For the point prevalence, all persons with a diagnosis before 01/01/2016 and in observation in the database at this date contributed to the numerator. The denominator was the total number of persons in observation on 01/01/2016.

The incidence rate of each disease was estimated over the total study period. Only newly diagnosed patients, diagnosed within the observation time and within the study period, contributed to the numerator. The denominator was the total number of person-years at risk, i.e. the sum across all subjects in the population of observation times falling within the study period and before a diagnosis of the disease of interest, if present.

Survival estimation:

Kaplan Meier survival probabilities for time since first diagnosis were estimated. Median survival, as a proxy for median disease duration, was extracted as the time point where the survival probability decreased below 50%.

For all analyses a minimum cell counts of 5 was used when reporting results, with any smaller counts reported as "<5". Counts of zero are reported.

From the incidence rate and median disease duration, the indirect prevalence was calculated, with 95% CI.

Results from the databases were combined using random effects meta-analysis.

Results

Direct estimated point prevalences were obtained, with results heterogeneous between the databases for most disease: for Cystic fibrosis between 12.1 and 21.2 per 100,000, for Haemophilia between 9.4 and 21.0 per 100,000, for Pancreatic cancer between 11.5 and 48.2 per 100,000, and for Sickle cell disease prevalences between 28.4 and 53.1 per 100,000. Point prevalences of PAH were consistent but could only be obtained in CPRD GOLD (39.2 per 100,000) and SIDIAP (38.2 per 100,000).

Incidence rates (IR) were also heterogeneous: for Cystic fibrosis between 0.5 and 1.6 per 100,000 PY, for Haemophilia between 1.1 and 2.1 per 100,000 PY, for PAH between 5.5 and 7.1 per 100,000 PY, for Pancreatic cancer between 10.6 and 20.8 per 100,000 PY, 95% CI 20.5 to 21.2), for Sickle cell disease, between 3.4 and 5.1 per 100,000 PY.

Median disease duration could only be estimated for some diseases. For Cystic fibrosis in IPCI it was estimated as 59.0 years (calculation of CI not possible). Median survival time for PAH estimated as 4.3 years in SIDIAP and 5.2 years in CPRD GOLD. For Pancreatic cancer, median survival time was between 0.33 years (4 months) and 1.10 years (13 months).

The indirect estimated prevalence for Cystic fibrosis in IPCI was 30.7 per 100,000 (no CI available), compared to the direct estimated prevalence of 12.1 (95% CI 10.2 to 14.2). For PAH, indirect estimated prevalences were lower than direct estimated prevalences: in SIDIAP 23.3 per 100,000 (95% CI 22.0 to 24.7) compared to 38.2 per 100,000 (95% CI 36.6 to 39.8) and in CPRD GOLD 36.8 per 100,000 (95% CI 35.1 to 38.7) compared to 39.2 per 100,000 (95% CI 37.5 to 41.0). For Pancreatic cancer the differences between indirect and direct estimated prevalence were even larger: in CPRD GOLD 3.5 per 100,000 (95% CI (3.4 to 3.7) compared to 11.5 per 100,000 (95% CI 10.6 to 12.5), in SIDIAP 17.5 per 100,000 (95% CI 16.9 to 18.1) compared to 42.1 per 100,000 (95% CI 40.5 to 43.8) and in IPCI 22.7 per 100,000 (95% CI 20.6 to 24.9) compared to 48.2 per 100,000 (95% CI 44.3 to 52.3).

Meta-analysis of the measures showed high heterogeneity. For this reason, no calculation of indirect estimated prevalence based on meta-analysis results of IR and median disease duration was done.

Discussion

Using the included EHR data sources, with limited longitudinal observation periods for individuals, it is not possible to estimate median disease duration for the rare diseases with relatively long-life expectancy included in this study. This hampered the calculation of indirect estimated prevalence from IR and median disease duration.

For diseases with shorter life expectancy, like in this study PAH and Pancreatic cancer, the indirect estimated prevalence may not always align with the direct estimated prevalence. For PAH in SIDIAP this might be related to the observed change in incidence over time. However, estimates of both incidence and median disease duration rely on the correct date of first diagnosis. This might be questionable because of the limited duration of observation periods of patients in the databases. The diseases regarded in this study are considered to be recorded for all patients, but the date of first recording might not always be the first diagnosis date. This will affect the estimation of both incidence and median disease duration. Recording of mortality during the observation periods of the patients is expected to be quite good, while a GP will know or be informed about date of death of the patients in the practice. However, complete recording cannot be assured. Moreover, for patients leaving the practice, independence with the time to death can be questionable, e.g. for patients moving to a nursing home.

4. LIST OF ABBREVIATIONS

Acronyms/terms	Description
CDM	Common Data Model
CI	Confidence Interval
CPRD GOLD	Clinical Practice Research Datalink GOLD
DARWIN EU®	Data Analysis and Real-World Interrogation Network
EHR	Electronic Health Records
EMA	European Medicines Agency
GP	General Practitioner
IP	Inpatient
IPCI	Integrated Primary Care Information Project
IR	Incidence rate
NA	Not applicable
NL	the Netherlands
OMOP	Observational Medical Outcomes Partnership
OP	Outpatient
RWD	Real-World Data
PAH	Pulmonary Arterial Hypertension
PY	Person years
SIDIAP	Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària
SP	Spain
TBC	To be confirmed
UK	United Kingdom

5. AMENDMENTS AND UPDATES

Number	Date	Section of study protocol	Amendment or update	Reason

6. MILESTONES

STUDY SPECIFIC DELIVERABLE	TIMELINE (planned)	TIMELINES (actual)
Draft Study Protocol		18/01/2024
Final Study Protocol		02/04/2024
Creation of Analytical code	19/02/2024	28/03/2024
Execution of Analytical Code on the data	26/02/2024	04/04/2024
Interim Study Report (if applicable)	Not applicable	Not applicable
Draft Study Report	28/03/2024	26/04/2024
Final Study Report	22/04/2024	
Draft Manuscript (if agreed on)		
Final Manuscript (if agreed on)		

7. RATIONALE AND BACKGROUND

Prevalence of a disease or condition is defined as the proportion of individuals in a population affected by a condition at a given point in time. Quantifying disease prevalence is important from a public health perspective, e.g. to understand the impact of diseases on the population, or to plan and allocate health care resources. Measuring disease prevalence is also important from a medicine regulatory viewpoint, as regulatory agencies grant incentives for the development of new therapies for rare diseases, thus diseases with low prevalence.

If the number of people with the disease is known in a population of known size, direct estimation of the prevalence proportion is straightforward. This holds for a sample of the population, provided the sample is representative of the population. For chronic diseases, complete prevalence, i.e. counting all individuals ever diagnosed with the disease, is typically of interest.

Disease prevalence depends on the rate of incidence of the disease in the population as well as on the average duration of the disease. Under the assumption that both the incidence of the disease and its average duration are stable over time, a well-known mathematical relationship between prevalence (P), incidence (I) and average duration (D) is(Rothman 2012):

$$\frac{P}{1-P} = I \cdot D$$

For diseases with relatively low prevalence, $(1-P) \approx 1$ and the above expression reduces to:

$$P = I \cdot D$$

Application of this formula can be useful for example when the prevalence is unknown but where the incidence can be estimated from diagnoses in hospitals, and using assumptions for the duration of disease (Kristjansdottir, Rafnsson et al. 2023), or where input from different sources is combined (Willey, Coppo et al. 2023).

This expression can be used to obtain an indirect estimation of disease prevalence from estimates of the disease incidence and mean (or median) duration, provided the following assumptions hold:

- The prevalence is relatively low
- Disease incidence is stable over time
- Disease duration is stable over time

For chronic diseases without cure, the value of D used can correspond to the median survival time after diagnosis. For non-chronic diseases, the value of D used can correspond to the median time from diagnosis to cure.

In recent years, real-world data (RWD) sources, particularly from primary care, have been used to estimate the prevalence of chronic diseases. The rationale behind this is that the population included in these databases can be considered a representative sample of the general population. The same reasoning has been used to produce incidence figures as well as estimations of disease duration (e.g. survival) using this type of sources.

There is uncertainty, however, around how direct and indirect methods to estimate prevalence agree with each other, both in situations where the assumptions underpinning the indirect method hold, the degree to which a chronic disease is truly life-long, as well as in settings where they can be more questionable (e.g. because incidence and or disease duration evolved over time). This study aims at addressing this question in the context of using RWD sources.

8. RESEARCH QUESTION AND OBJECTIVES

The objective of this study was to compare direct and indirect estimations of prevalence of some rare, chronic diseases. This comparison was done for the following diseases:

- Cystic fibrosis
- Haemophilia (A and B)
- Pulmonary arterial hypertension
- Pancreatic cancer
- Sickle cell disease

For each of these diseases the first diagnosis of each person in the population was used. In addition, a distinction was made between first diagnoses occurring during childhood (age below 18 years) and diagnoses with first occurrence within a person at adult age (age >= 18 years) (except for Pancreatic cancer, for which only diagnoses during adulthood are considered).

9. RESEARCH METHODS

14.1 Study Type and Study Design

Table 9-1 shows the study types and designs for this study, both for the population-level analyses (incidence and prevalence) and the patient-level analyses (disease duration in patients diagnosed).

Table 9-1. Description of Potential Study Types and Related Study Designs.

STUDY TYPE	STUDY DESIGN	STUDY CLASSIFICATION
Population-level descriptive epidemiology	Population-level cohort	Off the shelf
Patient-level characterisation	Cohort analysis	Off the shelf

Incidence rate and complete point prevalence will be estimated in the total population of the databases.

Disease duration (approximated with the median survival time from first diagnosis until death) will be estimated in patients with the disease.

14.2 Study Setting and Data Sources

For this study, we considered as suitable data sources those including individuals who can be considered as a representative sample of the general population and have the potential for long observation periods for subjects. Therefore, primary care data sources were used but no hospital data sources. Lifelong observation of patients is not available in any of the data sources within the DARWIN EU© network at the time of initiating this study. However, patients in primary care databases will often have several years of observation time. Also, history of diagnoses before subject's observation time start might be recorded. In addition, the selected data sources needed to systematically record mortality. For each new data release of IPCI, for each practice, patterns of mortality over time are checked and practices with unreliable data are excluded. For CPRD GOLD, mortality data in England is compared with national statistics to ensure reliability. For SIDIAP, there is linkage with the regional population register.

For eligible data sources within the DARWIN EU© network, counts of initially suggested diseases were produced. This resulted in selecting the three primary care databases presented in **Table 9-2**.

Table 9-2. Description of the selected Data Sources.

Country	Name of Database	Justification for Inclusion	Health Care setting	Type of Data	Number of active subjects during study period	Data lock for the last update
United Kingdom	CPRD GOLD	Ability to systematically record occurrence of the diseases of interest, good recording of mortality	Primary care	EHR	11,655,934	23/06/2023
the Netherlands	IPCI	Ability to systematically record occurrence of the diseases of interest, good recording of mortality	Primary care	EHR	2,743,512	30/06/2023
Spain	SIDIAP	Ability to systematically record occurrence of the diseases of interest, good recording of mortality	Primary care	EHR	8,065,563	30/06/2023

EHR: Electronic Health Records

The databases contain data of patients collected during the period they are registered in a GP practice. The median duration of the observation periods over all individuals in the database is less than 5 years in IPCI, 6 years in CPRD GOLD and 15 years in SIDIAP. In IPCI, diagnoses before observation time are present for part of the GP practices, but there is no systematic recording of historical diagnoses. For time after a patient's observation period, no information (e.g. death date) is available.

For all data sources in the DARWIN EU© network, the data partners are asked to describe their internal data quality process on the source data as part of the onboarding procedure. In addition, they are asked to share the results from three data quality assurance package: CdmOnboarding, Data Quality Dashboard (DQD) and DashboardExport. The latter exports a subset of analyses from the Achilles tool (https://github.com/OHDSI/Achilles), which systematically characterizes the data and presents it in a dashboard format to ease the detection of potential quality issues. The generated data characteristics such as age distribution, condition prevalence per year, data density, measurement value distribution can be compared against national healthcare data. CdmOnboarding creates a report with select characterisation of the clinical data within the database and with details on mapping coverage statistics that are closely

inspected upon onboarding. DQD provides more objective checks on conformance and plausibility, applied consistently across the data sources.

All data sources in the DARWIN EU© network use the OMOP Common Data Model format.

14.3 Study Period

The study period was from January 1st, 2010, to December 31st, 2022.

14.4 Follow-up

For all individuals in the databases, the observation period is recorded, i.e. the period during which the individual is monitored. In the primary care databases used in this study, it is the period the individual is registered in the GP practice.

- For the incidence estimation, the follow-up used was the patient's observation period that overlapped with the study period. Only first diagnoses which fell within this follow-up period contributed to the incidence numerator. The follow-up period contributed time at risk to the person-years denominator, restricted to the time before the first diagnosis of the relevant disease, if present.
- For the direct estimation of the point prevalence, a disease diagnosed before an individual's observation period, if present, was also captured. Diagnoses before the observation period could be present if a GP had received information about the patient's history from the former GP, or if a GP had entered historical information, e.g. the diagnosis date of an inherited disease. However, individuals only contributed to the point prevalence numerator and denominator if the timepoint used to assess the point prevalence was within their observation period. Allowing events to be included before the observation period would have introduced bias as individuals without historical events cannot contribute to the numerator or the denominator. Using only data within patients' observation period is in line with the data used for the calculation of incidence.
- In the survival analysis, the follow-up of patients started at the first diagnosis of the disease of interest and ended at patient's death, end of patient's observation period or end of the study period, whatever came first.

In the databases used, all recorded individuals have only one consecutive observation period, meaning that for each analysis they provided one follow-up period. If an individual had changed GP practice, and both practices contribute data to the database, he/she was registered with a different patient ID.

14.5 Study Population with in- and exclusion criteria

For the incidence and prevalence estimations, the complete database population was included.

For disease duration, all patients with the diagnosis of interest were included, without any exclusion criteria. Diagnoses could be recorded during the observation period of the patient in the database or in

history, and could be done in primary, inpatient or outpatient care setting. In addition to disease cohorts with patients with a diagnosis at any age, for each disease (except for Pancreatic cancer) we distinguished patients with first diagnosis at paediatric age (age at diagnosis below 18 years) and patients with first diagnosis at adult age (age at diagnosis at age 18 or older). For Pancreatic cancer only diagnoses during adulthood were considered.

The SNOMED codes used for selecting the diagnoses are given in **Appendix I - Table 1**.

14.6 Variables

9.1.1 Outcomes

The diseases in this study are all chronic. For each disease, the first recorded occurrence of the diagnosis within an individual was selected. The codes (concept ids) used for selection of the diagnoses are listed in **Appendix I - Table 1**.

Mortality was assessed from the death table in the OMOP CDM. Date of death was recorded by the GP or obtained from additional documents (e.g. letters from hospitals). Recording of death is limited to the observation time of the patient, or shortly (within some weeks) afterwards. Individuals without date of death are censored at the end date of their observation time.

For SIDIAP, there is linkage to the regional population register.

9.1.2 Other covariates

Age at diagnosis was calculated using the individual's date of birth and the date of diagnosis.

14.7 Study size

No sample size calculation was done.

14.8 Data transformation

Analyses of direct estimated prevalence, incidence and disease duration were conducted separately for each database. Before study initiation, test runs of the analytics were performed on a simulated set of patients and quality control checks were performed. After all the tests were passed, the final 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 results.

The study results of all data sources were checked after they were made available to the DARWIN EU® Coordination Centre. All results were locked and timestamped for reproducibility and transparency.

Indirect estimation of prevalence and meta-analysis of database results was done at the DARWIN EU® Coordination Centre.

14.9 Statistical Methods

9.1.3 Main Summary Measures

Characteristics of patients in the disease cohorts were described with counts and percentages or with median and inter-quartile range.

9.1.4 Main Statistical Methods

Patient characteristics, namely age at diagnosis and sex, were extracted using the DARWIN EU® R package PatientProfiles (Català, Guo et al. 2024).

Point prevalences were estimated at January 1st of calendar years 2010 up to 2022. At each time point the numerator was the number of individuals in observation with prevalent disease and the denominator was the total number of individuals in observation.

Figure 1 shows some examples of individuals with their observation period and the time point of first diagnosis of a disease. Focus is at the point prevalences on January 1st, 2010, and January 1st, 2016. The latter was used for the comparison with the indirect estimated prevalence.

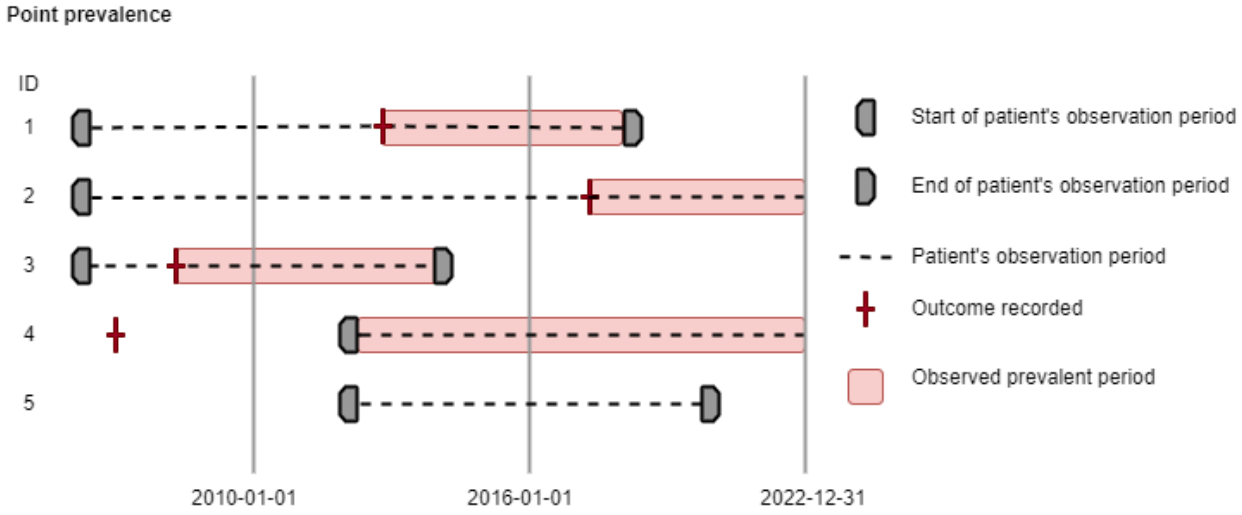


Figure 1. Example patients for point prevalence estimation.

For patient 1, observation time started before January 1st, 2010. Date of diagnosis was on January 1st, 2013. Observation time ended after January 1st, 2016. Because both January 1st, 2010, and January 1st, 2016, fall within the observation period, this patient contributed 1 observed person to the denominators of the point prevalence calculations of both time points. To the numerator for January 1st, 2010, 0 events are contributed, and to the numerator for January 1st, 2016, 1 event is contributed.

For patient 2, observation time started before January 1st, 2010. Date of diagnosis was on January 1st, 2017. Observation time was ongoing at the end of the study period. Both timepoints are within the observation period of this patient and no diagnosis was present. This means this patient contributed 0 events to both numerators and 1 observed person to both denominators.

For patient 3, observation time started before January 1st, 2010. Date of diagnosis was during the observation period and before January 1st, 2010. Observation time ended on January 1st, 2014. This patient contributed 1 event to the numerator and 1 observed person to the denominator for January 1st, 2010. This patient does not contribute to the calculation for January 1st, 2016.

Patient 4 had a diagnosis before January 1st, 2010. Observation time only started on January 1st, 2012, and was ongoing at the end of the study period. This patient did not contribute to the calculation for January 1st, 2010, and contributed 1 event to the numerator and 1 observed person to the denominator for January 1st, 2016.

For individual 5, observation time started on January 1st, 2012, and ended on January 1st, 2019. There was no diagnosis for this individual. This individual contributed 0 events to the numerator and 1 observed person to the denominator for January 1st, 2016.

Incidence rates were estimated for each calendar year and for the total study period. Denominators were the total number of person-years at risk, i.e. observation time of a person within the study period, until a diagnosis occurs, or observation time ends. To the numerator, only first diagnoses within the observation time of a patient (and within study period or in the respective calendar year) contributed.

Figure 2 shows some examples of individuals with their observation period and the time point of first diagnosis of a disease. Calendar years 2010 up to 2013 are shown and some period before 20210.

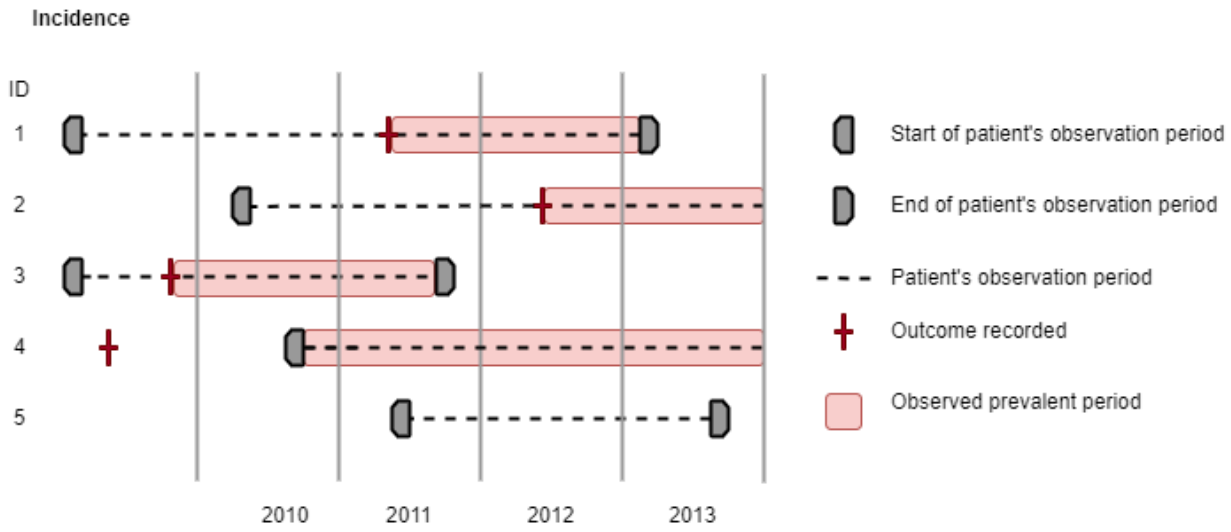


Figure 2. Example patients for incidence estimation.

For patient 1, observation time started before 2010. Date of diagnosis was in 2011. Observation time ended in 2013. This patient contributed a full year person time to the denominator for 2010 and part of a year, from January 1st until date of diagnosis, to the denominator for 2011. After diagnosis, this person is no longer at risk for incident diagnosis. The diagnosis contributed one event to the numerator for 2011. This patient did not contribute to incidence calculations of 2012 and later.

For patient 2, observation time started in 2010. Date of diagnosis was in 2012. Observation time was ongoing at the end of the study period. This patient contributed part of a year, from start of observation time until December 31st, to the denominator for 2010, a full year to 2011 and part of a year, from January 1st until date of diagnosis, to 2012. The diagnosis contributed one event to the numerator for 2012. This patient did not contribute to incidence calculations of 2013 and later.

For patient 3, observation time started before 2010 and date of diagnosis was also before 2010. This patient did not contribute to any of the incidence calculations for 2010 and later, because of the preceding diagnosis after which the patient is no longer at risk for a first diagnosis.

Patient 4 had a (historical) diagnosis before 2010 and observation time started in 2010. Again, the early diagnosis meant that the patient was no longer at risk at the start of the study period and consequently did not contribute to any of the incidence calculations.

Patient 5 had no diagnosis and observation time started in 2011 and ended in 2013. This patient did not contribute to the incidence calculation of 2010, contributed time from start of observation time until December 31st to the denominator for 2011, a full year to 2012 and part of a year, from January 1st until end of observation time, to 2013.

The estimations of direct estimated point prevalence and incidence were done as described in the DARWIN Complete Catalogue of Standard Data Analyses, using the DARWIN EU® R package IncidencePrevalence (Burn, Raventós et al. 2023, Raventós, Català et al. 2024). The specifications used in this study for these analyses are reported in **Table 9-3**.

Table 9-3. Specifications for standard analyses.

ANALYSIS	PARAMETER	SETTINGS
Point prevalence	Cohort date range	01/01/2010 to 31/12/2022
	Age	All
	Sex	Both
	Interval	Years
	Timepoint	Start
Incidence rates of diseases	Cohort date range	01/01/2010 to 31/12/2022
	Age	All
	Sex	Both
	Interval	Years, overall
	Outcome washout	Infinite
	Repeated events	False

To estimate the median disease duration, for each individual with a disease diagnosis, the time between diagnosis and end of follow-up was determined. For individuals with mortality, the end of follow-up was the date of death, and their survival status was 'event'. For individuals without mortality their survival status was ´censored´ at the end of follow-up, the date of the end of their observation period or the end of the study period. Censoring at the end of the study period is administrative censoring. Censoring at the end of observation period is administrative censoring if this end date represents the end of data contribution by

the GP, in this case end of observation time is the same for all patient in this practice. Censoring is non-administrative if the end date is the date at which the patients was leaving the practice. Median survival time, i.e. median disease duration, was determined as the timepoint where the Kaplan-Meier curve crossed the 50% survival rate. The limits of the 95% CI for the median were determined as the timepoints where curves of the 95% CI of survival rates crossed the 50% line (Brookmeyer and Crowley 1982). For the survival analysis, the R package `survival` was used.

The indirect estimated prevalence was calculated using the overall incidence (I) and median disease duration (D):

$$P = I \cdot D \tag{1}$$

The CI for this indirect estimated prevalence was generated by using the relation on the log scale:

$$\log P = \log(I \cdot D) = \log I + \log D \tag{2}$$

hence for the variance

$$Var(\log P) = Var(\log I) + Var(\log D) + 2 \cdot Cov(\log I, \log D) \tag{3}$$

The variance of $\log I$ was calculated as:

$$Var(\log I) = 1/C \tag{4}$$

where C is the total number of cases used for the calculation of the incidence I.

The variance of $(\log D)$ was approximated using the limits D_{low} and D_{upp} of the 95% CI for D:

$$Var(\log D) = \left((\log D_{upp} - \log D_{low})/(2 \cdot 1.96)\right)^2 \tag{5}$$

Incidence rate and median disease duration were assumed to be independent, hence.

$$Cov(\log I . \log D) = 0 \tag{6}$$

With the obtained variance of $Var(\log P)$ the 95% CI for $\log P$ was calculated and next converted to a CI for P.

For the indirect estimated prevalence, the incidence rate over the total study period was used.

For comparing direct and indirect estimated prevalence, the direct estimated point prevalence in the center of the study period (January 1st, 2016) was chosen.

9.1.5 Missing Values

In the database, being EHR from clinical practice, it might be that disease diagnoses during the observation time of an individual are missed. It might also be that a disease is not recorded at the date of first diagnosis, because this date was before patient's observation time. In the IPCI database it is possible to record historical data on disease diagnosis, but there is no direct instruction for the GP to enter the exact first date (if known). In SIDIAP and CPRD GOLD it is not possible to record a date of diagnosis outside observation time. The size of this missingness is unknown.

Death during or at the end of the observation period of a patient could be missing. As part of quality measures, the pattern of recording of death in a practice is investigated, and when obvious irregularities are seen, e.g. clearly too low mortality, or sudden change in mortality patterns, a practice is excluded from the database. However, this does not exclude that there might be some some patients (although limited

because otherwise would have been identified when assessing quality of reporting) for which death has not been recorded. This type of missingness cannot be detected.

The follow-up of patients for whom no date of death is recorded was either administratively or non-administratively censored. Administrative censoring happened at the end of the study period (December 31st, 2022). If the end of a patient's observation period is because of the end of the data supply period of the vendor, this is also administrative censoring. Reason for non-administrative censoring is the individual leaving the GP practice because of moving outside the area covered by the GP practice or changing to another GP practice. In SIDIAP, if a patient switches to another GP in the region, the two periods are recorded as one continuous observation period. In IPCI and CPRD GOLD, if an individual had left the GP practice, it is possible that he/she was registered in another GP practice who also contributes data to the database, but because of the limited coverage of practices in both databases, this is not very likely. If the new practice is also in the database, the patient re-entered the database. However, for practical and privacy reasons it is not possible to connect data of the same person from different GP practices. The individual will be registered with a different patient ID and the observation periods will be different by GP practice (i.e. thus no overlap).

In the databases it is not possible to know the reason for non-administrative censoring of a person. Incidence rates implicitly assume non-administrative censoring occurred completely at random. For the survival analysis all censoring was handled as censored survival time in the Kaplan-Meier estimations.

See 14.2 Limitations of the research methods, for further discussion.

9.1.6 Sensitivity Analysis

No sensitivity analyses were done.

9.1.7 Deviations from the Protocol

The results from the databases showed large heterogeneity for most estimated IRs and median disease durations. This hampered obtaining reliable meta-analysis results of these measures. Therefore, the calculation of prevalence using the meta-analysis result of an IR and the meta-analysis result of a median disease duration was not done.

10. DATA MANAGEMENT

10.1 Data management

All databases used in this study are mapped to the OMOP common data model. This enables the use of standardised analytics and tools across the network since the structure of the data and the terminology system is harmonised. The OMOP CDM is 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. Each data partner executed the study code against their database containing patient-level data and returned the results set 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 and used for the further analyses.

10.2 Data storage and protection

For this study, participants from various EU member states processed personal data from patients which is collected in national/regional electronic health record databases. 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 databases 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 are adhered to. In agreement with these regulations, rather than combining person level data and performing only a central analysis, local analyses were run, which generated non-identifiable aggregate summary results.

11. QUALITY CONTROL

General database quality control

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

Study specific quality control

The study code was based on R packages developed within DARWIN EU©. These packages include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R packages were made publicly available via GitHub.

The complete study code for this study was tested using test data to ensure that the correct output was created.

12. RESULTS

The numbers of cases are reported for all diseases, in total and split into paediatric and adult diagnoses, in **Table 12-1**. Other tables and figures in this section show only results for the disease cohorts with all diagnosed individuals and thus not categorized by age. Results split into paediatric and adult diagnoses are given in Appendix III.

12.1 Participants

Table 12-1 reports the total numbers of individuals in the databases, the numbers with observation time during the study period and for each of the diseases of interest, the number of cases diagnosed.

Table 12-1. Numbers of individuals and numbers of first diagnoses in the databases.

		CPRD GOLD		IPCI		SIDIAP	
Total		17,267,137		2,817,331		8,553,325	
Active during study period		11,655,934	(68%)	2,743,512	(97%)	8,065,563	(94%)
Cystic fibrosis	All diagnoses	3,369		326		1,446	
	Paediatric diagnoses	1,188	(35%)	165	(51%)	478	(33%)
	Adult diagnoses	2,181	(65%)	161	(49%)	968	(67%)
Haemophilia	All diagnoses	1,575		711		2,130	
	Paediatric diagnoses	506	(32%)	235	(33%)	470	(22%)
	Adult diagnoses	1,069	(68%)	476	(67%)	1,660	(78%)
PAH[1]	All diagnoses	8,118		NA		6,712	
	Paediatric diagnoses	281	(3%)	NA		374	(6%)
	Adult diagnoses	7,837	(97%)	NA		6,338	(94%)
Pancreatic cancer	All diagnoses	13,269		4,029		19,900	
Sickle cell disease	All diagnoses	5,753		2,142		5,253	
	Paediatric diagnoses	2,008	(35%)	1,045	(49%)	2,599	(49%)
	Adult diagnoses	3,745	(65%)	1,097	(51%)	2,654	(51%)

PAH: Pulmonary arterial hypertension. NA: Not available.

[1] Diagnoses of Pulmonary arterial hypertension (PAH) could not be extracted in IPCI because there is no disease code available in the disease vocabulary used in the source data.

In IPCI, part of the first diagnoses of the diseases was before the observation time of the patients in the databases. This applied for 73% (238 out of 326) of the diagnoses of Cystic fibrosis, for 69% (494 out of 711) of Haemophilia, for 23% (910 out of 4029) of Pancreatic cancer and for 73% (1554 out of 2142) of Sickle cell disease diagnoses. In the other databases, no diagnoses before observation time were present.

12.2 Descriptive Data

In **Table 12-2** the disease cohorts are described regarding sex and age at time of first recorded diagnosis.

Table 12-2. Descriptive statistics on age and sex for patients in the disease cohort.

Disease		CPRD GOLD	IPCI	SIDIAP
Cystic fibrosis	Total	3,369	326	1,446
	Female	2,170 (64.4%)	164 (50.3%)	771 (53.3%)
	Male	1,199 (35.6%)	162 (49.7%)	675 (46.7%)
	Age at diagnosis[1]	27 [9, 42]	17 [0, 31]	32 [9, 55]
Haemophilia	Total	1,575	711	2,130
	Female	483 (30.7%)	341 (48.0%)	907 (42.6%)
	Male	1,092 (69.3%)	370 (52.0%)	1,223 (57.4%)
	Age at diagnosis	32 [13, 54]	28 [10, 48]	38 [20, 58]
PAH	Total	8,118	NA	6,712
	Female	4,808 (59.2%)		4,150 (61.8%)
	Male	3,310 (40.8%)		2,562 (38.2%)
	Age at diagnosis	75 [64, 82]		77 [66, 83]
Pancreatic cancer	Total	13,269	4,029	19,900
	Female	6,666 (50.2%)	2,009 (49.9%)	9,433 (47.4%)
	Male	6,603 (49.8%)	2,020 (50.1%)	10,467 (52.6%)
	Age at diagnosis	73 [64, 80]	70 [62, 78]	73 [63, 81]
Sickle cell disease	Total	5,753	2,142	5,253
	Female	3,526 (61.3%)	1,288 (60.1%)	2,710 (51.6%)
	Male	2,227 (38.7%)	854 (39.9%)	2,543 (48.4%)
	Age at diagnosis	27 [7, 37]	19 [0, 33]	18 [4, 38]

[1] For Age at diagnosis, median and inter-quartile range are presented.
PAH: Pulmonary arterial hypertension. NA: Not available.

Inequal presence of sexes was seen for Cystic fibrosis in CPRD GOLD (more females), for Haemophilia in CPRD GOLD and SIDIAP (more males), for PAH (more males) and for Sickle cell disease (more females in all 3 databases). Sex distribution, especially for haemophilia, was not in line with literature where predominance in males described. Further interpretation is added to the interpretation section of the discussion.

The diagnoses of Cystic fibrosis, Haemophilia and Sickle cell disease are made at younger age (medians between 17 and 38 year) compared to the diagnoses of PAH and Pancreatic cancer (medians between 70 and 77 year).

Descriptive statistics regarding sex and age at diagnosis for the paediatric and adult diagnosis cohorts are given in **Table 1 in Appendix III**.

12.3 Main Results

12.3.1 Direct estimated prevalences

In Table 12-3 the point prevalences of the five selected diseases in the three databases are presented. These point prevalences are estimated for January 1st, 2016, i.e. centered within the study period, and are reported per 100,000 individuals.

Table 12-3. Direct estimated point prevalences on January 1st, 2016.

	Database	N of prevalent cases	N in population	Prevalence per 100,000	95% CI
Cystic fibrosis	CPRD GOLD	1,036	4,882,768	21.2	(20.0; 22.5)
	IPCI	139	1,152,605	12.1	(10.2; 14.2)
	SIDIAP	752	5,835,354	12.9	(12.0; 13.8)
Haemophilia	CPRD GOLD	459	4,882,768	9.4	(8.6; 10.3)
	IPCI	242	1,152,605	21.0	(18.5; 23.8)
	SIDIAP	987	5,835,354	16.9	(15.9; 18.0)
PAH	CPRD GOLD	1,915	4,882,768	39.2	(37.5; 41.0)
	SIDIAP	2,227	5,835,354	38.2	(36.6; 39.8)
Pancreatic cancer	CPRD GOLD	560	4,882,768	11.5	(10.6; 12.5)
	IPCI	555	1,152,605	48.2	(44.3; 52.3)
	SIDIAP	2,459	5,835,354	42.1	(40.5; 43.8)
Sickle cell disease	CPRD GOLD	1,385	4,882,768	28.4	(26.9; 29.9)
	IPCI	612	1,152,605	53.1	(49.1; 57.5)
	SIDIAP	1,955	5,835,354	33.5	(32.1; 35.0)

PAH: Pulmonary arterial hypertension. CI: confidence interval

Prevalence of Cystic fibrosis was comparable between IPCI and SIDIAP (12.1 per 100,000, 95% CI 10.2 to 14.2, and 12.9 per 100,000, 95% CI 12.0 to 13.8, respectively) and higher in CPRD GOLD (21.2 per 100,000, 95% CI 20.0 to 22.5). Prevalence of Haemophilia was lowest in CPRD GOLD (9.4 per 100,000, 95% CI 8.6 to 10.3), higher in SIDIAP (16.9 per 100,000, 95% CI 15.9 to 18.0) and highest in IPCI (21.0 per 100,000, 95% CI 18.5 to 23.8). Cases of PAH were not present in IPCI as there is not a disease code for PAH in the IPCI source data.

Prevalences in the other databases were comparable (CPRD GOLD 39.2 per 100,000, 95% CI 37.5 to 41.0, SIDIAP 38.2 per 100,000, 95% CI 36.6 to 39.8). The prevalence of Pancreatic cancer was lowest in CPRD GOLD (11.5 per 100,000, 95% CI 10.6 to 12.5), higher in SIDIAP (42.1 per 100,000, 95% CI 40.5 to 43.8) and highest in IPCI (48.2 per 100,000, 95% CI 44.3 to 52.3). For Sickle cell disease, prevalences varied from 28.4 per 100,000 in CPRD GOLD (95% CI 26.9 to 29.9), 33.5 per 100,000 in SIDIAP (95% CI 32.1 to 35.0) to 53.1 per 100,000 in IPCI (95% CI 49.1 to 57.5).

Direct estimated prevalences for the paediatric and adult diagnosis cohorts are provided in Table 2 in Appendix III. Note that these are not prevalences estimated in two different strata, but different types of diagnoses are distinguished. The numbers in the population at the date of the point prevalence estimation are equal for all diagnoses (**Table 12-3**), for paediatric and for adult diagnoses (**Table 2 in Appendix III**).

The point prevalences on January 1st of each calendar year during the study period are plotted in Figure 1 of Appendix III. Note the different scales of y-axis for the different diseases. An increase over time is seen for several of the diseases, for example in CPRD GOLD for PAH, in IPCI for Haemophilia and Sickle cell disease and in SIDIAP for all diseases except for PAH.

12.3.2 Incidences

In Table 12-4 the IRs of the five diseases in the three databases are presented, estimated over the complete study period and reported per 100,000 person years.

Table 12-4. Incidences.

	Database	N of persons	N of events	PY	IR per 100,000 PY	95% CI
Cystic fibrosis	CPRD GOLD	11,654,601	1,063	64,679,359	1.6	(1.5; 1.7)
	IPCI	2,743,270	73	14,029,555	0.5	(0.4; 0.7)
	SIDIAP	8,065,116	921	76,980,486	1.2	(1.1; 1.3)
Haemophilia	CPRD GOLD	11,655,388	713	64,686,677	1.1	(1.0; 1.2)
	IPCI	2,743,016	191	14,027,993	1.4	(1.2; 1.6)
	SIDIAP	8,065,168	1,637	76,977,171	2.1	(2.0; 2.2)
PAH	CPRD GOLD	11,654,125	4,622	64,669,199	7.1	(6.9; 7.4)
	SIDIAP	8,063,748	4,211	76,962,551	5.5	(5.3; 5.6)
Pancreatic cancer	CPRD GOLD	11,655,330	6,882	64,685,533	10.6	(10.4; 10.9)
	IPCI	2,742,595	2,878	14,024,284	20.5	(19.8; 21.3)
	SIDIAP	8,064,252	16,030	76,956,938	20.8	(20.5; 21.2)
Sickle cell disease	CPRD GOLD	11,653,987	2,169	64,674,429	3.4	(3.2; 3.5)
	IPCI	2,741,989	552	14,022,689	3.9	(3.6; 4.3)
	SIDIAP	8,064,740	3,948	76,963,345	5.1	(5.0; 5.3)

PAH: Pulmonary arterial hypertension. PY: person years. IR: incidence rate. CI: confidence interval

IRs of Cystic fibrosis were below 2 per 100,000 PY in all databases: in IPCI 0.5 per 100,000 PY (95% CI 0.4 to 0.7), in SIDIAP 1.2 per 100,000 PY (95% CI 1.1 to 1.3) and in CPRD GOLD 1.6 per 100,000 PY (95% CI 1.5 to 1.7). For Haemophilia, lowest IR was observed in CPRD GOLD (1.1 per 100,000 PY, 95% CI 1.0 to 1.2), higher in IPCI (1.4 per 100,000 PY, 95% CI 1.2 to 1.6) and highest in SIDIAP (2.1 per 100,000 PY, 95% CI 2.0 to 2.2). Diagnoses of PAH were not present in IPCI. IR of PAH in SIDIAP was 5.5 per 100,000 PY (95% CI 5.3 to 5.6) and in CPRD GOLD 7.1 per 100,000 PY (95% CI 6.9 to 7.4). IRs of Pancreatic cancer were highest among the five diseases of interest in this study: 10.6 per 100,000 PY (95% CI 10.4 to 10.9) in CPRD GOLD and almost twice at high in IPCI (20.5 per 100,000 PY, 95% CI 19.8 to 21.3) and SIDIAP (20.8 per 100,000 PY, 95% CI 20.5 to 21.2). For Sickle cell disease, IRs were 3.4 per 100,000 PY (95% CI 3.2 to 3.5) for CPRD GOLD, 3.9 per 100,000 PY (95% CI 3.6 to 4.3) for IPCI and 5.1 per 100,000 PY (95% CI 5.0 to 5.3) for SIDIAP.

IRs for paediatric and adult diagnoses are presented in Table 3 in Appendix III. Note that these are not IRs estimated in two different strata, but different types of diagnoses (i.e. diagnosed at pediatric or adult age) are distinguished. The numbers of persons at risk for the disease and their person years for all, for paediatric and for adult diagnoses are close but not equal, depending on how many persons and person years are excluded because of a previous diagnosis.

The IRs per calendar year during the study period are plotted in **Figure 2 of Appendix III**. Note the different scales of y-axis for the different diseases. Except for a decrease for PAH in SIDIAP, no important change over time is seen.

12.3.3 Median disease durations

Table 12-5. Median disease durations**Table 12-5** presents for each disease the numbers and percentages of cases who died. Mortality was highest for PAH with 54.4% and 70.6% dying in CPRD GOLD and SIDIAP respectively (no PAH cases in IPCI), and for Pancreatic cancer with 66.0% dying in IPCI, 77.5% in SIDIAP and 87.6% in CPRD GOLD. Only for these two diseases, median survival time with 95% CI could be estimated from the Kaplan-Meier curve. Median survival time from a diagnosis of PAH on was 4.3 years (95% CI 4.1 to 4.5) in SIDIAP and 5.2 years (95% CI 4.9 to 5.3) in CPRD GOLD. For Pancreatic cancer, in CPRD GOLD a median survival time of 0.33 years (4 months) was estimated (95% CI 0.32 to 0.34), in SIDIAP 0.84 years (10 months, 95% CI 0.81 to 0.87 years) and in IPCI 1.10 years (13 months, 95% CI 1.01 to 1.21 years). For Cystic fibrosis, in IPCI a median survival time of 59 years was estimated, but data were too scarce to construct a 95% CI.

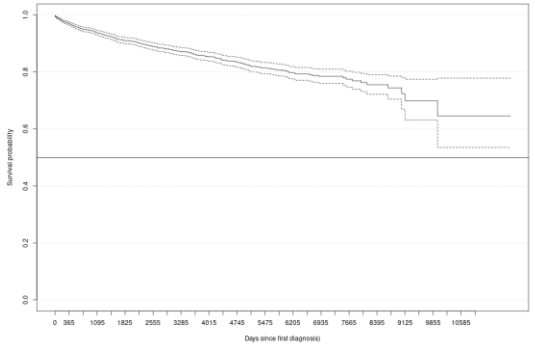
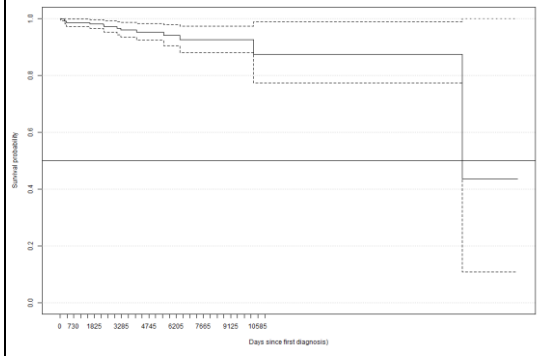
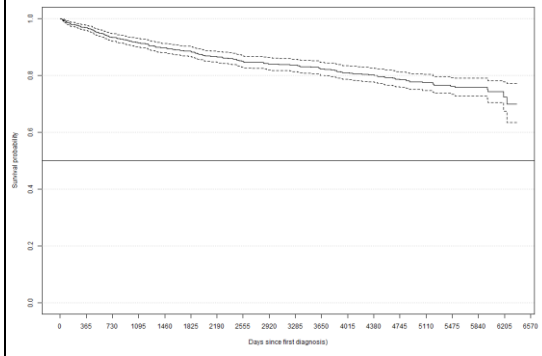
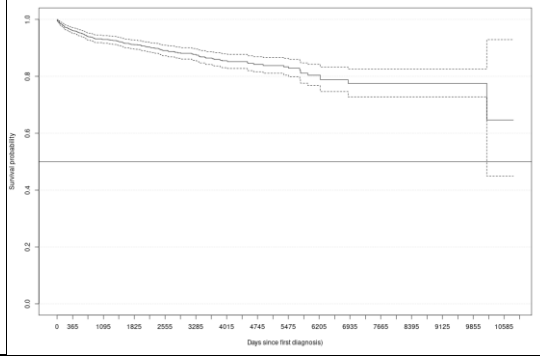
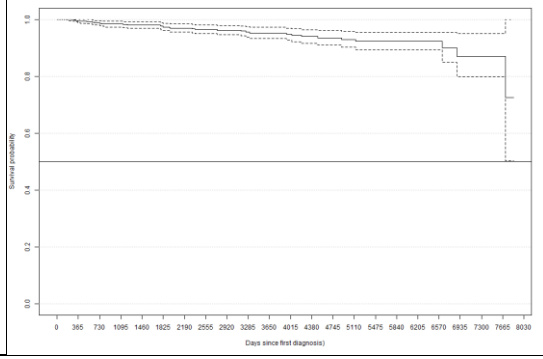
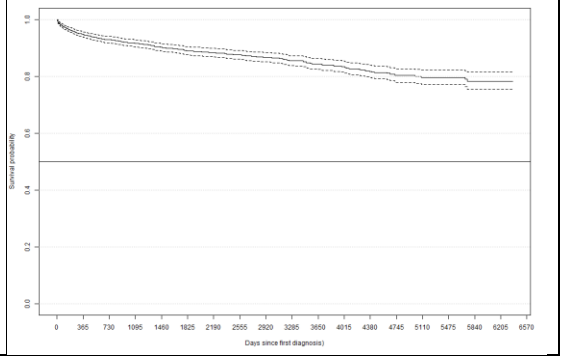
Table 12-5. Median disease durations.

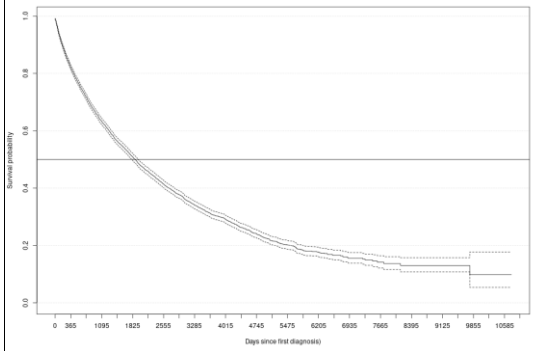
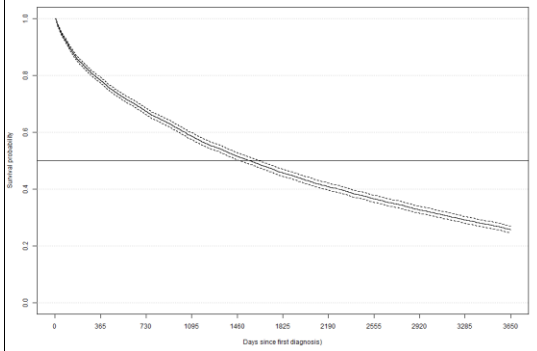
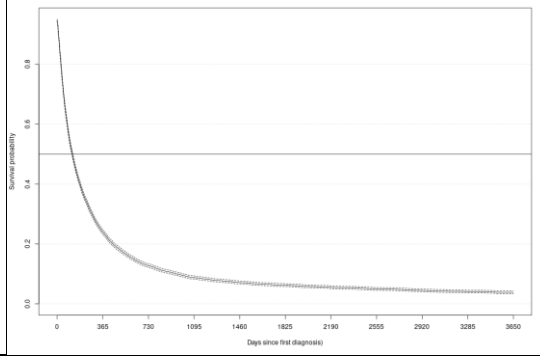
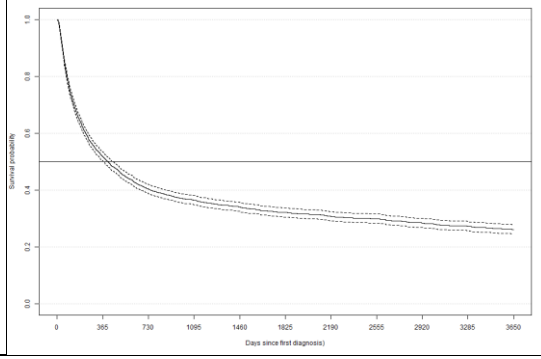
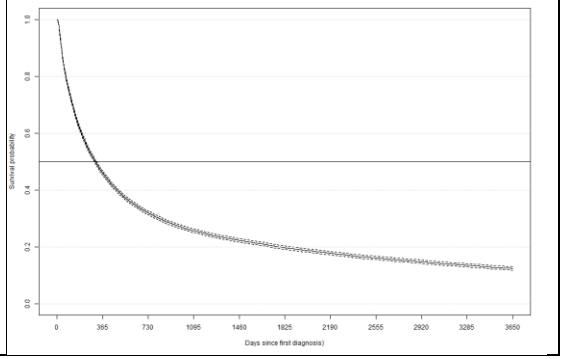
	Database	N diagnosed	Mortality	Median survival time (yr)	95% CI
Cystic fibrosis	CPRD GOLD	3,369	396 (11.8%)	NA	NA
	IPCI	326	15 (4.6%)	59.0	NA
	SIDIAP	1,446	237 (16.4%)	NA	NA
Haemophilia	CPRD GOLD	1,575	165 (10.5%)	NA	NA
	IPCI	711	33 (4.6%)	NA	NA
	SIDIAP	2,130	282 (13.2%)	NA	NA
PAH	CPRD GOLD	8,118	4,418 (54.4%)	5.2	(4.9;5.3)
	SIDIAP	6,712	4,740 (70.6%)	4.3	(4.1;4.5)
Pancreatic cancer	CPRD GOLD	13,269	11,624 (87.6%)	0.33	(0.32;0.34)
	IPCI	4,029	2,658 (66.0%)	1.10	(1.01;1.21)
	SIDIAP	19,900	15,419 (77.5%)	0.84	(0.81;0.87)
Sickle cell disease	CPRD GOLD	5,753	131 (2.3%)	NA	NA
	IPCI	2,142	25 (1.2%)	NA	NA
	SIDIAP	5,253	247 (4.7%)	NA	NA

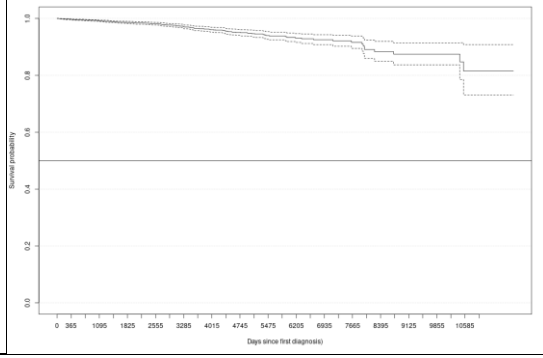
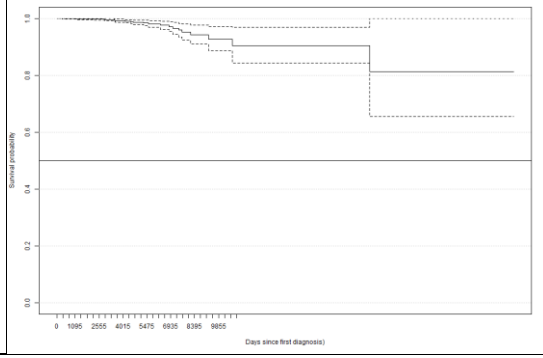
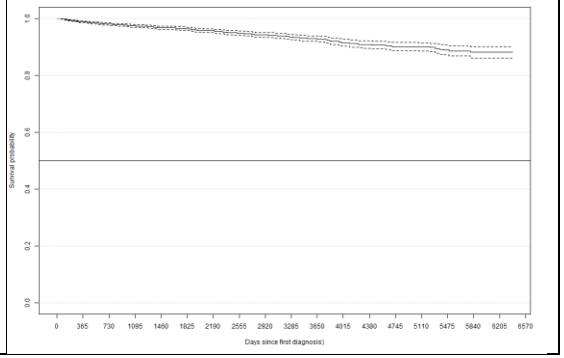
PAH: Pulmonary arterial hypertension. yr: years. CI: confidence interval. NA: Not available.

The Kaplan Meier curves with survival after diagnoses at all ages are shown in NA: Not available

Figure 3. The Kaplan Meier curves for diagnoses at paediatric age and at adult age are shown in **Figure 3 in Appendix III**.

	CPRD GOLD	IPCI	SIDIAP
Cystic fibrosis			
Haemophilia			

	CPRD GOLD	IPCI	SIDIAP
PAH		NA	
Pancreatric cancer			

	CPRD GOLD	IPCI	SIDIAP
Sickle cell disease			

NA: Not available

Figure 3. Kaplan-Meier curves for survival after diagnosis, all ages.

12.3.4 Indirect estimated prevalences

In **Table 12-6** results are given of the indirect estimation of prevalences. As for the other results shown in this chapter, these are results for the disease cohorts with diagnoses at all ages. Results are only shown for the combination of disease and data source for which the estimation was possible, i.e. where IR and median disease duration could be estimated for at least one of the databases. These are:

- Cystic fibrosis in IPCI
- PAH in CPRD GOLD and SIDIAP
- Pancreatic cancer in all three databases

Table 12-6. Results of the indirect prevalence estimations

		IR per 100,000 PY		Median disease duration (yrs)		Prevalence per 100,000 Indirect estimate		Prevalence per 100,000 Direct estimate	
	Database	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Cystic fibrosis	IPCI	0.52	(0.41; 0.65)	58.95	NA	30.67	NA	12.06	(10.21; 14.24)
PAH	CPRD GOLD	7.15	(6.94; 7.36)	5.15	(4.94; 5.35)	36.81	(35.06; 38.66)	39.22	(37.5; 41.02)
	SIDIAP	5.47	(5.31; 5.64)	4.27	(4.05; 4.47)	23.34	(22.04; 24.72)	38.16	(36.61; 39.78)
Pancreatic cancer	CPRD GOLD	10.64	(10.39; 10.89)	0.33	(0.32; 0.34)	3.53	(3.39; 3.67)	11.47	(10.56; 12.46)
	IPCI	20.52	(19.79; 21.29)	1.10	(1.01; 1.21)	22.66	(20.62; 24.89)	48.15	(44.31; 52.33)
	SIDIAP	20.83	(20.51; 21.15)	0.84	(0.81; 0.87)	17.46	(16.87; 18.08)	42.14	(40.51; 43.84)

PAH: Pulmonary arterial hypertension. PY: person years. IR: incidence rate. CI: confidence interval. yrs: years

The direct and indirect estimated prevalences with 95% CI of the three diseases are also plotted in Error! Reference source not found..

PAH: Pulmonary arterial hypertension.

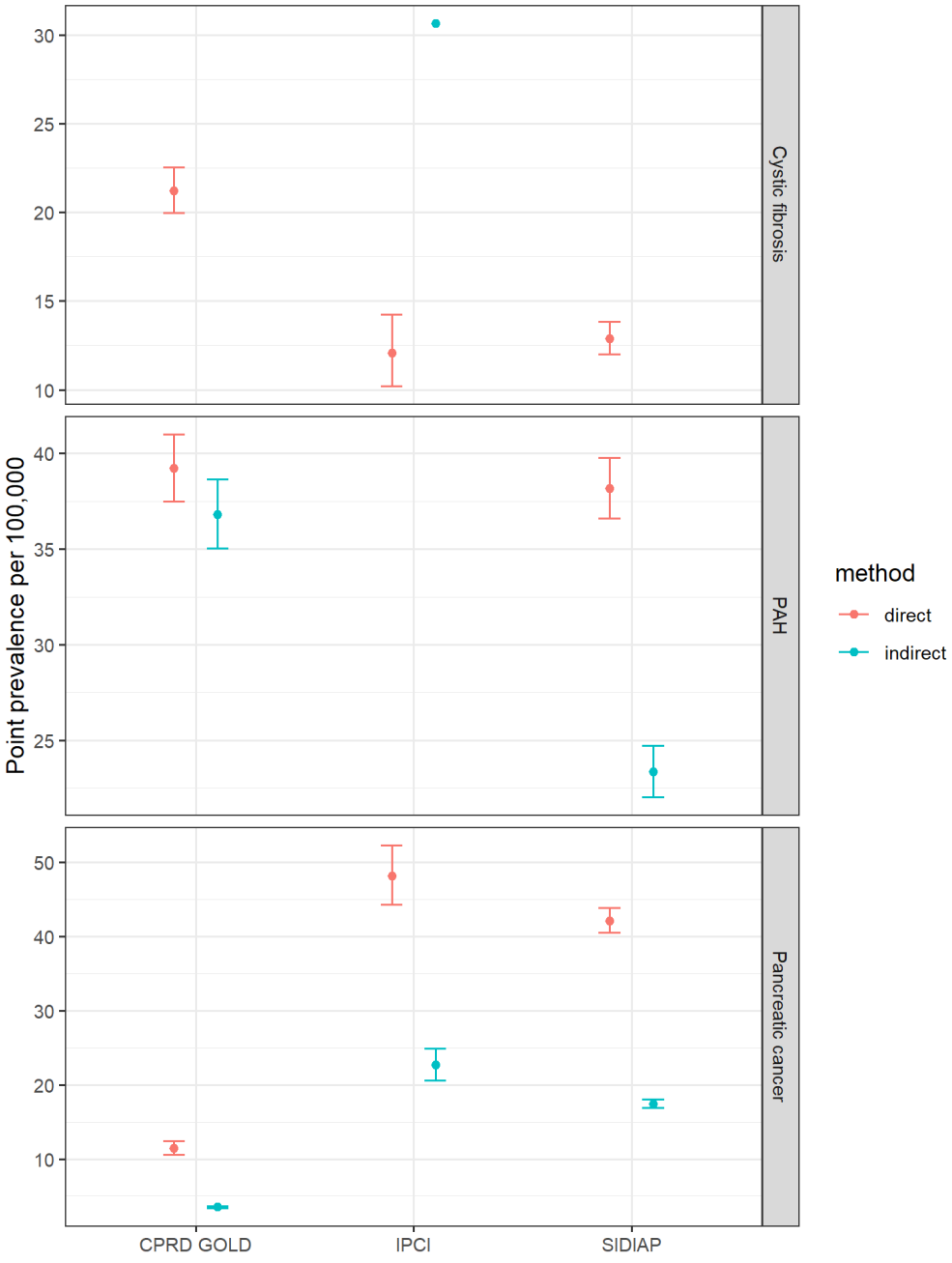


Figure 4. Direct and indirect estimated prevalences.

The indirect estimated prevalence of Cystic fibrosis in IPCI was higher than the direct estimated prevalence, namely 30.7 per 100,000 compared to 12.1 per 100,000 (95% CI 10.2 to 14.2). For the indirect estimation no CI could be calculated because a 95% CI for the median disease duration was not available.

For PAH, the indirect estimated prevalence in CPRD GOLD was some lower compared to the direct estimated prevalence, with overlapping 95% CIs: 36.8 per 100,000 (95% CI 35.1 to 38.7) versus 39.2 per 100,000 (95% CI 37.5 to 41.0). In SIDIAP the difference between indirect and direct estimation was larger. While the direct estimation was close to that in CPRD GOLD (38.2 per 100,000, 95% CI 36.6 to 39.8), the indirect estimation was 23.3 per 100,000 (95% CI 22.0 to 24.7).

For Pancreatic cancer, in all three databases the indirect estimated prevalence was lower compared to the direct estimation: for CPRD GOLD, indirect estimation was 3.5 per 100,000 (95% CI 3.4 to 3.7) compared to the direct estimation of 11.5 per 100,000 (95% CI 10.6 to 12.5), for IPCI, indirect estimation was 22.7 per 100,000 (95% CI 20.6 to 24.9) compared to the direct estimation of 48.2 per 100,000 (95% CI 44.3; 52.3), for SIDIAP, indirect estimation was 17.5 per 100,000 (95% CI 16.9 to 18.1) compared to the direct estimation of 42.1 per 100,000 (95% CI 40.5; 43.8).

The complete table with results of indirect estimated prevalences is given in **Table 5 in Appendix III**.

12.3.5 Meta-analysis of results

Forest plots reporting the meta-analyses of the direct estimated prevalences are presented in **Figure 4 in Appendix III.** For most prevalences, I^2 was above 90%, indicating that the differences between the direct estimated prevalences are likely due to true differences between the databases rather than sampling error. This was not the case for the prevalence of PAH, all diagnoses, which could be estimated in CPRD GOLD and SIDIAP and had an I^2 of 0%. The meta-analysis result was 38.7 per 100,000 persons (95% CI 37.5 to 39.8). For PAH, adult diagnoses, I^2 of direct estimated prevalences in CPRD GOLD and SIDIAP was 79%. Random effect meta-analysis result was 35.8 per 100,000 persons (95% CI 33.4 to 38.4).

Forest plots reporting the meta-analyses of the IR are presented in **Figure 5 in Appendix III**. Again, for most IRs, I^2 was above 90%. A lower I^2 was found for the IR of Haemophilia at paediatric age (I^2 = 71%, random effect meta-analysis result 0.42 per 100,000 PY, 95% CI 0.36 to 0.49).

Forest plots reporting the meta-analyses of the median disease durations are presented in **Figure 6 in Appendix III**. Median duration with standard error could only be estimated for three diagnoses: PAH at all ages, PAH at adult age and Pancreatic cancer. For all these outcomes, I^2 was 97% or higher.

Forest plots reporting the meta-analyses of the indirect estimated prevalences are presented in **Figure 7 in Appendix III**. Meta-analysis could only be done for three diagnoses: PAH at all ages, PAH at adult age and Pancreatic cancer. For all these outcomes, I^2 was 99% or higher.

Because of the large heterogeneity of most estimated IRs and median disease durations, reliable meta-analysis results of these measures were not available. Therefore, no calculation of indirect estimated prevalence using the meta-analysis result of the IR and the meta-analysis result of the median disease duration was done.

13. MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS

Adverse events/adverse reactions were not collected or analysed as part of this evaluation. The nature of this non-interventional evaluation, through the use of secondary data, does not fulfil the criteria for

reporting adverse events, according to module VI, VI.C.1.2.1.2 of the Good Pharmacovigilance Practices (https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-submission-reports_en.pdf).

14. DISCUSSION

14.1 Key Results

In this study, we attempted indirect estimation of the prevalence of five rare, chronic diseases, using IR and median disease duration, estimated in three EHR databases. Important finding was that for three of the diseases of interest (Cystic fibrosis, Haemophilia and Sickle cell disease) the observation time of the patients was not long enough to estimate the median disease duration in the included data sources. This hampered the calculation of indirect estimated prevalence.

Median disease duration could be estimated for PAH in two databases. In CPRD GOLD, the indirect and direct estimated prevalences were close: the indirect estimated prevalence was 36.8 per 100,000 (95% CI 35.1 to 38.7) and the direct estimated prevalence was 39.2 per 100,000 (95% CI 37.5 to 41.0). However, in SIDIAP the indirect estimated prevalence was much lower compared to the direct estimated prevalence: 23.3 per 100,000 (95% CI 22.0 to 24.7) versus 38.2 per 100,000 (95% CI 36.6 to 39.8).

Calculation of indirect estimated prevalence was also possible for pancreatic cancer in all three databases. The indirect estimated prevalences were lower than the direct estimated prevalence. For CPRD GOLD, the indirect estimation was 3.5 per 100,000 (95% CI 3.4 to 3.7) compared to the direct estimation of 11.5 per 100,000 (95% CI 10.6 to 12.5). For IPCI, the indirect estimation was 22.7 per 100,000 (95% CI 20.6 to 24.9) compared to the direct estimation of 48.2 per 100,000 (95% CI 44.3; 52.3). For SIDIAP, the indirect estimation was 17.5 per 100,000 (95% CI 16.9 to 18.1) compared to the direct estimation of 42.1 per 100,000 (95% CI 40.5; 43.8).

14.2 Limitations of the research methods

Data in EHR is collected for use in clinical practice and not necessarily for research purposes and so data quality issues must be considered, despite the data quality checks at time of DP onboarding. Relevant for this study is that a diagnosis recorded at a certain date might not reflect the first time a diagnosis is made. If a disease is already present when a patient registers within a GP practice, the GP might record the diagnosis at the entry date of the patient or at the date the patient visits the GP with specific complaints. In CPRD GOLD and SIDIAP no diagnosis dates before the start date of the observation time of the patient are present. This means that the first diagnosis dates may be missing. In IPCI entering a diagnosis date before observation time is possible, but it might vary between practices as it is unknown how GPs handle these dates of diagnoses in history. The medians of age at diagnoses of Cystic fibrosis, Haemophilia and Sickle cell disease (between 17 and 38 years) suggest that quite a large proportion of diagnoses might be recorded not at the first date of diagnosis.

Recording of mortality during the observation periods of the patients is expected to be quite good, while usually a GP will know or be informed about the date of death of patients in the practice. However, complete recording cannot be assured.

As mentioned before, the observation time of individuals in the databases appeared to be too short to estimate the median survival time for most diseases. In IPCI, the median duration of the observation periods over all individuals in the database is less than 5 years. In CPRD GOLD this median is 6 years and in SIDIAP 15 years. This hampers the estimation of median disease duration of diseases with long survival times.

Dates of death after a patient's observation time are not included in the databases.

Patients with no recorded date of death are censored at the latest date information is available. This can be the end of the study period for this study (December 31st, 2022) or the end of data transfer from the GP practice to the database for the database release used in this study. These types of censoring can be considered as non-informative censoring. Another reason for the end of observation time is that a patient is leaving the GP practice. For this type of non-administrative censoring, we cannot be sure it is unrelated to the (not recorded) time until death. Leaving the practice could be related to the patient's health, e.g. if the patient moves to a nursing home. The reason for the end of observation time is not recorded in the data.

Differences in direct prevalence estimates and incidence rates were observed between the databases for some of the diseases like a lower direct prevalence estimate and a lower incidence rate of pancreatic cancer in CPRD GOLD compared to SIDIAP and IPCI. These differences might be explained by true differences but might also be explained by the fact that conditions were identified based on predefined concept ids but no thorough phenotyping of the conditions of interest was done as this was a method research study and the aim was mainly to compare within databases and not across databases. Of the diseases of interest, more thorough phenotyping was done for PAH as this condition was one of the diseases of interest in a previous DARWIN EU study (https://catalogues.ema.europa.eu/node/3797/administrative-details).

The effect of changes of coding over time was observed for the incidence rate of Sickle Cell Disease which increased steeply in SIDIAP from 2021 on. SIDIAP has explored this in more detail and this higher incidence rate (compared to the beginning of the study) was still confirmed in more frequent years (2022 and 2023). According to the database this relates to optimizing coding (in the source data) of individuals with Sickle Cell Disease. In addition, this might also be the result of further implementation of the newborn screening program. (Delgado-Pecellin, Alvarez Rios et al. 2020)

We used complete point prevalence estimate of January 1st, 2016to compare the direct and indirect estimated prevalence. The point prevalence increased over time for several diseases and databases. However, in all cases the indirect estimated prevalence was outside the range of direct estimated prevalences over time.

Another limitation is that diagnoses were identified from primary care records and linkage with hospital records has not been performed and/or could not be confirmed (i.e. IPCI). However, it is unlikely that this would have affected the comparisons between methods. Also, as this was a methods study, we limited to a limited range of conditions in a limited range of data sources as at time of study start, only 3 primary care databases had been onboarded.

14.3 Interpretation

From the indirect estimated prevalences that could be calculated, for prevalence of PAH in CPRD GOLD, the results from both methods were close. For PAH in SIDIAP, indirect estimated prevalence was lower than direct estimated prevalence. However, the IR of PAH in SIDIAP did not fulfil the assumption of stability over time. The IR decreased from 13.5 per 100,000 PY (95% CI 12.6 to 14.5) in 2010 to 2.0 per 100,000 PY (95% CI 1.6 to 2.4) in 2022. This had been explored with the DP and a potential explanation could not be provided.

For Pancreatic cancer, indirect estimated prevalences were also much lower than direct estimated prevalences. In CPRD GOLD and IPCI, the IR of this disease was stable over time. In SIDIAP there was an increase from 17.3 per 100,000 PY (95% CI 16.2 to 18.4) in 2010 to 25.2 per 100,000 PY (95% CI 24.0 to 26.6) in 2022. We know that survival probabilities for patients with Pancreatic cancer improved over time, but this is not limited to Spain only. Other explanations could be changes in diagnosis strategies and/or disease coding over time. With regard to survival in individuals with pancreatic cancer, approximately 25% of individuals had a 10 year survival which is higher than published in literature where 10 year survival rates was less than 5%. (Tonini and Zanni 2021) The exact explanation of this high survival rate can not be provided but as mentioned under the limitation section, we might have underreported mortality.

Some characteristics of the individuals with the conditions of interest do not match with what is described in literature, e.g. a median age of 27 years (CPRD GOLD), 17 years (IPCI) and 32 years (SIDIAP) at time of CF diagnosis although currently neonatal screening of cystic fibrosis (“newborn screening programme”) is standard. Here as well, we do not know the exact reason for this discrepancy, but it might reflect an incorrect recording of date of diagnosis for historical cases. With regard to sex distribution, we reported a higher proportion of females with CF diagnosis whereas the ECFSPR report of 2021 reported slightly more males than females with CF.(Zolin, Orenti et al. 2023)

Also, a stronger male predominance would be expected for hemophilia than what we reported It is important to clarify that “hemophilia” as condition of interest, consisted of different concept identification codes and for IPCI no granularity in disease codes is available and in SIDIAP this mainly consisted of codes for “Hereditary factor VIII deficiency disease”. As described in literature if the type of type of hemophilia is unknown, distribution by sex can vary and not necessarily consists predominantly of males. (Hemophilia 2023, Statista 2024) Although no extensive phenotyping was done where conditions were selected on relevant concept ids, results on the direct estimate of prevalence and incidence were in line with literature for some of the conditions of interest. For example, with regard to the IR of pancreatic cancer we reported an IR of 20.5 (IPCI) and 20.8 (SIDIAP) per 100,000 PY which is in line with the data from the International Agency for Research on Cancer reporting an IR of 19.6 per 100,000 PY in Europe (17.2/100,000 for UK and 15.3/100,000 for pancreatic ductal adenocarcinoma (PDAC) in the Netherlands). (WHO , Latenstein, van der Geest et al. 2020). Also, the direct estimate of prevalence of cystic fibrosis is in line with literature. Orphanet reports a prevalence of 10-50 per 100,000 which is in line with the results from our study (range of 12-21 per 100,000). (Orphanet , Lopes-Pacheco 2019, Burgel, Burnet et al. 2023) For Sickle cell disease, we observed an increase of prevalence over time in IPCI and SIDIAP. This is in line with literature stating that the prevalence of Sickle cell disease is increasing due to global population movements and increasing life expectancy. Our direct estimates of the prevalence of Sickle cell disease are in line with what is

reported in literature referring to an overall prevalence of less than 50 per 100,000. (Manu Pereira, Colombatti et al. 2023)

Survival in patients with Cystic fibrosis in the UK is reported as median survival age increasing during the last decades from 43.5 to 56.1.(Naito, Charman et al. 2023) We estimated in CPRD GOLD a survival probability of 60% at 30 years after diagnosis. For median survival of patients with PDAC in the Netherlands 3.5 months (0.3 years) was reported (Latenstein, van der Geest et al. 2020) with large heterogeneity between patients with different treatments received. This is in line with the median survial we estimated in CPRD GOLD (0.33) but median survival in IPCI and SIDIAP were much higher.

14.4 Generalisability

Large heterogeneity was observed for measures of several of the diseases. This hampers generalisability of the results to other European countries.

15. CONCLUSION

Using data from three databases with EHR data, it was not possible to estimate median disease duration for the diseases with relatively long-life expectancy studied in this research. This.was due to the limited observation periods for individuals in the GP practices. This hampered the calculation of indirect estimated prevalence from IR and median disease duration.

For diseases with shorter life expectancy, like in this study PAH and Pancreatic cancer, the indirect estimated prevalencewas always lower compared to the direct estimated prevalence.

16. REFERENCES

Ali, M. S., K. Berencsi, K. Marinier, N. Deltour, S. Perez-Guthann, L. Pedersen, P. Rijnbeek, F. Lapi, M. Simonetti, C. Reyes, J. Van der Lei, M. Sturkenboom and D. Prieto-Alhambra (2020). "Comparative cardiovascular safety of strontium ranelate and bisphosphonates: a multi-database study in 5 EU countries by the EU-ADR Alliance." Osteoporos Int.

Berencsi, K., A. Sami, M. S. Ali, K. Marinier, N. Deltour, S. Perez-Gutthann, L. Pedersen, P. Rijnbeek, J. Van der Lei, F. Lapi, M. Simonetti, C. Reyes, M. Sturkenboom and D. Prieto-Alhambra (2020). "Impact of risk minimisation measures on the use of strontium ranelate in Europe: a multi-national cohort study in 5 EU countries by the EU-ADR Alliance." Osteoporos Int **31**(4): 721-755.

Braeye, T., H. D. Emborg, A. Llorente-García, C. Huerta, E. Martín-Merino, T. Duarte-Salles, G. Danieli, L. Tramontan, D. Weibel, C. McGee, M. Villa, R. Gini, M. Lehtinen, L. Titievsky and M. Sturkenboom (2020). "Age-specific vaccination coverage estimates for influenza, human papillomavirus and measles containing vaccines from seven population-based healthcare databases from four EU countries - The ADVANCE project." Vaccine **38**(16): 3243-3254.

Brookmeyer, R. and J. Crowley (1982). "A CONFIDENCE-INTERVAL FOR THE MEDIAN SURVIVAL-TIME." Biometrics **38**(1): 29-41.

Burgel, P. R., E. Burnet, L. Regard and C. Martin (2023). "The Changing Epidemiology of Cystic Fibrosis: The Implications for Adult Care." Chest **163**(1): 89-99.

Burn, E., B. Raventós and M. Català. (2023). "IncidencePrevalence." 2024, from https://darwin-eu.github.io/IncidencePrevalence/.

Burn, E., C. Tebé, S. Fernandez-Bertolin, M. Aragon, M. Recalde, E. Roel, A. Prats-Uribe, D. Prieto-Alhambra and T. Duarte-Salles (2021). "The natural history of symptomatic COVID-19 during the first wave in Catalonia." Nat Commun **12**(1): 777.

Carey, I. M., N. Nirmalananthan, T. Harris, S. DeWilde, U. A. R. Chaudhry, E. Limb and D. G. Cook (2023). "Prevalence of co-morbidity and history of recent infection in patients with neuromuscular disease: A cross-sectional analysis of United Kingdom primary care data." PLoS One **18**(3): e0282513.

Català, M., Y. Guo, M. Du, K. Lopez-Guell and E. Burn. (2024). "PatientProfiles." 2024, from https://darwin-eu-dev.github.io/PatientProfiles/.

de Ridder, M. A. J., M. de Wilde, C. de Ben, A. R. Leyba, B. M. T. Mosseveld, K. M. C. Verhamme, J. van der Lei and P. R. Rijnbeek (2022). "Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands." International Journal of Epidemiology.

Delgado-Pecellin, C., I. Alvarez Rios, M. D. A. Bueno Delgado, M. M. Jimenez Jambrina, M. E. Quintana Gallego, P. Ruiz Salas, I. Marcos Luque and E. Melguizo Madrid (2020). "[Results of the neonatal screening on Western Andalusia after a decade of experience.]." Rev Esp Salud Publica **94**.

Engelkes, M., E. J. Baan, M. A. J. de Ridder, E. Svensson, D. Prieto-Alhambra, F. Lapi, C. Giaquinto, G. Picelli, N. Boudiaf, F. Albers, L. A. Evitt, S. Cockle, E. Bradford, M. K. Van Dyke, R. Suruki, P. Rijnbeek, M. Sturkenboom, H. M. Janssens and K. M. C. Verhamme (2020). "Incidence, risk factors and re-exacerbation rate of severe asthma exacerbations in a multinational, multidatabase pediatric cohort study." Pediatr Allergy Immunol **31**(5): 496-505.

Fahmi, A., D. Wong, L. Walker, I. Buchan, M. Pirmohamed, A. Sharma, H. Cant, D. M. Ashcroft and T. P. van Staa (2023). "Combinations of medicines in patients with polypharmacy aged 65-100 in primary care: Large variability in risks of adverse drug related and emergency hospital admissions." PLoS One **18**(2): e0281466.

Hemophilia, W. f. o. (2023). Report on the Annual Global Survey 2022

Herrett, E., A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa and L. Smeeth (2015). "Data Resource Profile: Clinical Practice Research Datalink (CPRD)." Int J Epidemiol **44**(3): 827-836.

James, G., E. Collin, M. Lawrance, A. Mueller, J. Podhorna, L. Zaremba-Pechmann, P. Rijnbeek, J. van der Lei, P. Avillach, L. Pedersen, D. Ansell, A. Pasqua, M. Mosseveld, S. Grosdidier, U. Gungabissoon, P. Egger, R. Stewart, C. Celis-Morales, M. Alexander, G. Novak and M. F. Gordon (2020). "Treatment pathway analysis of newly diagnosed dementia patients in four electronic health record databases in Europe." Soc Psychiatry Psychiatr Epidemiol.

Kristjansdottir, A., V. Rafnsson and R. T. Geirsson (2023). "Comprehensive evaluation of the incidence and prevalence of surgically diagnosed pelvic endometriosis in a complete population." Acta Obstet Gynecol Scand **102**(10): 1329-1337.

Lane, J. C., K. L. Butler, J. L. Poveda-Marina, D. Martinez-Laguna, C. Reyes, J. de Bont, M. K. Javaid, J. Logue, J. E. Compston, C. Cooper, T. Duarte-Salles, D. Furniss and D. Prieto-Alhambra (2020). "Preschool Obesity Is Associated With an Increased Risk of Childhood Fracture: A Longitudinal Cohort Study of 466,997 Children and Up to 11 Years of Follow-up in Catalonia, Spain." J Bone Miner Res **35**(6): 1022-1030.

Latenstein, A. E. J., L. G. M. van der Geest, B. A. Bonsing, B. Groot Koerkamp, N. Haj Mohammad, I. de Hingh, V. E. de Meijer, I. Q. Molenaar, H. C. van Santvoort, G. van Tienhoven, J. Verheij, P. A. J. Vissers, J. de Vos-Geelen, O. R. Busch, C. H. J. van Eijck, H. W. M. van Laarhoven, M. G. Besselink, J. W. Wilmink and G. Dutch Pancreatic Cancer (2020). "Nationwide trends in incidence, treatment and survival of pancreatic ductal adenocarcinoma." Eur J Cancer **125**: 83-93.

Lopes-Pacheco, M. (2019). "CFTR Modulators: The Changing Face of Cystic Fibrosis in the Era of Precision Medicine." Front Pharmacol **10**: 1662.

Ly, N. F., C. Flach, T. S. Lysen, E. Markov, H. van Ballegooijen, P. Rijnbeek, T. Duarte-Salles, C. Reyes, L. H. John, L. Karimi, C. Reich, S. Salek and D. Layton (2023). "Impact of European Union Label Changes for Fluoroquinolone-Containing Medicinal Products for Systemic and Inhalation Use: Post-Referral Prescribing Trends." Drug Saf **46**(4): 405-416.

Manu Pereira, M. D. M., R. Colombatti, F. Alvarez, P. Bartolucci, C. Bento, A. L. Brunetta, E. Cela, S. Christou, A. Collado, M. de Montalembert, L. Dedeken, P. Fenaux, F. Galacteros, A. Glenthoj, V. Gutierrez Valle, A. Kattamis, J. Kunz, S. Lobitz, C. McMahon, M. Pellegrini, S. Reidel, G. Russo, M. Santos Freire, E. van Beers, P. Kountouris and B. Gulbis (2023). "Sickle cell disease landscape and challenges in the EU: the ERN-EuroBloodNet perspective." Lancet Haematol **10**(8): e687-e694.

Mata-Cases, M., J. Franch-Nadal, J. Real, M. Gratacòs, F. López-Simarro, K. Khunti and D. Mauricio (2018). "Therapeutic inertia in patients treated with two or more antidiabetics in primary care: Factors predicting intensification of treatment." Diabetes Obes Metab **20**(1): 103-112.

Monteagudo, M., A. Nuñez, I. Solntseva, N. Dhalwani, A. Booth, M. Barrecheguren, D. Lambrelli and M. Miravitlles (2021). "Treatment Pathways Before and After Triple Therapy in COPD: A Population-based Study in Primary Care in Spain." Arch Bronconeumol (Engl Ed) **57**(3): 205-213.

Naito, Y., S. Charman, J. Duckers and S. Clarke. (2023). "UK Cystic Fibrosis Registry 2022 Annual Data Report." from https://www.cysticfibrosis.org.uk/sites/default/files/2023-12/CFT_2022_Annual_Data_Report_Dec2023.pdf.

Orphanet. "Cystic Fibrosis." from https://www.orpha.net/en/disease/detail/586.

Ortega, Y., E. Aragonès, J. L. Piñol, J. Basora, A. Araujo and J. J. Cabré (2018). "Impact of depression and/or anxiety on the presentation of cardiovascular events in a cohort with metabolic syndrome. StreX project: Five years of follow-up." Prim Care Diabetes **12**(2): 163-171.

Ramos, R., M. Comas-Cufí, R. Martí-Lluch, E. Balló, A. Ponjoan, L. Alves-Cabratosa, J. Blanch, J. Marrugat, R. Elosua, M. Grau, M. Elosua-Bayes, L. García-Ortiz and M. Garcia-Gil (2018). "Statins for primary prevention of cardiovascular events and mortality in old and very old adults with and without type 2 diabetes: retrospective cohort study." Bmj **362**: k3359.

Raventós, B., M. Català, M. Du, Y. Guo, A. Black, G. Inberg, X. Li, K. López-Güell, D. Newby, M. de Ridder, C. Barboza, T. Duarte-Salles, K. Verhamme, P. Rijnbeek, D. Prieto Alhambra and E. Burn (2024).

"IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model." Pharmacoepidemiol Drug Saf **33**(1): e5717.
Recalde, M., V. Davila-Batista, Y. Díaz, M. Leitzmann, I. Romieu, H. Freisling and T. Duarte-Salles (2021). "Body mass index and waist circumference in relation to the risk of 26 types of cancer: a prospective cohort study of 3.5 million adults in Spain." BMC Med **19**(1): 10.
Recalde, M., C. Rodríguez, E. Burn, M. Far, D. García, J. Carrere-Molina, M. Benítez, A. Moleras, A. Pistillo, B. Bolíbar, M. Aragón and T. Duarte-Salles (2022). "Data Resource Profile: The Information System for Research in Primary Care (SIDIAP)." Int J Epidemiol **51**(6): e324-e336.
Rothman, K. (2012). Epidemiology, OUP.
Statista. (2024). "Distribution of people with bleeding disorders worldwide in 2022, by gender." from https://www.statista.com/statistics/495675/percentager-of-people-with-bleeding-disorders-in-worldwide-by-gender/#:~:text=In%202022%2C%20some%2091%20percent,with%20hemophilia%20A%20were%20male.
Tonini, V. and M. Zanni (2021). "Pancreatic cancer in 2021: What you need to know to win." World J Gastroenterol **27**(35): 5851-5889.
Troncoso-Mariño, A., A. Roso-Llorach, T. López-Jiménez, N. Villen, E. Amado-Guirado, S. Fernández-Bertolin, L. A. Carrasco-Ribelles, J. M. Borras and C. Violán (2021). "Medication-Related Problems in Older People with Multimorbidity in Catalonia: A Real-World Data Study with 5 Years' Follow-Up." J Clin Med **10**(4).
WHO. "International Agency for Research on Cancer." from https://gco.iarc.who.int/today/en/dataviz/tables?mode=population&cancers=13.
Wigglesworth, S., A. Neligan, J. M. Dickson, A. Pullen, E. Yelland, T. Anjuman and M. Reuber (2023). "The incidence and prevalence of epilepsy in the United Kingdom 2013-2018: A retrospective cohort study of UK primary care data." Seizure **105**: 37-42.
Willey, C. J., R. Coppo, F. Schaefer, M. Mizerska-Wasiak, M. Mathur and M. J. Schultz (2023). "The incidence and prevalence of IgA nephropathy in Europe." Nephrol Dial Transplant **38**(10): 2340-2349.
Zolin, A., A. Orenti, A. Jung and J. van Rens (2023). ECFSPR Annual Report 2021.

17. ANNEXES

17.1 **Appendix I:** Concept definitions

Appendix I - Table 1. Concept ids for diseases

Condition	**Included concept ids (also descendants of these concept ids were included)**
Cystic fibrosis	254320, 441267
Haemophilia (A and/or B)	4236898
Pulmonary arterial hypertension	4013643
Pancreatic cancer	199754, 432843, 434293, 440649, 4157459, 4178960, 4180793, 4209933, 36713362, 36713363
Sickle cell disease	22281, 24006, 25518, 315523, 321263, 443721, 443726, 443738, 40485018

17.2 **Appendix II**: Description of databases

CPRD GOLD, United Kingdom

The Clinical Practice Research Datalink (CPRD) GOLD is a database of anonymised electronic health records (EHR) from General Practitioner (GP) clinics in the UK that use the Vision® software system for their management (Herrett, Gallagher et al. 2015). The source population encompasses 98% of the UK, registered with GPs responsible for non-emergency care and referrals. Participating GPs provide CPRD EHR for all registered patients who did not specifically request to opt out of data sharing. Covering 4.6% of the current UK population, GOLD includes 4.9% of contributing GP practices, providing comprehensive information within its defined source population. GOLD contains data from all four UK constituent countries and the current regional distribution of its GP practices is 5.7% in England, 55.6% in Scotland, 28.4% in Wales, and 10.2% in Northern Ireland (May 2022).

GOLD data include patient's demographic, biological measurements, clinical symptoms and diagnoses, referrals to specialist/hospital and their outcome, laboratory tests/results, and prescribed medications. GOLD has been assessed and found broadly representative of the UK general population in terms of age, gender, and ethnicity. GOLD has been widely used internationally for observational research to produce nearly 3,000 peer-reviewed publications, making GOLD the most influential UK clinical database so far (Carey, Nirmalananthan et al. 2023, Fahmi, Wong et al. 2023, Wigglesworth, Neligan et al. 2023).
In 2019, CPRD launched AURUM and since then has encouraged practices from England to move from the software that feeds GOLD (Vision) to the one that feeds AURUM (Emis). GOLD data from 2019 therefore mainly represents Wales/Scotland/NI and AURUM represents England. However, GOLD data collected before 2019 fully represent the UK. CPRD provides for each build release an updated list of practices which moved from GOLD to AURUM. An overlap between GOLD and AURUM can occur, because historical data for these practices have been transferred from Vision/GOLD to Emis/AURUM. When DARWIN-EU® uses both databases the safest and easiest solution would be to disregard these practices in GOLD. The licence also covers HES/ONS data, which can be requested on a study-by-study basis as linked data. This data only covers England and is planned to be mapped to OMOP in the future.
In terms of quality checks, the integrity, structure and format of the data is reviewed. Collection-level validation ensures integrity by checking that data received from practices contain only expected data files and ensures that all data elements are of the correct type, length and format. Duplicate records are identified and removed.[1] Transformation-level validation checks for referential integrity between records ensure that there are no orphan records included in the database (for example, that all event records link to a patient), while research-quality-level validation covers the actual content of the data. CPRD provides a patient-level data quality metric in the form of a binary 'acceptability' flag (Herrett, Gallagher et al. 2015). This is based on recording and internal consistency of key variables including date of birth, practice registration date and transfer out date.

IPCI, the Netherlands

The Integrated Primary Care Information (IPCI) database is a longitudinal observational database containing routinely collected data extracted from computer-based patient records of a selected group of general practitioners (GPs) across the Netherlands (de Ridder, de Wilde et al. 2022). IPCI was started in 1992 by the department of Medical Informatics of the Erasmus University Medical Center in Rotterdam. The current database includes patient records from 2006 on, when the size of the database started to increase significantly. The demographic composition of the IPCI population mirrors that of the general Dutch

population in terms of age and sex. Although the geographical spread is limited, GP practices are located in urban and non-urban areas.
Patient-level data includes demographic information, patient's complaints and symptoms, diagnoses, laboratory test results, lifestyle factors and correspondence with secondary care, such as referral and discharge letters. For complaints, symptoms and diagnoses, Dutch GPs use International Classification of Primary Care (ICPC-1) coding, an international standard developed and updated by the World Organization of Family Doctors' (WONCA) International Classification Committee.
IPCI data quality has been previously documented and IPCI has proved valuable for epidemiological studies (Ali, Berencsi et al. 2020, Berencsi, Sami et al. 2020, Engelkes, Baan et al. 2020, James, Collin et al. 2020). In terms of quality control, extensive quality control steps are performed prior to each data release. These include comparison of 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, completeness of laboratory results, availability of hospital letters and prescriptions, proportion of patients with blood pressure measurement, etc) (de Ridder, de Wilde et al. 2022). Based on this information, two quality scores have been created. Practices with low scores have been excluded.
The IPCI database is registered on the European Medicines Agency (EMA) ENCePP resources database (http://www.encepp.eu).

Information System for Research in Primary Care (SIDIAP), Spain

The Information System for Research in Primary Care (SIDIAP) is a dynamic database of pseudo-anonymized electronic health records of the primary care patient population in Catalonia, Spain (Recalde, Rodríguez et al. 2022). It contains data of approximately 80% of the Catalan population registered in over 280 primary care practices throughout Catalonia since 2005.
The database contains data recorded in primary care centres on a daily basis. Additionally, it integrates data from external sources including biomarkers data from laboratories and records of drug prescription and dispensation. The dataset covers demographics, all-cause mortality, disease diagnoses classified under the International Classification of Diseases 10th revision (ICD-10), prescription and dispensation records of drugs, results of laboratory tests, socio-economic indicators, vaccination records, lifestyle information, parent–child linkage and various clinical parameters. Additional data from other data sources such as hospital discharges, mental health centres or specific disease registries can be obtained through diverse linkages. The demographic composition within SIDIAP closely mirrors that of the broader Catalan population, encompassing a representative spectrum of geographic distribution, age, and sex proportions. The database is updated every 6 months.
SIDIAP data quality has been previously documented and SIDIAP has proved valuable for epidemiological studies (Mata-Cases, Franch-Nadal et al. 2018, Ortega, Aragonès et al. 2018, Ramos, Comas-Cufí et al. 2018, Braeye, Emborg et al. 2020, Lane, Butler et al. 2020, Burn, Tebé et al. 2021, Monteagudo, Nuñez et al. 2021, Recalde, Davila-Batista et al. 2021, Troncoso-Mariño, Roso-Llorach et al. 2021, Ly, Flach et al. 2023). In terms of data integrity and reliability, SIDIAP has been subject to rigorous evaluation. Quality checks have been implemented including central identification of duplicate patient ID and visual inspection for temporal patterns in the registry of a certain variable. Furthermore, the data undergoes assessment for availability (longitudinality and reliability), plausibility (range checks and unusual values) and consistency using visualization tools. Specifically, for biochemistry data, consistency for measurements taken in different laboratories is assessed, and unit conversion is undertaken when needed.

17.3 **Appendix III**: Supplementary tables and figures

Appendix III is a separate document.