




# Study Protocol

## P2 C1-013


12/07/2024

Version 2.0

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	


## Table of Contents

<b>DOCUMENT HISTORY .....</b>	<b>3</b>
<b>LIST OF ABBREVIATIONS.....</b>	<b>5</b>
<b>1. TITLE .....</b>	<b>6</b>
<b>2. RESPONSIBLE PARTIES – STUDY TEAM.....</b>	<b>6</b>
<b>3. ABSTRACT (Stand alone summary of the study protocol) .....</b>	<b>6</b>
<b>4. AMENDMENTS AND UPDATES.....</b>	<b>8</b>
<b>5. MILESTONES.....</b>	<b>9</b>
<b>6. RATIONALE AND BACKGROUND .....</b>	<b>9</b>
<b>7. RESEARCH QUESTION AND OBJECTIVES .....</b>	<b>10</b>
<b>8. RESEARCH METHODS.....</b>	<b>11</b>
8.1 Study type and Study Design .....	11
8.2 Study Setting and Data Sources.....	11
8.3 Follow-up .....	12
8.4 Study Population with inclusion and exclusion criteria.....	15
8.5 Variables .....	16
8.6.1. Exposure/s (where relevant).....	16
8.6.2. Outcome/s (where relevant) .....	16
8.6.3. Other covariates, including confounders, effect modifiers and other variables (where relevant).....	17
8.6 Study size .....	17
8.7 Analysis .....	17
8.8.1. Outcome cohorts.....	17
8.8.2. Descriptive statistics.....	17
8.8.3. Point prevalence.....	18
8.8.4. Incidence rates .....	19
8.8.5. Median disease duration.....	19
8.8.6. Indirect estimated prevalence.....	19
8.9 Evidence synthesis.....	20
<b>9 DATA MANAGEMENT .....</b>	<b>20</b>
<b>10 QUALITY CONTROL .....</b>	<b>21</b>
<b>11 LIMITATIONS OF THE RESEARCH METHODS .....</b>	<b>21</b>
<b>12 MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS.....</b>	<b>22</b>
<b>13 GOVERNANCE BOARD ASPECTS.....</b>	<b>22</b>
<b>14 PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS.....</b>	<b>22</b>
<b>15 OTHER ASPECTS.....</b>	<b>22</b>
<b>16 REFERENCES .....</b>	<b>22</b>
<b>17 ANNEXES.....</b>	<b>23</b>


	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
		<b>Dissemination level: Public</b>

## DOCUMENT HISTORY

VERSION	DATE	DESCRIPTION
1.0	19/01/2024	First draft
1.1	14/02/2024	Updated version for archiving purposes
1.2	02/04/2024	Updated version for archiving purposes
2.0	12/07/2024	Final version uploaded in the HMA-EMA Catalogue


	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

<b>Study Title</b>	DARWIN EU® - Comparing direct and indirect methods to estimate prevalence of chronic diseases using real-world data.
<b>Protocol version identifier</b>	2.0
<b>Date of last version of protocol</b>	12/07/2024
<b>EU PAS register number</b>	EUPAS1000000088
<b>Active substance</b>	NA
<b>Medicinal product</b>	NA
<b>Research question and objectives</b>	<p>In the context of chronic diseases with relatively low prevalence, how do direct and indirect RWD-based estimates of prevalence compare with each other?</p> <p>The specific objectives of this study are to:</p> <ol style="list-style-type: none"> <li>1) Estimate the disease prevalence (direct estimate based on the proportion of individuals with the condition).</li> <li>2) Estimate the disease incidence rate.</li> <li>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.</li> <li>4) Produce an indirect estimation of prevalence as the product of incidence and median survival.</li> </ol> <p>for the following diseases:</p> <ul style="list-style-type: none"> <li>• Cystic fibrosis</li> <li>• Haemophilia (A and/or B)</li> <li>• Pulmonary arterial hypertension</li> <li>• Pancreatic cancer</li> <li>• Sickle cell disease</li> </ul> <p>Results will be provided overall and where possible stratified by age group: paediatrics (0-17 years old) and adults (18 years old and above).</p>
<b>Country(ies) of study</b>	United Kingdom, the Netherlands, Spain
<b>Author</b>	Maria de Ridder

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## LIST OF ABBREVIATIONS

<b>Acronyms/terms</b>	<b>Description</b>
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
NA	Not applicable
NL	the Netherlands
OMOP	Observational Medical Outcomes Partnership
OP	Outpatient
RWD	Real-World Data
SIDIAP	Sistema d'Informació per al Desenvolupament de la Investigació en Atenció Primària
SP	Spain
TBC	To be confirmed
UK	United Kingdom

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## 1. TITLE

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

## 2. RESPONSIBLE PARTIES – STUDY TEAM

STUDY TEAM ROLE	NAMES	ORGANISATION
Study Project Manager/Principal Investigator	Maria de Ridder	Erasmus MC
Data Scientist	Ross Williams Cesar Barbosa Maarten van Kessel	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

\*Data partners' role is only to execute code at their data source, review and approve their results. These people do not have an investigator role. Data analysts/programmers do not have an investigator role and thus declaration of interests (DOI) for these people is not needed.


## 3. ABSTRACT (STAND ALONE SUMMARY OF THE STUDY PROTOCOL)

### 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.

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
		<b>Dissemination level: Public</b>

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 prevalence will be compared using real-world data sources.

### **Research question and objective**

The objective of this study is to compare direct and indirect estimations of prevalence of some rare, chronic diseases.

This will be done considering all patients with the disease 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

A retrospective cohort design to estimate disease point prevalence and incidence.

A 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 will be used.

#### Population

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

All patients with the disease will be used 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.

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

#### Data sources

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

#### Sample size

No sample size has been calculated. All patients with the disease will be 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, will be calculated as well as incidence rate over the total study period.

For each patient only the first diagnosis of a disease will be considered. The date of this diagnosis can be before the observation period of the patient, i.e. before the period during which the patient is monitored in the database. A GP can have entered historical events, for example the diagnosis of an inherited disease. Also, information received from a patient's former GP might be imported. For this study, it is important to have as much patient's history as possible. The disease is considered to stay present during patient's observation period (and beyond). For point prevalence, all persons with a diagnosis before 01/01/2016 and in observation in the database at this date contribute to the numerator. The denominator is the total number of persons in observation on 01/01/2016. For the calculation of the incidence rate, only newly diagnosed patients (diagnosed within the observation time and within the study period) contribute to the numerator. The denominator is the total number of person-years at risk, i.e. the sum across all subjects included in the cohort of the observation time within the study period or until a diagnosis occurs.

*Survival estimation:* Kaplan Meier estimates for survival probabilities for time since first diagnosis. Median survival, as a proxy for disease duration, is the time point where the survival probability decreases to below 50%.

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


From the incidence rate and median disease duration, the indirect prevalence will be calculated.

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

## **4. AMENDMENTS AND UPDATES**

NUMBER	DATE	SECTION OF STUDY PROTOCOL	AMENDMENT OR UPDATE	REASON
None				



	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## 5. MILESTONES

STUDY SPECIFIC DELIVERABLE	TIMELINE
Draft Study Protocol	18/01/2024
Final Study Protocol	02/04/2024
Creation of Analytical code	19/02/2024
Execution of Analytical Code on the data	26/02/2024
Interim Study Report (if applicable)	NA
Draft Study Report	28/03/2024
Final Study Report	22/04/2024

## 6. 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, 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[1]:

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

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 [2], or where input from different sources is combined [3].

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

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.


## 7. RESEARCH QUESTION AND OBJECTIVES

Table 1 shows the objective of this study with some specifications.

Table 1. Primary and secondary research questions and objective.

### A. Primary research question and objective.

<b>Objective:</b>	To compare direct and indirect estimations of prevalence of some rare, chronic diseases
<b>Hypothesis:</b>	No hypothesis is tested
<b>Population:</b>	Total population in the databases
<b>Diseases:</b>	Cystic fibrosis Haemophilia (A and/or B) Pulmonary arterial hypertension Pancreatic cancer Sickle cell disease
<b>Setting:</b>	3 databases with Electronic Healthcare Records from primary care
<b>Main outcome:</b>	Direct and indirect estimated prevalences

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## 8. RESEARCH METHODS

### 8.1 Study type and Study Design

Table 2 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 2. 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.

### 8.2 Study Setting and Data Sources

For this study, suitable data sources should include individuals who can be considered as a representative sample of the general population and have the potential for long observation periods for subjects. Therefore, we focus on primary care data sources and hospital data sources are not included. Lifelong observation of patients is not available in any of the data sources within the DARWIN EU network, however, patients in primary care databases will often have several years of observation time. Also history of diagnoses before subject's observation time might be recorded. In addition, data sources should have a good recording of mortality.

When it comes to assessing the reliability of data sources, the data partners were 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 the national healthcare data. CdmOnboarding creates a report with select characterisation of the clinical data within the database and 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.

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 3.


	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

Table 3. Description of the selected Data Sources.

Country	Name of Database	Justification for Inclusion	Health Care setting	Type of Data	Number of active subjects	Feasibility count of disease (if relevant)	Data lock for the last update
United Kingdom	CPRD GOLD	Sufficient feasibility counts, sufficiently long median duration of observation periods, good recording of mortality	Primary care	EHR	3.0 million		Source data 01/07/2023
the Netherlands	IPCI	Sufficient feasibility counts, sufficiently long median duration of observation periods, good recording of mortality	Primary care	EHR	1.3 million		Source data 23/09/2023
Spain	SIDIAP	Sufficient feasibility counts, sufficiently long median duration of observation periods, good recording of mortality	Primary care	EHR	5.8 million		Source data 30/06/2023


#### Study Period

Study period is from January 1<sup>st</sup>, 2010, to December 31<sup>st</sup>, 2022.

### 8.3 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 is the part of patient's observation period that overlaps with the study period. Only first diagnoses which fall within this follow-up period will contribute to

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
		<b>Dissemination level: Public</b>

the incidence numerator. The follow-up period will contribute 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, will also be captured. Diagnoses before the observation period can be present if a GP has received information about the patient's history from the former GP, or if a GP enters historical information, e.g. the diagnosis date of an inherited disease. However, individuals can only contribute to the point prevalence numerator and denominator if the time point used to assess the point prevalence is within their observation period.
- In the survival analysis, the follow-up of patients starts at the first diagnosis of the disease of interest and ends at patient's death, end of patient's observation period or end of the study period, whatever comes first.


Further information on follow-up time is given in [Table 4](#).

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

**Table 4.** Operational Definition of Time 0 (index date) and other primary time anchors.

Study population name(s)	Time Anchor Description	Number of entries	Type of entry	Washout window	Care Setting <sup>1</sup>	Code Type <sup>2</sup>	Incident with respect to...
Total population	Start of FU: latest of: - entry in database - start of study period (01/01/2010)  End of FU: earliest of: - date of mortality - end of patient's observation period, i.e. deregistration of patient or end of GP data contribution - end of study period (31/12/2022)	1	NA	NA	PC, IP and OP	NA	NA
Patients with disease diagnosis	Start of FU: first diagnosis of disease  End of FU: earliest of: - date of mortality - end of patient's observation period, i.e. deregistration of patient or end of GP data contribution - end of study period (31/12/2022)	1	Incident	[-Inf, -1]	PC, IP and OP	SNOMED codes	Disease of interest

<sup>1</sup> IP = inpatient, OP = outpatient, NA = not applicable, PC = primary care

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## 8.4 Study Population with inclusion and exclusion criteria

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

For disease duration, all patients with the diagnosis of interest will be included, without any exclusion criteria.


Inclusion criteria are described in **Table 5**.

**Table 5.** Operational Definitions of Inclusion Criteria.

Criterion	Details	Assessment window	Care Settings <sup>1</sup>	Code Type	Applied to study populations:
Patients with a disease diagnosis	Earliest recording of disease diagnosis will be identified, this can be before or during observation time	Complete history and observation period of patient	PC, IP and OP	SNOMED code	Patients with disease diagnosis
Pediatric patients	Earliest recording of disease is at age 0-17	Complete history and observation period of patient	PC, IP and OP	SNOMED code	Patients with disease diagnosis at paediatric age
Adult patients	Earliest recording of disease is at age 18+	Complete history and observation period of patient	PC, IP and OP	SNOMED code	Patients with disease diagnosis at adult age

<sup>1</sup>IP = inpatient, OP = outpatient, NA = not applicable, PC = primary care

No exclusion criteria are applied

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## 8.5 Variables

### 8.6.1. Exposure/s (where relevant)

No exposures are considered in this study.

### 8.6.2. Outcome/s (where relevant)


While the diseases in this study are all chronic, for each disease, only the first occurrence of the diagnosis will be selected. Details are provided in **Table 6**. The codes (concept ids) used for selection of the diagnoses are listed in **Appendix I – Concept definitions**. As well as the occurrence of disease mortality is also an outcome.

**Table 6.** Operational Definitions of Outcome.

Outcome name	Details	Type of outcome	Washout window	Care Settings <sup>1</sup>	Code Type	Applied to study populations
Cystic fibrosis	Preliminary code list provided in Table 1 in Appendix I	Incidence and prevalence	[-Inf, -1]	PC, IP and OP	SNOMED	Total population
Haemophilia	Preliminary code list provided in Table 1 in Appendix I	Incidence and prevalence	[-Inf, -1]	PC, IP and OP	SNOMED	Total population
Pulmonary arterial hypertension	Preliminary code list provided in Table 1 in Appendix I	Incidence and prevalence	[-Inf, -1]	PC, IP and OP	SNOMED	Total population
Pancreatic cancer	Preliminary code list provided in Table 1 in Appendix I	Incidence and prevalence	[-Inf, -1]	PC, IP and OP	SNOMED	Total population
Sickle cell disease	Preliminary code list provided in Table 1 in Appendix I	Incidence and prevalence	[-Inf, -1]	PC, IP and OP	SNOMED	Total population
Mortality		Time-to-event				Patients with disease

<sup>1</sup> IP = inpatient, OP = outpatient, PC = primary care



	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

### 8.6.3. Other covariates, including confounders, effect modifiers and other variables (where relevant)

For all diseases, next to a cohort including all diagnosed patients, a paediatric (age 0-17) and an adult (age 18 and older) cohort will be generated. To generate these two cohorts, age at first diagnosis will be calculated.

## 8.6 Study size

No sample size has been calculated.

## 8.7 Analysis

The analyses which will be done are given in **Table 7**. Details are provided in Sections 8.8.3 – 8.8.6.

**Table 7.** Description of Study Types and Type of analysis.

STUDY TYPE	STUDY CLASSIFICATION	TYPE OF ANALYSIS
Population-level descriptive epidemiology	Off-the-shelf	- Incidence rates of the condition of interest - Prevalence rates of the condition of interest
Patient-level characterisation	Off-the-shelf	- Survival time

Analyses will be done and reported in each database separately. A meta-analysis of results will be carried out, which is described in section 8.9.

### 8.8.1. Outcome cohorts

The outcome cohorts for the diseases will include only the first diagnosis date of each patient. This diagnosis date is cohort start date. Cohort end date is the end of the observation period of the patient.

Based on birth date of the patient and date of first diagnosis, age at diagnosis will be determined. Age will be categorized into 0 – 17 year (paediatric) and 18 year and older (adult). For each disease, if feasible, three outcome cohorts will be generated:


- patients with first diagnosis at any age
- patients with first diagnosis at paediatric age
- patients with first diagnosis at adult age

All analyses described below will be done for each of these cohorts, under restriction of the minimum cell count criterion, i.e. for outcome cohorts with less than 5 subjects no analyses will be done.

### 8.8.2. Descriptive statistics

An attrition table will be given to report the number of subjects used in the analysis.

For the disease cohorts, sex and age at diagnosis will be described.

	D2.2.3 - Study Protocol for P2-C1-013	
	Author(s): Maria de Ridder, Katia Verhamme	Version: 2.0
	Dissemination level: Public	

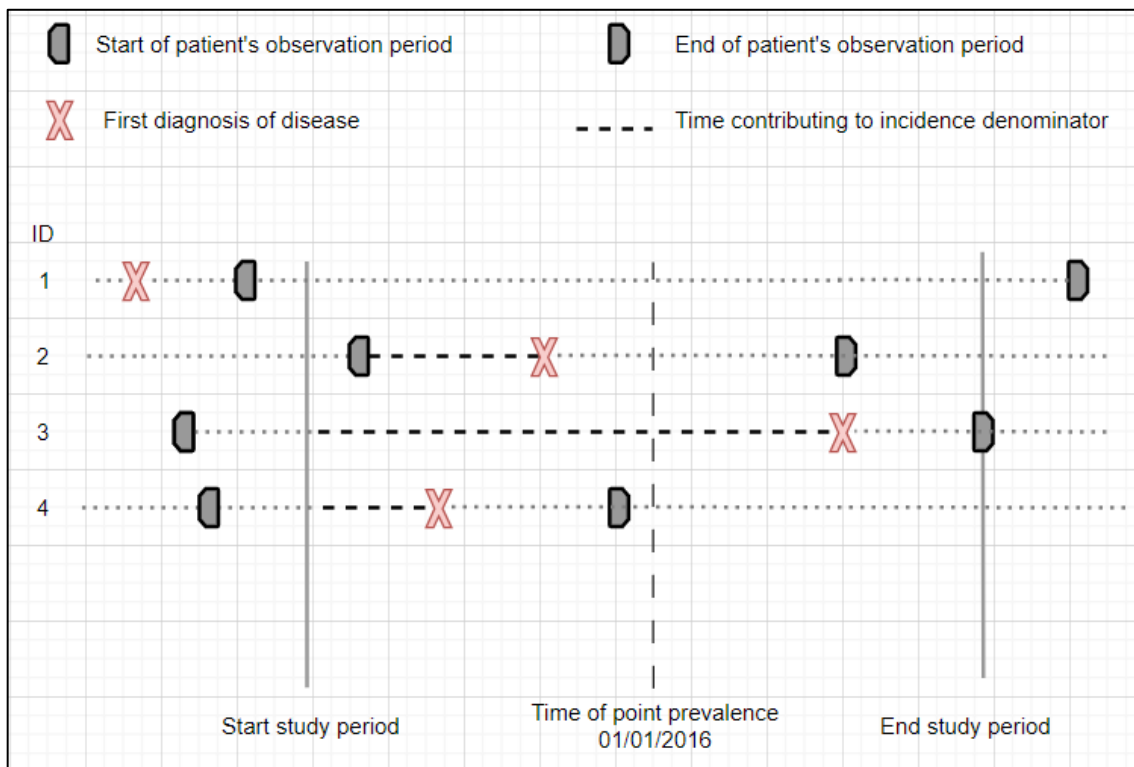
### 8.8.3. Point prevalence

The R package *IncidencePrevalence* will be used to estimate complete point prevalence.


Point prevalence will be estimated for January 1 of each calendar year during the study period. Denominator is the total number of persons in observation at the specific time point. All persons with a diagnosis before the time point and in observation in the database at this time point contribute to the numerator. Point prevalences will be plotted against calendar year. The point prevalence at 01/01/2016 will be used for the comparison with the indirect estimated prevalence.

In **Figure 1** some patients are shown with their observation period and time of first diagnosis relative to the study period. The observation period of patient 1 includes the time point for the point prevalence at 01/01/2016. Therefore, this patient contributes to the denominator for the point prevalence. First diagnosis of this patient is before 01/01/2016 (and even before the study period and before the patient's observation period), so this patient also contributes to the numerator. Likewise, patient 2 is in observation at 01/01/2016 and is diagnosed before, so this patient also contributes to both the denominator and numerator. For patient 3, 01/01/2016 falls within the observation period and therefore this patient contributes to the denominator only. The date of first diagnosis for patient 3 is after 01/01/2016, so this patient does not contribute to the numerator for point prevalence. The observation period of patient 4 ends before 01/01/2016. Hence this patient does not contribute to the denominator nor the numerator.

The estimate of the point prevalence will be reported with 95% confidence intervals (CI) using Wilson score method.



**Figure 1** Diagram with patient data to illustrate incidence and prevalence calculations.

	D2.2.3 - Study Protocol for P2-C1-013	
	Author(s): Maria de Ridder, Katia Verhamme	Version: 2.0
	Dissemination level: Public	

While the complete point prevalence estimated at the mid-point of the study period (January 1<sup>st</sup> 2016) will be used to compare with the indirect estimate of prevalence, estimates of complete point prevalence on January 1<sup>st</sup> of every year of the study period will also be reported to describe potential trends over time.

#### 8.8.4. Incidence rates

Incidence rates will be estimated over the complete study period and by calendar year. Denominator is 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) will contribute. The incidence rate by calendar year will be plotted to check stability over time. This will be used in the discussion about the validity of the indirect estimated prevalence.

Patient 1 in **Figure 1** is diagnosed before the study period. Therefore, this patient is not at risk for incident disease during the study period, so s/he does not contribute with person time to the denominator for the incidence rate (i.e. patients with a diagnosis before the start of follow-up will be excluded from the analysis for incidence). Therefore, the diagnosis also does not contribute to the numerator. Patient 2 contributes person time from start of observation period to the time of diagnosis. This diagnosis, within the study period and within the observation period of the patient, contributes to the numerator. Both patient 3 and patient 4 contribute person time from start of study period to time of diagnosis, and again their diagnoses contribute to the numerator.

The estimate of the incidence rate will be reported with 95% CI using exact Poisson method.

#### 8.8.5. Median disease duration

Survival time will be analysed with the R package *survival*. Time zero for the survival analysis is the date of first diagnosis. Patients for whom no date of death is recorded will be censored at the date of end of their observation period or administratively censored at the end of study period.

Median survival time, i.e. disease duration, will be estimated with 95% CI using Brookmeyer-Crowley method [4].

#### 8.8.6. Indirect estimated prevalence

From the overall incidence ( $I$ ) and median disease duration ( $D$ ) the indirect prevalence will be calculated with formula

$$P = I \cdot D \quad (1)$$

A CI for this indirect estimated prevalence will be generated by using the relation on the log scale:

$$\log P = \log(I \cdot D) = \log I + \log D \quad (2)$$

hence for the variance


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

This will be calculated using

$$Var(\log I) = 1/C \quad (4)$$

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

An approximation of  $Var(\log D)$  will be generated using the limits  $D_{low}$  and  $D_{upp}$  of the 95% CI for  $D$ :

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
		<b>Dissemination level: Public</b>

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

We will assume incidence rate and disease duration are independent, hence

$$Cov(\log I . \log D) = 0 \quad (6)$$

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

## 8.9 Evidence synthesis

Estimated direct prevalence, incidence and median survival will be reported separately for each database. In addition, the results will be combined across data sources. To this end, a random-effect meta-analyses model will be fitted for all the results described below:

1. The direct estimated point prevalence. This will be done using the function `metaprop` of the R *meta* package using as input the numerators and denominators utilised to calculate the point prevalence in each database.
2. The incidence rate. For this meta-analysis, function `metarate` of the *meta* R package will be used with input the total number of events and the total number of person years in each database.
3. The median disease duration. For this meta-analysis, the function `metagen` of the *meta* R package will be used. Depending on the skewness of the CIs of the medians  $D$ , meta-analysis will be performed on  $\log D$ . Input will be the estimates and the approximated standard errors calculated as described above in 8.8.6. equation (5).
4. The indirect estimated prevalence. For this meta-analysis, function `metagen` of the *meta* R package will be used. Meta-analysis will be performed on  $\log P$ . Input will be the estimates and the approximated standard errors calculated as described above in 8.8.6. equation (3).


Forest plots and measures of statistical heterogeneity between data sources will be reported in addition to the combined results of the quantities of interest listed above.

For each of the results in 1) to 4) above, it might be that not all databases provide an estimate, for example, because the number in a numerator might be below the minimum value of 5 or the median disease duration might not be observed in the available data. Meta-analysis will be done with estimates available.

Finally, we will also calculate an indirect prevalence using the combined results for incidence and median disease duration.

## 9 DATA MANAGEMENT

All databases 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>.

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

The analytic code for this study will be written in R. Each data partner will execute the study code against their database containing patient-level data and will then return the results set which will only contain aggregated data.

The results from each of the contributing data sites will then be combined in tables and figures and the meta-analyses will be executed.

R packages used are *IncidencePrevalence*, *survival* and *meta*.

## 10 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>). In particular, it is expected that data partners will have run the OHDSI Data Quality Dashboard tool (<https://github.com/OHDSI/DataQualityDashboard>). This tool provides numerous checks relating to the conformance, completeness and plausibility of the mapped data. Conformance focuses on checks that describe the compliance of the representation of data against internal or external formatting, relational, or computational definitions, completeness in the sense of data quality is solely focused on quantifying missingness, or the absence of data, while plausibility seeks to determine the believability or truthfulness of data values. Each of these categories has one or more subcategories and are evaluated in two contexts: validation and verification. Validation relates to how well data align with external benchmarks with expectations derived from known true standards, while verification relates to how well data conform to local knowledge, metadata descriptions, and system assumptions.

### Study specific quality control

Because this is a methodological study, no thorough phenotyping for the diseases will be performed. Preliminary to finalising the study code, some diagnostics will be run in the databases. This will involve generation of the disease cohorts, checking the distribution of age at diagnosis and of survival time for those who die. This information will be used to decide about the final definition of the disease cohorts, for example to exclude diagnoses at unrealistic ages.


The study code will use two R packages developed by the DARWIN EU® CC. Package *IncidencePrevalence* has been developed to estimate incidence and prevalence of conditions and drug use. This package includes numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package is publicly available in CRAN.

The R packages *survival* and *meta*, to be used for survival analysis and meta-analysis respectively, are both available in CRAN, widely used and well-documented.

## 11 LIMITATIONS OF THE RESEARCH METHODS

The study will be informed by routinely collected healthcare data and so data quality issues must be considered. Also, the period during which information about the patient is present will often be limited because it might be restricted to the period the patient is registered in the GP practice and to the period the GP data is made available for the database. A GP can have entered historical events, for example the diagnosis of an inherited disease, and/or information received from a patient's former GP might be imported. However, this is not standard practice for all GPs.

This means diagnoses might be missing if falling outside the monitored period. It can also be the case that the first recorded diagnosis of the disease is not the correct date of onset. For example, this can be the case

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

when the first diagnosis was before the patient's observation period in the database. Also, there might be patients with asymptomatic disease, which results in a late diagnosis, sometimes only shortly before death. These incorrect dates of diagnosis can give bias in the estimation of incidence and prevalence, especially in the estimation of the median disease duration. Moreover, a missing first diagnosis might result in an incorrect classification of the disease being diagnosed at adult age, while it should be paediatric. However, this would not affect the comparison between direct and indirect estimation methods.

## 12 MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS

Adverse events/adverse reactions will not be 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](https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-submission-reports_en.pdf)).

Only in case of prospective data collection, there is a need to describe the procedures for the collection, management and reporting of individual cases of adverse events/adverse reactions.

## 13 GOVERNANCE BOARD ASPECTS

All data sources require approval from their respective IRB boards.

## 14 PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS


A study report including an executive summary, tables, figures and meta-analysis results, will be submitted to EMA by the DARWIN EU® CC upon completion of the study.

## 15 OTHER ASPECTS

None.

## 16 REFERENCES


1. Rothman, K., *Epidemiology*. 2012: OUP.
2. Kristjansdottir, A., V. Rafnsson, and R.T. Geirsson, *Comprehensive evaluation of the incidence and prevalence of surgically diagnosed pelvic endometriosis in a complete population*. *Acta Obstet Gynecol Scand*, 2023. **102**(10): p. 1329-1337.
3. Willey, C.J., et al., *The incidence and prevalence of IgA nephropathy in Europe*. *Nephrol Dial Transplant*, 2023. **38**(10): p. 2340-2349.
4. Brookmeyer, R. and J. Crowley, *A CONFIDENCE-INTERVAL FOR THE MEDIAN SURVIVAL-TIME*. *Biometrics*, 1982. **38**(1): p. 29-41.

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
		<b>Dissemination level: Public</b>

## 17 ANNEXES

**Appendix I** – Concept definitions

**Appendix II** - ENCePP checklist for study protocols


	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## Appendix I – Concept definitions

Table 1 Concept definitions for diseases

<b>Condition</b>	<b>Included concept ids, incl. descendants</b>
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



	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

## Appendix II - ENCePP checklist for study protocols

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

EU PAS Register® number:  
Study reference number (if applicable):


<b><u>Section 1: Milestones</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
1.1 Does the protocol specify timelines for				Overview and 5
1.1.1 Start of data collection <sup>1</sup>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
1.1.2 End of data collection <sup>2</sup>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
1.1.3 Progress report(s)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
1.1.4 Interim report(s)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
1.1.5 Registration in the EU PAS Register®	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
1.1.6 Final report of study results.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Comments:

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

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

<sup>2</sup> Date from which the analytical dataset is completely available.

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	


Comments:

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

Comments:

<b><u>Section 4: Source and study populations</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
4.1 Is the source population described?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.2
4.2 Is the planned study population defined in terms of:				8.3, 8.4, 8.5, 8.6
4.2.1 Study time period	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4.2.2 Age and sex	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4.2.3 Country of origin	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4.2.4 Disease/indication	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4.2.5 Duration of follow-up	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4.3 Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.5

Comments:

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

<b>Section 5: Exposure definition and measurement</b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
5.1 Does the protocol describe how the study exposure is defined and measured? (e.g. operational details for defining and categorising exposure, measurement of dose and duration of drug exposure)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
5.2 Does the protocol address the validity of the exposure measurement? (e.g. precision, accuracy, use of validation sub-study)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
5.3 Is exposure categorised according to time windows?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
5.4 Is intensity of exposure addressed? (e.g. dose, duration)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
5.5 Is exposure categorised based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
5.6 Is (are) (an) appropriate comparator(s) identified?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	


Comments:

--

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

Comments:

--

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	<b>Author(s): Maria de Ridder, Katia Verhamme</b>	<b>Version: 2.0</b>
	<b>Dissemination level: Public</b>	

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

Comments:


--

<b><u>Section 8: Effect measure modification</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
8.1 Does the protocol address effect modifiers? (e.g. collection of data on known effect modifiers, sub-group analyses, anticipated direction of effect)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

Comments:

--

<b><u>Section 9: Data sources</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
9.1 Does the protocol describe the data source(s) used in the study for the ascertainment of:				
9.1.1 Exposure? (e.g. pharmacy dispensing, general practice prescribing, claims data, self-report, face-to-face interview)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.2
9.1.3 Covariates and other characteristics?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.2
9.2 Does the protocol describe the information available from the data source(s) on:				
9.2.1 Exposure? (e.g. date of dispensing, drug quantity, dose, number of days of supply prescription, daily dosage, prescriber)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.6
9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.6

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	Author(s): Maria de Ridder, Katia Verhamme	Version: 2.0
	Dissemination level: Public	

<b>Section 9: Data sources</b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
9.3 Is a coding system described for:				
9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.6
9.3.3 Covariates and other characteristics?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
9.4 Is a linkage method between data sources described? (e.g. based on a unique identifier or other)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

Comments:


--

<b>Section 10: Analysis plan</b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
10.1 Are the statistical methods and the reason for their choice described?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.8
10.2 Is study size and/or statistical precision estimated?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.8
10.3 Are descriptive analyses included?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.8
10.4 Are stratified analyses included?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8.8
10.5 Does the plan describe methods for analytic control of confounding?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
10.6 Does the plan describe methods for analytic control of outcome misclassification?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
10.7 Does the plan describe methods for handling missing data?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
10.8 Are relevant sensitivity analyses described?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

Comments:

--

<b>Section 11: Data management and quality control</b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
11.1 Does the protocol provide information on data storage? (e.g. software and IT environment, database maintenance and anti-fraud protection, archiving)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	9
11.2 Are methods of quality assurance described?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	10
11.3 Is there a system in place for independent review of study results?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	Author(s): Maria de Ridder, Katia Verhamme	Version: 2.0
	Dissemination level: Public	

Comments:

<b><u>Section 12: Limitations</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
12.1 Does the protocol discuss the impact on the study results of:				
12.1.1 Selection bias?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	11
12.1.2 Information bias?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
12.1.3 Residual/unmeasured confounding? (e.g. anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
12.2 Does the protocol discuss study feasibility? (e.g. study size, anticipated exposure uptake, duration of follow-up in a cohort study, patient recruitment, precision of the estimates)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	11


Comments:

<b><u>Section 13: Ethical/data protection issues</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	13
13.2 Has any outcome of an ethical review procedure been addressed?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
13.3 Have data protection requirements been described?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	9

Comments:

<b><u>Section 14: Amendments and deviations</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
14.1 Does the protocol include a section to document amendments and deviations?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4

Comments:

	<b>D2.2.3 - Study Protocol for P2-C1-013</b>	
	Author(s): Maria de Ridder, Katia Verhamme	Version: 2.0
	Dissemination level: Public	

<b><u>Section 15: Plans for communication of study results</u></b>	<b>Yes</b>	<b>No</b>	<b>N/A</b>	<b>Section Number</b>
15.1 Are plans described for communicating study results (e.g. to regulatory authorities)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	14
15.2 Are plans described for disseminating study results externally, including publication?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

Comments:

--

Name of the main author of the protocol: Maria de Ridder

Date: 10/01/2024

Signature: *Maria de Ridder*