Study Protocol

P3-C3-011

DARWIN EU® - Incidence, period prevalence, and characterisation of individuals with paediatric pulmonary arterial hypertension

Authors: N. Hunt, K. Verhamme

19/08/2025

Version 7.0

Public

CONTENTS

Study title	DARWIN EU® - Incidence, period prevalence, and characterisation of individuals with paediatric pulmonary arterial hypertension
Protocol version	V7.0
Date	19/08/2025
EUPAS number	EUPAS1000000716
Active substance	*Bosentan, ambrisentan, macitentan, sildenafil, tadalafil, riociguat*, *treprostinil, epoprostenol, Iloprost, selexipag, and ralinepag*
Medicinal product	All medicinal products with the ingredients listed within the classes: endothelin receptor antagonists (ERAs, *bosentan, ambrisentan, macitentan*), phosphodiesterase type 5 inhibitors (PDE5-is, *sildenafil, tadalafil*), soluble guanylate cyclase stimulators (sGC, *riociguat*), and prostacyclin receptor agonists (*treprostinil, epoprostenol, iloprost, selexipag, ralinepag*)
Research question and objectives	1. Estimate the yearly incidence and period prevalence of pulmonary arterial hypertension (PAH) in the paediatric population, stratified by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years) 2. Characterise paediatric patients newly diagnosed with PAH: a) Describe the number and proportions of individuals by sex and age at index date b) Within 180-days prior to index date and then within the first five years after index date, within sequential 90-day periods, potential aetiology (congenital heart disease, bronchopulmonary dysplasia, congenital diaphragmatic hernia, persistent pulmonary hypertension of the newborn) and comorbidities (right heart failure, ascites, arrythmia, haemoptysis, lung-heart transplant, atrial septostomy or Pott shunt, syncope) c) Within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag), or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists d) Describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag), or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + Prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years) at time of prescription/dispensing e) Within 180-days prior to index date and then within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals with at least one record for each of the following measures: 6-minute walk distance (6MWD) test, echocardiography, NT-proBNP, WHO functional class, right heart catheterisation, and cardiovascular MRI (Magnetic resonance imaging). f) Within the first five-years after index date in sequential 90-day periods, describe the number of and proportion of individuals who were admitted to hospital or died
Countries of study	Denmark, Finland, France, Germany, Norway, Sweden

Authors	Nicholas Hunt, n.hunt@darwin-eu.org Katia Verhamme, k.verhamme@darwin-eu.org

LIST OF ABBREVIATIONS

Acronyms/terms	Description
CC	Coordinating centre
CDM	Common Data Model
DARWIN EU®	Data Analysis and Real World Interrogation Network
DK-DHR	Danish Data Health Registries
DRE	Digital Research Environment
DQD	Data Quality Dashboard
ED	Emergency department
EHR	Electronic health records
EMA	European Medicines Agency
ERA	Endothelin receptor antagonist
FinOMOP - THL	Consortium of the Finnish OMOP data partners - Finnish Institute for Health and Welfare
GDPR	General Data Protection Regulation
GP	General practitioner
HI-SPEED	Health Impact - Swedish Population Evidence Enabling Data-linkage
InGef RDB	Institut für angewandte Gesundheitsforschung research database
IP	Inpatient
IR	Incidence rate
MRI	Magnetic resonance imaging
NLHR	Norwegian Linked Health Registry
NT-proBNP	N-terminal prohormone of brain natriuretic peptide
OMOP	Observational Medical Outcomes Partnership
OP	Outpatient
PAH	Pulmonary arterial hypertension
PDCO	Paediatric Committee
PDE-5i	Phosphodiesterase type 5 inhibitor
PIP	Paediatric Investigation Plan
PPHN	Persistent pulmonary hypertension of the newborn
SMPA-GU	Swedish Medical Products Agency – Gothenburg University
sGC	Soluble guanylate cyclase
WHO	World Health Organisation

1. TITLE

DARWIN EU® - Incidence, period prevalence, and characterisation of individuals with paediatric pulmonary arterial hypertension

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator	Nicholas Hunt Katia Verhamme	Erasmus MC
Data Scientist	Ioanna Nika Cesar Barboza	Erasmus MC
Study Manager	Natasha Yefimenko	Erasmus MC
Data Partner*	**Names**	**Organisation**
CDWBordeaux	Guillaume Verdy Romain Griffer	Bordeaux University Hospital
DK-DHR	Claus Møldrup Elvira Bräuner Susanne Bruun	Danish Medicines Agency
FinOMOP-THL	Tiina Wahlfors Gustav Klingstedt Toni Lehtonen	Finnish Institute for Health and Welfare
HI-SPEED	Fredrik Nyberg Huiqi Li	University of Gothenburg
InGef RDB	Josephine Jacob Raeleesha Norris Alexander Harms Annika Vivirito	Institut für Angewandte Gesundheitsforschung
NLHR	Saeed Hayati Nhung Trinh Hedvig Nordeng Maren Mackenzie Olson	University of Oslo

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

3. ABSTRACT

Title

DARWIN EU® - Incidence, period prevalence, and characterisation of individuals with paediatric pulmonary arterial hypertension

Rationale and background

The intention of the study is to investigate the occurrence of pulmonary arterial hypertension (PAH) in paediatric patients by age group and to understand the size of this population in various EU countries. Additionally, the study aims to characterise the disease and treatments use in the paediatric PAH population in a real-world setting. This study can therefore be used in any future paediatric PAH related regulatory procedures.

Research objectives

1. Estimate the yearly incidence and period prevalence of pulmonary arterial hypertension (PAH) in the paediatric population, stratified by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years)
2. Characterise paediatric patients newly diagnosed with PAH:
    a) Describe the number and proportions of individuals by sex and age at index date
    b) Within 180-days prior to index date and then within the first five years after index date, within sequential 90-day periods, potential aetiology (congenital heart disease, bronchopulmonary dysplasia, congenital diaphragmatic hernia, persistent pulmonary hypertension of the newborn) and comorbidities (right heart failure, ascites, arrythmia, haemoptysis, lung-heart transplant, atrial septostomy or Potts shunt, syncope)
    c) Within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag) or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + Prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists
    d) Describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag) or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + Prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years) at time of prescription/dispensing
    e) Within 180-days prior to index date and then within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals with at least one record for each of the following measures: 6-minute walk distance (6MWD) test, echocardiography, NT-proBNP, WHO functional class, right heart catheterisation, and cardiovascular MRI (Magnetic resonance imaging).
    f) Within the first five-years after index date in sequential 90-day periods, describe the number of and proportion of individuals who were admitted to hospital or died

Study design

Cohort study of population-level descriptive disease epidemiology and characterisation.

Study period

The study period will be from 01/01/2014 to 31/12/2024.

Population

For objective 1, all individuals (<18 years) will be included from the first day in each calendar year. For incidence calculation they should have no prior PAH diagnosis at this index date and for period prevalence this does not apply. For objective 2, the index date is the date of first diagnosis of PAH (with no prior occurrence). For both objectives, individuals with less than 365 days continuous observation prior to the index date will be excluded (does not apply to individuals aged less than one year old, or those followed in CDWBordeaux).

Variables

The outcome for objective 1 is PAH diagnosis. For objective 2, covariates will be used to characterise the included individuals with PAH: Demographic factors include sex, age and age group; conditions include right heart failure, ascites, arrythmia, haemoptysis, and syncope, congenital heart disease, bronchopulmonary dysplasia, congenital diaphragmatic hernia, persistent pulmonary hypertension of the newborn; procedures or measurements include lung-heart transplant, atrial septostomy or Pott shunt, 6 minute walking test, echocardiography, NT-proBNP test, WHO functional class, right heart catheterisation, cardiovascular MRI; and drug treatments, including mono- and combination therapies of endothelin receptor antagonists, phosphodiesterase type 5 inhibitors, soluble guanylate cyclase stimulators, and prostacyclin receptor agonists.

Data sources

1. Clinical Data Warehouse of Bordeaux University Hospital (CDWBordeaux), France
2. Danish Health Registries (DK-DHR), Denmark
3. Consortium of the Finnish OMOP data partners (FinOMOP-THL), Finland
4. Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED), Sweden
5. Institut für angewandte Gesundheitsforschung Berlin GmbH (InGef RDB), Germany
6. Norwegian Linked Health Registry data (NLHR), Norway

Statistical analysis

The calculation of PAH incidence rates with 95% confidence intervals per data source will be stratified by year and age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years). Period prevalence will be reported as the percentage of patients with new or ongoing PAH of the total population at risk in each data source, stratified by year and age group. We will characterise individuals with PAH in terms of demographics at index date, as well as the number and proportion of records of comorbidities, procedures, treatments, hospitalisation in sequential 90-day windows after index date, and at baseline [-180,0] for procedures, measurements, and acute conditions. Chronic conditions will be measured at index date and any time prior [-inf,0]. Characterisation by number and proportion of treatment records will also be stratified by age group at time of prescribing/dispensing.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study milestones and deliverables	Planned dates*
Final Study Protocol	To be confirmed by EMA
Creation of Analytical code	5th June 2025
Execution of Analytical Code on the data	25th August 2025
Draft Study Report	30th September 2025
Final Study Report	To be confirmed by EMA

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

6. RATIONALE AND BACKGROUND

This study was triggered by discussions in the paediatric committee (PDCO) about the Paediatric Investigation Plan (PIP) for the endothelin receptor blocker Ambrisentan. The scope of the study was extended beyond describing the use of Ambrisentan, aiming to investigate the occurrence of pulmonary arterial hypertension (PAH) in paediatric patients and to characterise the disease and treatment use in a real-world setting. This study can therefore be used in any future PAH related regulatory procedures.

The course of PAH will lead to hypertrophy and remodelling of the right ventricle, and, if untreated, may lead to death.(1) Pulmonary hypertension (PH) is divided into 5 groups.(2) Group 1 is PAH, and in this group, there are different subgroups. The most prevalent groups in the paediatric population are idiopathic PAH, PAH associated with congenital heart disease, and persistent pulmonary hypertension of the newborn. Group 2 includes pulmonary hypertension due to left-sided heart disease, group 3 includes pulmonary hypertension due to lung diseases and/or hypoxia, group 4 includes chronic thromboembolic pulmonary hypertension, and group 5 includes pulmonary hypertension with unclear, multifactorial, or other uncommon causes. Two common causes of pulmonary hypertension in children are bronchopulmonary dysplasia and congenital diaphragmatic hernia, these two are included in group 3.

In the general adult population, as estimated in French hospital data sources, there was an annual incidence of PAH of two to five cases per million.(3) However, in children it is rare and therefore it is difficult to enrol this specific population in clinical trials. Treatment patterns and disease course in the paediatric population remains understudied. The use of real-world data from across the DARWIN EU® network will be leveraged to cover wider populations across diverse data source to better understand the incidence, prevalence, and characterisation including treatment course for the European population. Building on the evidence from a previous DARWIN EU® study, that study specifically investigated co-prescribing the endothelin receptor antagonists (ERAs) and phosphodiesterate-5 inhibitors (PDE-5is) in individuals with PAH.(4)

7. RESEARCH QUESTION AND OBJECTIVES

1. Estimate the yearly incidence and period prevalence of pulmonary arterial hypertension (PAH) in the paediatric population, stratified by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years)
2. Characterise paediatric patients newly diagnosed with PAH:
    a) Describe the number and proportions of individuals by sex and age at index date
    b) Within 180-days prior to index date and then within the first five years after index date, within sequential 90-day periods, potential aetiology (congenital heart disease, bronchopulmonary dysplasia, congenital diaphragmatic hernia, persistent pulmonary hypertension of the newborn) and comorbidities (right heart failure, ascites, arrythmia, haemoptysis, lung-heart transplant, atrial septostomy or Potts shunt, syncope)
    c) Within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag) or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists
    d) Describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag) or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + Prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years) at time of prescription/dispensing
    e) Within 180-days prior to index date and then within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals with at least one record for each of the following measures: 6-minute walk distance (6MWD) test, echocardiography, NT-proBNP, WHO functional class, right heart catheterisation, and cardiovascular MRI (Magnetic resonance imaging
    f) Within the first five-years after index date in sequential 90-day periods, describe the number of and proportion of individuals who were admitted to hospital or died

Description of the proposed objectives to be achieved in the study (**Table 1**).

Table 1. Objectives.

Objective 1

Objective:	Estimate the yearly **incidence** and **period prevalence** of pulmonary arterial hypertension, in the paediatric population, stratified by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years)
Hypothesis:	No hypothesis testing
Population*:*	**Incidence population** Inclusion criteria • Aged <18 years at index date • at least 365 days observation (minimum observation does not apply to those aged <1 year or individuals followed in hospital setting) Exclusion criteria • Occurrence of pulmonary arterial hypertension prior to index date **Prevalence population** Inclusion criteria • Aged <18 years at index date • at least 365 days observation (minimum observation does not apply to those aged <1 year or individuals in the hospital setting, CDWBordeaux)
Exposure:	n/a
Comparator:	n/a
Outcome:	Pulmonary arterial hypertension (first diagnosis for incidence estimation, or ongoing disease from a diagnosis in a prior period is included for prevalence)
Time *(when follow up begins and ends):*	The index date is the first date in the study period, where an individual satisfies the eligibility criteria for each calendar year. Individuals will be followed until the first of: occurrence of the outcome, the date of their 18th birthday, end of the study period (31/12/2024, or latest data), end of the study period calendar year, death, or disenrollment from the data source.
Setting:	Primary and secondary care
Main measure of effect:	Incidence rates and period prevalence (percentage)

Objective 2

Objective:	Characterise paediatric patients newly diagnosed with PAH: a) Describe the number and proportions of individuals by sex and age at index date b) Within 180-days prior to index date and then within the first five years after index date, within sequential 90-day periods, potential aetiology (congenital heart disease, bronchopulmonary dysplasia, congenital diaphragmatic hernia, persistent pulmonary hypertension of the newborn) and comorbidities (right heart failure, ascites, arrythmia, haemoptysis, lung-heart transplant, atrial septostomy or Potts shunt, syncope) c) Within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors

	(PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag) or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + Prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists d) Describe the number and proportion of individuals treated with monotherapy of the following treatments, including endothelin receptor antagonists (ERAs, bosentan, ambrisentan, macitentan), phosphodiesterase type 5 inhibitors (PDE5-is, sildenafil, tadalafil), soluble guanylate cyclase stimulators (sGC, riociguat), prostacyclin receptor agonists (treprostinil, epoprostenol, iloprost, selexipag, ralinepag) or combination therapy of these classes, including ERAs + PDE5-I, ERAs + sGC, PDE5-i + Prostacyclin receptor agonists, and ERA + PDE5-I + prostacyclin receptor agonists by age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years) at time of prescription/dispensing e) Within 180-days prior to index date and then within the first five-years after index date, within sequential 90-day periods, describe the number and proportion of individuals with at least one record for each of the following measures: 6-minute walk distance (6MWD) test, echocardiography, NT-proBNP, WHO functional class, right heart catheterisation, and cardiovascular MRI (Magnetic resonance imaging) f) Within the first five-years after index date in sequential 90-day periods, describe the number of and proportion of individuals who were admitted to hospital or died
Hypothesis:	No hypothesis testing
Population*:*	Inclusion criteria • Recorded diagnosis of pulmonary arterial hypertension at index date • Aged <18 years at index date • Observation in the data source of 365 days prior to the index date (except for those aged less than one year old and except for individuals in the hospital setting, CDWBordeaux) Exclusion criteria • Occurrence of pulmonary arterial hypertension prior to index date • Occurrence of right sided heart failure prior to index date (for characterisation of individuals in terms of right sided heart failure incidence)
Exposure:	n/a
Comparator:	n/a
Outcome:	n/a
Time *(when follow up begins and ends):*	The index date is the date of first pulmonary arterial hypertension diagnosis in the study period. Individuals will be followed until the date of their 18th birthday, end of the study period (31/12/2024, or latest data), death, disenrollment from the data source after index date or five-years after index date.
Setting:	Primary and secondary care
Main measure of effect:	Counts and percentage

8. RESEARCH METHODS

8.1. Study type and study design

For the first objective, we will perform a disease epidemiology study to estimate the incidence and period prevalence of PAH.

For the second objective, we will perform a characterisation study to describe the demographic, conditions, procedures, hospital admissions, and treatments for paediatric individuals newly diagnosed with PAH. We will also describe the number and proportion of individuals with a record of PAH drug treatment by age group.

An overview of the study design for objective 1 can be seen in **Figure 1** and for objective 2 in **Figure 2**.

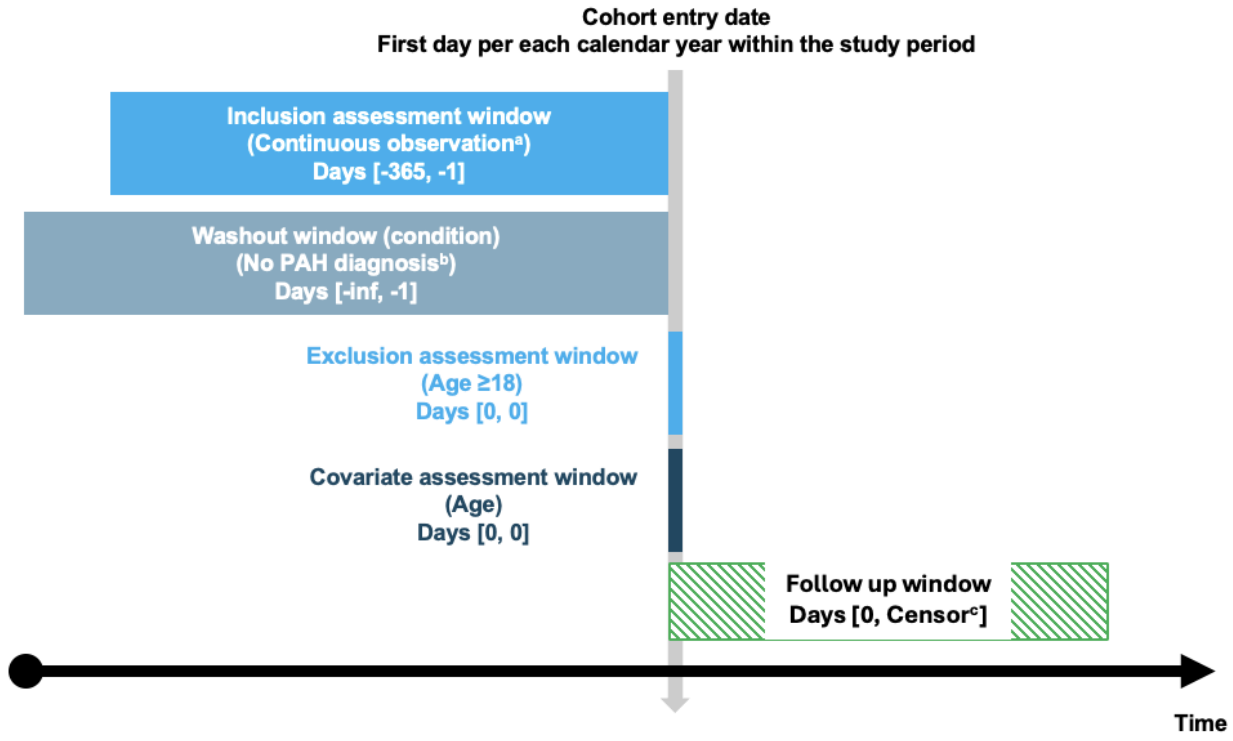

Figure 1. A graphical depiction of the study design for objective 1.

a. Applies to individuals older than one year
b. Applies for incidence estimation only
c. First of death, disenrollment, end of data source availability, occurrence of PAH (incidence estimation only), end of each calendar year (i.e., 31st December), or end of the study period (31/12/2024)

PAH = pulmonary arterial hypertension

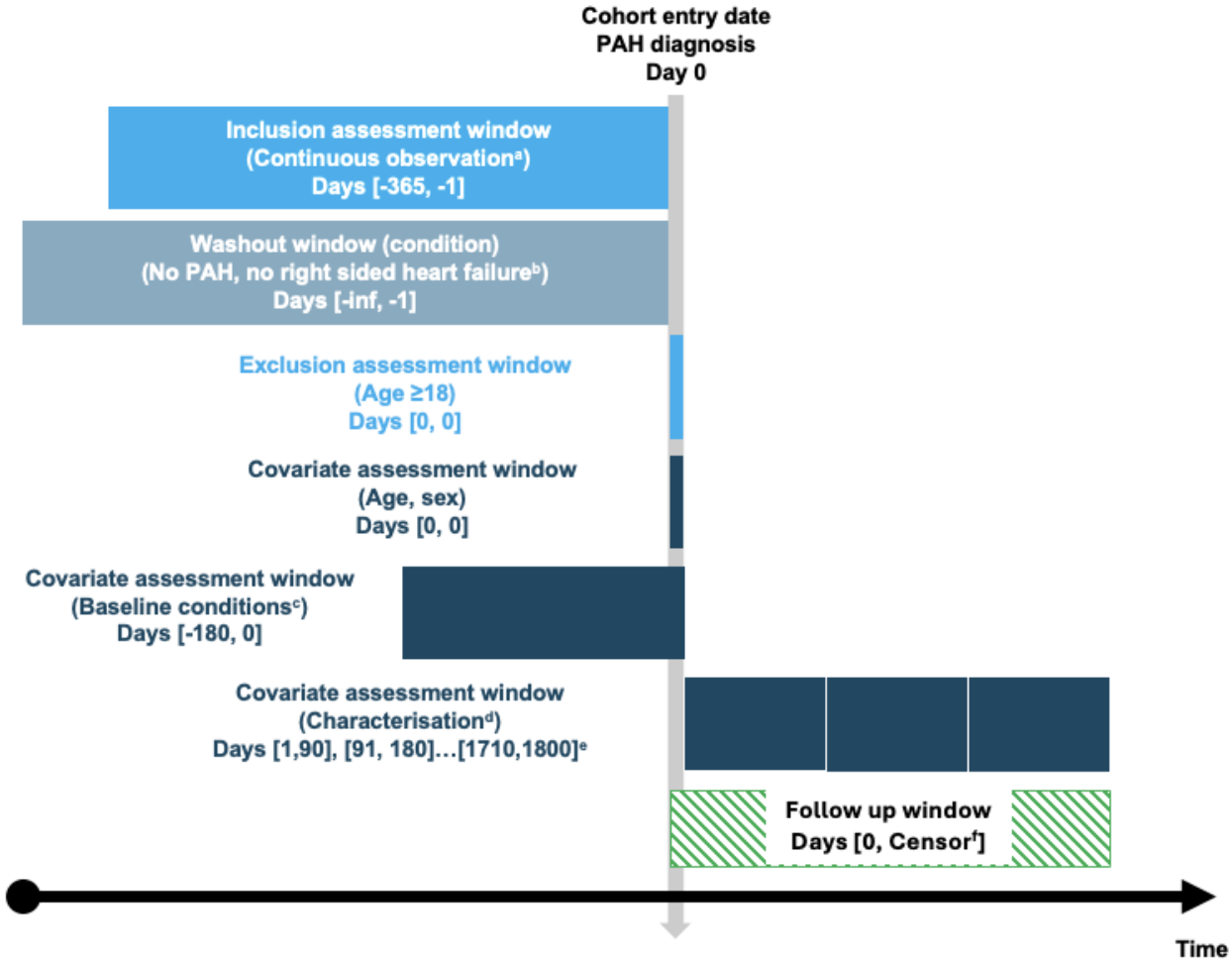

Figure 2. A graphical depiction of the study design for objective 2.

a. Applies to individuals older than one-year and those in CDWBordeaux
b. "No right sided heart failure" only applies in the characterisation of individuals in terms of right sided heart failure incidence
c. Comorbidities (defined in 8.6.3)
d. Comorbidities, drugs, tests, occurrence of death, occurrence of hospitalisation or emergency department visit (all defined in section 8.6.3)
e. Assessed within sequential 90-day intervals until five years (twenty 90-day intervals)
f. First of death, disenrollment, end of data source availability, five years, or end of the study period (31/12/2024)

PAH = pulmonary arterial hypertension

8.2. Study setting and data sources

This study will be conducted using routinely collected data from six data sources in six European countries, five of which are EU member states. All data were a priori mapped to the OMOP CDM. The data sources will include:

1. Clinical Data Warehouse of Bordeaux University Hospital (CDWBordeaux), France
2. Danish Health Registries (DK-DHR), Denmark
3. Consortium of the Finnish OMOP data partners (FinOMOP-THL), Finland
4. Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED), Sweden
5. Institut für angewandte Gesundheitsforschung Berlin GmbH (InGef RDB), Germany
6. Norwegian Linked Health Registry data (NLHR), Norway

For this study, six data sources from the DARWIN EU® Database Catalogue were selected. In general, the selection process was based on the number of individuals with the diagnosis of interest, geographical spread, and the experience gained from databases that participated in other similar DARWIN EU® studies. For this specific study, the selection process also extended to ensure that some of the included data sources had coverage of both the inpatient and outpatient settings, as individuals with PAH are often treated over both healthcare settings. Based on the feasibility assessment performed, the suggested databases are considered fit for purpose for at least part of the objectives.

When it comes to assessing the reliability of data sources, the data partners are asked to describe their internal data quality process on the source data as part of the DARWIN EU® onboarding procedure. To further ensure data quality, we utilise the Achilles tool, which systematically characterises the data and generates data characteristics such as age distribution, condition prevalence per year, data density, measurement value distribution which can be compared against expectations for the data. Additionally, the data quality dashboard (DQD) provides more objective checks on plausibility consistently across the data sources. In terms of relevance, more general-purpose diagnostic tools, CohortDiagnostics and DrugExposureDiagnostics, were developed. The CohortDiagnostics package provides additional insights into cohort characteristics, record counts, and index event misclassification. The DrugExposureDiagnostics package assesses ingredient specific diagnostics for drug exposure records. Furthermore, data is maintained up to date by extracting the release dates for each dataset in the network and monitoring when data is out-of-date with the expected refresh cycle (typically quarterly or half-yearly). In addition, it is important to have a clear understanding of the time period covered by each released database, as this can vary across different domains. To facilitate this, the CDMOnboarding (and Achilles) packages contain a ‘data density’ plot. This plot displays the number of records per OMOP domain on a monthly basis. This allows getting insights when data collection started, when new sources of data were added, and until when data was included.

CDWBordeaux does not have continuous follow-up of individuals contained within the data source. As such, this data source will not be used to calculate incidence or period prevalence (objective 1) but only characterisation (objective 2). In addition, for CDWBordeaux, there will be no requirement of 365 days observation prior to the index date to be included.

Clinical Data Warehouse of Bordeaux University Hospital (CDWBordeaux), France

The clinical data warehouse of the Bordeaux University Hospital comprises electronic health records on more than 2 million patients with data collection starting in 2005. The hospital complex is made up of three main sites and comprises a total of 3,041 beds (2021 figures). The data source currently holds information about the person (demographics), visits (inpatient and outpatient), conditions and procedures (billing codes), drugs (outpatient prescriptions and inpatient orders and administrations), measurements (laboratory tests and vital signs), and dates of death (in or out-hospital death).(5)

Danish Data Health Registries (DK-DHR), Denmark

Danish health data is collected, stored, and managed in national health registers at the Danish Health Data Authority, and covers the entire population, which makes it possible to study the development of diseases and their treatment over time. There are no gaps in terms of gender, age, and geography in Danish health data due to mandatory reporting on all patients from cradle to grave, in all hospitals and medical clinics. Personal identification numbers enable linking of data across registers, so it captures data on all Danes throughout their lives, regardless of whether they have moved around the country. High data quality due to standardisation, digitisation, and documentation means that Danish health data is not based on interpretation. The Danish Health Data Authority is responsible for the national health registers and for maintaining and developing standards and classifications in the Danish healthcare system. Legislation ensures balance between personal data protection and use. The current data release includes data on the entire Danish population of 5.9 million persons from 1995. It includes data from the following registries: The central Person Registry, The National Patient Registry, The Register of Pharmaceutical Sales, The National Cancer Register, The Cause of Death registry, The Laboratory Database (including Coronavirus disease 2019 test results), and The Vaccination Registry (including COVID-19 vaccinations).

Finnish Care Register for Health Care (FinOMOP-THL), Finland

The THL database covers both public and private, primary and specialised inpatient and outpatient health care encounters in Finland starting from 2011. The entire public sector and private inpatient encounters have been included since 2011, while private outpatient encounters, including occupational care, are included since 2020. The main content of the THL CDM is The Finnish Care Register for Health Care (fi:Hoitoilmoitusrekisteri, HILMO). It is a continuation of the former Hospital Discharge Register, which originally gathered data on patients discharged from hospitals. The Care Register has comprehensive data on the use of services and service users from Finnish public inpatient and outpatient primary and specialised care nationwide. Since 1998 the register has covered both public outpatient and inpatient specialised care and private inpatient care (TerveysHilmo). From 2011 the register has covered public primary care (AvoHilmo). From 2020 the register has covered private outpatient care and occupational care. In addition, the CDM also contains the vaccination data from the Finnish National Vaccination Register, and positive COVID-19 test results from the Finnish National Infectious Diseases Register, which is maintained by THL. The CDM is currently produced from the above-mentioned and limited to observation periods commencing after 1/1/2011. The National Population registry is also used as a source for the CDM database. The National Population registry data forms the basis for forming the patient population. This ensures up-to-date location (municipality of residence) of patients as well as complete death occurrences (although not the cause of death). Using the complete population as a basis for the person table also serves to facilitate calculations on a population level, e.g. incidence rates. The current CDM population comprises all persons having been alive and residing in Finland since the beginning of 2011.

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

The Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED) study is a nationwide linked multi-register, regularly updated, observational study for timely response over time to scientific questions around effectiveness and safety of approved drugs that can arise suddenly, requiring rapid evidence for timely regulatory action - to protect patients' health and lives. The study data covers the whole Swedish population (about 10 million), with data on specialist care (National Patient Register), drug use (Prescribed Drug Register), cause of death (Cause-of-Death Register), sociodemographic data, and selected clinical data. Most data start from 2015; prescription drug data is available from 2018. The study population and all data is updated quarter yearly. HI-SPEED builds on the predecessor project SCIFI-PEARL (Swedish COVID-19 Investigation for Future Insights - a Population Epidemiology Approach using Register Linkage) that was initiated in 2020 to conduct research on Covid-19 and pandemic-related questions.

Institut für angewandte Gesundheitsforschung Berlin GmbH (InGef RDB), Germany

The InGef database comprises anonymized longitudinal claims data of about 10 million individuals across approximately 50 statutory health insurance providers (SHIs) throughout Germany. Data is longitudinally linked over a period of currently ten years. Patients can be traced across health care sectors. All patient-level and provider-level data in the InGef research database are anonymised to comply with German data protection regulations and German federal law. German SHI claims data available in the InGef database includes information on demographics (year of birth, gender, death date if applicable, region of residence on administrative district level); hospitalisations; outpatient services (diagnoses, treatments; specialties of physicians); dispensing of drugs; dispensing of remedies and aids; and sick leave and sickness allowance times. In addition, costs or cost estimates from SHI perspective are available for all important cost elements. All diagnoses in Germany are coded using the International Classification of Diseases, version 10 in the German Modification (ICD-10-GM). The persistence (membership over time) is rather high in the InGef database: During a time period of 5 years (2009 to 2013), 70.6% of insurance members survived and remained insured with the same SHI without any gap in their observational time. Persons leaving one of the participating SHIs and entering another participating SHI, can be linked during yearly database consistency updates and are thus not lost over time. The InGef database is dynamic in nature, i.e. claims data are updated in an ongoing process and new SHIs may join or leave the database. By law, only the last 10 years of data are allowed to be used. At every new release this window shifts, dropping older data and adding new data.

Norwegian Linked Health Registry data (NLHR), Norway

Norway has a universal public health care system consisting of primary and specialist health care services covering a population of approximately 5.4 million inhabitants. Many population-based health registries were established in the 1960s with use of unique personal identifiers facilitating linkage between registries. Data in these health registries are used for health analysis, health statistics, improving the quality of healthcare, research, administration, and emergency preparedness. We harmonised data from the following registries: the Medical Birth Registry of Norway (MBRN), the Norwegian Prescription Registry (NorPD), the Norwegian Patient Registry (NPR), Norway Control and Payment of Health Reimbursement (KUHR), the Norwegian Surveillance System for Communicable Diseases (MSIS), the Norwegian Immunisation Registry (SYSVAK), the National Death Registry, and the National Registry (NR). Linkage between the registries was facilitated using project-specific person ID generated from unique personal identification assigned at birth or immigration for all legal residents in Norway. In brief: MBRN stores information about the pregnancy, the mother, father and child; NPR records diagnosis in secondary care (e.g., hospital); KUHR contains information about diagnosis and contact in primary care (e.g., GPs and outpatient specialists) – to be included in third release; NorPD recorded all medications dispensed outside of hospitals; MSIS collects test results of communicable diseases (e.g., Sars-Cov-2); and SYSVAK recorded vaccinations.

Information on data sources planned to be used with a justification for their choice in terms of ability to capture the relevant data is described in **Table** .

Table 2. Description of the selected data sources.

Country	Name of Database	Justification for Inclusion	Objectives	Health Care setting	Type of Data	Number of active subjects (total subjects)	Feasibility count of disease (PAH)	Data lock for the last update
France	CDWBordeaux	Adequate number of individuals with the disease of interest (PAH) Contributes to the geographical diversity of data sources included Adequate coverage of PAH treatments	2	Hospital care	EHR	246k (2.3m)	4000	01/09/2024
Denmark	DK-DHR	Adequate number of individuals with the disease of interest (PAH) Coverage of in- and outpatient prescribing data Nationwide denominator Contributes to the geographical diversity of data sources included Adequate coverage of PAH treatments	1, 2	Primary care and hospital	Registry	5.9m (8.6m)	16100	18/01/2025
Finland	FinOMOP-THL	Adequate number of individuals with the disease of interest (PAH) Contributes to the geographical diversity of data sources included Adequate coverage of PAH treatments	1, 2	Primary and secondary care	Registry	5.7m (6.6m)	1500	01/10/2024
Sweden	HI-SPEED	Adequate number of individuals with the disease of interest (PAH) Nationwide denominator. Contributes to the geographical diversity of data sources included Adequate coverage of PAH treatments	1, 2	Secondary care with linkage to primary care (primary care available for 40% of population)	Registry	10.6m (11.7m)	9700	01/08/2024

Country	Name of Database	Justification for Inclusion	Objectives	Health Care setting	Type of Data	Number of active subjects (total subjects)	Feasibility count of disease (PAH)	Data lock for the last update
Germany	InGef RDB	Adequate number of individuals with the disease of interest (PAH) Nation representative denominator Contributes to the geographical diversity of data sources included Adequate coverage of PAH treatments	1, 2	Primary and secondary care (in-and outpatient for objective 1 and characterisation on objective 2), secondary care (inpatient only objective 2 index date)	Claims	7.7m (10.5m)	9500	18/04/2025
Norway	NLHR	Adequate number of individuals with the disease of interest (PAH) Nationwide denominator. Contributes to the geographical diversity of data sources included Adequate coverage of PAH treatments	1, 2	Primary and secondary care	Registry	6.9m (7.3m)	1300	01/12/2024

Abbreviations: pulmonary arterial hypertension, PAH; electronic health records, EHR; Clinical Data Warehouse of Bordeaux, CDWBordeaux; Danish Data Health Registries, DK-DHR; Consortium of the Finnish OMOP data partners - Finnish Institute for Health and Welfare, FinOMOP-THL, Institut für angewandte Gesundheitsforschung, InGef; research database, RDB; NLHR, Norwegian Linked Health Registry; Health Impact - Swedish Population Evidence Enabling Data-linkage, HI-SPEED.

8.3. Study period

The study period will be from 1st January 2014 to 31st December 2024.

8.4. Follow-up

For objective 1, the index date is the first day of each year within the study period which satisfies the eligibility criteria (specified in section 8.5). For objective 2, the index date is the date of first diagnosis of pulmonary arterial hypertension in the study period, with no prior diagnosis of pulmonary arterial hypertension. For InGef RDB, the outpatient data source will not be used for the diagnosis defining the index date of objective 2 (characterisation) due to potential misclassification of diagnosis dates.

Individuals will be followed up until the first of: death, end of observation in data, end date of each sequential year within the study period, or end of the study period (31/12/2024). For objective 1 incidence estimation, follow-up will stop at the date of first PAH diagnosis. For estimating period prevalence, all condition records of PAH within the study period will be used.

The operational definitions of the index dates are described in **Table 3** and the concept set definitions of pulmonary arterial hypertension are listed in **Table S1**.

Table 3. Operational definition of time 0 (index date) and other primary time anchors.

Study population names	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type	Diagnosis position	Incident with respect to…
Incidence estimation	Time 0 (index date)	Multiple	n/a	[-inf,-1]	IP, OP	SNOMED	n/a	Pulmonary arterial hypertension
Prevalence estimation (PAH)	Time 0 (index date)	Multiple	n/a	none	IP, OP	n/a	n/a	n/a
Patient characterisation	Time 0 (index date)	Single	Incident	[-inf,-1]	IP, OP[2]	SNOMED	any	Pulmonary arterial hypertension

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable.
[2]Not for InGef RDB

8.5. Study population with inclusion and exclusion criteria

For objective 1 (incidence of individuals with PAH):

Inclusion criteria

- Aged <18 years at index date
- Observation in the data source of 365 days prior to the index date (except for those aged less than one year old and individuals in hospital settings, CDWBordeaux)

Exclusion criteria

- Prior occurrence of pulmonary arterial hypertension

For objective 1 (period prevalence of individuals with PAH):

Inclusion criteria

- Aged <18 years at index date
- Observation in the data source of 365 days prior to the index date (except for those aged less than one year old and individuals in hospital settings, CDWBordeaux)

For objective 2 (characterisation of individuals with PAH)

Inclusion criteria

- Recorded diagnosis of pulmonary arterial hypertension
- Aged <18 years at index date
- Observation in the data source of 365 days prior to the index date (except for those aged less than one year old and individuals in hospital settings, CDWBordeaux)

Exclusion criteria

- Occurrence of pulmonary arterial hypertension prior to index date
- Occurrence of right sided heart failure any time prior to index date (for characterisation of individuals in terms of right sided heart failure incidence)

The operational definitions of the inclusion and exclusion criteria are presented by means of **Table 4** and **Table 5**, respectively.

Table 4. Operational definitions of inclusion criteria.

Criterion	Details	Assessment window	Care Settings	Code Type	Applied to study populations:
PAH diagnosis	Pulmonary arterial hypertension diagnosis	[0,0]	IP[1], OP[2]	SNOMED	Objective 2
Paediatric	Aged <18 years at index date	[0,0]	n/a	n/a	Objective 1 and 2
365 days observation	Data source start date >365 days prior to index date in individuals older than one-year and not in CDWBordeaux	[-365,-1]	n/a	n/a	Objective 1 and 2 (except CDWBordeaux)

1) All included data sources will use information obtained from the inpatient setting to identify incident PAH
2) All included data sources will use information obtained from the outpatient setting to identify incident PAH except for InGef RDB

Abbreviations: PAH = pulmonary arterial hypertension, IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

Table 5. Operational definitions of exclusion criteria.

Criterion	Details	Assessment window	Care Settings[1]	Code Type	Applied to study populations:
No PAH diagnosis before index date[2]	Prior diagnosis record of PAH	[-inf,-1]	IP, OP, ED, OT	SNOMED	Objective 1 (when estimating incidence) and objective 2
No right sided heart failure before index date	Prior diagnosis record of right sided heart failure	[-inf,-1]	IP, OP, ED, OT	SNOMED	Objective 2 (for characterisation of individuals by right sided heart failure only)

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] PAH = pulmonary arterial hypertension

8.6. Variables

8.6.1. Exposures

Characterisation of treatment of PAH is described under other covariates.

8.6.2. Outcomes

The outcome is a record of a pulmonary arterial hypertension diagnosis (for objective 1). The operational definition of this outcome is presented in the **Table 6**. For InGef, we will use diagnoses from both the in- and outpatient data sources for this outcome.

Table 6. Operational definitions of outcome

Outcome name	Details	Type of outcome	Washout window	Care Settings[1]	Code Type	Applied to study populations
Pulmonary arterial hypertension	Diagnosis record of pulmonary arterial hypertension occurring during follow-up	Binary	Yes	IP, OP, ED	SNOMED	Objective 1

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

8.6.3. Other covariates

Age and sex will be measured at index date [0,0]. Individuals will be categorised at index date into age groups (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years).

Occurrence of hospitalisation and death will be assessed within sequential 90-day periods after index date until five-years post index date (e.g., from day one until day 90, from day 91 to day 180… day 1711 to day 1800), or end of follow-up (**Table 7**).

Conditions including right heart failure, ascites, arrythmia, haemoptysis, syncope, and bronchopulmonary dysplasia will be measured at index date minus 180 days [-180,0] and within sequential 90-day periods after index date until five-years post index date (e.g., from day one until day 90, from day 91 to day 180… day 1711 to day 1800), or end of follow-up (**Table 7**). There are several conditions, including congenital heart disease, congenital diaphragmatic hernia, and persistent pulmonary hypertension of the newborn, which will be measured for at any time prior to the index date and index date itself [-inf,0]. For ascites,

cardiac arrythmia, haemoptysis, and syncope, we will apply a washout window of 90-days after each diagnosis record to account for any potential repeated diagnosis records of the same condition. For right heart failure only the first diagnosis record in follow-up will be considered. For InGef, we will use diagnoses from both the in- and outpatient data sources for these covariates.

Cardiac arrhythmia will be assessed as an overall category and further categorised (if counts allow) into type of cardiac arrhythmia, namely atrial flutter, atrial fibrillation, SVT (supraventricular tachycardia), VT (ventricular tachycardia), VF (ventricular fibrillation), sick sinus syndrome, and AV block.(6)

Procedures including lung-heart transplant, atrial septostomy or Pott shunt, 6 minute walking test, echocardiography, NT-proBNP test, WHO functional class, right heart catheterisation, cardiovascular MRI will be measured at index date minus 180 days [-180,0] and within sequential 90-day periods after index date until five-years post index date (e.g., from day one until day 90, from day 91 to day 180… day 1711 to day 1800), or end of follow-up (**Table 7**). For lung-heart transplant and atrial septostomy or Potts shunt, only the first procedure record in follow-up will be considered.

Drug prescription/dispensing records of monotherapy and combination therapy of endothelin receptor antagonists (ERAs, *bosentan, ambrisentan, macitentan*), phosphodiesterase type 5 inhibitors (PDE5-is, *sildenafil, tadalafil*), soluble guanylate cyclase stimulators (*riociguat*), prostacyclin receptor agonists (*treprostinil, epoprostenol, iloprost, selexipag, ralinepag*) will be assessed within sequential 90-day periods after index date until five-years post index date for objective 2c (e.g., from day one until day 90, from day 91 to day 180… day 1711 to day 1800), or end of follow-up (**Table 7**). For objective 2d, drug prescription/dispensing records will be assessed until day 1800.

The operational definition of the covariates is described in the **Table 7** and the associated concept set definitions are listed in **Table S2**.

Table 7. Operational definitions of covariates.

Characteristic	Details	Type of variable	Assessment windows	Care Settings [1]	Code Type	Diagnosis Position[2]	Applied to study populations
Sex	-	Binary	[0,0]	n/a	n/a	n/a	Objective 2
Age	Age at index date	Continuous	[0,0]	n/a	n/a	n/a	Objective 2
Age group	Age group at index date: 0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years	Categorical	[0,0]	n/a	n/a	n/a	Objective 1 and 2
Hospitalisation	Record of hospitalisation or emergency department visit	Binary	90-day sequential intervals[3]	IP, ED	n/a	n/a	Objective 2
Death	Death record	Binary	90-day sequential intervals[3]	IP, ED	n/a	n/a	Objective 2
Right sided heart failure	Condition record of right sided heart failure	Binary	[-inf,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	Any	Objective 2

Characteristic	Details	Type of variable	Assessment windows	Care Settings [1]	Code Type	Diagnosis Position[2]	Applied to study populations
Ascites	Condition record of ascites	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	Any	Objective 2
Cardiac Arrythmia*	Condition record of Cardiac arrhythmia (for cardiac arrhythmia overall) and by type of cardiac arrhythmia*	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	Any	Objective 2
Haemoptysis	Condition record of haemoptysis	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	Any	Objective 2
Syncope	Condition record of syncope	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	Any	Objective 2
Congenital heart disease	Condition record of congenital heart disease	Binary	[-inf,0]	IP, OP, ED	SNOMED	Any	Objective 2
Bronchopulmonary dysplasia	Condition record of bronchopulmonary dysplasia	Binary	[-180,0]	IP, OP, ED	SNOMED	Any	Objective 2
Congenital diaphragmatic hernia	Condition record of congenital diaphragmatic hernia	Binary	[-inf,0]	IP, OP, ED	SNOMED	Any	Objective 2
Persistent pulmonary hypertension of the newborn	Condition record of persistent pulmonary hypertension of the newborn	Binary	[-Inf, 0]	IP, OP, ED	SNOMED	Any	Objective 2
Endothelin receptor antagonists	Monotherapy of either bosentan, ambrisentan, or macitentan	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
Phosphodiesterase type 5 inhibitors	Monotherapy of either sildenafil or tadalafil	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
Soluble guanylate cyclase stimulators	Monotherapy of riociguat	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
Prostacyclin receptor agonists	Monotherapy of either treprostinil, epoprostenol, iloprost, selexipag, or ralinepag	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2

Characteristic	Details	Type of variable	Assessment windows	Care Settings [1]	Code Type	Diagnosis Position[2]	Applied to study populations
Endothelin receptor antagonists and phosphodiesterase type 5 inhibitors	Combination therapy of ERAs + PDE5-I	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
Endothelin receptor antagonists and soluble guanylate cyclase stimulators	Combination therapy of ERAs + sGC	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
Phosphodiesterase type 5 inhibitors and prostacyclin receptor agonists	Combination therapy of PDE5-i + prostacyclin receptor agonists	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
Endothelin receptor antagonists and phosphodiesterase type 5 inhibitors and prostacyclin receptor agonists	Combination therapy of ERA + PDE5-i + prostacyclin receptor agonists	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
No PAH treatment	No treatment by one of the PAH drug classes (ERAs, PDE5-is, prostacyclin receptor agonists, or sGCs)	Binary	90-day sequential intervals[3], [1,1800]	IP, OP, ED	RxNorm	n/a	Objective 2
6 minute walking test	Measurement or procedure record of 6 minute walking test	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED, LOINC	n/a	Objective 2
echocardiography	Measurement or procedure record of echocardiography	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED, LOINC	n/a	Objective 2
NT-proBNP	Measurement record of NT-proBNP test	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	LOINC	n/a	Objective 2
WHO functional class	Measurement record of WHO functional class	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	LOINC	n/a	Objective 2

Characteristic	Details	Type of variable	Assessment windows	Care Settings [1]	Code Type	Diagnosis Position[2]	Applied to study populations
Right heart catheterisation	Procedure record of right heart catheterisation	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	n/a	Objective 2
Cardiovascular MRI	Procedure record of cardiovascular MRI	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	n/a	Objective 2
Lung-heart transplant	Procedure record of lung-heart transplant	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	n/a	Objective 2
Atrial septostomy	Procedure record of atrial septostomy	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	n/a	Objective 2
Potts shunt	Procedure record of Potts shunt	Binary	[-180,0], 90-day sequential intervals[3]	IP, OP, ED	SNOMED	n/a	Objective 2

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)
[3] Sequential 90-day assessment windows from one day after the index date (e.g., from day one until day 90) until five-years after index date
*Cardiac arrhythmia will be classified as "cardiac arrhythmia overall" and further categorized into atrial flutter, atrial fibrillation, SVT (supraventricular tachycardia), VT (ventricular tachycardia), VF (ventricular fibrillation), sick sinus syndrome, and AV block

8.7. Study size

No sample size has been calculated for this study, as it is an exploratory descriptive disease epidemiology study in which the entire available data from each source will be used. A feasibility assessment has been undertaken.

8.8. Analysis

R packages

The analysis will be conducted on data mapped OMOP-CDM using the standardised R packages developed for DARWIN EU® which are described per objective:

Objective 1: Incidence and prevalence of pulmonary arterial hypertension

Calculation of population-based incidence rates and prevalence proportions is part of DARWIN EU® pipelines for population-level descriptive epidemiology. This analysis will be performed using the *IncidencePrevalence* R package and will provide incidence rates, stratified by age and for all age groups combined.(7) To ensure the reliable estimation of incidence rate using an appropriate denominator population, this objective will be investigated only within the data sources with continuous observation: DK-DHR, FinOMOP-THL, HI-SPEED, NLHR, and InGef RDB.

Objective 2: Characterisation of individuals diagnosed with pulmonary arterial hypertension

Characterisation of patient demographics, potential aetiologies, comorbidities, risk factors, and treatments before and after diagnosis will be conducted using the *CohortCharacteristics* R packages.(8) Characterisation by the proportion of individuals with PAH treatment within age groups of treatment will be conducted using the *IncidencePrevalence* R package.(7) To construct combination treatment cohorts, the *CohortConstructor* R package will be utilised.(9) To estimate the number of untreated (by one of the four drug classes), the *DrugUtilisation* R package will be used.(10)

Methods planned to obtain point estimates with confidence intervals of measure of occurrence

Objective 1: Incidence and period prevalence of pulmonary arterial hypertension

Annual incidence rates of pulmonary arterial hypertension will be calculated as the number of newly diagnosed pulmonary arterial hypertension per 100,000 person-years of the population at risk of the condition during the period for each calendar year. Individuals enter the denominator population at the start of each calendar year within the study period, or once the eligibility criteria are fulfilled. Those study participants who enter the denominator population will then contribute time at risk up until their first diagnosis during each calendar year. If they do not have the condition of interest, they will contribute time at risk until the end of the calendar year. Time-at-risk of subjects who die or are diagnosed with PAH will be censored at the time of death or diagnosis date, respectively. Similarly, time at risk of subjects who are lost to follow-up will be censored at the time of loss to follow-up (last contact). Subjects with data until the end of the calendar period without a record of the condition will be administratively censored at the end of each calendar period. Incidence rates will be given together with 95% Poisson confidence intervals. **Figure 3** represents an example of incidence rate estimation. Period prevalence will be calculated by counting the number of individuals with a PAH diagnosis per calendar year, as well as ongoing disease from prior to each study period (**Figure 4**). These counts will be divided by the denominator (the number of persons at risk in the period) to calculate a proportion. Period prevalence will be reported as a percentage with 95% confidence intervals, as estimated by the Wilson Score method.

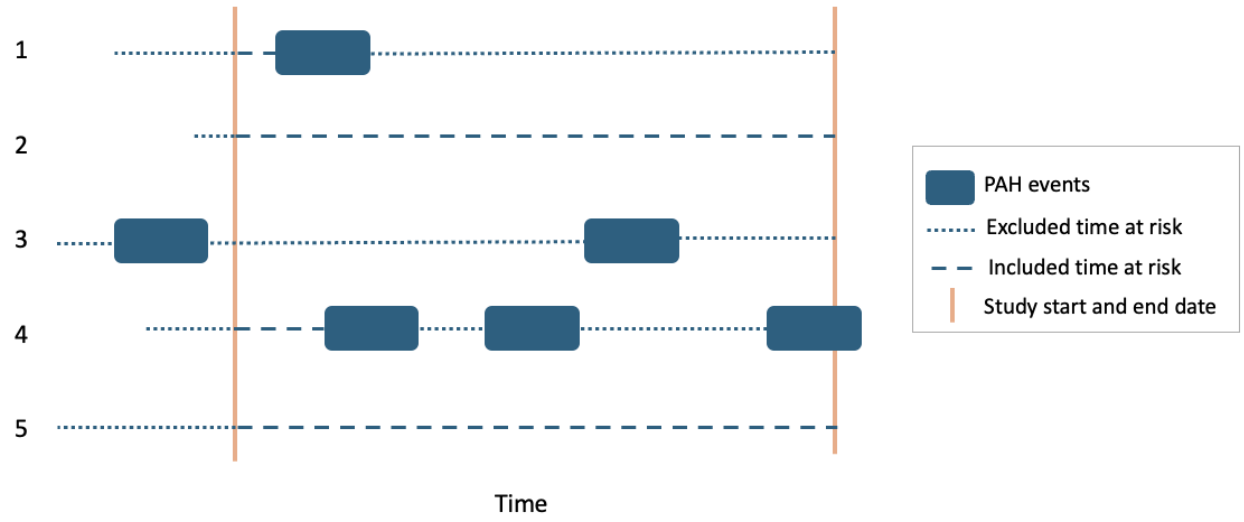

Figure 3. Example of incidence rate estimation. Patient IDs 1 and 4 contribute time at risk between the study start until they have an incident outcome of interest. Patient IDs 2 and 5 contribute time at risk between the study start and end date, as no outcome of interest is observed between this period nor before the study start date. Infinite wash out is applied because only the first outcome after follow-up will be included.

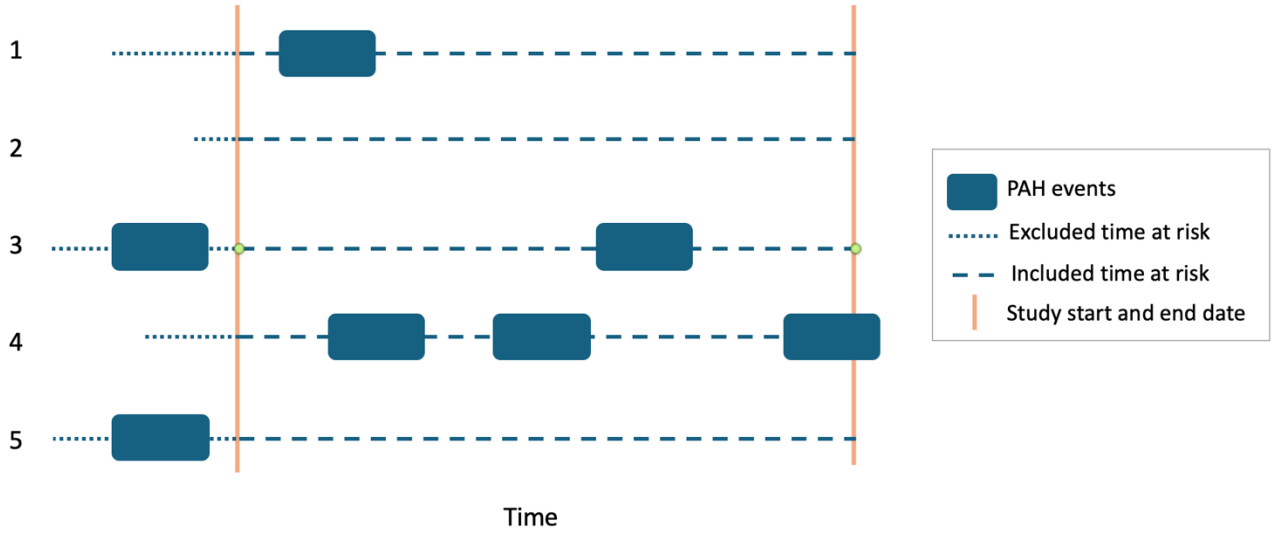

Figure 4. Example of period prevalence estimation. Patients 1, 3 , 4, and 5 contribute to the numerator population, therefore the prevalence would be 80%.

Objective 1 output

- Table 1. Number of participants, number of incident cases, total time at risk, and incidence rate, stratified by age group
- Figures 1-5. Incidence rates of pulmonary arterial hypertension for each calendar year, pooled over all age groups and stratified by age group, per data partner (except CDWBordeaux)
- Figures 6-10. Period prevalence (%) of pulmonary arterial hypertension for each calendar year, pooled over all age groups and stratified by age group, per data partner (except CDWBordeaux)
- Supplementary Tables 1-10. Numbers reported in figures 1-10

Objective 2: Characterisation of individuals diagnosed with pulmonary arterial hypertension

For each patient characteristic (as defined using a list of concepts seen in the appendix) the number and proportion of individuals with a record within each specified time window will be presented. Sex and median age will be measured at index date (i.e., date of diagnosis of incident PAH). For the characterisation of individuals in terms of comorbidities, only the first occurrence of chronic conditions will be considered. For acute conditions (cardiac arrhythmia, ascites, haemoptysis, syncope), we will apply a 90-day washout window prior to each diagnosis to ensure there is no overestimation of conditions due to repeated records. Any repeated records of the same diagnosis within 90-days prior will be disregarded. With the exception of the 90-day window, acute conditions will be measured repeatedly over the course of follow-up. We will exclude individuals who have a prior occurrence of right heart failure at any time prior to index date to ensure it is the incident diagnosis. With one-time procedures (lung-heart transplant and atrial septostomy or Potts shunt), we will only consider the first event during follow-up.

In objective 2c, the number and proportion of individuals prescribed or dispensed PAH drug treatments, including ERAs, PDE5-is, sGCs, or prostacyclin receptor agonists, will be calculated over follow-up, which is divided into 90-day windows from the index date until five-years after index date. Drug eras will be constructed with a maximum gap between two sequential prescriptions of the same drug class (i.e., the end date of one to the start date of the next) of 30 days, adding seven days to the end of last prescription to account for surplus supply.

Combination therapy will include the following: ERAs + PDE5-I, ERAs + sGC, PDE5-i + prostacyclin receptor agonists, and ERA + PDE5-i + prostacyclin receptor agonists. Using drug eras of each individual drug class (ERAs, PDE5-is, sGCs, or prostacyclin receptor agonists), drug combinations will be identified as overlaps in individual drug eras. When one of the components of a combination therapy ends (i.e., end of that individual drug era), the combination ends, and the individual enters monotherapy or another combination therapy (if switching from triple to dual therapy). See **Figure 5** for a graphical depiction of a treatment assignment example.

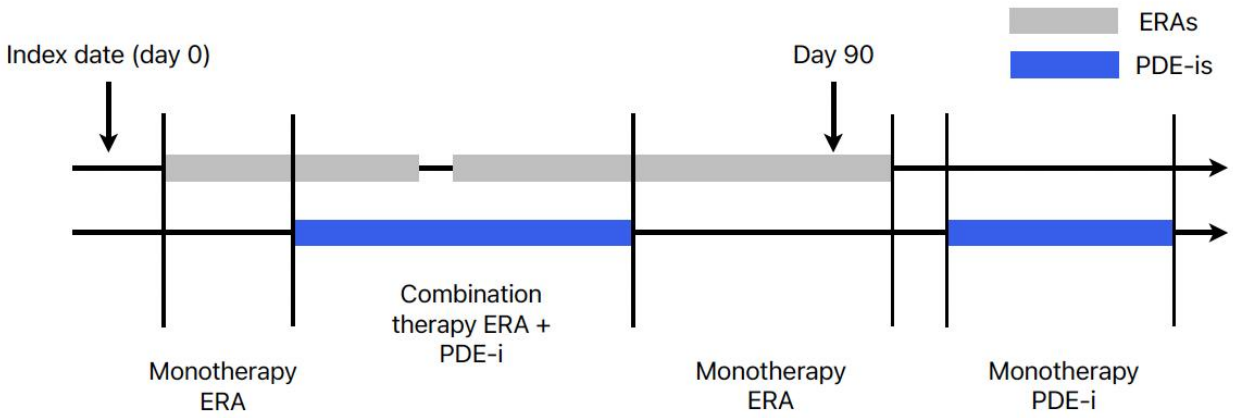

Figure 5. An example of an individual treatment trajectory. The individual is considered to be treated with monotherapy ERA and combination therapy ERA +PDE-I in the day 1 to 90 window. In the 91 to 180 window, the individual considered to be treated with monotherapy ERAs and monotherapy PDE-is.

Combination therapy will be assigned in a hierarchal manner due to the likely event that an individual could fall into multiple predefined groups at one time. The full algorithm can be seen in **Appendix II**. In short, the algorithm first selects the time during triple overlap combination therapy (ERA + PDE5-i + prostacyclin receptor agonists) and excludes it from the time of dual or monotherapy that are subsets of the drugs included in the triple combination. Likewise, this hierarchy applies for the time on the dual therapies (ERAs + PDE5-I, ERAs + sGC, PDE5-i + prostacyclin receptor agonists) over the monotherapies. To ensure any data artifacts, such as treatment overlaps after the algorithm has been applied or overlaps from one therapy line to the next therapy line, do not contribute towards the numerator of the prevalence calculation, we will apply a minimum length of treatment of 30-days for all monotherapies and combination therapies. This minimum length requirement is applied after the assignment of the monotherapy and combination therapy time, during the process of the algorithm.

In objective 2d, the number and proportion of individuals using PAH monotherapies or combination therapies will be estimated within each age group (0 to 1 year, 1 to 2-years, 2 to 5-years, 5 to 12-years, and 12 to 17-years). The follow-up time after the index date until day 1800 for each individual will be divided into these age groups. When a drug cohort (monotherapy or combination therapy) intersects with the age group period in the follow-up, that drug class will be counted and included as a numerator.

A minimum cell counts of 5 will be used when reporting results, with any smaller counts reported as "<5" for privacy protection reasons.

Objective 2 output

- Table 2 Baseline characterisation in terms of sex, age, specified comorbidities, procedures, and measurements, within the six data sources (column)
- Tables 3-8. Characterisation within each sequential three-month follow-up period until five-years post index date in terms of comorbidities, procedures, treatments, measurements, death, and hospitalisation
- Tables 9-14. The number and percentage (prevalence) of individuals with a prescription/dispensing record of each of the PAH treatments within each age group during follow-up
- Figure 11-16. Line plot for treatments. Each line is a different mono therapy, combination therapy, or none of the prespecified therapies. The X-axis is the time at each window, and the Y-axis is the proportion (%).

Methods to deal with missing data

For the disease epidemiology studies we assume that the absence of a diagnosis record means that the person did not receive the diagnosis.

8.9. Evidence synthesis

Estimated incidence of the outcomes will be reported separately for each of the data sources included. No pooling of data will take place in this federated analysis, and we will not perform a meta-analysis of the estimates.

9. DATA MANAGEMENT

9.1. Data Management

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

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

9.2. Data storage and protection

For this study, participants from various EU member states will process personal data from individuals which is collected in national/regional electronic health record 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 are already used for pharmaco-epidemiological research and have a well-developed mechanism to ensure that European and local regulations dealing with ethical use of the data and adequate privacy control are adhered to. In agreement with these regulations, rather than combining

person level data and performing only a central analysis, local analyses will be run, which generate non-identifiable aggregate summary results.

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

10. QUALITY CONTROL

10.1. 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). 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.

10.2. Study specific quality control

Phenotyping of the condition of interest, as well as covariates will be done by two clinical epidemiologists. The quality of the phenotype will be assessed by running CohortDiagnostics on the involved data partners.

11. LIMITATIONS OF THE RESEARCH METHODS

11.1. General limitations

The study will be informed by routinely collected health care data and so data quality issues must be considered. In particular, a recording of a prescription or dispensation does not mean that the patient actually took the drug. In addition, the recording of events used for patient characterisation may vary across databases and may be incomplete.

11.2. Study-specific limitations

Pulmonary arterial hypertension has several aetiologies which would be of interest to measure the incidence of including: pulmonary hypertension due to left heart disease, pulmonary hypertension due to lung diseases or hypoxia, pulmonary hypertension due to pulmonary artery obstructions, and pulmonary hypertension with unclear and/or multifactorial mechanisms. However, as these are complex phenotypes to construct and given the difference in granularity of the data sources, the aetiologies will not be investigated directly by use of aetiology-specific concepts. Some of the concepts used for PAH diagnosis and the characterisation by baseline comorbidities will be used to estimate aetiology. Some paediatric PAH cases can be misclassified as pulmonary hypertension due to lung diseases and/or hypoxia. To countenance this, we will characterise individuals with diagnosis of PAH by diagnosis records of bronchopulmonary dysplasia or congenital diaphragmatic hernia within 180 days prior to index date.

We will not apply the inclusion criteria of a minimum of 365 days observation prior to the index date in the included hospital data source (CDWBordeaux). This may result in misclassification of the index date, as the individuals may have a diagnosis of PAH that falls before observation of the individual in the data source, and therefore the index date is not the incident PAH diagnosis. Furthermore, these hospital data sources cannot be used for incidence or period prevalence calculations, as there is no denominator population available with the constant follow-up as in the registry data sources. The InGef RDB outpatient component of the data source has the date of each diagnosis set to the end of each quarterly period (three-months) in which the diagnosis takes place. This means that date of condition diagnoses recorded in the outpatient setting are likely misclassified. This is particularly important for diagnoses, which define the index date. For objective 2, PAH will therefore only be identified from the inpatient setting in InGef. This will under-report the number of individuals included in the characterisation of PAH. In addition, the incident population in objective 1 will not exactly match the characterisation population in objective 2. Conditions used for characterisation in the InGef outpatient data source will also be misclassified, as the diagnosis date may fall into the incorrect characterisation window.

Individuals will be characterised in terms of use of PAH drugs by drug eras constructed by estimation of the end date of the treatment episode of sequential prescriptions. Prescription end date is often not recorded in RWD, so it is likely imputed. This imputation can differ in terms of how accurately it represents the actual prescription length, depending on medicinal product, indication, and data source. Drug eras in the standard analytics pipeline also do not account for oversupply (i.e., overlapping prescriptions). Furthermore, it will not be possible to accurately estimate possible transitions from one drug class to another drug class, without a study focussing on treatment patterns specifically, such as duration of treatment, treatment overlap, and gaps between treatment. Descriptions of overall prescribing trends through the use of multiple sequential assessment windows will give an understanding of general treatment course.

12. MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS

In agreement with the new guideline on good pharmacovigilance practice (EMA/873138/2011), there will be no requirement for expedited reporting of adverse drug reactions as only secondary data will be used in this study.

13. GOVERNANCE BOARD ASPECTS

The protocol will be submitted to the institutional review boards of FinOMOP-THL, HI-SPEED, NLHR, and CDWBordeaux. DK-DHR and InGef RDB have blanket IRB approval to conduct DARWIN EU® studies.

14. PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

A PDF report including an executive summary, and the specified tables and or figures will be submitted to EMA by the DARWIN EU® CC upon completion of the study. An interactive dashboard incorporating all the results (tables and figures) will be provided alongside the pdf report. The full set of underlying aggregated data used in the dashboard will also be made available if requested. The full report will be adapted to a manuscript to be published in a peer-reviewed scientific journal.

15. REFERENCES

1. Hassoun PM. Pulmonary Arterial Hypertension. N Engl J Med. 2021;385(25):2361-76.

2. Ruopp NF, Cockrill BA. Diagnosis and Treatment of Pulmonary Arterial Hypertension: A Review. Jama. 2022;327(14):1379-91.

3. Humbert M, Sitbon O, Chaouat A, Bertocchi M, Habib G, Gressin V, et al. Pulmonary arterial hypertension in France: results from a national registry. Am J Respir Crit Care Med. 2006;173(9):1023-30.

4. Arinze J, Verhamme, K. DARWIN EU® - Co-prescribing of endothelin receptor antagonists (ERAs) and phosphodiesterate-5 inhibitors (PDE-5is) in pulmonary arterial hypertension (PAH) HMA-EMA Catalogue: EMA; 2023 [Available from: https://catalogues.ema.europa.eu/node/3797/administrative-details.

5. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109.

6. Yinadsawaphan T, Suppah M, Ravi SN, Farina JM, Scott RL, Sorajja D. Epidemiology and Clinical Outcomes of Cardiac Arrhythmias in Pulmonary Arterial Hypertension. CHEST Pulmonary. 2025;3(1):100132.

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

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

9. Burn E CM, Mercade-Besora N, Alcalde-Herraiz M, Du M, Guo Y, Chen X, Lopez-Guell K, Rowlands E CohortConstructor: Build and Manipulate Study Cohorts Using a Common Data Model. R package version 0.4.0 2025 [Available from: https://ohdsi.github.io/CohortConstructor/.

10. Burkard T, López-Güell K, Gorbachev A, Bellas L, Jödicke AM, Burn E, et al. Calculating daily dose in the Observational Medical Outcomes Partnership Common Data Model. Pharmacoepidemiol Drug Saf. 2024;33(6):e5809.

16. ANNEXES

Appendix I: concept definitions (conditions & drugs)

Table S1. Preliminary list of pulmonary arterial hypertension newborn concept IDs (SNOMED)

Concept	Concept IDs (including descendants)	Excluded
Pulmonary arterial hypertension	4013643, 604305, 604306,604307, 715991, 4119611, 4121620, 4124831, 36715093, 40482858, 40493243, 44782560, 44782561, 44782562, 44783618, 44783619, 44783620, 44783621, 44783622, 44783623, 44783624, 44783625, 44783626	4121462

Table S2. Preliminary list of covariate concepts

Concept	Domain	Concept IDs (including descendants)	Ontology
Right heart failure	Condition	312927,317000,608954,4014159,4030258,4079296, 4193236,4195785,4195892,4215446,4233424,42426 69,4273632,4284562,4307356,35615055,37309625, 44782713	SNOMED
Ascites	Condition	200528	SNOMED
Cardiac Arrythmia - Atrial Flutter* - Atrial Fibrillation* - Supraventricular tachycardia* - Ventricular Tachycardia* - Ventricular Fibrillation* - Sick Sinus Syndrome* - AV block*	Condition	44784217 314665 313217 4275423 4103295 437894 4261842 316135	SNOMED
Haemoptysis	Condition	261687	SNOMED
Lung-heart transplant	Procedure	4137127, 4337138	CPT4, SNOMED
Atrial septostomy or Potts shunt	Procedure	4312891, 4336893	CPT4, OPCS4, SNOMED
Syncope	Condition	135360	SNOMED
Congenital heart disease	Condition	312723	SNOMED
Bronchopulmonary dysplasia	Condition	4283942	SNOMED
Congenital diaphragmatic hernia	Condition	4066010, 45881158	SNOMED, LOINC
Persistent pulmonary hypertension of the newborn	Condition	4121462	SNOMED
Endothelin receptor antagonists	Drug	1321636, 1337068, 44507580	RxNorm
Phosphodiesterase type 5 inhibitors	Drug	1316262, 1336926	RxNorm
Soluble guanylate cyclase stimulators	Drug	44506752	RxNorm

Concept	Domain	Concept IDs (including descendants)	Ontology
Prostacyclin receptor agonists	Drug	1327256, 35604848, 36856411, 1354118, 1344992	RxNorm
6 minute walking test	Procedure	606289, 40766814	SNOMED, LOINC
Echocardiography	Procedure	4230911, 45876598	CPT4, SNOMED
NT-proBNP	Procedure, Measurement	3029187, 3029435, 42529224, 42529225, 42870364, 46236287, 46236288	LOINC
WHO functional class	Procedure	715984, 715985, 715986, 715987, 715988	OMOP Extension
Right heart catheterisation	Procedure	4223626	HCPCS, SNOMED, CPT4
Cardiovascular MRI	Procedure	43020548, 4082987	CPT4, SNOMED, HCPCS, ICD10PCS

*Cardiac arrhythmia will be classified as “cardiac arrhythmia overall” and further categorized into atrial flutter, atrial fibrillation, SVT (supraventricular tachycardia), VT (ventricular tachycardia), VF (ventricular fibrillation), sick sinus syndrome, and AV block

Appendix II: Algorithm for identifying monotherapies, combination therapies and any treatment cohorts

Algorithm 1 Input definitions and overall workflow

```
# Construct drug eras from exposures for ERAs, PDE-5i, PRAs, and SGCs
cdm[[allTreatmentsTable]] ← getDrugEras(cdm, drugs, gapEra)

# Define treatment instructions to construct monotherapies and combination therapies. Inclusion rules correspond to individual drug (eras) participating in a combination. Exclusion rules correspond to combination therapies that are supersets: are constructed using all drugs (eras) needed to construct the new combination therapy or monotherapy.
treatmentInstructions ← list of treatment definitions:
    name: ERA+PDE-5i+PRA
    inclusion cohorts: [ERA drug eras, PDE-5i drug eras, PRA drug eras]
    exclusion cohorts: []

    name: ERA+PDE-5i
    inclusion cohorts: [ERA drug eras, PDE-5i drug eras]
    exclusion cohorts: [ERA+PDE-5i+PRA]

    name: PDE-5i+PRA
    inclusion cohorts: [PDE-5i drug eras, PRA drug eras]
    exclusion cohorts: [ERA+PDE-5i+PRA]

    name: ERA+SGC
    inclusion cohorts: [ERA drug eras, SGC drug eras]
    exclusion cohorts: []

    name: SGC monotherapy
    inclusion cohorts: [SGC drug eras]
    exclusion cohorts: [ERA+SGC]

    name: PRA monotherapy
    inclusion cohorts: [PRA drug eras]
    exclusion cohorts: [ERA+PDE-5i+PRA, PDE-5i+PRA]

    name: PDE-5i monotherapy
    inclusion cohorts: [PDE-5i drug eras]
    exclusion cohorts: [ERA+PDE-5i, PDE-5i+PRA, ERA+PDE-5i+PRA]

    name: ERA monotherapy
    inclusion cohorts: [ERA drug eras]
    exclusion cohorts: [ERA+PDE-5i, ERA+SGC, ERA+PDE-5i+PRA]

# Construct drug combinations and monotherapies
cdm ← generateCombinationDependentCohorts(cdm, treatmentInstructions, allTreatmentTable, minLength = 30 days)

# Construct cohort corresponding to individuals that are treated with any of the combinations and monotherapies
cdm ← getAnyTreatmentCohort(cdm , allTreatmentsTable, treatmentInstructions$names)

# Get percentage of treated (treatment-specific) and untreated individuals per time-window. A person is untreated in a window if no treatment is taken in that time-window. A person is considered treated with a specific treatment, if the treatment intesects the specific time-window.
results ← geCharacterizationResultsPerWindow(cdm, allTreatmentsTable, pahTable, windows)
```

Algorithm 2 Generate combination dependent cohorts

```
function GENERATECOMBINATIONDEPENDENTCOHORTS(cdm, treatmentInstructions, allTreatmentTable, minLength)
    for instruction in treatmentInstructions do
        Extract name, inclusionRules, and exclusionRules
        # Build inclusion cohort
        if more than one inclusionRules exists then
            Intersect these cohorts → inclusion cohort
        else
            Use single cohort → inclusion cohort
        if exclusionRules are not empty then
            if more than one exclusionRules exists then
                Unite these cohorts → exclusion cohort
            else
                Use single cohort → exclusion cohort
            # Apply exclusion logic
            Intersect inclusion and exclusion cohorts
            Keep cohort segments exclusive to the inclusion cohort
        # Remove artifacts
        Filter out resulting cohort entries with duration <= minLength
        # Save resulting cohort
        Name resulting cohort using name provided
        Append resulting cohort to allTreatmentTable
    return cdm
```

Algorithm 3 Construct cohort of individuals that are on any combination therapy or monotherapy

```
function GETANYTREATMENTCOHORT(cdm, allTreatmentsTable, monotherapiesAndCombinationsNamesList)
    Extract monotherapies and combination therapies from allTreatmentTable
    Unite all monotherapy and combination therapy into a single cohort
    Append resulting cohort to allTreatmentTable
    return cdm
```

Appendix III: ENCePP checklist for study protocols

Study title:
DARWIN EU® - Incidence, period prevalence, and characterisation of individuals with paediatric pulmonary arterial hypertension

EU PAS Register® number: EUPAS1000000716
Study reference number (if applicable): P3-C3-011

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

Comments:

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

Comments:

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

Section 3: Study design		**Yes**	**No**	**N/A**	**Section Number**
3.1	Is the study design described? (e.g., cohort, case-control, cross-sectional, other design)	☒	☐	☐	8.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	8.2
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.8
3.4	Does the protocol specify measure(s) of association? (e.g., risk, odds ratio, excess risk, rate ratio, hazard ratio, risk/rate difference, number needed to harm (NNH))	☐	☐	☒	
3.5	Does the protocol describe the approach for the collection and reporting of adverse events/adverse reactions? (e.g. adverse events that will not be collected in case of primary data collection)	☐	☐	☒	

Comments:

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

Comments:

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.1	Does the protocol describe how the study exposure is defined and measured? (e.g. operational details for defining and categorizing exposure, measurement of dose and duration of drug exposure)	☐	☐	☒	
5.2	Does the protocol address the validity of the exposure measurement? (e.g., precision, accuracy, use of validation sub-study)	☐	☐	☒	
5.3	Is exposure categorized according to time windows?	☐	☐	☒	
5.4	Is intensity of exposure addressed? (e.g., dose, duration)	☐	☐	☒	
5.5	Is exposure categorized based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?	☐	☐	☒	

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.6	Is (are) (an) appropriate comparator(s) identified?	☐	☐	☒	

Comments:

Section 6: Outcome definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
6.1	Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated?	☐	☐	☒	
6.2	Does the protocol describe how the outcomes are defined and measured?	☐	☐	☒	
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)	☐	☐	☒	
6.4	Does the protocol describe specific outcomes relevant for Health Technology Assessment? (e.g. HRQoL, QALYs, DALYS, health care services utilization, burden of disease or treatment, compliance, disease management)	☐	☐	☒	

Comments:

Section 7: Bias		**Yes**	**No**	**N/A**	**Section Number**
7.1	Does the protocol address ways to measure confounding? (e.g., confounding by indication)	☐	☐	☒	
7.2	Does the protocol address selection bias? (e.g. healthy user/adherer bias)	☒	☐	☐	8.5, 8.8
7.3	Does the protocol address information bias? (e.g. misclassification of exposure and outcomes, time-related bias)	☒	☐	☐	8.5, 8.8

Comments:

Section 8: Effect measure modification		**Yes**	**No**	**N/A**	**Section Number**
8.1	Does the protocol address effect modifiers? (e.g., collection of data on known effect modifiers, sub-group analyses, anticipated direction of effect)	☒	☐	☐	8.8

Comments:

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

Comments:

Section 10: Analysis plan		**Yes**	**No**	**N/A**	**Section Number**
10.1	Are the statistical methods and the reason for their choice described?	☒	☐	☐	8.8
10.2	Is study size and/or statistical precision estimated?	☒	☐	☐	8.7
10.3	Are descriptive analyses included?	☒	☐	☐	8.8
10.4	Are stratified analyses included?	☒	☐	☐	8.8
10.5	Does the plan describe methods for analytic control of confounding?	☐	☐	☒	
10.6	Does the plan describe methods for analytic control of outcome misclassification?	☐	☐	☐	8.8
10.7	Does the plan describe methods for handling missing data?	☒	☐	☐	8.8
10.8	Are relevant sensitivity analyses described?	☐	☒	☐	

Comments:

Section 11: Data management and quality control		**Yes**	**No**	**N/A**	**Section Number**
11.1	Does the protocol provide information on data storage? (e.g., software and IT environment, database maintenance and anti-fraud protection, archiving)	☒	☐	☐	9.2
11.2	Are methods of quality assurance described?	☒	☐	☐	10.0
11.3	Is there a system in place for independent review of study results?	☐	☐	☒	

Comments:

Section 12: Limitations		**Yes**	**No**	**N/A**	**Section Number**
12.1	Does the protocol discuss the impact on the study results of:				
	12.1.1 Selection bias?	☒	☐	☐	
	12.1.2 Information bias?	☒	☐	☐	
	12.1.3 Residual/unmeasured confounding? (e.g., anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).	☐	☐	☒	11
12.2	Does the protocol discuss study feasibility? (e.g. study size, anticipated exposure uptake, duration of follow-up in a cohort study, patient recruitment, precision of the estimates)	☒	☐	☐	8.2

Comments:

Section 13: Ethical/data protection issues		**Yes**	**No**	**N/A**	**Section Number**
13.1	Have requirements of Ethics Committee/ Institutional Review Board been described?	☒	☐	☐	13
13.2	Has any outcome of an ethical review procedure been addressed?	☒	☐	☐	

Section 13: Ethical/data protection issues	**Yes**	**No**	**N/A**	**Section Number**
13.3 Have data protection requirements been described?	☒	☐	☐	9.2

Comments:

Section 14: Amendments and deviations	**Yes**	**No**	**N/A**	**Section Number**
14.1 Does the protocol include a section to document amendments and deviations?	☒	☐	☐	4

Comments:

Section 15: Plans for communication of study results	**Yes**	**No**	**N/A**	**Section Number**
15.1 Are plans described for communicating study results (e.g., to regulatory authorities)?	☒	☐	☐	14
15.2 Are plans described for disseminating study results externally, including publication?	☒	☐	☐	14

Comments:

Name of the main author of the protocol: Nicholas Hunt

Date: 24/06/2025

Signature: