Study Protocol P3-C1-005

11/07/2024

Version 2.1

Table of Contents

DOCUMENT HISTORY

VERSION	DATE	DESCRIPTION
V1.0	03/05/2024	**First version of protocol submitted to EMA**
V2.0	04/06/2024	**Archiving version submitted to EMA**
V2.1	11/07/2024	**Final version uploaded in the HMA-EMA Catalogue**

Study Title	DARWIN EU® - Characterising interstitial lung disease in Europe
Protocol version identifier	2.1
Date of last version of protocol	11/07/2024
EU PAS register number	EUPAS1000000172
Active substance	None
Medicinal product	None
Research question and objectives	1. To measure the incidence of newly diagnosed Interstitial Lung Disease (ILD) across different European countries, overall and stratified by risk factors such as age, sex and calendar time. The incidence will be assessed for the overall group of ILD as well as for the following subtypes of ILD: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD. 2. To describe the characteristics of patients diagnosed with ILD in 4 European countries. Characteristics include demographics, most common comorbidities, prespecified risk factors and treatments taken before diagnosis. Characteristics will be assessed for the overall group of ILD as well as for the following subtypes of ILD: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD 3. To measure survival rates in patients diagnosed with ILD across different European countries overall and stratified by risk factors such as age, sex and calendar time. Survival rates will be assessed for the overall group of ILD as well as for the following subtypes of ILD: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD
Country(ies) of study	Spain, United Kingdom, France and Germany
Author	Nicholas Hunt n.hunt@darwin-eu.org Katia Verhamme k.verhamme@darwin-eu.org

LIST OF ABBREVIATIONS

Acronyms/term	Description
CDM	Common Data Model
CC	Coordinating centre
DA	Disease Analyzer
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DRE	Digital Research Environment
DI-ILD	Drug-induced interstitial lung disease
DQD	Data Quality Dashboard
EEA	European Economic Area
EHR	Electronic Health Records
EMA	European Medicines Agency
ED	Emergency department
EU	European Union
GDPR	General Data Protection Regulation
ICD	International Classification of Diseases
ID	Index date
ILD	Interstitial lung disease
IP	Inpatient
IMI	Immune checkpoint inhibitors
LPD	Longitudinal Patient Database
MA	Marketing Authorisation
mTOR	Mammalian target of rapamycin
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OP	Outpatient
SD	Standard deviation
SNOMED	Systematized Nomenclature of Medicine
TKI	Tyrosine kinase inhibitors
TNF	Tumour necrosis factor
WHO	World Health Organisation

1. TITLE

DARWIN EU® - Characterising interstitial lung disease in Europe

2. RESPONSIBLE PARTIES – STUDY TEAM

STUDY TEAM ROLE	NAMES	ORGANISATION
Principal Investigator/Epidemiologist	Katia Verhamme Nicholas Hunt	Erasmus MC Erasmus MC
Data Scientist	Adam Black Ross Williams	Erasmus MC Erasmus MC
Data Partner*	**Names**	**Organisation**
CPRD GOLD	Antonella Delmestri	University of Oxford
BIFAP	Patricia García-Poza Miguel-Angel Macia-Martinez Ana Llorente-Garcia	Agencia Española de Medicamentos Y Productos Sanitarios (AEMPS)
CDW Bordeaux	Romain Griffier Guillaume Verdy Vianney Jouhet	Bordeaux University Hospital
IQVIA Germany Disease Analyser	Hugo Vernooij James Brash	IQVIA

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

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

Title

DARWIN EU® - Characterising interstitial lung disease in Europe

Rationale and background

Interstitial lung disease (ILD) is a heterogenous group of respiratory disorders affecting the interstitium of the lungs. Drug-induced ILD are adverse drug reactions from a wide range of drugs, many of which can be life-threatening diseases. Measuring the incidence of ILD and characterising its population in Europe may guide signal detection validation discussions for drug-induced ILD.

Research questions

What are the incidence, the characteristics and overall survival of patients diagnosed with ILD and ILD-subtypes in four European countries in the period 2010-2022?

Objectives

The objectives are to measure:

(i) the incidences of ILD as well as for the following subtypes of ILD: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD. Incidences will be stratified by age, sex and calendar time

(ii) to characterise the patients with ILD by age, sex, comorbidities and concomitant medication. This analysis will be done in the overall ILD group and in the following ILD-subtypes (alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD)

(iii) to measure the survival rates in patients diagnosed with ILD and within the following ILD-subtypes (alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD. Survival rates will be stratified by age, sex and calendar time.

Research methods

Study design

Cohort study with population-level descriptive epidemiology to estimate the incidence of ILD, and patient-level characterisation to estimate the median survival time and to characterise ILD patients by prior comorbidities and drugs used.

Study Population

All patients in the databases newly diagnosed with ILD in the period 1st January 2010 to 31st December 2022 with at least 365 days of data visibility prior to the date of first ILD diagnosis (except CDWBordeaux).

Outcome

Death

Variables

Condition of interest

ILD and ILD-subtypes namely i) alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), ii) lung fibrosis and iii) drug induced ILD

Data sources

1. Clinical Practice Research Datalink (CPRD) GOLD, United Kingdom
2. Base de datos para la Investigacion Farmacoepidemiologica en el Ambito Publico (BIFAP), Spain
3. Clinical Data Warehouse (CDW) Bordeaux, France
4. IQVIA Disease Analyzer Germany (IQVIA DA Germany), Germany

Sample size

Based on a preliminary feasibility assessment, the expected number of record counts for the selected condition of interest (i.e. ILD) ranged from 3,900 (CDWBordeaux) to 8,800 (IQVIA DA Germany).

Analysis

Calculation of the incidence of ILD and ILD subtypes by means of the *IncidencePrevalence* R package which will be stratified by age, sex and calendar year. Characterisation of patients newly diagnosed with ILD and for the ILD subtypes of interest (see above) using the *CohortCharacteristics* and *CohortDiagnostics* R packages and finally calculating the survival rates in patients with newly diagnosed ILD. This analysis will also be repeated in individuals from the respective ILD subtypes of interest (see above)

4. AMENDMENTS AND UPDATES

NUMBER	DATE	SECTION OF STUDY PROTOCOL	AMENDMENT OR UPDATE	REASON

5. MILESTONES

Additional study specific deliverables (e.g. results of validation of new disease concepts, statistical analysis plan, etc.) might be requested but this will be on an individual basis.

STUDY SPECIFIC DELIVERABLE	Estimative TIMELINE
Draft Study Protocol	3rd May 2024
Final Study Protocol	4th of June
Creation of Analytical code	June 2024
Execution of Analytical Code on the data	June 2024
Interim Study Report (if applicable)	NA
Draft Study Report	30th July 2024
Final Study Report	TBD

TBD=To be determined

6. RATIONALE AND BACKGROUND

Interstitial lung disease (ILD) describes a heterogenous group of respiratory disorders affecting the interstitium of the lungs. (Kaul, Cottin et al. 2021) ILD encompasses many different pathological processes including drug-induced ILDs. (Spagnolo, Bonniaud et al. 2022) Drug-induced interstitial lung disease (DI-ILD) is also a large and very heterogeneous group of adverse drug reactions, ranging from mild to progressive and life-threatening disease. The number of drugs associated with the development of ILD continues to rise, mainly due to the use of novel monoclonal antibodies and biologics for neoplastic and rheumatologic diseases, many of which are associated with lung toxicity, and includes, among others, chemotherapeutics, molecular targeting agents, immune checkpoint inhibitors, antibiotics, antiarrhythmics, and conventional or biologic disease-modifying antirheumatic drugs. (Paolo, Philippe et al. 2022)

There is a geographical variation in the reporting of ILD as an adverse drug reaction in spontaneous case reports, with cases from Japan being more frequently reported.(Pinheiro, Blake et al. 2016, Kaul, Cottin et al. 2021) The assessment of spontaneous case reports requires knowledge of patient characteristics that commonly co-occur with either the disease or the suspected reaction in order to make an informed assessment as to whether such occurrences are drug induced or may simply be due to confounding factors.

A better understanding of the population diagnosed with ILD in Europe and its incidence (including characterising comorbidities, risk factors, medications administered around diagnosis date) may guide signal detection validation discussions for drug-induced ILD within the regulatory network. The EMA has already performed in-house study measuring the incidence of ILD in the IMRD UK, IQVIA France and IQVIA Germany databases (EUPAS50623).(De Jong 2023) This study also included ILD-subtypes allocated by aligned code groups (e.g. codes representing pneumonitis, codes representing pulmonary fibrosis) that may vary in either incidence or recording, and which could also be relevant for patient characterisation. In performing disease epidemiology and patient characterisation study of ILD and ILD-subtypes in Europe, signal detection and validation for potential drug-induced ILD will be better guided.

7. RESEARCH QUESTION AND OBJECTIVES

Research questions

What was the incidence, what were the characteristics and what was the overall survival of patients newly diagnosed with ILD in four European countries in the period 2010-2022. These research questions are explored in the whole population of individuals newly diagnosed with ILD as well as for the two most common ILD subtypes (alveolitis/pneumonitis and lung fibrosis)

Objectives

Objective 1: to measure the incidences of ILD and the incidence of ILD subtypes (alveolitis/pneumonitis excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD), stratified by age, sex and calendar time.

Objective 2: to characterise the patients with ILD by age, sex, comorbidities and concomitant medication. This analysis will be done for the overall ILD group and for the following ILD-subtypes (alveolitis/pneumonitis excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD).

Objective 3: to measure the survival rates in patients diagnosed with ILD and to measure survival rates within the following ILD-subtypes (alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD). Survival rates will be stratified by age, sex and calendar time

Description of the proposed objectives to be achieved in the study is displayed in **Table 1**.

Table 1. Primary and secondary research questions and objective.

A. Primary research question and objective.

Objective:	Objective 1: To measure the incidence of diagnosed ILD in four European countries, overall and stratified by age, sex and calendar time. This analysis will be repeated for the following ILD subtypes: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD). Objective 2: To describe the characteristics of patients diagnosed with ILD in four European countries in terms of demographics (age and sex), comorbidities, and medications used in the one year prior to diagnosis. This analysis will be repeated for the following ILD subtypes: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD). Objective 3: To measure survival rates in patients diagnosed with ILD in four European countries overall and stratified by age, sex and calendar time. This analysis will be repeated for the following ILD subtypes: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD).
Hypothesis:	Not applicable
Population ***(mention key inclusion-exclusion criteria):***	All patients should be present in the databases with at least 365 days of prior history (except for CDWBordeaux) and be newly diagnosed with ILD (i.e. no prior diagnosis of ILD).
Exposure:	None

Comparator:	None
Outcome:	Overall survival (objective 3)
Time *(when follow up begins and ends):*	Follow-up starts after study start date (1st January 2010) or on the date where there is 1 year of prior history. (one year of history not needed for CDWBordeaux) whatever comes last. For objective 2 and 3, the follow-up starts upon the first diagnosis of ILD. The End of follow-up will be defined as the earliest of loss to follow-up, end of data availability, death, or end of study period (31st of December 2022), whatever comes first.
Setting:	Primary care data (CPRD GOLD, BIFAP, IQVIA DA Germany) and secondary care data (CDW Bordeaux).
Main measure of effect:	Incidence rates, expressed as number of individuals newly diagnosed with ILD per 100,000 person-years. Incidence rates will be calculated overall and stratified by sex, age (age category) and calendar time (year). Incidence rates will be calculated for ILD overall and for the following ILD subtypes: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD. Median survival time in patients with newly diagnosed ILD stratified by sex, age-category, and calendar time. If numbers are sufficient, survival rates will not only be done for the overall group but also for the following ILD subtypes: alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD.

8. RESEARCH METHODS

8.1 Study type and Study Design

The Study Types with related Study Designs are described in the **Table 2** below and are selected from the Catalogue of Data analytics.

A cohort study will be conducted using routinely collected health data from four databases. The study will comprise two consecutive parts:

- To address objective 1: A population-based cohort study including newly diagnosed ILD and ILD-subtype patients, to assess the incidence rates of ILD, stratified by different age categories, sex and calendar date. The denominator population includes all present in the database with at least one year of data availability (not for CDWBordeaux).
- To address objective 2: A cohort of newly-diagnosed ILD patients to characterise them with regards to the most common comorbidities, concomitant medications, age at first diagnosis, sex and calendar year. This analysis will be done in the overall group of individuals newly diagnosed with ILD and by ILD subtypes (see above). .

- To address objective 3: A cohort of newly-diagnosed ILD and ILD subtype patients to calculate survival time from the date of first diagnosis of ILD to date of death, stratified by age category (at first diagnosis), sex and calendar date (year at index date).

Table 2. Description of Potential Study Types and Related Study Designs.

STUDY TYPE	STUDY DESIGN	STUDY CLASSIFICATION
Population-level descriptive epidemiology	Population-level cohort	Off the shelf (C1)
Patient-level characterisation	Cohort analysis	Off the shelf (C1)

8.2 Study Setting and Data Sources

This study will be conducted using routinely collected data from four databases in four European countries. All databases were previously mapped to the OMOP Common Data Model (CDM).

1. Clinical Practice Research Datalink (CPRD) GOLD, United Kingdom
2. Base de datos para la Investigacion Farmacoepidemiologica en el Ambito Publico (BIFAP), Spain
3. Clinical Data Warehouse (CDW) Bordeaux, France
4. IQVIA Disease Analyzer Germany (IQVIA DA Germany), Germany

For this study, four databases in the DARWIN EU® Database Catalogue were considered fit for purpose. The selection process was based on the size of the databases, the number of individuals with the diagnosis of interest, the suitability of denominator population for population-level rates, information regarding mortality, geographical spread and the experience gained from databases that participated in other similar DARWIN EU® studies. Based on the feasibility assessment performed, the suggested databases are considered fit for purpose for at least part of the objectives.

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 3**.

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 utilize the Achilles tool, which systematically characterizes the data and generates data characteristics such as age distribution, condition prevalence per year, data density, measurement value distribution 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 are out-of-date with the expected refresh cycle (typically quarterly or half-yearly). In addition, it is important to have 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 to get insights when data collection started, when new sources of data were added and when until when data was included.

Table 3. Description of the selected Data Sources.

Country	Name of Database	Justification for Inclusion	Health Care setting	Type of Data	Number of active subjects	Feasibility count of disease (ILD)*	Data lock for the last update
United Kingdom	CPRD GOLD	Adequate number of individuals with the diagnosis of interest, Suitable denominator population for population-level rates, Adequate information regarding mortality is captured Contribute to geographical diversity of data sources included	Primary care	EHR	17m	6,900	04/11/2023
Spain	BIFAP	Adequate number of individuals with the diagnosis of interest Suitable denominator population for population-level rates Adequate information regarding mortality is captured	Primary care	EHR	22m	7,900	31/03/2023

Country	Name of Database	Justification for Inclusion	Health Care setting	Type of Data	Number of active subjects	Feasibility count of disease (ILD)*	Data lock for the last update
		Contribute to geographical diversity of data sources included					
France	CDW Bordeaux	Adequate number of individuals with the diagnosis of interest Adequate information regarding mortality is captured Contribute to geographical diversity of data sources included	Secondary care	EHR	2.4m	3,900	16/11/2023
Germany	IQVIA DA Germany	Adequate number of individuals with the diagnosis of interest Suitable denominator population for population-level rates Contribute to geographical diversity of data sources included	Primary care and outpatient secondary care	Claims	43m	8,800	23/01/2024

*Counts for ILD overall (not for the different subtypes). The feasibility assessment with counts for the different conceptids has been added in the appendix. Rounded up to the nearest hundred.

Clinical Practice Research Datalink GOLD, United Kingdom (University of Oxford)

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

GOLD data include patient's demographic, biological measurements, clinical symptoms and diagnoses, referrals to specialist/hospital and their outcome, laboratory tests/results, and prescribed medications. GOLD has been assessed and found broadly representative of the UK general population in terms of age, gender, and ethnicity.(Herrett, Gallagher et al. 2015) GOLD has been widely used internationally for observational research to produce nearly 3,000 peer-reviewed publications, making GOLD the most influential UK clinical database so far.(Carey, Nirmalananthan et al. 2023, Fahmi, Wong et al. 2023, Wigglesworth S 2023).

In terms of quality checks, the integrity, structure and format of the data is reviewed. Collection-level validation ensures integrity by checking that data received from practices contain only expected data files and ensures that all data elements are of the correct type, length and format. Duplicate records are identified and removed.[1] Transformation-level validation checks for referential integrity between records ensure that there are no orphan records included in the database (for example, that all event records link to a patient), while research-quality-level validation covers the actual content of the data. CPRD provides a patient-level data quality metric in the form of a binary 'acceptability' flag.[1] This is based on recording and internal consistency of key variables including date of birth, practice registration date and transfer out date.

Base de datos para la Investigacion Farmacoepidemiologica en el Ambito Publico (BIFAP), Spain

BIFAP is a longitudinal population-based data source of medical patient records of the Spanish National Health Service (SNS) from 10 participating Regions throughout Spain out of the 17 Spanish Regions. Population currently included represents 36% of the total Spanish population. Spain has a SNS that provides universal access to health services through the Regional Healthcare Services. Primary care physicians (PCPs), both general practitioners and paediatricians, have a central role. They act as gatekeepers of the system and also exchange information with other levels of care to ensure the continuity of care. Most (98.9%) of the population is registered with a PCP and, in addition, most drug prescriptions are written at the primary care level. BIFAP includes a collection of databases linked at individual patient level. The main one is the Primary care Database given the central role of PCPs in the SNS. Linked, there are additional important structural databases like the medicines dispensed at community pharmacies and the patients' hospital diagnosis at discharge. Linkage to SARS-CoV-2 diagnostics test and COVID-19 vaccination registries are also included. Additional databases are also linked for a subset of patients (hospital pharmacy, cause of death registry). BIFAP program is a non-profit program financed by the Spanish Agency of Medicines and Medical Devices (AEMPS), a government agency belonging to the Ministry of Health in collaboration with the Regional health authorities. The main use of BIFAP is for research purposes in order to evaluate the adverse and beneficial effects of drugs and drug utilization patterns in the general population under real conditions of use.

Clinical Data Warehouse, 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 database currently holds information

about the person (demographics), visits (inpatient and outpatient), conditions and procedures (billing codes), drugs (outpatient prescriptions and inpatient orders and administrations), measurements (laboratory tests and vital signs) and dates of death (in or out-hospital death).

IQVIA Disease Analyser (DA) Germany, Germany

IQVIA Disease Analyzer (DA) Germany is a database of de-identified electronic medical records from specialised and general primary practices (GP) in Germany since 1992. This dataset encompasses approximately 3% of all outpatient practices within Germany, ensuring a substantial representation of the national healthcare landscape.(Rathmann, Bongaerts et al. 2018, Zappacosta, Cascarano et al. 2022) The sampling methods used for practice selection, taking into account physician's demographics, specialty focus, community size category and federal state location, was instrumental in constructing a database that accurately mirrors the diverse spectrum of healthcare providers in the country.(Rathmann, Bongaerts et al. 2018) Consequently, data within IQVIA DA Germany database has been demonstrated to be representative of general and specialised practices throughout Germany.
The database contains demographics records, basic medical data, disease diagnosis according to International Classification of Diseases, 10th revision (ICD-10), and prescription records.(Zappacosta, Cascarano et al. 2022) While the database partly records information on deaths and procedures, it currently does not support linkage with external data sources and therefore information on mortality is incomplete. Routine updates are conducted at regular intervals. The quality of data is assessed based on several criteria including completeness of information and correctness (e.g. linkage between diagnosis and prescriptions).
IQVIA DA Germany is suitable for pharmacoepidemiologic and pharmacoeconomic studies as previously demonstrated.(Tanislav, Rosenbauer et al. 2022, Zappacosta, Cascarano et al. 2022, Ly, Flach et al. 2023).

8.3 Study Period

The study period will be from 1st January 2010 to 31st December 2022.

8.4 Follow-up

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

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Incident with respect to…
All individuals from the respective databases with at least 1 year of valid database history (not for CDWBordeaux).	Study entry date	Single entry	Incident	Anytime prior to study entry date	IP, OP, OT	SNOMED	Diagnosis of ILD
Patients newly diagnosed with ILD	Date of incident ILD	Single entry	Incident	Anytime prior to ILD	IP, OP, OT	SNOMED	Diagnosis of ILD

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

To calculate incidence rates, it is important to have an appropriate denominator population and their contributed observation time. Study participants in the denominator population will begin contributing person time on the respective date of the latest of the following: 1) study start date (1st January 2010), 2) date at which they have 1 year of prior history. Participants will stop contributing person time at the earliest date of the following: 1) end of available data in each of the data sources (date of last data extraction), 2) death, 3) study end date (31st December 2022) or 4) date at which the observation period of the specific person ends (individual leaving practice or being diagnosed with ILD whatever comes first)

An example of entry and exit into the denominator population is shown in **Figure 1**. In this example, person ID 1 has already sufficient prior history before the study start date and observation period ends after the study end date, so will contribute during the complete study period. Person ID 2 and 4 enter the study only when they have sufficient prior history. Person ID 3 leaves when exiting the database (the end of observation period). Lastly, person ID 5 has two observation periods in the database. The first period contributes time from study start until end of observation period, the second starts contributing time again once sufficient prior history is reached and exits at study end date.

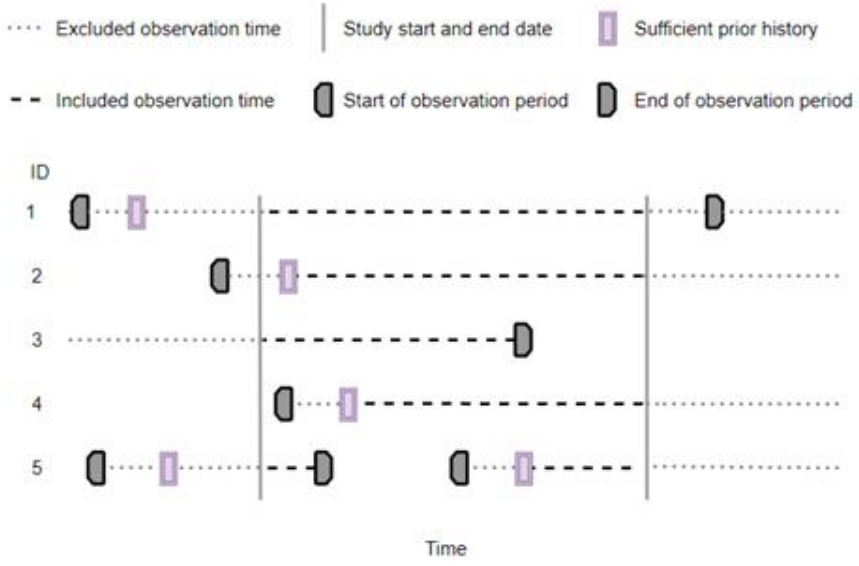


Figure 1. Included observation time for the denominator population.

For the survival analysis, patients will start follow up at ILD diagnosis index date and will be censored at the time of loss to follow-up, at end of data availability, study end date or date of death whatever comes first.

8.5 Study Population with inclusion and exclusion criteria

Population-level descriptive epidemiology

All patients with an incident diagnosis of ILD in the period 1st of January 2010 to 31st of December 2022 (or latest date available). Notably, all patients need to have at least 365 days of data visibility (except for CDWBordeaux) prior to the date of their first diagnosis and no prior diagnosis of ILD. This requirement of at least 1 year of prior data history will not hold for children <1 year of age.

Patient-level characterisation

The study cohort will include all individuals registered in the database between 1st of January 2010 and 31st of December 2022, with at least 365 days of data visibility (except for CDWBordeaux) and a new ILD diagnosis.

The concept definition of ILD and ILD-subtypes is described in Table S1, **Appendix I**. The operational definitions of the inclusion and exclusion criteria are presented by means of **Table 5** and **Table 6**, respectively.

Table 5. Operational Definitions of Inclusion Criteria.

Criterion	Details	Order of application	Assessment window	Care Settings [1]	Code Type	Diagnosis position[2]	Applied to study populations:
Observation period in the database during the period 2010-2022 (or the latest date available)	All individuals present in the period 2010-2022 (or the latest date available)	After	n/a	IP, OP, OT	n/a	n/a	All individuals within selected databases
Prior database history	Study participants will be required to have 365 days of prior history observed before contributing observation time	Prior	[-365, -1]	OP	n/a	n/a	All individuals within CPRD GOLD, BIFAP and IQVIA DA Germany
Washout period	Individuals newly diagnosed with ILD will be required not to have a diagnosis of ILD any time prior to the diagnosis of ILD	Prior	[-inf, -1]	OP	n/a	n/a	All individuals diagnosed with ILD

[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 Variables

8.6.1 Exposure/s (where relevant)

There is no exposure in this study.

8.6.2 Outcome/s (where relevant)

The operational definition of the outcomes is presented in the **Table 6**.

Table 6. Operational Definitions of Outcome.

Outcome name	Details	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Type	Diagnosis Position[2]	Applied to study populations
Death	Death events which occur prior to the end of the study period or the end of individual patient follow-up.	Yes	Time-to-event	No	Primary and secondary	N/A	N/A	Patients with an ILD diagnosis

[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, including confounders, effect modifiers and other variables

The study covariates are described conceptually, and the context or rationale for the choices are provided in this section. The operational definition of the covariates is described in the **Table 7**.

Covariates for the population-level descriptive epidemiology study (incidence and prevalence)

To characterize ILD and the most common ILD-subtypes (i) alveolitis/pneumonitis excluding aspiration and infectious pneumonitis ii) lung fibrosis and iii) drug induced ILD, the covariates for stratification in population-level descriptive epidemiology study will include age category (<18, 18-39, 40-59, 60-79 and >=80 years at index date), sex and calendar time (year) at start of follow-up.

Covariates for the patient-level characterisation study

As part of patient-level characterisation study, comorbidities will be measured in the period of 365 days prior to the index date (i.e. date of diagnosis of ILD). Concomitant medications will be measured within 30 and 365 days prior to the index date and reported for both time periods.

Covariates used to describe Patient characterisation consist of the following:

- Sex
- Age at index date (i.e. date of diagnosis of ILD). Age will be presented in age categories namely:
    - <18
    - 18-39
    - 40-59
    - 60-79
    - >=80 years.
- ILD subtypes:
    - alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis)
    - lung fibrosis
    - drug-induced

    - Unspecified ILD

- description of the top 10 of the most frequent comorbidities and concomitant medications using large-scale characterisation, as specified in the CohortCharacteristics R package and assessed in the period of 365 days prior to the index date.
- Comorbidities of interest (Gaillet, Favelle et al. 2015, Choi, Dauti et al. 2018, Joy, Arbiv et al. 2023), in particular:
  - COPD/emphysema
  - Connective tissue disease namely rheumatoid arthritis, sclerodermia, systemic lupus erythematosus (SLE), polymyositis/dermatomyositis, granulomatosis with polyangiitis (GPA)
  - GERD
  - Hepatitis B/C
  - Cancer excluding non-melanoma skin cancer
- Use of concomitant medication (assessed in the 30 days and 365 days prior to the index date) in particular:
  - chemotherapy
  - amiodarone
  - immune checkpoint inhibitors (ICIs)
  - tyrosine kinase inhibitors (TKIs including crizotinib, EGFR inhibitors, erlotinib, gefitinib)
  - mammalian target rapamycin (mTOR) inhibitors
  - rituximab
  - statins
  - methotrexate
  - nitrofurantoin
  - tumor necrosis factor (TNF)-α antagonists (Paolo, Philippe et al. 2022).

The operational definition of the covariates is described in **Table 7**. Index date is the start of the incident ILD diagnosis during the study period. The preliminary concepts for prespecified conditions and drugs of interest are described in **Appendix I**.

Table 7. Operational Definitions of Covariates.

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type	Diagnosis Position [2]	Applied to study populations
Age category	Assessed at start of follow-up or at date of diagnosis of ILD and categorised to either <18, 18-39, 40-59, 60-79 and >=80 years	Categorical	0	n/a	n/a	n/a	Applied to the general cohort and the cohort of individuals with newly diagnosed ILD.
Sex		Categorical	0	n/a	n/a	n/a	Applied to the general cohort and the cohort of individuals with newly diagnosed ILD
Large scale characterisation	Large-scale patient characterisation with regard to underlying comorbidity and use of concomitant medication	Binary	[-365, 0]	IP, OP, OT	SNOMED	n/a	Individuals with a new diagnosis of ILD
Comorbidity	Specified comorbidities	Binary	[-365, 0]	IP, OP, OT	SNOMED	n/a	Individuals with a new diagnosis of ILD
Concomitant medications	Specified concomitant medication	Binary	[-30, 0], [-365,0]	IP, OP, OT	RxNorm	n/a	Individuals with a new diagnosis of ILD

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

8.7 Study size

No sample size has been calculated for this disease epidemiology descriptive study, as our primary focus is to examine incidence of ILD, irrespective of the sample size. Based on a preliminary feasibility assessment, the expected number of record counts for ILD in the databases included in this study ranged from 3,900 (CDWBordeaux) to 8,800 (IQVIA DA Germany).

8.8 Analysis

In principle the type of analysis by study type is fixed as can be observed from **Table 8**.

Table 8. Description of Study Types and Type of analysis.

STUDY TYPE	STUDY CLASSIFICATION	TYPE OF ANALYSIS
Population-level descriptive epidemiology	Off-the-shelf (C1)	- Incidence rates of ILD and ILD-subtypes
Patient-level characterisation	Off-the-shelf (C1)	- Patient-level characteristics - Median survival of patients diagnosed with ILD and ILD-subtypes

8.8.1 Federated Network Analyses

Analyses will be conducted separately for each database. Before study initiation, test runs of the analytics are performed on a subset of the data sources or on a simulated set of patients and quality control checks are performed. Once all the tests are passed, the final package is released in the version-controlled Study Repository for execution against all the participating data sources.

The data partners locally execute the analytics against the OMOP-CDM in RStudio and review and approve the by default aggregated results before returning them to the Coordination Centre. Sometimes multiple execution iterations are performed, and additional fine tuning of the code base is needed. A service desk will be available during the study execution for support.

The study results of all data sources are checked after which they are made available to the team in the Digital Research Environment (DRE) and the Study Dissemination Phase can start. All results are locked and timestamped for reproducibility and transparency.

8.8.2 Patient privacy protection

Cell suppression will be applied as required by databases to protect people's privacy. Cell counts <5 will be masked.

8.8.3 Statistical model specification and assumptions of the analytical approach considered

Objective 1. Incidence of Diagnosed ILD and Common Subtypes: Calculation of population-based incidence rates is part of our pipelines for population-level descriptive epidemiology. This analysis will be performed using the *IncidencePrevalence* R package and will provide overall incidence rates, stratified by age, sex, and calendar year. (Ed 2024) Additionally, it can be repeated for prespecified ILD subtypes (alveolitis/pneumonitis (excluding aspiration and infectious pneumonitis), lung fibrosis and drug induced ILD). To ensure the reliable estimation of incidence rate of ILD using an appropriate denominator population, this objective will be investigated within the primary care databases: BIFAP, CPRD GOLD, and IQVIA DA Germany.

Objective 2. Characterization of Patients Diagnosed with ILD: Description of patient characteristics is part of our pipelines for patient-level characterisation. Characterization of patient demographics, comorbidities, risk factors, and treatments before diagnosis, as well as characterizing common ILD subtypes at diagnosis, will be conducted using the *CohortCharacteristics* and *CohortDiagnostics* R packages. (Marti 2024) This objective can be explored in all participating databases.

Objective 3. Survival Rates in Patients Diagnosed with ILD: Estimation of survival rates is part of our pipelines for patient-level characterisation. *CohortSurvival* R package can estimate survival rates stratified by age, sex, and year. Analysis can be repeated for the prespecified ILD subtypes, subject to sufficient data availability. (Ed 2024) This objective will be investigated in databases with complete mortality data such as BIFAP, CPRD GOLD and CDWBordeaux.

8.8.4 Methods to derive parameters of interest

All methods to derive parameters are predefined using the R packages specified in 8.8.3.

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

Disease epidemiology study

Incidence calculations of ILD

Annual incidence rates of ILD will be calculated as the number of newly diagnosed ILD per 100,000 person-years of the population at risk of the condition during the period for each calendar year. Those study participants who enter the denominator population will then contribute time at risk up to their first diagnosis during the study period. If they do not have the condition of interest, they will contribute time at risk up as described above. Time-at-risk of subjects who die will be censored at the time of death. 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 study period without a record of the condition will be administratively censored at the end of the study period. Incidence rates will be given together with 95% Poisson confidence intervals.

Output

- Table 1. Number of participants, total number of incident cases and total time at risk in each data source during the study period. Number of participants per pre-specified strata will be included where necessary/applicable.
- Figure 1. Incidence rate/s of disease over calendar time (year) overall
- Figure 2. Incidence rate/s of disease over calendar (year) stratified by sex and age
- Table 2. Numbers reported in figures 1 and 2

Patient-level characterisation study

Patient-level characteristics on index date

For each concept extracted at index date, the number of individuals (N, %) with a record of the drug and condition of interest within the pre-specified time windows will be provided. For all analyses n and % will be reported. A minimum cell counts of 5 will be used when reporting results, with any

smaller counts reported as "<5". All analyses will be reported by country/database, overall and stratified by age group, sex when possible (minimum cell count reached).

Overall survival on index date

The number and % of patients with all-cause mortality will be reported from time of ILD diagnosis to the date of death. Survival estimates for each subject following diagnosis of ILD and each ILD-subtype will be calculated. These will be presented as median survival and 95% confidence intervals overall, per ILD-subtype and per strata (age category at first diagnosis and sex). Individuals who are lost to follow-up will be censored at the time of loss of follow-up.

Output

- Table 1. Baseline characteristics (prespecified comorbidities and drugs) of newly diagnosed ILD patients.
- Table 2. Number and % of individuals who died following ILD diagnosis.
- Figure 1. Kaplan-Meier or Cumulative Incidence Function plots of the probability of a pre-specified outcome (overall survival) following index diagnosis of the condition of interest.

8.8.6 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

Results from analyses described in section 8.8 Data analysis will be presented separately for each database and no meta-analysis of results will be conducted.

9. DATA MANAGEMENT

9.1 Data management

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

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

9.2 Data storage and protection

For this study, participants from various EU member states will process personal data from patients which is collected in national/regional electronic health record databases. Due to the sensitive nature of this

personal medical data, it is important to be fully aware of ethical and regulatory aspects and to strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy.

All databases used in this study 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 Digital Research Environment. These output files do not contain any data that allow identification of subjects included in the study. The DRE 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

General database quality control

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

Study specific quality control

The study code will be based on several R packages including the *CohortCharacteristics*, *CohortSurvival*, *CohortDiagnostics* and the *IncidencePrevalence* package. These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via the DARWIN EU® GitHub repository.

11. LIMITATIONS OF THE RESEARCH METHODS

General limitations:

Characterisation/Indication: The ILD phenotype utilises codes recorded within databases and will not include imaging or test data. The accuracy and consistency of pre-defined condition recording, crucial for patient characterisation may vary across the databases included in the study. ILD is a complex condition and characterising the condition by subtype might be limited if ILD is only recorded as a general term without possibility to further categorise. Apart from CDWBordeaux, it is expected that some types of

medications related to ILD primarily administered in hospital will not be routinely captured within primary care databases. With regard to characterisation of individuals with ILD, previous exposure to smoking is of interest, however as smoking is not systematically recorded and thus not available in all databases, we have decided not to include smoking status as one of the covariates of interest. No linkage to hospitalisation diagnoses will be performed which may lead to an underestimation for certain types of covariates.

The analysis may suffer from small cell counts for some subgroup strata. If numbers are too low, counts will not be disclosed for governance reasons.

The incidence rate of ILD might be an underestimate of the actual incidence rate as ILD is a diagnosis made by the specialist. If information is not well transposed from secondary care to primary care, we might have an underestimate of the individuals with ILD.

Setting:

For this study, we included data from four data sources (CDWBordeaux, BIFAP, CPRD GOLD and IQVIA DA Germany). Results of these databases may not necessarily reflect management of individuals with ILD in other countries/databases.

Overall survival in individuals with ILD:

As mortality is not completely documented in IQVIA Germany, this database will not be used to investigate survival in individuals with ILD.

12. MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS

Adverse events/adverse reactions will not be collected or analysed as part of this evaluation. The nature of this non-interventional evaluation, through the use of secondary data, does not fulfil the criteria for reporting adverse events, according to module VI, VI.C.1.2.1.2 of the Good Pharmacovigilance Practices (https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-submission-reports_en.pdf).

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

13. GOVERNANCE BOARD ASPECTS

All data sources (except for IQVIA DA Germany) require approval from their respective IRB board.

14. PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

A study 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 study report. The full set of underlying aggregated data used in the dashboard will also be made available if requested.

15. OTHER ASPECTS

Not applicable

16. REFERENCES

1. Carey, I. M., N. Nirmalananthan, T. Harris, S. DeWilde, U. A. R. Chaudhry, E. Limb and D. G. Cook (2023). "Prevalence of co-morbidity and history of recent infection in patients with neuromuscular disease: A cross-sectional analysis of United Kingdom primary care data." PLoS One **18**(3): e0282513.
2. Choi, W. I., S. Dauti, H. J. Kim, S. H. Park, J. S. Park and C. W. Lee (2018). "Risk factors for interstitial lung disease: a 9-year Nationwide population-based study." BMC Pulm Med **18**(1): 96.
3. De Jong, V. (2023). Background incidence rates of Interstitial Lung Disease. HMA-EMA RWD Catalogue, EMA.
4. Ed, B. (2024). CohortSurvival.
5. Ed, B. (2024). IncidencePrevalence.
6. Fahmi, A., D. Wong, L. Walker, I. Buchan, M. Pirmohamed, A. Sharma, H. Cant, D. M. Ashcroft and T. P. van Staa (2023). "Combinations of medicines in patients with polypharmacy aged 65–100 in primary care: Large variability in risks of adverse drug related and emergency hospital admissions." PLOS ONE **18**(2): e0281466.
7. Gaillet, G., O. Favelle, L. Guilleminault, A. de Muret, E. Lemarie, T. Lecomte, S. Marchand-Adam and P. Diot (2015). "Gastroesophageal reflux disease is a risk factor for severity of organizing pneumonia." Respiration **89**(2): 119-126.
8. Herrett, E., A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. v. Staa and L. Smeeth (2015). "Data Resource Profile: Clinical Practice Research Datalink (CPRD)." International Journal of Epidemiology **44**(3): 827-836.
9. Joy, G. M., O. A. Arbiv, C. K. Wong, S. D. Lok, N. A. Adderley, K. M. Dobosz, K. A. Johannson and C. J. Ryerson (2023). "Prevalence, imaging patterns and risk factors of interstitial lung disease in connective tissue disease: a systematic review and meta-analysis." Eur Respir Rev **32**(167).
10. Kaul, B., V. Cottin, H. R. Collard and C. Valenzuela (2021). "Variability in Global Prevalence of Interstitial Lung Disease." Front Med (Lausanne) **8**: 751181.
11. Ly, N. F., C. Flach, T. S. Lysen, E. Markov, H. van Ballegooijen, P. Rijnbeek, T. Duarte-Salles, C. Reyes, L. H. John, L. Karimi, C. Reich, S. Salek and D. Layton (2023). "Impact of European Union Label Changes for Fluoroquinolone-Containing Medicinal Products for Systemic and Inhalation Use: Post-Referral Prescribing Trends." Drug Saf **46**(4): 405-416.
12. Marti, C. (2024). CohortCharacteristics.
13. Paolo, S., B. Philippe, R. Giulio, S. Nicola and C. Vincent (2022). "Drug-induced interstitial lung disease." European Respiratory Journal **60**(4): 2102776.
14. Pinheiro, L., K. Blake, J. Januskiene, Q. Y. Yue and P. Arlett (2016). "Geographical variation in reporting Interstitial Lung Disease as an adverse drug reaction: findings from an European Medicines Agency analysis of reports in EudraVigilance." Pharmacoepidemiol Drug Saf **25**(6): 705-712.

15. Rathmann, W., B. Bongaerts, H. J. Carius, S. Kruppert and K. Kostev (2018). "Basic characteristics and representativeness of the German Disease Analyzer database SEP" Int J Clin Pharmacol Ther **56**(10): 459-466.
16. Spagnolo, P., P. Bonniaud, G. Rossi, N. Sverzellati and V. Cottin (2022). "Drug-induced interstitial lung disease." Eur Respir J **60**(4).
17. Tanislav, C., J. Rosenbauer, R. Zingel and K. Kostev (2022). "No increased incidence of venous thrombosis or pulmonary embolism after SARS-CoV-2 vaccination in Germany." Public Health **207**: 14-18.
18. Wigglesworth S, N. A., Dickson JM, Pullen A, Yelland E, Anjuman T, Reuber M. (2023). "The incidence and prevalence of epilepsy in the United Kingdom 2013-2018: A retrospective cohort study of UK primary care data." Seizure **105**: 37-42.
19. Zappacosta, S., A. Cascarano, M. Konrad, C. Tanislav and K. Kostev (2022). "Association between COVID-19 and subsequent vascular events in primary care patients in Germany." Public Health **213**: 107-113.

17. ANNEXES

Appendix I: List of Stand-Alone documents (e.g. lists with concept definitions (conditions & drugs), validation procedures, questionnaires etc.)

Concept IDs related to ILD were identified within the Darwin Atlas and each concept ID was reviewed to decide whether it could be included as part of ILD (table S1.A) or whether it should be excluded (Table S1.B). Amongst the included concept IDs we next differentiated between alveolitis/pneumonitis, lung fibrosis, pneumoconiosis, PAP (pulmonary alveolar proteinosis) and sarcoidosis. Amongst the subclasses, 2 broad categories could be identified namely alveolitis/pneumonitis and lung fibrosis. There are some other classes too but with presumably with limited number of individuals.

Table S1.A. Concept definitions of interstitial lung disease (ILD) and ILD-subtypes

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
Parents						
4197819	51615001	Fibrosis of lung		Yes		
4119786	233703007	Interstitial lung disease			Yes	
253506	205237003	Pneumonitis	Yes			
435853	10501004	Pulmonary alveolar proteinosis	Yes			
Descendants						
37312199	789574002	Acute exacerbation of idiopathic pulmonary fibrosis		Yes		
4027868	13274008	Atrophic fibrosis of lung		Yes		
4032314	14700006	Bauxite fibrosis of lung		Yes		
4275496	36599006	Chronic fibrosis of lung		Yes		
3655115	846637007	Chronic pulmonary fibrosis caused by chemical vapors		Yes		
4112681	196028003	Chronic pulmonary fibrosis due to chemical fumes		Yes		
4025168	196125002	Diffuse interstitial pulmonary fibrosis		Yes		
37017059	713244007	Drug induced pulmonary fibrosis		Yes	Yes	
4140134	426437004	Familial idiopathic pulmonary fibrosis		Yes		

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
4197819	51615001	Fibrosis of lung		Yes		
4322799	71193007	Fibrosis of lung caused by radiation		Yes		
4066407	17385007	Graphite fibrosis of lung		Yes		
36675042	771306007	Hereditary fibrosing poikiloderma, tendon contractures, myopathy, pulmonary fibrosis syndrome		Yes		
45763750	700250006	Idiopathic pulmonary fibrosis		Yes		
600563	1017197007	Interstitial pulmonary fibrosis due to inhalation of drug		Yes		Yes
600562	1017196003	Interstitial pulmonary fibrosis due to inhalation of substance		Yes		
4236182	90610005	Interstitial pulmonary fibrosis of prematurity		Yes		
4209871	56841008	Massive fibrosis of lung		Yes		
4230447	40640008	Massive silicotic fibrosis of lung		Yes		
37208102	4.60561E+14	PF-ILD-progressive fibrosing interstitial lung disease		Yes		
253797	266368002	Post-inflammatory pulmonary fibrosis		Yes		
45769389	708030004	Pulmonary emphysema co-occurrent with fibrosis of lung		Yes		
4236725	405570007	Pulmonary fibrosis due to and following radiotherapy		Yes		
45768903	707434003	Pulmonary fibrosis due to Hermansky-Pudlak syndrome		Yes		
37109889	723829000	Pulmonary fibrosis, hepatic hyperplasia, bone marrow hypoplasia syndrome		Yes		
4025216	10613001	Acute berylliosis	Yes			
4110182	196052005	Acute drug-induced interstitial lung disorder	Yes			Yes
37312199	789574002	Acute exacerbation of idiopathic pulmonary fibrosis		Yes		
4341520	236302005	Acute interstitial pneumonia	Yes			
4112678	196021009	Acute pneumonitis due to chemical fumes	Yes			
4110644	196046009	Acute pulmonary radiation disease	Yes			
260434	196047000	Acute radiation pneumonitis	Yes			
45768986	707541006	Acute respiratory distress in newborn with surfactant disorder	Yes			
4124544	233760007	Acute silicosis	Yes			

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
4236624	90623003	Aluminosis of lung				
4017691	115666004	Animal handlers' lung	Yes			
4142782	33548005	Anthracosilicosis		Yes		
4239466	58691003	Antimony pneumoconiosis		Yes		
256450	22607003	Asbestosis		Yes		
45768909	707443007	Autoimmune pulmonary alveolar proteinosis	Yes			
4049753	15708009	Benign pneumoconiosis		Yes		
4221139	8247009	Berylliosis	Yes			
439853	69339004	Bird-fanciers' lung	Yes			
439298	196019004	Bronchitis and pneumonitis due to chemical fumes	Yes			
4026217	12088005	Budgerigar-fanciers' disease	Yes			
4291799	37711000	Cadmium pneumonitis	Yes			
4302900	78723001	Cannabinosis	Yes			
4223637	40218008	Carbon electrode makers' pneumoconiosis		Yes		
3173010	5.92E+15	Carmustine pulmonary toxicity	Yes			
4223637	233754007	Cerium pneumoconiosis		Yes		
4232596	404807005	Cheese-washers' lung	Yes			
4119795	233728004	Cholesterol pneumonia	Yes			
4111455	18121009	Chronic berylliosis	Yes			
4112814	196053000	Chronic drug-induced interstitial lung disorders	Yes			Yes
605242	1010670004	Chronic endogenous lipoid pneumonia	Yes			
45768998	707553005	Chronic exogenous lipoid pneumonia	Yes			
762964	4.34301E+14	Chronic interstitial lung disease			Yes	
45767051	704345008	Chronic interstitial pneumonia	Yes			
45769386	708026002	Chronic pneumonitis of infancy	Yes			

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
3655115	846637007	Chronic pulmonary fibrosis caused by chemical vapors		Yes		
4112681	196028003	Chronic pulmonary fibrosis due to chemical fumes		Yes		
37204512	783182004	Chronic respiratory distress with surfactant metabolism deficiency	Yes			
4119447	233762004	Chronic silicosis		Yes		
252946	29422001	Coal workers' pneumoconiosis		Yes		
4313425	86263001	Cobaltosis		Yes		
4040982	16623004	Coffee-workers' lung	Yes			
4216394	72270005	Collagenous pneumoconiosis		Yes		
4119442	233749003	Complicated pneumoconiosis		Yes		
4177951	49840000	Complicated silicosis		Yes		
45772936	707442002	Congenital pulmonary alveolar proteinosis	Yes			
36714118	719218000	Cryptogenic organizing pneumonia	Yes			
4120262	233692000	Cryptogenic pulmonary eosinophilia	Yes			
4311555	8549006	Desquamative interstitial pneumonia	Yes			
4028118	10713006	Diffuse interstitial rheumatoid disease of lung	Yes			
4116317	302913000	Diffuse pulmonary calcinosis		Yes		
4120270	233717003	Diffuse pulmonary neurofibromatosis		Yes		
4112813	196051003	Drug-induced interstitial lung disorder	Yes			Yes
4140472	427046006	Drug-induced pneumonitis	Yes			Yes
4177385	42680007	Endogenous lipoid pneumonia	Yes			
1340380	OMOP5166035	Exacerbation of interstitial pneumonia	Yes			
45768914	707449006	Exogenous lipoid pneumonia	Yes			
444084	37471005	Extrinsic allergic alveolitis	Yes			
4140134	426437004	Familial idiopathic pulmonary fibrosis		Yes		
435298	18690003	Farmers' lung	Yes			

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
4002903	11944003	Feather-pickers' disease	Yes			
4249023	73448002	Fish-meal workers' lung	Yes			
4027669	13151001	Flax-dressers' disease	Yes			
4285279	68333005	Furriers' lung	Yes			
46273640	3.28641E+14	Genetic disorder of surfactant dysfunction	Yes			
4051339	233625007	Giant cell pneumonia	Yes			
4103099	192658007	GIP - Giant cell interstitial pneumonitis	Yes			
4338389	87909002	Hard metal pneumoconiosis		Yes		
42599152	3.47351E+14	Hypersensitivity pneumonitis due to inhalation of Micropolyspora faeni spores	Yes			
37110291	724499007	Idiopathic acute eosinophilic pneumonia	Yes			
37110292	724500003	Idiopathic chronic eosinophilic pneumonia	Yes			
45769390	708031000	Idiopathic eosinophilic pneumonia	Yes			
45763749	700249006	Idiopathic interstitial pneumonia	Yes			
36712839	1.2381E+13	Idiopathic pneumonia syndrome	Yes			
45763750	700250006	Idiopathic pulmonary fibrosis		Yes		
438782	40527005	Idiopathic pulmonary hemosiderosis	Yes			
4119786	233703007	Interstitial lung disease			Yes	
4140605	427123006	Interstitial lung disease due to collagen vascular disease			Yes	
46272927	711379004	Interstitial lung disease due to connective tissue disease			Yes	
42539687	737182002	Interstitial lung disease due to granulomatous disease			Yes	
3655634	866103007	Interstitial lung disease due to juvenile polymyositis			Yes	
42537658	737183007	Interstitial lung disease due to metabolic disease			Yes	
42537657	737181009	Interstitial lung disease due to systemic disease			Yes	
46270493	3.28661E+14	Interstitial lung disease of childhood			Yes	
42539090	737184001	Interstitial lung disease with systemic vasculitis			Yes	

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
4273378	64667001	Interstitial pneumonia	Yes			
3655969	870573008	Interstitial pneumonia with autoimmune features	Yes			
600563	1017197007	Interstitial pulmonary fibrosis due to inhalation of drug		Yes		Yes
600562	1017196003	Interstitial pulmonary fibrosis due to inhalation of substance		Yes		
4218175	416916004	Lipoid pneumonitis	Yes			
4124539	233730002	Lupus pneumonia	Yes			
45768902	707433009	Lymphangioleiomyomatosis due to tuberous sclerosis syndrome		Yes		
4195014	44274007	Lymphoid interstitial pneumonia	Yes			
434975	25897000	Malt-workers' lung	Yes			
4229303	88687001	Manganese pneumonitis	Yes			
438175	86638007	Maple-bark strippers' lung	Yes			
4230447	40640008	Massive silicotic fibrosis of lung		Yes		
4121294	233751004	Metal pneumoconiosis		Yes		
4119445	233758005	Mica pneumoconiosis		Yes		
4137769	32139003	Mixed dust pneumoconiosis		Yes		
4119446	233759002	Mixed mineral dust pneumoconiosis		Yes		
433233	52333004	Mushroom workers' lung	Yes			
4006973	111292008	Necrotizing sarcoid granulomatosis			Yes	
45771019	707435002	Neuroendocrine cell hyperplasia of infancy			Yes	
4121295	233755008	Nickel pneumoconiosis		Yes		
4044215	129452008	Nonspecific interstitial pneumonia	Yes			
4124537	233698001	Paprika splitters' lung	Yes			
258564	60125001	Perinatal interstitial emphysema			Yes	
37208102	4.60561E+14	PF-ILD-progressive fibrosing interstitial lung disease		Yes		
259044	40122008	Pneumoconiosis		Yes		

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
37205802	785345002	Pneumoconiosis caused by sisal dust		Yes		
254389	17996008	Pneumoconiosis due to inorganic dust		Yes		
442125	805002	Pneumoconiosis due to silica		Yes		
256146	426853005	Pneumoconiosis due to silicate		Yes		
4249010	73144008	Pneumoconiosis due to talc		Yes		
253506	205237003	Pneumonitis	Yes			
3655110	846629004	Pneumonitis caused by fumes	Yes			
3655088	840728005	Pneumonitis caused by inhalation of oil	Yes			
3655111	846630009	Pneumonitis caused by vapors	Yes			
256721	64030005	Pneumonitis due to inhaled solid	Yes			
4187218	415126001	Pneumonitis due to inhaled substance	Yes			
4226132	405569006	Post-radiotherapy pneumonitis	Yes			
4215594	80614003	Prolonged pulmonary alveolitis	Yes			
435853	10501004	Pulmonary alveolar proteinosis (PAP)	Yes			
37109889	723829000	Pulmonary fibrosis, hepatic hyperplasia, bone marrow hypoplasia syndrome		Yes		
45768996	707551007	Pulmonary interstitial glycogenosis			Yes	
4174275	277844007	Pulmonary lymphangioleiomyomatosis			Yes	
4086243	24369008	Pulmonary sarcoidosis			Yes	
4266525	62371005	Pulmonary siderosis		Yes		
4221865	84004001	Radiation pneumonitis	Yes			
4045227	129451001	Respiratory bronchiolitis associated interstitial lung disease	Yes			
437313	7548000	Rheumatic pneumonia	Yes			
4184896	54867000	Rheumatoid fibrosing alveolitis		Yes		
4162539	398640008	Rheumatoid pneumoconiosis		Yes		
4120265	233700005	Rodent handlers' lung	Yes			

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
4093002	187233002	Sarcoidosis of lung with sarcoidosis of lymph nodes			Yes	
45771023	707510005	Secondary pulmonary alveolar proteinosis	Yes			
4050564	23315001	Sequoiosis	Yes			
4120266	233701009	Sewage workers' lung	Yes			
4144984	34004002	Siderosilicosis		Yes		
4247524	61233003	Silo-fillers' disease	Yes			
4179182	50589003	Silver polishers' lung disease	Yes			
4119441	233748006	Simple pneumoconiosis		Yes		
4243523	59773008	Simple pulmonary alveolitis	Yes			
4277596	64936001	Simple pulmonary eosinophilia	Yes			
4166078	47515009	Simple silicosis		Yes		
4124546	233767005	Stage 1 pulmonary sarcoidosis			Yes	
4124671	233768000	Stage 2 pulmonary sarcoidosis	Yes			
4119448	233769008	Stage 3 pulmonary sarcoidosis		Yes		
4119935	233770009	Stage 4 pulmonary sarcoidosis		Yes		
4196950	51277007	Stannosis		Yes		
4119444	233753001	Subacute berylliosis			Yes	
4124545	233761006	Subacute silicosis			Yes	
443890	13394002	Suberosis		Yes		
4119428	233702002	Summer-type hypersensitivity pneumonitis	Yes			
4121296	233756009	Thorium pneumoconiosis		Yes		
4119796	233733000	Toxic pneumonitis	Yes			
4027411	10785007	Vinyard sprayers' lung	Yes			
4305873	38729007	Wheat weevil disease	Yes			
4124543	233757000	Zirconium pneumoconiosis		Yes		

Concept ID	Concept Code	Concept Name	Alveolitis/ Pneumonitis	Lung fibrosis	Unspecified ILD	Drug-induced ILD*
1340517	OMOP5166172	Progression of massive fibrosis of lung		Yes		

PAP= Pulmonary Alveolar Proteinosis, * Drug induced ILD can contribute to multiple categories as there are drug induced alveolitis/pneumonitis as well as drug induced pulmonary fibrosis.

Table S1.B. Excluded concept ids for ILD

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
42598655	3.36791E+14	Atypical interstitial pneumonia of cattle	Condition	SNOMED Veterinary
4146744	30042003	Confluent fibrosis of lung	Condition	SNOMED
4262577	361196000	Idiopathic hilar fibrosis	Condition	SNOMED
4120272	233726000	Localized pulmonary fibrosis	Condition	SNOMED
36716112	721977007	Lung fibrosis, immunodeficiency, 46,XX gonadal dysgenesis syndrome	Condition	SNOMED
4173590	50196008	Perialveolar fibrosis of lung	Condition	SNOMED
4148685	3514002	Peribronchial fibrosis of lung	Condition	SNOMED
258335	90117007	Tuberculous fibrosis of lung	Condition	SNOMED
36674196	770760006	16q24.1 microdeletion syndrome	Condition	SNOMED
4195694	67782005	Acute respiratory distress syndrome	Condition	SNOMED
3661406	6.74814E+17	Acute respiratory distress syndrome due to disease caused by Severe acute respiratory syndrome coronavirus 2	Condition	SNOMED
42598655	3.36791E+14	Atypical interstitial pneumonia of cattle	Condition	SNOMED Veterinary
37116655	733453005	Congenital nephrotic syndrome, interstitial lung disease, epidermolysis bullosa syndrome	Condition	SNOMED
3655347	860890006	Fetal interstitial neoplasm of lung	Condition	SNOMED
440748	77690003	Interstitial emphysema of lung	Condition	SNOMED
4084955	240629003	Malarial shock lung	Condition	SNOMED
440431	46970008	Mycoplasma pneumonia	Condition	SNOMED
4294404	76090006	Pittsburgh pneumonia	Condition	SNOMED
4110506	195896004	Pneumonia due to pleuropneumonia-like organism	Condition	SNOMED
4232327	89687005	Postimmersion-submersion syndrome	Condition	SNOMED
4148529	35037009	Primary atypical interstitial pneumonia	Condition	SNOMED
1340509	OMOP5166164	Progression of acute respiratory distress syndrome	Condition	OMOP Extension
438791	196115007	Pulmonary congestion and hypostasis	Condition	SNOMED
44783638	697923008	Pulmonary hypertension in lymphangioleiomyomatosis	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4024112	196116008	Pulmonary hypostasis	Condition	SNOMED
4025173	196151000	Pulmonary insufficiency following shock	Condition	SNOMED
4024117	196152007	Pulmonary insufficiency following surgery	Condition	SNOMED
4024118	196153002	Pulmonary insufficiency following trauma	Condition	SNOMED
4050874	233621003	Rickettsial pneumonia	Condition	SNOMED
4119431	233713004	Seasonal cryptogenic organizing pneumonia with biochemical cholestasis	Condition	SNOMED
4025165	196112005	Abscess of lung with pneumonia	Condition	SNOMED
4267135	61884008	Achromobacter pneumonia	Condition	SNOMED
4112822	195908008	Actinomycotic pneumonia	Condition	SNOMED
3200502	1.283E+16	Active tuberculosis with risk for contagion	Condition	Nebraska Lexicon
46270510	3.5031E+13	Acute aspiration pneumonia	Condition	SNOMED
4048517	123587001	Acute bronchopneumonia	Condition	SNOMED
605225	1010650005	Acute endogenous lipoid pneumonia	Condition	SNOMED
45768997	707552000	Acute exogenous lipoid pneumonia	Condition	SNOMED
4240452	58890000	Adenoviral bronchopneumonia	Condition	SNOMED
254677	41207000	Adenoviral pneumonia	Condition	SNOMED
252548	195902009	Anthrax pneumonia	Condition	SNOMED
4309106	422588002	Aspiration pneumonia	Condition	SNOMED
4308451	83608006	Aspiration pneumonia due to inhalation of milk	Condition	SNOMED
4169796	42004004	Aspiration pneumonia due to inhalation of vomitus	Condition	SNOMED
4248154	72854003	Aspiration pneumonia due to near drowning	Condition	SNOMED
4327820	75426006	Aspiration pneumonia due to regurgitated food	Condition	SNOMED
4233319	40786001	Aspiration pneumonia due to regurgitated gastric secretions	Condition	SNOMED
4195452	44549008	Aspiration pneumonia resulting from a procedure	Condition	SNOMED
4050869	233606009	Atypical pneumonia	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
257315	53084003	Bacterial pneumonia	Condition	SNOMED
4223032	420544002	Bacterial pneumonia associated with AIDS	Condition	SNOMED
37019058	713544008	Bacterial pneumonia co-occurrent with human immunodeficiency virus infection	Condition	SNOMED
4116487	300999006	Basal pneumonia	Condition	SNOMED
4138769	425996009	Bilateral basal pneumonia	Condition	SNOMED
4245006	396286008	Bilateral bronchopneumonia	Condition	SNOMED
4236311	407671000	Bilateral pneumonia	Condition	SNOMED
42593423	2.81191E+14	Bronchoalveolar pneumonia	Condition	SNOMED Veterinary
256722	396285007	Bronchopneumonia	Condition	SNOMED
46269707	1.0625E+16	Bronchopneumonia due to Achromobacter	Condition	SNOMED
46269708	1.0625E+16	Bronchopneumonia due to anaerobic bacteria	Condition	SNOMED
46269709	1.06251E+16	Bronchopneumonia due to bacteria	Condition	SNOMED
46269710	1.06251E+16	Bronchopneumonia due to Escherichia coli	Condition	SNOMED
46269711	1.06252E+16	Bronchopneumonia due to Group A Streptococcus	Condition	SNOMED
46269712	1.06252E+16	Bronchopneumonia due to Group B Streptococcus	Condition	SNOMED
46269713	1.06252E+16	Bronchopneumonia due to Haemophilus influenzae	Condition	SNOMED
46269714	1.06253E+16	Bronchopneumonia due to Human metapneumovirus	Condition	SNOMED
46269715	1.06253E+16	Bronchopneumonia due to Klebsiella pneumoniae	Condition	SNOMED
46269716	1.06254E+16	Bronchopneumonia due to methicillin resistant Staphylococcus aureus	Condition	SNOMED
46269717	1.06254E+16	Bronchopneumonia due to methicillin susceptible Staphylococcus aureus	Condition	SNOMED
46269718	1.06254E+16	Bronchopneumonia due to Mycoplasma pneumoniae	Condition	SNOMED
46269719	1.06255E+16	Bronchopneumonia due to Proteus mirabilis	Condition	SNOMED
46269720	1.06255E+16	Bronchopneumonia due to Pseudomonas	Condition	SNOMED
46269721	1.06256E+16	Bronchopneumonia due to respiratory syncytial virus	Condition	SNOMED
46269722	1.06256E+16	Bronchopneumonia due to Staphylococcus	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
46269723	1.06256E+16	Bronchopneumonia due to Staphylococcus aureus	Condition	SNOMED
46269724	1.06257E+16	Bronchopneumonia due to Streptococcus	Condition	SNOMED
46269725	1.06257E+16	Bronchopneumonia due to Streptococcus pneumoniae	Condition	SNOMED
46269726	1.06258E+16	Bronchopneumonia due to virus	Condition	SNOMED
42598991	3.44291E+14	Brooder pneumonia	Condition	SNOMED Veterinary
4175598	50804000	Catarrhal pneumonia	Condition	SNOMED
43020558	471272001	Cavitary pneumonia	Condition	SNOMED
4110510	195911009	Chickenpox pneumonia	Condition	SNOMED
45757644	2.8791E+13	Chronic coccidioidomycotic pneumonia	Condition	SNOMED
46269693	1.02361E+14	Chronic pneumonia	Condition	SNOMED
4221767	417688002	Chronic progressive coccidioidal pneumonia	Condition	SNOMED
4293463	385093006	Community acquired pneumonia	Condition	SNOMED
4048518	123588006	Confluent bronchopneumonia with abscess formation	Condition	SNOMED
4048519	123591006	Confluent pneumonia	Condition	SNOMED
4174309	276693005	Congenital bacterial pneumonia	Condition	SNOMED
4070540	206289001	Congenital chlamydial pneumonia	Condition	SNOMED
4048148	206286008	Congenital Escherichia coli pneumonia	Condition	SNOMED
4048147	206284006	Congenital group A hemolytic streptococcal pneumonia	Condition	SNOMED
4071611	206285007	Congenital group B hemolytic streptococcal pneumonia	Condition	SNOMED
255084	78895009	Congenital pneumonia	Condition	SNOMED
4048149	206287004	Congenital pseudomonal pneumonia	Condition	SNOMED
4071610	206283000	Congenital staphylococcal pneumonia	Condition	SNOMED
4174308	276692000	Congenital viral pneumonia	Condition	SNOMED
42572881	3.44871E+14	Contagious bovine pleuropneumonia	Condition	SNOMED Veterinary
42599199	3.48481E+14	Contagious caprine pleuropneumonia	Condition	SNOMED Veterinary

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
3180888	2.184E+16	Cystic fibrosis related bronchopneumonia	Condition	Nebraska Lexicon
252949	7678002	Cytomegaloviral pneumonia	Condition	SNOMED
3189306	3.021E+16	Enterobacter cloacae pneumonia	Condition	Nebraska Lexicon
42599060	3.45621E+14	Enzootic mycoplasmal pneumonia of swine	Condition	SNOMED Veterinary
42598979	3.44131E+14	Enzootic pneumonia of calves	Condition	SNOMED Veterinary
42573179	3.55251E+14	Enzootic pneumonia of sheep	Condition	SNOMED Veterinary
1340436	OMOP5166091	Exacerbation of pneumonia	Condition	OMOP Extension
1340437	OMOP5166092	Exacerbation of pneumonia caused by SARS-CoV-2	Condition	OMOP Extension
42573349	4.0991E+13	Exudative pneumonia	Condition	SNOMED Veterinary
4046011	123590007	Focal pneumonia	Condition	SNOMED
4274981	65141002	Foreign body pneumonia	Condition	SNOMED
4049965	233613009	Fungal pneumonia	Condition	SNOMED
4322625	7063008	Gangrenous pneumonia	Condition	SNOMED
45757250	1.08706E+15	Gonococcal pneumonia	Condition	SNOMED
252655	195886008	Group B streptococcal pneumonia	Condition	SNOMED
260754	70036007	Haemophilus influenzae pneumonia	Condition	SNOMED
4248807	408680002	Healthcare associated bacterial pneumonia	Condition	SNOMED
4111119	181007	Hemorrhagic bronchopneumonia	Condition	SNOMED
4051335	233617005	Hemorrhagic pneumonia	Condition	SNOMED
4051338	233624006	Herpes simplex pneumonia	Condition	SNOMED
4143092	425464007	Hospital acquired pneumonia	Condition	SNOMED
4135197	31561003	Hypostatic bronchopneumonia	Condition	SNOMED
4310964	85469005	Hypostatic pneumonia	Condition	SNOMED
4052547	233622005	Infectious mononucleosis pneumonia	Condition	SNOMED
443410	312342009	Infective pneumonia	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4215807	71926009	Infective pneumonia acquired prenatally	Condition	SNOMED
37394479	1.03311E+15	Influenza with pneumonia due to seasonal influenza virus	Condition	SNOMED
4186568	41269000	Influenzal bronchopneumonia	Condition	SNOMED
4116488	301000005	Left lower zone pneumonia	Condition	SNOMED
4114030	301002002	Left upper zone pneumonia	Condition	SNOMED
4112820	195889001	Legionella pneumonia	Condition	SNOMED
4080435	240635003	Leishmanial pneumonia	Condition	SNOMED
4145369	426696003	Lingular pneumonia	Condition	SNOMED
4133224	278516003	Lobar pneumonia	Condition	SNOMED
4052546	233608005	Meningococcal pneumonia	Condition	SNOMED
3170862	3.115E+16	Methicillin resistant Staphylococcus aureus pneumonia	Condition	Nebraska Lexicon
4051336	233618000	Mycobacterial pneumonia	Condition	SNOMED
42573218	3.59921E+14	Mycoplasmal pneumonia	Condition	SNOMED Veterinary
4048052	123589003	Necrotizing bronchopneumonia	Condition	SNOMED
35622404	763888005	Necrotizing pneumonia caused by Panton-Valentine leukocidin producing Staphylococcus aureus	Condition	SNOMED
4080883	276695003	Neonatal aspiration pneumonia	Condition	SNOMED
4051333	233610007	Neonatal chlamydial pneumonia	Condition	SNOMED
4051337	233619008	Neonatal pneumonia	Condition	SNOMED
4110509	195909000	Nocardial pneumonia	Condition	SNOMED
4174281	277869007	Non-tuberculous mycobacterial pneumonia	Condition	SNOMED
4284985	68409003	Organized pneumonia	Condition	SNOMED
252552	81164001	Ornithosis with pneumonia	Condition	SNOMED
4274802	64880000	Parainfluenza virus bronchopneumonia	Condition	SNOMED
439857	64917006	Parainfluenza virus pneumonia	Condition	SNOMED
4205578	55679008	Peribronchial pneumonia	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4245499	60485005	Pleurobronchopneumonia	Condition	SNOMED
42600167	4.2721E+13	Pleuropneumonia	Condition	SNOMED Veterinary
4141619	266350000	Pneumococcal lobar pneumonia	Condition	SNOMED
258785	233607000	Pneumococcal pneumonia	Condition	SNOMED
4221503	420787001	Pneumococcal pneumonia associated with AIDS	Condition	SNOMED
4190647	415125002	Pneumocystosis pneumonia	Condition	SNOMED
255848	233604007	Pneumonia	Condition	SNOMED
256723	195878008	Pneumonia and influenza	Condition	SNOMED
4228277	421671002	Pneumonia associated with AIDS	Condition	SNOMED
605209	1010634002	Pneumonia caused by Acinetobacter	Condition	SNOMED
37119233	724498004	Pneumonia caused by Chlamydia pneumoniae	Condition	SNOMED
254066	233609002	Pneumonia caused by Chlamydiaceae	Condition	SNOMED
759821	1.0311E+13	Pneumonia caused by Enterobacter	Condition	SNOMED
759817	1.0291E+13	Pneumonia caused by Enterococcus	Condition	SNOMED
759816	1.0281E+13	Pneumonia caused by Enterococcus faecalis	Condition	SNOMED
759815	1.0271E+13	Pneumonia caused by Enterococcus faecium	Condition	SNOMED
37116366	733051000	Pneumonia caused by Gram positive bacteria	Condition	SNOMED
37016927	713084008	Pneumonia caused by Human coronavirus	Condition	SNOMED
36676238	772839003	Pneumonia caused by Influenza A virus	Condition	SNOMED
3661408	8.82785E+17	Pneumonia caused by SARS-CoV-2	Condition	SNOMED
759818	1.0301E+13	Pneumonia caused by Serratia	Condition	SNOMED
607087	1149093006	Pneumonia caused by vancomycin resistant Enterococcus	Condition	SNOMED
4256236	409665004	Pneumonia due to aerobic bacteria	Condition	SNOMED
257908	409664000	Pneumonia due to anaerobic bacteria	Condition	SNOMED
45757206	1.08272E+15	Pneumonia due to Ascaris	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
46269954	1.09295E+15	Pneumonia due to Bordetella parapertussis	Condition	SNOMED
260430	51530003	Pneumonia due to Escherichia coli	Condition	SNOMED
258180	430395005	Pneumonia due to Gram negative bacteria	Condition	SNOMED
46270121	1.42931E+14	Pneumonia due to H1N1 influenza	Condition	SNOMED
40481839	442094008	Pneumonia due to Histoplasma	Condition	SNOMED
4299862	38699009	Pneumonia due to Histoplasma capsulatum	Condition	SNOMED
40482061	445096001	Pneumonia due to Human metapneumovirus	Condition	SNOMED
40481335	441942006	Pneumonia due to infection by Streptococcus pyogenes	Condition	SNOMED
46270318	1.6311E+13	Pneumonia due to influenza	Condition	SNOMED
763012	4.34931E+14	Pneumonia due to Influenza A virus subtype H1N1	Condition	SNOMED
253790	64479007	Pneumonia due to Klebsiella pneumoniae	Condition	SNOMED
42573178	3.55231E+14	Pneumonia due to Mannheimia haemolytica	Condition	SNOMED Veterinary
4110039	195900001	Pneumonia due to measles	Condition	SNOMED
46270027	1.24691E+14	Pneumonia due to methicillin resistant Staphylococcus aureus	Condition	SNOMED
46274035	1.28711E+14	Pneumonia due to methicillin susceptible Staphylococcus aureus	Condition	SNOMED
4050872	233620002	Pneumonia due to parasitic infestation	Condition	SNOMED
4193964	39172002	Pneumonia due to Proteus mirabilis	Condition	SNOMED
252351	41381004	Pneumonia due to Pseudomonas	Condition	SNOMED
436145	195881003	Pneumonia due to respiratory syncytial virus	Condition	SNOMED
45768961	707508008	Pneumonia due to Schistosoma haematobium	Condition	SNOMED
45768960	707507003	Pneumonia due to Schistosoma japonicum	Condition	SNOMED
45771022	707503004	Pneumonia due to Schistosoma mansoni	Condition	SNOMED
40479642	441590008	Pneumonia due to Severe acute respiratory syndrome coronavirus	Condition	SNOMED
40480033	441658007	Pneumonia due to Staphylococcus aureus	Condition	SNOMED
261324	34020007	Pneumonia due to Streptococcus	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
261053	111900000	Pneumonia in aspergillosis	Condition	SNOMED
253235	59475000	Pneumonia in pertussis	Condition	SNOMED
4137435	32286006	Pneumonia in Q fever	Condition	SNOMED
259048	84753008	Pneumonia in systemic mycosis	Condition	SNOMED
4166072	45312009	Pneumonia in typhoid fever	Condition	SNOMED
4110507	195904005	Pneumonia with coccidioidomycosis	Condition	SNOMED
42573020	3.49781E+14	Porcine contagious pleuropneumonia	Condition	SNOMED Veterinary
260028	191727003	Post measles pneumonia	Condition	SNOMED
4153356	371072008	Postobstructive pneumonia	Condition	SNOMED
4203846	438764004	Postoperative aspiration pneumonia	Condition	SNOMED
4200891	314978007	Postoperative pneumonia	Condition	SNOMED
4112655	195888009	Proteus pneumonia	Condition	SNOMED
4334649	430969000	Recurrent aspiration pneumonia	Condition	SNOMED
37017277	713525001	Recurrent bacterial pneumonia	Condition	SNOMED
37017278	713526000	Recurrent bacterial pneumonia co-occurrent with human immunodeficiency virus infection	Condition	SNOMED
44782989	699014000	Recurrent pneumonia	Condition	SNOMED
4117114	301001009	Right lower zone pneumonia	Condition	SNOMED
4102253	301003007	Right middle zone pneumonia	Condition	SNOMED
4114031	301004001	Right upper zone pneumonia	Condition	SNOMED
45770900	1.09236E+15	Rubella pneumonia	Condition	SNOMED
258333	2523007	Salmonella pneumonia	Condition	SNOMED
4204819	308906005	Secondary bacterial pneumonia	Condition	SNOMED
3178885	1.30001E+14	Secondary pneumonia	Condition	Nebraska Lexicon
259852	22754005	Staphylococcal pneumonia	Condition	SNOMED
4276663	64703005	Terminal bronchopneumonia	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4119436	233731003	Traumatic pneumonia	Condition	SNOMED
254266	80003002	Tuberculous pneumonia	Condition	SNOMED
42599561	3.57061E+14	Tuberculous pneumonia of animals	Condition	SNOMED Veterinary
4280213	66429007	Unresolved lobar pneumonia	Condition	SNOMED
4212120	57702005	Unresolved pneumonia	Condition	SNOMED
259992	429271009	Ventilator associated pneumonia	Condition	SNOMED
261326	75570004	Viral pneumonia	Condition	SNOMED
4225318	421508002	Viral pneumonia associated with AIDS	Condition	SNOMED
260125	5505005	Acute bronchiolitis	Condition	SNOMED
4052545	233603001	Acute bronchiolitis due to adenovirus	Condition	SNOMED
254058	195739001	Acute bronchiolitis due to respiratory syncytial virus	Condition	SNOMED
4035960	15199004	Acute bronchiolitis with bronchospasm	Condition	SNOMED
4215773	718004	Acute bronchiolitis with obstruction	Condition	SNOMED
4112524	195737004	Acute exudative bronchiolitis	Condition	SNOMED
4243668	59903001	Acute obliterating bronchiolitis	Condition	SNOMED
4082065	240741002	Acute pulmonary African histoplasmosis	Condition	SNOMED
4052544	233602006	Acute viral bronchiolitis	Condition	SNOMED
4045589	13089009	Adenoviral bronchiolitis	Condition	SNOMED
42573322	3.9231E+13	Air sacculitis due to Aspergillus spp.	Condition	SNOMED Veterinary
42573323	3.9241E+13	Air sacculitis due to aspiration	Condition	SNOMED Veterinary
42599383	3.51301E+14	Airsacculitis	Condition	SNOMED Veterinary
257583	37981002	Allergic bronchopulmonary aspergillosis	Condition	SNOMED
45757063	1.03781E+14	Allergic bronchopulmonary mycosis	Condition	SNOMED
4244542	38534008	Aspiration of stomach contents after anesthesia AND/OR sedation in labor AND/OR delivery	Condition	SNOMED
4306082	155597006	Aspiration pneumonitis	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4112809	196040003	Aspiration pneumonitis due to anesthesia during labor and delivery	Condition	SNOMED
37018497	715069001	Aspiration pneumonitis of fetus	Condition	SNOMED
4170900	41997000	Asthmatic pulmonary alveolitis	Condition	SNOMED
252942	233691007	Asthmatic pulmonary eosinophilia	Condition	SNOMED
437588	67242002	Bagassosis	Condition	SNOMED
4176128	50076003	Baritosis	Condition	SNOMED
4295710	76157009	Bituminosis	Condition	SNOMED
42573181	3.55281E+14	Bovine respiratory disease complex	Condition	SNOMED Veterinary
4165112	4120002	Bronchiolitis	Condition	SNOMED
46269741	1.06292E+16	Bronchiolitis caused by influenza virus	Condition	SNOMED
40482069	445102008	Bronchiolitis due to Human metapneumovirus	Condition	SNOMED
4179634	52409006	Bronchiolitis exudativa	Condition	SNOMED
42538810	762618008	Bronchiolitis obliterans syndrome due to and after lung transplantation	Condition	SNOMED
4008726	111901001	Bronchocentric granulomatosis	Condition	SNOMED
4311814	85761009	Byssinosis	Condition	SNOMED
4050731	233672007	Byssinosis grade 3	Condition	SNOMED
260041	3487004	Candidiasis of lung	Condition	SNOMED
4222062	421047005	Candidiasis of lung associated with AIDS	Condition	SNOMED
4196400	44547005	Chalicosis	Condition	SNOMED
44790797	2.43001E+14	Chemical inhalation injury	Condition	SNOMED
46274046	1.06258E+16	Chemical pneumonitis caused by anesthesia	Condition	SNOMED
42539089	737180005	Chronic bronchiolitis	Condition	SNOMED
4144107	266404004	Chronic chemical respiratory disease	Condition	SNOMED
4166517	47938003	Chronic obliterative bronchiolitis	Condition	SNOMED
3654836	840350008	Chronic obliterative bronchiolitis due to chemical fumes	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
3654837	840351007	Chronic obliterative bronchiolitis due to vapor	Condition	SNOMED
4084973	240742009	Chronic pulmonary African histoplasmosis	Condition	SNOMED
4080753	240747003	Chronic pulmonary blastomycosis	Condition	SNOMED
42598908	3.42641E+14	Chronic viral encephalomyelitis of sheep	Condition	SNOMED Veterinary
4188480	47082005	Congenital rubella pneumonitis	Condition	SNOMED
4309805	85438006	Diatomaceous earth disease	Condition	SNOMED
4328679	430476004	Diffuse panbronchiolitis	Condition	SNOMED
4123255	233694004	Dog house disease	Condition	SNOMED
4051465	233673002	Drug-induced bronchiolitis obliterans	Condition	SNOMED
4119427	233695003	Dry rot lung	Condition	SNOMED
4124672	233771008	Endobronchial sarcoidosis	Condition	SNOMED
4279553	367542003	Eosinophilic asthma	Condition	SNOMED
42599654	3.57991E+14	Eosinophillic bronchopneumonitis	Condition	SNOMED Veterinary
42599221	3.49001E+14	Equine allergic pneumonitis	Condition	SNOMED Veterinary
3187037	2.607E+16	Exacerbation of chronic bronchiolitis	Condition	Nebraska Lexicon
1340452	OMOP5166107	Exacerbation of pulmonary tuberculosis	Condition	OMOP Extension
42599228	3.49061E+14	Feline pneumonitis	Condition	SNOMED Veterinary
4294182	385479009	Follicular bronchiolitis	Condition	SNOMED
4273372	64631008	Fullers' earth disease	Condition	SNOMED
42573321	3.9221E+13	Fungal air sacculitis	Condition	SNOMED Veterinary
36674821	770674007	Ghon complex	Condition	SNOMED
4196622	79958002	Grain fever	Condition	SNOMED
4000159	19274004	Grain-handlers' disease	Condition	SNOMED
4248029	72656004	Granulomatous pneumonia	Condition	SNOMED
4256894	408681003	Healthcare associated Legionnaires' disease	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
256036	31920006	Hemorrhagic varicella pneumonitis	Condition	SNOMED
4026139	197367007	Hepatic granulomas in berylliosis	Condition	SNOMED
4119929	233744008	Hilar lymph node sarcoidosis	Condition	SNOMED
37110889	725415009	House allergic alveolitis	Condition	SNOMED
4121299	233774000	Humidifier fever	Condition	SNOMED
434670	48347002	Humidifier lung	Condition	SNOMED
4243675	59940009	Hypersensitivity alveolitis in lungworm infection	Condition	SNOMED
4176753	428697002	Inactive tuberculosis of lung	Condition	SNOMED
260936	186175002	Infiltrative lung tuberculosis	Condition	SNOMED
3168004	1.773E+16	Inhalation injury due to anhydrous ammonia	Condition	Nebraska Lexicon
3170305	9.55E+15	Inhalation lung injury due to chlorine	Condition	Nebraska Lexicon
765178	1.29211E+14	Isoniazid resistant tuberculosis of lung	Condition	SNOMED
4272230	36696005	Kaolinosis	Condition	SNOMED
4250618	7343008	Liparitosis	Condition	SNOMED
4081069	238676008	Lofgrens syndrome	Condition	SNOMED
4120263	233696002	Lycoperdonosis	Condition	SNOMED
4315046	86649001	Meconium pneumonitis	Condition	SNOMED
4309159	213223003	Mendelson's syndrome as a complication of care	Condition	SNOMED
4188630	47386001	Mendelson's syndrome resulting from a procedure	Condition	SNOMED
4151351	28295001	Middle lobe syndrome	Condition	SNOMED
4052548	233623000	Mononuclear interstitial pneumonia	Condition	SNOMED
45768849	707370001	Multiple hyalinizing granuloma of lung	Condition	SNOMED
4227290	87695000	Necrotizing bronchiolitis	Condition	SNOMED
4120264	233697006	New Guinea lung	Condition	SNOMED
255175	80602006	Nodular tuberculosis of lung	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4021760	105977003	Non-infectious pneumonia	Condition	SNOMED
4222731	40100001	Obliterative bronchiolitis	Condition	SNOMED
4112832	195989002	Pituitary snuff-takers' disease	Condition	SNOMED
45768851	707373004	Plasma cell granuloma of lung	Condition	SNOMED
4112676	196017002	Pneumoconiosis associated with tuberculosis	Condition	SNOMED
254561	38976008	Pneumonic plague	Condition	SNOMED
444099	46207001	Pneumonitis due to acquired toxoplasmosis	Condition	SNOMED
42536542	735466008	Pneumonitis due to aspiration of blood	Condition	SNOMED
36712850	1.2571E+13	Pneumonitis due to Herpes zoster	Condition	SNOMED
4311410	86294001	Pneumonitis due to inhalation of essence	Condition	SNOMED
4112683	196034005	Pneumonitis due to inhalation of milk	Condition	SNOMED
4112839	196033004	Pneumonitis due to inhalation of regurgitated food	Condition	SNOMED
4110180	196035006	Pneumonitis due to inhalation of vomitus	Condition	SNOMED
442297	57463004	Pneumonitis due to inhaled liquid	Condition	SNOMED
255735	35339003	Primary pneumonic plague	Condition	SNOMED
42572644	3.36161E+14	Pulmonary abscess due to Rhodococcus	Condition	SNOMED Veterinary
4138244	32204007	Pulmonary actinobacillosis	Condition	SNOMED
258354	187052004	Pulmonary African histoplasmosis	Condition	SNOMED
4077129	17993000	Pulmonary arteritis	Condition	SNOMED
4296039	76846002	Pulmonary endarteritis	Condition	SNOMED
4102140	28122003	Pulmonary eosinophilic granuloma	Condition	SNOMED
4345699	240387006	Pulmonary glanders	Condition	SNOMED
46269691	1.01401E+14	Pulmonary granuloma	Condition	SNOMED
4119928	233742007	Pulmonary hyalinizing granuloma	Condition	SNOMED
44783637	697921005	Pulmonary hypertension in sarcoidosis	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4345215	240391001	Pulmonary melioidosis	Condition	SNOMED
4051334	233614003	Pulmonary mucormycosis	Condition	SNOMED
253954	154283005	Pulmonary tuberculosis	Condition	SNOMED
258061	45556008	Pulmonary tularemia	Condition	SNOMED
4119785	233699009	Pyrethrum alveolitis	Condition	SNOMED
40490351	446946005	Reinfection pulmonary tuberculosis	Condition	SNOMED
40490814	447006007	Relapse pulmonary tuberculosis	Condition	SNOMED
4237921	57089007	Respiratory syncytial virus bronchiolitis	Condition	SNOMED
3175074	2.47E+16	Right lower lobe pneumonitis	Condition	Nebraska Lexicon
4119936	233772001	Sarcoid pulmonary calcification	Condition	SNOMED
4053655	1259003	Schistosis	Condition	SNOMED
442637	67525007	Secondary pneumonic plague	Condition	SNOMED
4077734	19076009	Sick building syndrome	Condition	SNOMED
4119932	233763009	Silicotuberculosis	Condition	SNOMED
45768848	707369002	Single hyalinizing granuloma of lung	Condition	SNOMED
4330286	22482002	Subacute obliterative bronchiolitis	Condition	SNOMED
3654572	836478002	Subacute obliterative bronchiolitis due to chemical fumes	Condition	SNOMED
3654573	836479005	Subacute obliterative bronchiolitis due to vapor	Condition	SNOMED
4112837	196027008	Toxic bronchiolitis obliterans	Condition	SNOMED
4159649	371043007	Toxic inhalation injury	Condition	SNOMED
4089507	187196002	Toxoplasma pneumonitis	Condition	SNOMED
4008132	12181002	Tropical pulmonary alveolitis	Condition	SNOMED
4081994	278484009	Tropical pulmonary eosinophilia	Condition	SNOMED
4251772	74387008	Tuberculosis of hilar lymph nodes	Condition	SNOMED
260315	186177005	Tuberculosis of lung with cavitation	Condition	SNOMED

Concept ID	Concept Code	Concept Name	Domain	Vocabulary
4304867	81554001	Tuberculosis of lung with involvement of bronchus	Condition	SNOMED
260630	186204008	Tuberculosis of lung, bacteriological and histological examination not done	Condition	SNOMED
4091167	186203002	Tuberculosis of lung, bacteriologically and histologically negative	Condition	SNOMED
253121	186194007	Tuberculosis of lung, confirmed by culture only	Condition	SNOMED
255454	186193001	Tuberculosis of lung, confirmed by sputum microscopy with or without culture	Condition	SNOMED
256018	186195008	Tuberculosis of lung, confirmed histologically	Condition	SNOMED
42600053	4.0621E+13	Bronchointerstitial pneumonia	Condition	SNOMED Veterinary
42599152	3.47351E+14	Hypersensitivity pneumonitis due to inhalation of Micropolyspora faeni spores	Condition	SNOMED Veterinary

Table S2 Feasibility person counts for ILD per database

concept_id	concept_name	BIFAP	CDW Bordeaux	CPRD GOLD	IQVIA - DA Germany
4119786	Interstitial lung disease	7900	3900	6900	8800
253506	Pneumonitis	100	100	2200	100
253797	Post-inflammatory pulmonary fibrosis	10600		1000	
256036	Hemorrhagic varicella pneumonitis	200			
258564	Perinatal interstitial emphysema	100	100		100
260434	Acute radiation pneumonitis	300		300	600
435853	Pulmonary alveolar proteinosis	100		100	
437313	Rheumatic pneumonia	100			
438782	Idiopathic pulmonary hemosiderosis	100		100	
439298	Bronchitis and pneumonitis due to chemical fumes	300	100	100	1200
440748	Interstitial emphysema of lung	800		200	
444084	Extrinsic allergic alveolitis	2300	100	1900	5000
3655088	Pneumonitis caused by inhalation of oil	200		100	
3655110	Pneumonitis caused by fumes	100			
3655111	Pneumonitis caused by vapors	200			
4025168	Diffuse interstitial pulmonary fibrosis	300		7300	
4028118	Diffuse interstitial rheumatoid disease of lung			200	
4044215	Nonspecific interstitial pneumonia	600			
4045227	Respiratory bronchiolitis associated interstitial lung disease	100		200	
4089507	Toxoplasma pneumonitis				100
4110180	Pneumonitis due to inhalation of vomitus			500	
4110182	Acute drug-induced interstitial lung disorder	200	100	100	200
4112678	Acute pneumonitis due to chemical fumes			100	
4112681	Chronic pulmonary fibrosis due to chemical fumes			100	
4112809	Aspiration pneumonitis due to anesthesia during labor and delivery				100
4112813	Drug-induced interstitial lung disorder	100		200	200
4112814	Chronic drug-induced interstitial lung disorders	100		100	100

concept_id	concept_name	BIFAP	CDW Bordeaux	CPRD GOLD	IQVIA - DA Germany
4112839	Pneumonitis due to inhalation of regurgitated food			100	
4119796	Toxic pneumonitis	200			
4140472	Drug-induced pneumonitis	100			
4148529	Primary atypical interstitial pneumonia	100			
4187218	Pneumonitis due to inhaled substance	49500	12200	500	12400
4195014	Lymphoid interstitial pneumonia	100		100	
4218175	Lipoid pneumonitis			100	
4221865	Radiation pneumonitis	200			
4226132	Post-radiotherapy pneumonitis	100			
4236182	Interstitial pulmonary fibrosis of prematurity			100	
4236725	Pulmonary fibrosis due to and following radiotherapy	100			
4273378	Interstitial pneumonia	200		1300	
4306082	Aspiration pneumonitis	3100		7800	
4311555	Desquamative interstitial pneumonia	100			
4341520	Acute interstitial pneumonia	300			
36714118	Cryptogenic organizing pneumonia	800		500	
37208102	PF-ILD-progressive fibrosing interstitial lung disease				
45763749	Idiopathic interstitial pneumonia	1300			
45763750	Idiopathic pulmonary fibrosis	9100		4300	
45771019	Neuroendocrine cell hyperplasia of infancy	100			
46272927	Interstitial lung disease due to connective tissue disease			100	
46274046	Chemical pneumonitis caused by anesthesia	100	200		100

Table S3. Preliminary codes for comorbidities

Concept name	Concept ID (SNOMED)	Descendants
COPD	255573	yes
Emphysema	261325	yes
Rheumatoid arthritis	80809	yes
Sclerodermia	40352976	yes
Systemic Lupus Erythematosus	257628	yes
Polymyositis/dermatomyositis	80800, 80182	yes
Granulomatosis with polyangiitis	313223	yes
Gastro-oesophageal reflux disease	318800	yes
Hepatitis B/C	40483136	yes
Cancer/malignancy	443392	yes

Table S4. Preliminary codes for medications as risk factor for ILD

Concept name	Concept ID (RxNorm)	Descendants	Excluded
Chemotherapy	TBC		
Amiodarone	1309944	yes	
Immune checkpoint inhibitors (ICIs)	40238188, 741851, 45892628, 45775965, 42629079, 1593273, 1594034, 35200783, 1536789	yes	
Tyrosine kinase inhibitors (TKIs including EGFR inhibitors)	40242675, 1325363, 1319193, 43533090, 35605522, 36853448, 43009062, 35884401, 36851588, 701915, 36861628	yes	
Mammalian target rapamycin (mTOR) inhibitors	9011440, 19092845, 36851551, 19034726	yes	
Rituximab	1309944	yes	
Statins	1545958, 1539403, 1592085, 1551860, 1549686, 1592180, 1510813, 40165636	yes	
Methotrexate	1305058	yes	
Nitrofurantoin	920293	yes	
Tumor necrosis factor (TNF)-α antagonists	1119119, 912263, 1151789, 937368, 19041065, 36855655, 36853282	yes	

TBC= to be confirmed

Appendix II: ENCePP checklist for study protocols

Study title:
DARWIN EU® - Characterising interstitial lung disease in Europe

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

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)	☐	☐	☒	

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.5	Is exposure categorized based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?	☐	☐	☒	
5.6	Is (are) (an) appropriate comparator(s) identified?	☐	☐	☒	

Comments:

Section 6: Outcome definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
6.1	Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated?	☒	☐	☐	8.6
6.2	Does the protocol describe how the outcomes are defined and measured?	☒	☐	☐	
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)	☐	☐	☒	
7.3	Does the protocol address information bias? (e.g. misclassification of exposure and outcomes, time-related bias)	☐	☐	☒	

Comments:

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

Comments:

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

Section 10: Analysis plan		**Yes**	**No**	**N/A**	**Section Number**
10.7	Does the plan describe methods for handling missing data?	☐	☐	☒	
10.8	Are relevant sensitivity analyses described?	☒	☐	☐	8.8

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?	☒	☐	☐	
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

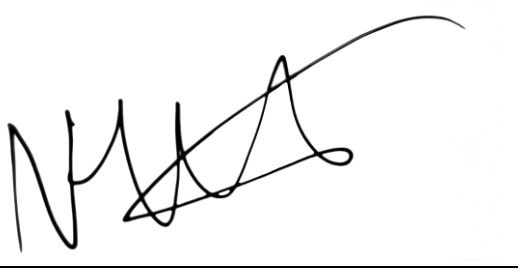
Comments:

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: 04th June 2024

Signature: 

Appendix III: Additional Information