

Burden of Idiopathic Pulmonary Fibrosis in the United Kingdom Using Electronic Health Records

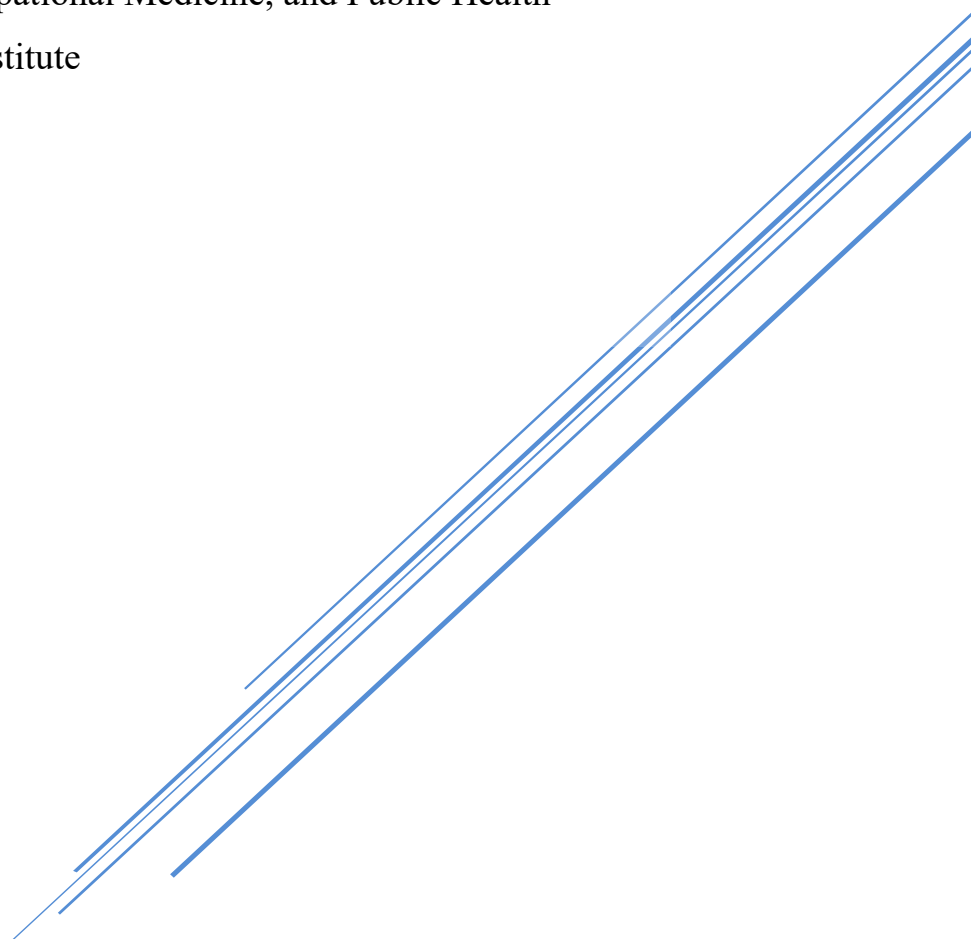
Rikisha Dhimant Shah, MPH, B.Pharm

Respiratory medicine, Occupational Medicine, and Public Health

National Heart and Lung Institute

Imperial College London

4 December 2024



THESIS SUBMITTED TO IMPERIAL COLLEGE LONDON IN ACCORDANCE WITH THE
REQUIREMENTS FOR THE DOCTOR OF PHILOSOPHY

DECLARATION OF ORIGINALITY

I, Rikisha Dhimant Shah, hereby certify that the work presented in this thesis is my own and all information presented from other works is properly referenced.

Rikisha D. Shah

COPYRIGHT DECLARATION

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non-Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

ACKNOWLEDGEMENTS

Jenni, thank you for believing in me and giving me the opportunity to pursue my PhD—this was a dream come true. What a journey it's been, from COVID right around when I started to the move and the arrival of little Rishaan. Your unwavering confidence in me made all the difference, and words cannot fully express how grateful I am. Thank you for everything, Jenni.

Ann, I truly can't imagine this journey without you. Thank you for being more than just a colleague—your friendship, advice, encouragement, and willingness to listen have meant so much to me. I've cherished every one of our conversations and the laughter we've shared along the way. Your support has been invaluable, and I've learned so much from you. Who knew I would go from a nervous receiver of what Ann's review of my work would look like to looking forward to how you found my writing and thoughts. You've helped me unlock my inner academic and researcher, and for that I am thankful.

Pete, thank you for your kindness and patience throughout this journey. I will always remember my first few weeks as a PhD student, seeing you in the clinic and witnessing the direct impact on patients. That experience truly shaped my desire to contribute to this field and make a difference in helping these patients.

I would also like to thank my friends **Laura and Phil**—thank you for always encouraging me and the endless chats that helped me fuel through this journey.

Mom, Dad and Hetuk – thank you for always believing in me. Thank you for making me feel I could do anything.

To my little Superman, **Rishaan** – Mom can do it! Everything is achievable.

I don't believe my journey would have even started without my partner in crime, **Rohit**. He knew this was my dream. He pushed, supported, and stood by me through all my ups, downs, and craziness on each day of this journey. Thank you for your unconditional love, encouragement, trust, and unwavering support.

Thank you, Rohit – this one is for you (we did it!).

Some of the work was funded through an investigator-initiated grant from Boehringer Ingelheim awarded to Professor Quint. Boehringer Ingelheim had no input into the methodology, interpretation, or decision to publish any of the work.

ABSTRACT

This thesis aims to deepen the understanding of the burden of ILD, with a particular focus on its most aggressive subtype, IPF, by addressing gaps in the current epidemiological knowledge. The primary objectives were to estimate the incidence, prevalence, and mortality of IPF and its subtypes in England, and to investigate potential changes in the disease's epidemiology over the past decade, considering advances in diagnostic techniques and the availability of treatment options. This research includes an assessment of the reliability of IPF diagnostic coding in UK primary care, an update on the epidemiology of IPF from 2008 to 2018, and an exploration of trends in survival and causes of death in patients, comparing the pre- and post-antifibrotic eras. The study found that diagnostic coding for IPF in primary care has become more focused over time, with a noticeable increase in the incidence of IPF defined using diagnostic codes in primary care settings, suggesting improved management in primary setting. However, hospital-based data showed a gradual decrease in cases, potentially reflecting better management and earlier diagnosis in primary care, reducing the need for intervention for IPF related complications. In terms of IPF-related mortality, the data revealed an initial increase, followed by a stable rate towards the end of the study period, though the lack of follow-up data makes it difficult to draw firm conclusions regarding its impact. Similarly, survival rates for IPF patients showed no significant improvement, which could be due to factors such as eligibility and access of antifibrotic treatments, delays in initiating treatment, or the advanced stage of disease at the time of diagnosis. Overall, the findings of this thesis contribute to a clearer understanding of the evolving burden of IPF in England, with implications for guiding more effective care strategies for patients with this chronic and debilitating disease.

TABLE OF CONTENTS

List of Figures	8
List of Tables	11
List of Abbreviations	14
1. Introduction	17
1.1.What is ILD and why ILD?	
1.2.Epidemiology of ILD	
1.3.Why IPF	
1.4.Rationale and hypothesis	
1.4.1. Specific aims and objectives	
1.4.2. Structure and organization of this thesis	
2. Systematic Literature Review	
Incidence and Prevalence of ILD Worldwide.....	30
2.1.Introduction	
2.2.Methods	
2.3.Results	
2.4.Discussion	
3. Data sources	67
3.1.What is CPRD and what does it do?	
3.2.How is healthcare managed in the UK?	
3.2.1. Primary care	
3.2.1.1.What is the role of GP practice in the UK?	
3.2.1.2.How does CPRD work with GP practices?	
3.2.2. Secondary care	
3.3.CPRD datasets and which data is suitable for my research?	
3.4.What information is available in Aurum data?	
3.5.What are the strengths and weakness of Aurum data?	
3.5.1. Strengths	
3.5.2. Weakness	
3.6.What linked data are available, and which did I use for my research?	
3.6.1. Hospital episodes statistics	
3.6.2. Office of national statistics	
3.7.What are the steps I followed to get access to CPRD data?	
3.7.1. ISAC approval and data access	
3.7.2. Two-step data download data	
3.7.3. Constructing code list	
4. Validation of the recording of IPF in routinely collected electronic healthcare records in England	91
4.1.Introduction	

4.2.Methods	
4.3.Results	
4.4.Discussion	
5. Updated epidemiology of IPF in England from 2008 to 2018	122
5.1.Introduction	
5.2.Methods	
5.3.Results	
5.4.Discussion	
6. Impact of antifibrotic treatments on survival: Changes in survival and causes of death in people with IPF in England, 2010 to 2022.....	151
6.1.Introduction	
6.2.Methods	
6.3.Results	
6.4.Discussion	
7. General Discussion: Summary of Main Findings Conclusion.....	173
8. References	188
9. Appendices	197
9.1.Appendix 1. Supplementary material to Chapter 2	
9.2.Appendix 2. Supplementary material to Chapter 4	
9.3.Appendix 3. Supplementary material to Chapter 5	
9.4.Appendix 4. Supplementary material to Chapter 6	
9.5.Appendix 5. Publications Associated with Thesis	

LIST OF FIGURES

CHAPTER 1

Figure 1.1. Aetiological Classification of ILDs.

Figure 1.1. Schematic overview of the research undertaken to explain how the various components contribute to the project's overall aim.

CHAPTER 2

Figure 2.1. Search strategies for EMBASE.

Figure 2.2. PRISMA Flow Diagram to Demonstrate Studies Identified, Screened, and Included.

Figure 2.3. Number of publications for ILD for the study period, 2015 to 2021.

Figure 2. 1. Geographical distribution of publications included.

CHAPTER 3

Figure 3. 1 CPRD Aurum dataset structure

Figure 3. 2 Data linkage process schema

Figure 3.3. Steps followed to find IPF patients in NHS universe.

CHAPTER 4

Figure 4.1. Eligibility criteria for Aurum-HES APC – ONS linked datasets.

Figure 4.2. Sensitivity analysis: effect of extending the HES APC cohort to include those who had IPF listed in the third diagnostic position on the concordance in the recording of IPF between ONS, Aurum and HES APC (restricted to patients who died during the study period 2008–2018)

Figure 4.3. Sensitivity analysis: effect of restricting the ONS cohort to those who had IPF coded as the underlying cause of death in the concordance in the recording of IPF between ONS (the underlying cause of death as IPF), Aurum and HES APC (restricted to patients who died during the study period 2008–2018)

Figure 4.4. Method for positive predictive value and sensitivity.

Figure 4.5. Frequency of IPF medcode ids over time (Aurum data)

Figure 4.6. Trends in the use of ‘broad’ and ‘narrow’ IPF clinical codes in CPRD Aurum primary care data, 2008–2018.

Figure 4.7. Patient Attrition in the three datasets: Aurum, HES-APC and ONS.

Figure 4.8. Concordance in the recording of IPF across ONS, Aurum and HES-APC. ONS, study eligible population with any cause of death as IPF in their death record during the study period; Aurum, patients with one clinical code denoting IPF in Aurum (using a broad set of codes) and had died for any reason during the study period; HES-APC, patients admitted to the hospital for IPF at some point before their death for any reason (during the study period)

CHAPTER 5

Figure 5.1. Incidence rate per 100,000 person-years in England, from 2008 to 2018 for men(all ages)

Figure 5.2. Incidence rate per 100,000 person-years in England, from 2008 to 2018 for women (all ages)

Figure 5.3. Incidence rate per 100,000 person-years in England, from 2008 to 2018 for men aged ≥ 70 years.

Figure 5.4. Incidence rate per 100,000 person-years in England, from 2008 to 2018 for women aged ≥ 70 years.

Figure 5.5. Prevalence per 100,000 persons- in England, from 2008 to 2018 for men (all ages)

Figure 5.6. Prevalence per 100,000 persons- in England, from 2008 to 2018 for women (all ages)

Figure 5.7. Prevalence per 100,000 persons- in England, from 2008 to 2018 for men aged ≥ 70 years.

Figure 5.8. Prevalence per 100,000 persons- in England, from 2008 to 2018 for women aged ≥ 70 years.

Figure 5.9. Adjusted and Crude Mortality rates per 100,000 person-years.

CHAPTER 6

Figure 6.1. Flow chart of patient inclusion

Figure 6.2.Crude and ASMR for Aurum narrow and Aurum broad cohorts

Figure 6.3. Crude and ASMR for HES only and HES plus Aurum cohorts

Figure 6.4. Cumulative survival for incident cases for 0 to 3 years for the four study cohorts

Figure 6.4. Cumulative survival for 3 to 5 years in incident cases for the four study cohorts

Figure 6.5. Cumulative survival for 5 to 7 years in incident cases for the four study cohorts

LIST OF TABLES

CHAPTER 2

Table 2.1. List of ILD subgroups

Table 2.2. Literature search terms for IPF

Table 2.3. Distribution of ILD conditions, sorted by distribution of publications (descending)

Table 2.1. Top researched region and country for each of the ILD subgroups

Table 2.2. Studies reporting IPF incidence and sorted by case definition (primary general, overall, broad and narrow)

Table 2.6. Studies reporting ILD incidence are grouped by ILD subgroups and sorted into ILD subgroups.

Table 2.7. List of studies reporting IPF prevalence, sorted by region.

Table 2.8. Studies reporting non-IPF prevalence, population level prevalence, by ILD condition, and year of publication.

Table 2.9. Studies reporting prevalence for Autoimmune ILDs, cohort-specific prevalence, sorted by autoimmune condition.

Table 2.10. Studies reporting prevalence for exposure ILDs, cohort specific prevalence, sorted by ILD condition.

Table 2.11. Assessment for studies reporting incidence.

Table 2.12. Assessment for studies reporting prevalence.

CHAPTER 3

Table 3.1. List of linked datasets and the ones used in my research.

Table 3.2. List my applications to CPRD to request Aurum data for my analyses.

CHAPTER 4

Table 4.1. Key Considerations for using the primary and secondary data sources

Table 4.2. List of Medcode ids included in the broad and narrow categories.

Table 4.3. Frequency of use of IPF clinical codes in UK primary care (CPRD Aurum data set), 2008–2018, ascending order

Table 4.4. Clinical and demographic characteristics of base cohorts: Aurum, HES-APC and ONS

Table 4.5. Positive Predictive Value of the Code-Based IPF Algorithms, sorted based on high PPV and sensitivity.

Table 4.6. Demographic and clinical characteristics for the patients found using algorithms DA1, DA3, DA6 and DA7

CHAPTER 5

Table 5.1. Incidence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum narrow cohort definition

Table 5.2. Incidence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum broad cohort definition

Table 5.3. Incidence rate Poisson regression modelling of idiopathic pulmonary fibrosis for HES-APC cohort definition

Table 5.4. Incidence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum plus HES APC cohort definition

Table 5.5. Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum narrow cohort definition

Table 5.6. Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum broad cohort definition

Table 5.7. Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for HES-only cohort definition

Table 5.8. Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum plus HES APC cohort definition

Table 5.9. Age standardised mortality rates and Poisson regression modelling of deaths from idiopathic pulmonary fibrosis

CHAPTER 6

Table 6.1. Clinical and demographic characteristics for the four prevalent study cohorts

Table 6.2. Crude mortality rate per 100,000 persons for four study cohorts

Table 6.3. Age and sex-standardised mortality rate per 100,000 persons for the four study cohorts

Table 6.4. Median survival in years (95% CI) among all incident patients for the four study cohorts

Table 6.5. Relative effects of calendar year of diagnosis, sex, and age group on the risk of mortality for the four study cohorts

LIST OF ABBREVIATION

A&E – Accident and emergency
ALAT – Latin American Thoracic Association (Asociación Latinoamericana de Tórax)
APC – Admitted patient care
ATS - American Thoracic Society
ASMR – Age- and sex – standardised rate
BAL - Bronchoalveolar lavage
BHD - Birt-Hogg-Dubé syndrome
BTS - British thoracic society
CI – Confidence interval
COP – Cryptogenic organizing pneumonia
COPD – Chronic obstructive pulmonary disorder
COVID-19 – Coronavirus disease 2019
CPRD – Clinical practice research datalink
CRD – Current registration date
CTD – Connective tissue disorder
CT – Computed tomography
CVD – Cardiovascular disorder
DA – Diagnostic algorithm
DID – Diagnostic imaging dataset
DIP - Desquamative Interstitial Pneumonia
EMIS -Egton Medical Information Systems
EHR – Electronic health records
ERS – European respiratory society
EU – European Union
FCEs – Finished consultant episodes
FEV – Forced expiratory volume
FVC – Forced vital capacity
GBD – Global burden of disease
GP – General practitioner
HES – Hospital episode statistics
HP – Hypersensitivity pneumonitis
HR – Hazard ratio
HRCT - High-resolution computed tomography
ICD – International classification of disease
IIP – Idiopathic interstitial pneumonia

ILD – Interstitial lung disease

IPF – Idiopathic pulmonary fibrosis

IQR - Interquartile range

ISAC – Independent Scientific Advisory Committee

JRS – Japan respiratory society

LAM - Lymphangioliomyomatosis

LCD – Last collection date

LIP - Lymphocytic Interstitial Pneumonia

MCTD – Mixed connective tissue disorder

MHRA - Medicines and Healthcare products Regulatory Agency

MI – Myocardial infarction

NHS – National health service

NICE - National Institute for Health and Care Excellence

NIHR - National Institute for Health and Care Research

NSIP - Nonspecific Interstitial Pneumonia

ONS – Office of national statistics

OP – Outpatient

PF – Progressive fibrosis

PICO – Patient, intervention, comparison, and outcome

PLCH - Pulmonary Langerhans Cell Histiocytosis

PPV – Positive predictive value

PRISMA - Preferred Reporting Items for Systematic reviews and Meta-Analyses

PROSPERO - Prospective Register of Systematic Reviews

PY – Person years

RA - Rheumatoid Arthritis

RCT – Randomised controlled trials

ROBINS-I - Risk of Bias In Non-randomized Studies

SD – Standard deviation

SLB – Surgical lung biopsy

SNOMED - Systematised Nomenclature of Medicine

SSc – Systemic sclerosis

SUS – Secondary uses services

TTP – Trusted third party

TLB - transbronchial lung biopsy

TOD -Transfer out date

UK – United Kingdom

USA -United States of America

UTS – Up-to-standard

Chapter 1

Introduction to Interstitial Lung Disease and Idiopathic Pulmonary Fibrosis

Overview

This introductory chapter begins with a brief description of interstitial lung disease (ILD) and the complex conditions that fall under its umbrella. It discusses the various changes and challenges in the field of ILD, including its diagnostic criteria, clinical implications on patient diagnosis and management, and understanding the burden of ILDs. Idiopathic pulmonary fibrosis (IPF), the most prevalent and aggressive type of ILD, remains idiopathic despite advances in the understanding of ILD and IPF. I explain the rationale for the overarching aim of my thesis, which is to contribute to the understanding and evidence base to understand the ILD burden better; in particular, I set out to shed further light on the most aggressive subtype of ILD, IPF, in England. The latter part of this chapter offers a preview of what is to come, illustrating my research narrative and outlining the aim addressed by each chapter. Finally, I wrap up with an overview of the overall structure of my thesis.

1.1. What is ILD, and why is ILD?

Interstitial lung disease (ILD) encompasses various lung disorders that affect the interstitium, the connective tissue matrix surrounding the alveoli, and the adjacent small blood vessels. This interstitial tissue plays a crucial role in facilitating the exchange of oxygen and carbon dioxide during respiration. Physiologically, oxygen from the air passes from the alveoli across a single-cell thick membrane into the surrounding capillary bed and into our bloodstream before being delivered to the vital organs of our body. At the same time, carbon dioxide moves from the capillaries into our alveoli and is exhaled. However, in ILDs, this delicate balance is disrupted as the interstitium becomes thickened, stiffened, or scarred, impairing its ability to transfer gases efficiently. ILD results in abnormal accumulation of inflammatory cells in lung tissue, causing shortness of breath and cough. This impairment is exacerbated by inflammation; if inflammation persists, scarring (fibrosis) may replace the normal lung tissue, eventually resulting in hypoxaemia and respiratory failure¹. Symptoms such as shortness of breath and cough stem from the abnormal accumulation of inflammatory cells in the lung tissue. While ILDs encompass various disorders with different etiologies, they exhibit similarities in imaging, presenting features such as decreased oxygen transfer, lung stiffness, and architectural distortion^{1,2}.

Moreover, ILDs can extend beyond the interstitium to affect other lung structures, including airspaces, peripheral airways, and vessels, along with their respective epithelial and endothelial linings. Common imaging findings include non-infective infiltrates, predominantly in the pulmonary interstitium and alveoli, which may progress to irreversible fibrosis in certain cases¹⁻³. Understanding ILD's complex pathophysiology and manifestations is critical for accurate diagnosis and effective management of these challenging conditions.

Evolution in the terminology and classification of ILD

The terminology and classification of ILDs have evolved significantly over the past decades, driven by advancements in understanding the pathophysiology, imaging, and histopathology of these complex disorders. Initially, ILDs were broadly categorised under terms like "diffuse parenchymal lung diseases," which often led to ambiguous diagnoses and management strategies. Historically, the broad categorisation of ILDs posed challenges due to overlapping pathophysiological and morphological features². The advent of high-resolution

computed tomography (HRCT) provided a more detailed view of lung parenchyma, enabling more precise differentiation between various ILD patterns⁴. Concurrently, molecular biology and genetics advances uncovered specific biomarkers and genetic mutations associated with particular ILD subtypes^{5,6}, facilitating a more nuanced classification system. This shift is exemplified by the refined diagnostic criteria and classifications outlined by organisations such as the American Thoracic Society (ATS) and the European Respiratory Society (ERS), which emphasise a multidisciplinary approach incorporating clinical, radiological, and pathological data⁷⁻¹². The aim was to have a systematic and evidence-based classification of ILD to improve diagnostic accuracy and guide targeted therapeutic interventions. Over time, clinicians, radiologists, pathologists, and researchers have used emerging data to disentangle ILDs arising from different causes, each with distinct prognoses and treatment approaches. Advancements in understanding underlying mechanisms, partly driven by ongoing investigations into treatment options, have refined these classifications^[1,2]. For example, identifying specific phenotypes of pulmonary fibrosis has improved the comprehension of which phenotypes respond to corticosteroids². This progress in understanding disease mechanisms has coincided with investigating diverse treatment avenues for ILD. The primary focus until recently has been on IPF due to its aggressive nature and poor survival time¹³⁻¹⁵. Proper classification, therefore, allows for accurate diagnosis and effective management, improving patient outcomes². Moreover, research to date has shown that epidemiology, prognoses, and survival rates vary by ILD type, and proper classification leads to more accurate recording of patient types, eventually enhancing the quality of research in this field.

Despite rigorous diagnostic guideline criteria, there remain a group of patients who exhibit nonspecific or conflicting clinical or histopathological findings and so cannot be classified within a specific ILD subgroup – so called unclassifiable, these patients exhibit nonspecific or conflicting clinical or histopathological findings. As these patients exhibit a particular disease-behaviour, with not much known about their prognosis and management, they were classified as ‘unclassifiable’ ILD.

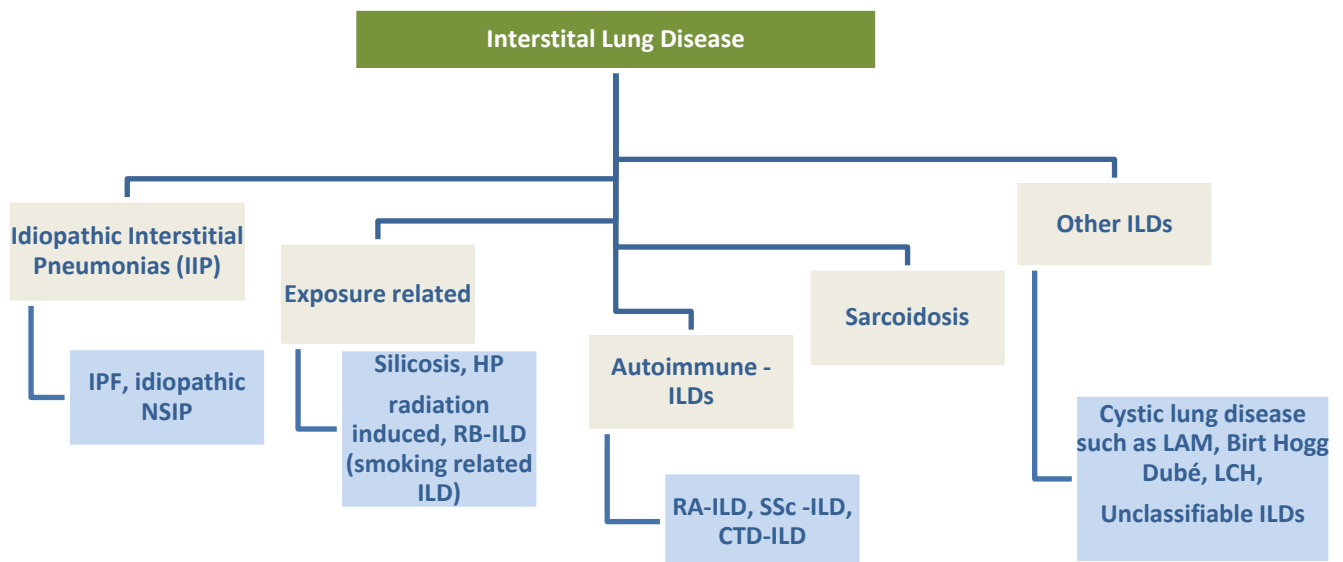
A major challenge in determining the actual ILD burden is the variations in diagnostic criteria and ILD classification over time and across regions, leading to inconsistencies in epidemiological data¹⁶. There is a

pressing need for standardised diagnostic criteria to improve data comparability and reliability. Up-to-date epidemiological data on ILD will inform public health planning, ensuring effective distribution of resources for research, healthcare services, and patient support. By understanding ILD prevalence and impact, policymakers and healthcare providers can allocate resources more efficiently, focusing on areas with the highest need. Accurate epidemiological data underscores the need for developing new diagnostic methods and treatments. Recognising the 'true' burden of ILD attracts research funding and fosters collaborations to discover innovative solutions. This is particularly crucial for diseases like IPF, where early diagnosis and targeted treatments can significantly improve patient outcomes ¹⁷.

In summary, studying the epidemiology of ILD is essential for understanding its true burden, improving diagnostic and treatment strategies, and guiding public health initiatives. These will ultimately contribute to better patient outcomes and quality of life.

For my research, I have grouped 200-300 individual ILD conditions into 5-6 major subtypes according to their aetiology (Figure 1.1).

Figure 1. 2 Aetiological Classification of ILDs



IIP, idiopathic interstitial pneumonia; IPF, idiopathic pulmonary fibrosis; NSIP, nonspecific interstitial pneumonia; RB-ILD, respiratory bronchiolitis interstitial lung disease; HP, hypersensitivity pneumonia; RA-ILD, rheumatoid arthritis interstitial lung disease; SSc-ILD, systemic sclerosis interstitial lung disease; LAM, lymphangioleiomyomatosis.;LCH, Langerhans cell histiocytosis.

1.2. ILD patient care pathway and the key pillars of comprehensive management

ILD patient care begins with the initial presentation, where a patient presents with symptoms such as shortness of breath, chest discomfort, crackling sounds, clubbing of fingers or persistent cough, and diagnosis is confirmed through clinical evaluation, imaging (e.g., high-resolution CT), pulmonary function tests, and now rarely lung biopsy. There are ATS/ERS guidelines on diagnosis of ILD conditions, mainly for IPF [7-12,21]. Historically, the treatment options, disease awareness and diagnostic guidelines for IPF were limited. As the research for this complex group of diseases continued, its awareness increased. Increased awareness and research in this field supported release of evidence-based guidelines for IPF overtime [7-12,21]. The key pillars of comprehensive ILD management include early diagnosis and accurate classification to guide treatment, multidisciplinary care involving various specialists, and a combination of pharmacologic (e.g., antifibrotic drugs, corticosteroids) and non-pharmacologic interventions (e.g., oxygen therapy, pulmonary rehabilitation)[2,28-31, 9,10,25,31,122,146,166-169].

To ensure comprehensive management and appropriate treatment plan, it is essential to have an accurate diagnosis of the ILD condition, for example whether IPF, rheumatoid arthritis ILD, or unclassifiable ILDs. The severity of the disease is assessed through pulmonary function tests (FEV, FEV1/FVC), 6-minute walk tests, patterns on the scans, and presence of comorbidities like pulmonary hypertension, lung cancer. Treatment landscape for ILDs includes antifibrotic drugs for IPF and other fibrosing ILDs ('IPF-like'), immunosuppressants for autoimmune ILDs, oxygen therapy or lung transplantation. Additionally, there are also nonpharmacological therapies to support and provide a holistic patient care. These include pulmonary rehabilitation and palliative care. Disease monitoring and follow-up are crucial for tracking disease progression. Disease monitoring allows for treatment adjustments if needed. But keeping in mind that these treatments are not curative. Initiation of treatment may slow disease progression and manage symptoms to maintain the quality of life for as long as possible. Specially, in advanced stages where palliative care plays a pivotal role in improving one's quality of life. When considering the patient care for this chronic condition it's

important to also consider the patient perspective, which differs at patient level, for example how a patient feels, what's the goal for a patient's care - whether one's priority to manage symptoms. The goal may vary depending on the ILD condition and its severity.

1.3. Epidemiology of ILD

As described in the above section, the ILD condition's pathophysiology varies, and studying this diverse group of chronic conditions is complex but critical due to its significant epidemiological impact, variability in incidence, and profound influence on patient outcomes. Epidemiological studies reveal that ILDs have a global incidence ranging from 1 to 32 per 100,000 person-years. Geographical, demographic, and environmental factors influence this comprehensive range. For example, developed countries with advanced diagnostic capabilities report higher prevalence and incidence rates¹⁶. Kaul et al. reported that the prevalence of ILD ranged from 6.7 to 71.0 per 100,000 persons in the EU, varying among ILD conditions and by sex. The progression and prognosis of ILD vary significantly among its subtypes¹⁵. In the EU and UK, the incidence of ILD for men was reported at 7.22 per 100,000, compared to 4.34 per 100,000 for women, with an increasing trend over the years. Age-standardised death rates (ASDRs) for ILDs also show significant variations. For instance, in the UK, the ASDR for men increased by 1.6% from 2001 to 2017, reaching 5.27 per 100,000¹⁸. IPF, has an untreated historic median survival rate of just three to five years¹³⁻¹⁵, often worse than many cancers¹⁴.

In contrast, other types, like sarcoidosis, have better survival rates but still pose significant health risks¹⁹.

Different ILD types have varying prevalence and mortality rates. Sarcoidosis and IPF are most prevalent in Europe, but IPF has the highest mortality rate in contrast to sarcoidosis^{16,19,20}. These epidemiological insights are vital for healthcare planning and prioritising medical research. Understanding the number of people living with this chronic condition and its survival rates helps prioritise research and healthcare resources for the most affected populations.

1.4. Why IPF?

IPF is characterised by a dismal prognosis, a progressive fibrosing process limited to the lungs, exhibiting a histopathologic and/or radiologic pattern of usual interstitial pneumonia (UIP)²¹. Our understanding of IPF

has evolved significantly over time. However, ongoing research remains crucial due to the disease's complex nature and the evolving diagnostic guidelines from organisations like ATS and ERS^{7-12,21}. Despite extensive efforts, IPF remains idiopathic, poor quality of life²², highlighting the urgency for further exploration. Predominantly afflicting elderly White males, IPF is associated with a grim prognosis, with an untreated historic three to five-year survival rate worse than many cancers^{14,23}. As the most prevalent idiopathic interstitial pneumonia (IIP)^{15,16,20}, IPF imposes a significant public health burden, necessitating research to find new treatments and improve existing ones to reduce mortality rates. Strongman and colleagues reported an incidence of IPF in the UK from 2000 to 2012, ranging from 2.85 to 8.65 per 100,000 patient-years, with a 78% increase in incidence over the study period, with a median survival of three years for incident cases diagnosed during this period with no observed improvement in survival rates¹³. Another study using UK death records reported the age-standardised mortality rates (per 100,000 person-years) in the UK increased from 1.66 in 1966 to 8.29 in 2016; the annual increase in the number of deaths of IPF was approximately 5%²⁴. Historically, treatment options, disease awareness, diagnostic guidelines, and management strategies for IPF were limited. However, significant advancements, including the approval of antifibrotic treatments like pirfenidone and nintedanib, along with the release of evidence-based ATS-ERS IPF guidelines^{7-11,25}, have been made over the past decade. Despite these developments, it remains unclear whether there have been changes in the burden of IPF over time. There is a lack of recent data assessing the impact of these improvements on IPF burden after 2012 in the UK^{13,24}, particularly concerning incidence, prevalence, and survival following the availability of antifibrotic treatments.

The study of establishing IPF disease burden is essential for several reasons: i) the insights gained from epidemiological studies, for example, how IPF is coded in the real world and exploring patient profiles can support breakthroughs in understanding other fibrotic diseases, ii) given the similarities with other ILDs, IPF can be difficult to diagnose. Epidemiological research can support improving diagnostic accuracy, facilitating earlier detection and intervention. iii) understanding the factors influencing disease progression and patient responses to treatment can enhance improved patient outcomes, iv) to-date literature shows IPF has a high mortality rate, with many patients surviving only a few years post-diagnosis. Research aimed at finding and improving treatments is essential for reducing this mortality rate, v) Disease burden (incidence, prevalence,

quality of life) can inform public health policies and strategies aimed at the management of IPF, vi) Despite several advancements in ILD and notably IPF, there remain many unanswered questions remain about IPF. Additional research to address and fill these knowledge gaps will contribute to a comprehensive understanding of the disease, vii) Lastly, quantifying the disease burden will drive research that enhances patient education, support resources, and advocacy efforts, thereby empowering patients, and their families to manage the disease more effectively.

Overall, deriving the actual disease burden is crucial to IPF research, as it can improve patient care for this debilitating disease and promote medical advances in this field.

Diagnosis, guidelines, and management for IPF

As our knowledge of various ILD behaviours, particularly IIPs, expanded, it became clear that IPF constituted a distinct disease category. Diagnosing IPF is a complex process requiring meticulous evaluation and collaboration among a multidisciplinary team involving radiologists, pulmonologists, and pathologists. The 2002 ATS/ERS statement [2000 ATS/ETS guidelines] marked a significant milestone in IPF diagnosis and treatment guidelines ^{7,26}. It delineated four major and minor diagnostic criteria for identifying UIP patterns, emphasising the absence of known causes of lung fibrosis alongside the presence of UIP as pivotal factors. However, recognising UIP on HRCT scans poses challenges due to its nuanced radiologic features, such as reticular opacities and honeycombing, often accompanied by traction bronchiectasis. These criteria were updated in 2011 to provide simplified, evidence-based guidelines, aiming for a more confident diagnosis of IPF ^{8,9}. The revised guidelines integrate HRCT findings with histopathological data to assign diagnostic likelihood, enhancing diagnostic accuracy. Despite these advancements, ongoing research is imperative to refine diagnostic criteria further and optimise patient care.

Further updates to these guidelines have major implications for clinical practice for the diagnosis and, thus, management of IPF patients ¹⁰⁻¹². These guidelines represent a significant advancement in clinical practice, offering a structured framework for IPF diagnosis previously lacking in the field. Based on a comprehensive review of the literature, these guidelines provide a disciplined approach to diagnosing and managing IPF ^{12,20}. Therefore, it is essential to have up-to-date information for this chronic disease. More research is required to

learn more about the conditions and disease burden. This structured approach to IPF diagnosis outlined in guidelines has profound implications for clinical practice, providing clinicians with clear criteria and guidance for accurately diagnosing and managing this challenging disease.

1.5. Rationale and hypothesis

Aetiological factors contributing to ILD are not limited to exposure to environmental hazards, smoking, or autoimmune conditions. While specific ILD subtypes, such as IPF and occupationally induced ILDs, have garnered considerable attention in research and clinical practice, autoimmune ILDs have recently emerged as a focal point of investigation²⁷.

Understanding the burden of ILD is crucial for effective disease management in England. Despite advancements, recent research is needed to clarify ILD's prevalence, incidence, and impact on public health. The evolving characterisation and management of IPF complicate accurate burden assessment, highlighting the need for standardised definitions and methods in epidemiological studies. I plan to use large population-level datasets to understand more about this population. The idea is to understand how these patients perform in the real world. However, keeping in mind the use of large population datasets, there is the problem of the validity of IPF cases found and whether these are 'true' IPF cases. When utilizing large secondary datasets to identify IPF cases, it is crucial to consider potential limitations related to the reliability and validity of the identified cases. One key concern is whether the data sources accurately capture true IPF cases. Large population-based datasets often face challenges such as data quality, consistency, and accuracy. These datasets may be incomplete, outdated, or collected for purposes other than the intended research. Such limitations can introduce errors related to reliability, validity, and misclassification, potentially impacting the accuracy of case identification. Herrett and colleagues^[133] conducted a systematic literature review to summarise findings across range of methods used to validate diagnoses in the CPRD. Of 357 validations investigated, 303 (85%) utilised data from outside the CPRD to validate diagnoses. The median proportion of cases with a confirmed diagnosis was 89% (range 24-100%). Overall, the estimates of the validity were high, which tells us CPRD can be used for large scale studies by adopting a robust study framework to enhance the reliability of my findings and better understand how IPF patients perform in the real-world settings. Another study validated

the recording of COPD in CPRD. These patients with COPD can be accurately identified from UK primary care records using specific diagnostic codes. Requiring spirometry or COPD medications only marginally improved accuracy. The high accuracy applies since the introduction of an incentivised disease register for COPD as part of Quality and Outcomes Framework in 2004 ^[152].

Recent therapeutic advancements, including introducing antifibrotic agents ²⁸⁻³¹, have shown promise in slowing disease progression and improving survival among patients with IPF. However, the hypothesis regarding the survival benefit of these treatments warrants further investigation, as robust evidence supporting this assertion remains elusive. A key question for IPF and other fibrosing ILDs ('IPF-like') is whether antifibrotic treatments have effectively improved patient outcomes including survival. Measuring the true impact of these treatments on survival presents several challenges. Several factors must be considered: time of diagnosis, early versus late diagnosis, presence of other chronic conditions, disease prognosis, access to the medications, adherence to treatment and overall patient health. The time of diagnosis—whether early or late—plays a crucial role in treatment efficacy. Patients diagnosed at an advanced stage may not experience the same benefits as those diagnosed earlier, as the disease may have already progressed to a point where intervention is less effective. Additionally, the presence of other chronic conditions complicates disease management and may interfere with the effectiveness of antifibrotic therapy.

Beyond patient-specific considerations, other challenges also affect treatment outcomes. Access to antifibrotic therapy is not uniform, with many patients unable to receive treatment due to healthcare system limitations, for example according to NICE guideline in the UK until recently only patients with a severe condition ($FEV_1/FVC > 80\%$) were eligible for treatment ^[146]. Moreover, the uptake of antifibrotics has been hindered by the 'wait-and-watch' approach adopted by some physicians. The side effects associated with antifibrotics can discourage both physicians and patients from pursuing or continuing treatment, further complicating their impact on survival.

As mentioned in the above paragraph, until recently, the NICE guidelines restricted antifibrotic eligibility based on lung function test such as FEV/FVC [9,10,25,31,122,146,166-169]. This policy effectively limited access to treatment to only the most severe cases, potentially diminishing the observable benefits of antifibrotics. By the time these patients received treatment, disease progression may have been advanced to observe meaningful survival benefits. Furthermore, patients with multiple comorbidities face an increased ‘pill burden,’ making adherence to a non-curative antifibrotic regimen even more challenging.

Another critical limitation in evaluating the survival benefits of antifibrotics is the lack of sufficient long-term data. The slow uptake of these treatments, coupled with limited follow up time, hampers the ability to draw definitive conclusions about their effectiveness. While clinical trials have demonstrated that antifibrotics slow disease progression [28,29,171], their impact in real-world on survival remains uncertain. More extensive, robust, longitudinal studies are needed to assess the true benefit of these treatments and refine strategies for their optimal use in clinical practice.

In conclusion, while antifibrotics represent an important advancement in the treatment of IPF, their impact on survival is not straightforward. A range of factors related to patient, access and behavior complicates the evaluation of their true efficacy. Until data with longer follow-up is evaluated, the question of whether antifibrotics significantly improve survival remains open to further investigation. However, it is important to note that antifibrotics although not curative may improve patient’s quality of life. It is important to consider the impact of the antifibrotics beyond survival.

A comprehensive understanding of disease trends, including the impact of recent developments on patient outcomes in England, still needs to be discovered.

This research addresses these gaps by consolidating data on various ILDs, especially IPF, to discern trends within the UK population. Through rigorous analysis, I seek to evaluate the hypothesis that recent disease

characterisation and management advancements have translated into tangible improvements in patient outcomes. By elucidating the current landscape of ILD burden and outcomes in the UK, this research endeavours to inform future disease management and resource allocation strategies.

1.5.1. Specific aims and objectives

The overarching aim of this thesis was to contribute to the understanding and evidence base to better understand the ILD burden by addressing the gap in current knowledge of this disease epidemiology. In particular, I shed further light on the most aggressive subtype of ILD, IPF, and its burden (prevalence, incidence, and mortality) in England. Taking in to account improved disease diagnosis and the availability of treatment options, it's important to determine if there has been a change in IPF epidemiology over the past decade. I envisage that the findings presented here will help guide efforts toward holistic care for these patients.

Within the above overall study aim, I identified several current knowledge gaps and devised a series of individual studies to address them. My specific objectives in conducting this research project were thus as follows:

Aim 1: To estimate the incidence and prevalence of overall ILD and its subtypes.

The aim was to appraise the published literature on the incidence and prevalence of various ILDs over the study period (2015 to 2021) and provide a comprehensive understanding of the global incidence and prevalence of overall ILD and its subtypes. I sought to identify areas with more data, whether the available data is robust, and whether certain subtypes of conditions have more data available. This allowed me to appreciate the burden of ILD conditions and comprehend the areas where there remains a need for further studies.

Aim 2: To assess the reliability of recording the diagnostic entity IPF in the UK primary care database.

Here, I aim to validate the most effective method for identifying patients with IPF. I explored whether relying solely on a primary care database is adequate for identifying IPF cases by comparing the use of different sets

of SNOMED codes, including specific codes for IPF versus generic codes such as fibrosis. Additionally, this part of the research investigated whether a list of clinical diagnostic codes alone is sufficient for identifying IPF cases in the UK primary care database or if a more complex algorithm incorporating additional data such as radiology or hospital admissions is necessary for improved accuracy. The research also aimed to observe potential changes in general practitioners' coding practices over time.

Aim 3: To establish the incidence and prevalence of IPF in England from 2008 to 2018 and investigate IPF-related mortality.

There is a lack of recent data estimating the epidemiology of IPF in England. Several studies in the past have employed routinely collected healthcare data to describe IPF trends in the UK. However, these studies predated the introduction of antifibrotics and updates to the diagnostic guidelines. Using the clinical codes validated in the above aim (Aim2), I estimated the incidence and prevalence for algorithms that showed a high positive predicted value. I used the Office of National Statistics (ONS) data, which has the registered deaths in the UK, to compute IPF-specific mortality for the same period. How IPF is defined has changed over time, owing to updates in clinical guidelines - in an attempt to standardise the diagnosis. Therefore, in this aim, my focus was to update the epidemiology of IPF in England and post key updates in this field with the hypothesis that I may identify a rise in IPF burden over time, given the increased awareness, which may have resulted in early diagnosis of the disease.

Aim 4: To investigate changes in survival and causes of death in people diagnosed with IPF over the last 10 years in England.

As far as we know, there is no recent data to verify whether there has been an improvement in the survival and mortality of patients with IPF since the availability of the two antifibrotic treatments pirfenidone and nintedanib in the UK. Whether developments in the field of IPF over the past decade have translated to better patient management, improved disease outcomes, and increased quality of life is still unknown. Using the clinical codes validated in Aim 2, I estimated survival and all-cause mortality in people diagnosed with IPF in the primary and secondary databases. Here, I aimed to understand the changes in the causes of death and

survival over time in people with IPF to understand whether I saw any influence of the antifibrotic treatments in patients living with this chronic condition.

1.6. Structure and organisation of this thesis

The thesis is organised as follows (Figure 1.2). This opening chapter (**Chapter 1**) provides an introductory overview of ILD, its high-level subtypes, disease epidemiology, disease diagnosis and management, and its evolution over the past decade. It highlights the critical updates in the ILD and, more specifically, in the IPF. Given the focus on IPF, a short description of IPF is also included in this section. The findings of a systematic review appraising the published literature are presented in **Chapter 2**. This systematic review was conducted to appreciate the burden of ILD overall and by its subtypes and thus inform the subsequent analyses. The main results of the systematic review have since been published and are presented in Chapter 2.

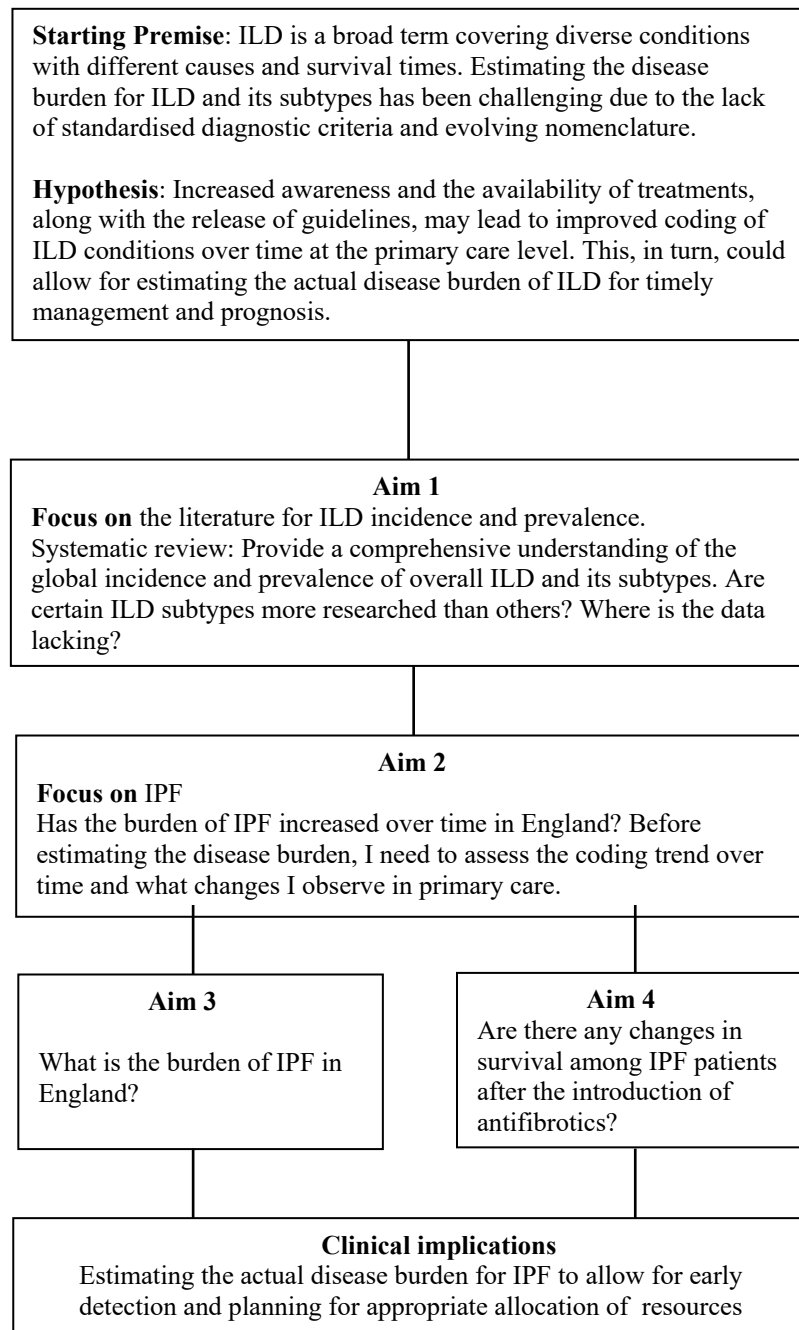
Chapter 3 provides a comprehensive description of the healthcare system in the UK and how routine healthcare records are collected. Primary and secondary health-related data are used as the source of patient data for the data analyses in my research. These include electronic health records (EHR) from primary care and administrative data from secondary care, and both datasets are available from NHS Digital under the Imperial College London license. In addition, mortality statistics use registered deaths in the UK compiled by the ONS. This chapter also provides a more general discussion of some of the inherent limitations of these datasets for epidemiological research.

Chapter 4 describes the IPF validation study, which validated the primary and secondary health care records in England to find IPF cases. The findings of this validation study were published, and they formed the foundation for Chapters 5 and 6, which determined the burden of IPF in the UK.

Chapters 5 and 6 detail my research project's two primary data analyses. Each chapter presents an update on the disease burden in England, incorporating the findings from Aims 1 and 2. These chapters describe in depth the study design and methods used to derive the findings presented here. In the discussion section, I explain the results in context to other work and make suggestions on how analysis could be further advanced, building on the existing work.

Chapter 7 is a general discussion focusing on the implications of the collective findings on the disease burden of IPF in England. This work concludes with a section on suggestions for future research.

Figure 1. 3 Schematic overview of the research undertaken to explain how the various components contribute to the project's overall aim.



Chapter 2

Incidence and prevalence of interstitial lung diseases worldwide: A systematic literature review

Overview

ILD encompasses a broad spectrum of over 150 different conditions classified as ILD. The diverse nature of these conditions has posed challenges in accurately determining ILD's overall incidence and prevalence and understanding specific diagnostic subgroups. Research has predominantly focused on certain conditions, such as IPF, while less common or difficult-to-diagnose subgroups, such as environmental or occupational exposures, have received limited attention. Furthermore, the lack of global and country-level standardisation in data collection methods for non-IPF ILDs has further complicated the assessment of its actual incidence. Although some evidence suggests a potential increase in the incidence of ILD, particularly for non-IPF ILDs, in the absence of standardised data collection tools, gaps in published literature have hindered the confirmation of this hypothesis. This chapter aims to provide a comprehensive review of published literature on a global landscape to gain deeper insights into ILD incidence and prevalence, which is crucial in determining if there is indeed a rise in ILD rates. The varying number of studies published on different ILDs underscores the discrepancies in research focus; IPF, for example, is extensively studied, while other conditions such as autoimmune-related ILDs have until fairly recently been less of a research focal point. The method and results of this chapter are published - *Incidence and prevalence of interstitial lung diseases worldwide: A systematic literature review*: <https://doi.org/10.1136/bmjresp-2022-001291>

2.1. Introduction

Chapter 1 outlined that various ILDs vary significantly in their aetiology, severity, prognosis, and clinical pathways, adding complexity to epidemiological studies. The challenges in studying ILD epidemiology are manifold, as discussed previously (relative rarity of individual ILDs, sampling biases, inconsistent methods for case ascertainment, lack of standard case definitions, the necessity of invasive diagnostic procedures, and evolving nomenclature and diagnostic criteria). Despite these obstacles, the field of ILD research is advancing, driven by greater awareness through the increased use of thoracic CT ^{5,6}, diagnostic guidelines based on evidence ⁷⁻¹² and the development of novel therapeutic agents, such as Pirfenidone and Nintedanib ²⁸⁻³¹.

Data from the Global Burden of Disease (GBD) study showed an increasing trend for ILD from 5.6 per 100,000 persons in 2007 to 6.6 per 100,000 in 2017, an increase of around 18%³². What we do not know is the reason for this rise in cases. When examined by subgroups or at a condition level, IPF has been extensively studied, with prevalence ranging significantly across age groups and geographies^{16,32}. However, for non-IPF ILDs, there is a notable gap in consistent epidemiological data. This gap is exacerbated by limited large-scale, population-based studies, making it challenging to estimate the burden of non-IPF ILDs accurately. Several factors, for example, environmental factors (e.g., pollution), occupational exposures (e.g., mining), and healthcare discrepancies (e.g., availability of resources to conduct a thorough examination), contribute to the broad spectrum of variation for non-IPF ILD estimates¹⁶.

In the UK, the incidence of IPF has previously been estimated at 2.85 to 8.65 per 100,000 person-years¹³. Similarly, Olson *et al.* reported overall ILD prevalence ranging from 6.3 – 76 per 100,000 (extracted from four studies) and for IPF from 1.3 – 63 per 100,000 persons in the USA³³ and Europe³⁴; what is imprecise is whether this is an 'actual' disparity in the numbers across regions or a result of the inconsistencies in how IPF is ascertained and recorded. It is also ambiguous whether the rise in ILD, particularly IPF, indicates a genuine increase or is merely related to an ageing population, given that the increase in awareness has translated to improved detection methods and, thus, changes in coding practices over time. Therefore, the increase may be due to improved case ascertainment of IPF rather than an actual increase in the disease cases. In this chapter, I seek to provide a detailed understanding of ILD epidemiology, highlighting consistent data and thus identifying areas that require further research.

My preliminary search shows a lack of consistent epidemiological information, which remains a significant challenge for ILD researchers. This chapter aims to review systematically and meta-analyse studies reporting key epidemiological outcomes—prevalence, incidence, and all-cause mortality—for ILD (by subtypes presented in Chapter 1, Figure 1.1].

2.2. Methods

2.2.1.1. Scoping Review

To define the scope of this review, I conducted a preliminary search to assess the existing literature and inform the direction of this study—this scoping exercise aimed to guide the planning and focus of the review. Initially, I considered all ILDs and their subtypes. In the past decade, diagnostic criteria for ILDs, particularly for IPF, have been established to standardise IPF diagnostic classifications and, thus, its recording. Therefore, I chose to focus on studies published within the last seven years (2015-2021) to ensure that the findings reflect the adoption of these updated practices. I excluded drug induced ILDs due to a recent SLR conducted summarizing this information³⁵⁻³⁸.

While the original goal was to synthesise data on all-cause mortality in ILD patients or mortality specifically associated with ILDs, the sheer volume of papers identified (over 15,000) reporting the incidence and prevalence of ILDs and their subgroups led me to narrow the scope of the review. Consequently, this research is focused solely on studies reporting the incidence and/or prevalence of ILDs.

2.2.2. Study Registration

As the first step, I drafted a protocol for this systematic review. The protocol followed the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA) guidelines³⁹ and was registered with the International Prospective Register of Systematic Reviews, PROSPERO (CRD42020203035).

2.2.3. Eligibility Criteria

The Population, Intervention or Exposure, Comparator, Outcomes, and Study Design (PICOS) framework guided the eligibility criteria (for example, types of studies and population criteria) and the data extraction process. In this systematic review, there was no comparator arm. The PICOS criteria used here included:

- **Population:** descriptive data relating to the study population, including the source population from which it is drawn (hospitalised patients, patients with predetermined comorbidities, patients undergoing a lung transplant, and/or patients with specific work history or exposure). For example, the study population may be limited to coal mine workers for the incidence of pneumoconiosis. Additional population characteristics drawn were age, sex distribution, population size, data source, collection method, and inclusion or exclusion criteria of the study.
- **Intervention or Exposure:** The primary exposure of interest was specific or unspecified ILD, and diagnosis of ILD was confirmed depending on the type of study; this can include confirmation by HRCT/biopsy or diagnosis codes or a combination of clinical findings to support ILD diagnosis. ILDs were classified into a specific or unspecified category depending on the type of exposure.
- **Outcomes:** The primary outcomes extracted were i) prevalence, defined as the number of new and existing cases in the study population during a prespecified period, and ii) incidence, the number of new cases in the study population. These outcomes were reported in various units, such as percentages, person-years, and rates.
- **Study design:** The study setting, design, period, aims, and objectives were extracted.

2.2.3.1.Types of Studies

Abiding by the PICOS framework, this review included all observational studies that employ either a case-control, cohort, or cross-sectional study design, excluding randomised controlled trials, case reports, reviews, and conference abstracts. Additional exclusions were -animal studies and participating patients under 18. Studies published within the past 7 years at the time of this review conducted (2015-2021) from any world region and reported in English were eligible⁴⁰. For additional studies, references were searched manually.

2.2.3.2.Types of Participants

Patients diagnosed with a specific subtype of ILD or broadly labelled as having ILD were included. These patients can arise from a general population (i.e. identified in claims data or electronic health records), a hospitalised setting and/or disease registry (IPF registry), and/or an occupational setting (coal workers, mine workers). For some individual ILDs, the incidence was analysed in patients diagnosed with a relevant pre-existing comorbidity; for example, I have considered the proportion of ILD in patients diagnosed with an existing autoimmune disorder (e.g. rheumatoid arthritis)

2.2.3.3.Types of Exposure

For this chapter, the exposure of interest is ILD. Thus, studies conducted in patients diagnosed with a specific ILD or given a diagnosis of "ILD" that is not otherwise classified by subtype were included. For studies drawn from real-world data such as claims and electronic health records, evidence of ILD, that is, diagnosis codes for overall ILD or any of the subgroups, was considered a confirmed diagnosis. However, for studies conducted at centres using patient charts or active data collection, a confirmed diagnosis included a clinical assessment of ILD with a supporting constellation of typical findings such as lung biopsy, multidisciplinary decision, or radiology. A complete list of ILD diagnoses is provided in Table 2.1.

Table 2. 3 List of ILD subgroups

ILD Subgroups	Individual ILDs
Occupational or exposure related ILD	Asbestosis or Silicosis
	Pneumoconiosis
	Respiratory bronchiolitis (RB-ILD) and other smoking related ILD
	Hypersensitivity pneumonitis (HP)
IIP	Idiopathic pulmonary fibrosis (IPF)
	Idiopathic Nonspecific interstitial pneumonia (NSIP)
	Desquamative interstitial pneumonia (DIP)
	Cryptogenic Organising pneumonia (COP)
Autoimmune ILD	Rheumatoid arthritis associated ILD (RA-ILD)
	Systemic sclerosis associated ILD (SSc-ILD)
	Sjogren's syndrome associated ILD (SS ILD)
	Inflammatory Myositis associated ILDs and antisynthetase syndromes
	Other autoimmune ILDs (example connective tissue disorder)
Cystic lung disease	Birt Hogg-Dube (BHD)
	Lymphangioleiomyomatosis (LAM)
	Pulmonary Langerhans cell histiocytosis (PLCH)
	Lymphoid interstitial pneumonia (LIP)

Other ILDs	Obliterative bronchiolitis
Pulmonary sarcoidosis	Pulmonary sarcoidosis
Unclassifiable ILD	Unclassifiable ILD

2.2.3.4.Outcomes

Studies that reported the prevalence and/or incidence of ILD were included. Studies (not limited to) exploring risk factors for ILD and/or the effectiveness of various treatments in ILD patients or those that describe the clinical characteristics of ILD patients were excluded.

2.2.4. Information Sources and Search Strategy

I searched two databases, MEDLINE and EMBASE, through the OVID interface for potentially relevant articles. The searches were conducted using predefined search strategies. PROSPERO was monitored periodically for ongoing or completed systematic reviews about any ILD. Additionally, a manual search was conducted for the reference lists of all included studies to check for other relevant articles. With the librarian's support, the search criteria for each ILD subgroup were developed (Figure 2.1.) This figure contains mortality, initially intended as part of this review. However, as mentioned, the papers reporting mortality were excluded due to a high count. Table 2.2 is an example of a medical subject heading and text words used to identify idiopathic pulmonary fibrosis. Relevant concepts combined using the Boolean operator 'AND' to search for potentially disease-relevant terms from the literature were included.

Table 2. 4 Literature search terms for IPF

Concept	Medical subject heading and text words
IPF	*idiopathic pulmonary fibrosis.mp
	*lung fibrosis.mp
	IPF.mp
	*fibrosing alveolitis.mp
	idiopathic pulmonary fibrosis*.mp
Epidemiology (incidence and/ or prevalence)	epidemio*.mp
	*epidemiology/
	*public health/
	inciden*.mp.
	prevalen*.mp.

	prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
--	---

Note: “.mp” describes a text word. Text followed by “/” describes medical subject headings. “*” allows truncated search words.

Figure 2. 2 Search Strategies for EMBASE

Overall ILD or cryptogenic fibrosing alveolitis search term epidemio*.mp. *epidemiology/
public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
*mortality/ or *mortality risk/ or *mortality rate/
*prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
5 or 6 or 7 or 8 4 or 9
ILD*.mp.
interstitial lung disease/ or interstitial lung diseas.mp.
11 or 12 10 and 13 limit 14 to human limit 15 to english language
limit 16 to yr="2009 - 2019"

Asbestosis or Silicosis search terms epidemio*.mp. *epidemiology/
public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
*mortality/ or *mortality risk/ or *mortality rate/
*prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
5 or 6 or 7 or 8 4 or 9 asbestosis.mp. or *asbestosis/ silicosis.mp. or *silicosis/ *asbestosis/ or *silicosis/
11 or 12 or 13 10 and 14 limit 15 to human limit 16 to english language limit 17 to yr="2009 - 2019"

Pneumoconiosis search term epidemio*.mp. *epidemiology/
public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
*mortality/ or *mortality risk/ or *mortality rate/
*prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
5 or 6 or 7 or 8 4 or 9
ILD*.mp.
interstitial lung disease/ or interstitial lung diseas.mp.
11 or 12 10 and 13
pneumoconiosis.mp. or exp pneumoconiosis/
10 and 15 limit 16 to human limit 17 to english language
limit 18 to yr="2009 - 2019"

Respiratory bronchiolitis search term epidemio*.mp. *epidemiology/
public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
*mortality/ or *mortality risk/ or *mortality rate/
*prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
5 or 6 or 7 or 8 4 or 9
ILD*.mp.
interstitial lung disease/ or interstitial lung diseas.mp.
11 or 12 10 and 13 exp bronchiolitis/ or Respiratory bronchiolitis.mp.
10 and 15 limit 16 to human limit 17 to english language limit 18 to yr="2009 - 2019"

Hypersensitivity pneumonitis or Extrinsic alveolar alveolitis search terms epidemio*.mp. *epidemiology/
public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
*mortality/ or *mortality risk/ or *mortality rate/
*prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
5 or 6 or 7 or 8
hypersensitivity pneumonitis.mp. or exp allergic pneumonitis/
*hypersensitivity pneumonitis/ or *extrinsic alveolar alveolitis/ or *EAA/ or *HP/
10 or 11
4 or 9 12 and 13 limit 14 to human limit 15 to english language limit 16 to yr="2009 - 2019"

Cystic lung disease epidemio*.mp. *epidemiology/
public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
*mortality/ or *mortality risk/ or *mortality rate/
*prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
5 or 6 or 7 or 8 4 or 9 cystic lung diseas*.mp.
cystic lung diseas*.mp. or exp lymphangioleiomyomatosis/ or exp Birt Hogg Dube syndrome/
*birt hogg dube/ or *LAM/ or *PLCH/ or *LIP/ or *lymphangioleiomyomatosis/ or *pulmonary Langerhans cell histiocytosis/ or
*lymphoid interstitial pneumonia/ or *Birt-Hogg-Dube/
11 or 12 or 13 10 and 14 limit 15 to human limit 16 to english language limit 17 to yr="2009 - 2019"

Sarcoidosis search term epidemio*.mp. *epidemiology/
 public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
 *mortality/ or *mortality risk/ or *mortality rate/
 *prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
 5 or 6 or 7 or 8
 4 or 9
 *lung sarcoidosis/ or *sarcoidosis/ or sarcoidosis*.mp.
 10 and 11 limit 12 to human limit 13 to english language limit 14 to yr="2009 - 2019"

Idiopathic interstitial lung disease epidemio*.mp. *epidemiology/
 public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
 *mortality/ or *mortality risk/ or *mortality rate/
 *prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/ 5 or 6 or 7 or 8
 4 or 9 idiopathic interstitial pneumon*.mp.
 *idiopathic interstitial pneumonia/ or *IIP/ or *idiopathic interstitial/
 idiopathic interstitial pneumonia/ or idiopathic interstitial pneumon.mp. or *Idiopathic Interstitial Pneumonias/ 11 or 12 or 13 10
 and 14 limit 15 to human limit 16 to english language limit 17 to yr="2009 - 2019"

Idiopathic pulmonary fibrosis epidemio*.mp. *epidemiology/
 public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
 *mortality/ or *mortality risk/ or *mortality rate/
 *prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
 5 or 6 or 7 or 8
 4 or 9
 *idiopathic pulmonary fibrosis/
 *lung fibrosis/ or *fibrosing alveolitis/ or idiopathic pulmonary fibrosis*.mp.
 11 or 12 10 and 13 limit 14 to human limit 15 to english language limit 16 to yr="2009 - 2019"

Autoimmune ILDs epidemio*.mp. *epidemiology/
 public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
 *mortality/ or *mortality risk/ or *mortality rate/
 *prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
 5 or 6 or 7 or 8 4 or 9
 ILD*.mp.
 interstitial lung disease/ or interstitial lung diseases.mp.
 11 or 12
 10 and 13 exp autoimmune disease/ or autoimmune*.mp.
 *autoimmune/ or *RA/ or *Rheumatoid arthritis/ or *rheumatoid arthritis/ or *scleroderma/ or *Scleroderma/ or *systemic sclerosis/
 or *Systemic Sclerosis/ 15 or 16 10 and 13 and 17 limit 18 to human limit 19 to English language limit 20 to yr="2009 - 2019"

Other ILDs epidemio*.mp. *epidemiology/
 public health/ 1 or 2 or 3 inciden.mp. prevalen*.mp.
 *mortality/ or *mortality risk/ or *mortality rate/
 *prevalence/ or epidemiological data/ or epidemiology/ or statistical concepts/ or statistical parameters/
 5 or 6 or 7 or 8 4 or 9
 ILD*.mp.
 interstitial lung disease/ or interstitial lung diseases.mp.
 11 or 12 cryptogenic organising pneumonia.mp. exp bronchiolitis obliterans organising pneumonia/ or exp cryptogenic organising
 pneumonia/ or cryptogenic organising pneumonia*.mp. exp bronchiolitis obliterans organising pneumonia/ or exp bronchiolitis
 obliterans/ or bronchiolitis obliterans*.mp.
 exp antidepressant agent/ or antidepressive*.mp.
 exp hyperthermic intraperitoneal chemotherapy/ or exp maintenance chemotherapy/ or exp chemotherapy/ or exp induction
 chemotherapy/ or exp cancer chemotherapy/ or exp adjuvant chemotherapy/ or Chemotherapy*.mp. or neoadjuvant chemotherapy/ or
 exp combination chemotherapy/ *Amiodarone/ or *Fluoxetine/ or *Nitrofurantoin/ or *amiodarone/ or *fluoxetine/ or *nitrofurantoin/
 14 or 15 or 16 or 17 or 18 or 19
 10 and 13 and 20 limit 21 to human limit 22 to english language limit 23 to yr="2009 - 2019"

2.2.5. Study records

Data management

EndNote (V.X8) was used to upload and maintain the literature search results. Duplicate results were removed.

2.2.6. Selection process

The first step of the selection process was an independent screening of titles and abstracts of all records identified in the databases. The primary reviewer reviewed all records; the second reviewer independently screened 10%. Both reviewers checked the titles and abstracts against the predefined eligibility criteria. Disputed records between the two reviewers were carried forward for the full-text screening. As part of the full-text screening, I obtained complete reports of the potential articles shortlisted in the screening stage, including any online supplementary material with insufficient information in the main article, to determine the fulfilment of the inclusion criteria. The primary reviewer reviewed all articles shortlisted for full-text screening, and the secondary reviewer independently screened 50% of all articles shortlisted for full-text screening. In case of discrepancy, it was resolved through a discussion with a third reviewer. A comprehensive record of reasons for excluding articles during the full-text screening was maintained.

2.2.6.1.Data items

The results were reported as unadjusted and adjusted estimates, depending on how the authors presented them. If any adjustment was applied, this was noted in the data extraction grid.

2.2.7. Data extraction

After the full-text screening, the primary reviewer extracted data items from the included publications using a predefined data extraction form. The second reviewer independently extracted data from 10% of these included articles using the same data extraction form. Similar to screening, any discrepancy between the primary and secondary reviewers was resolved by consulting a third reviewer. Data items were extracted using an Excel extraction tool, which included all information listed in the above section (items of interest will likely include author, study ID, study population, setting of the study population/recruitment country, type of ILD, outcome, study design, sample size, study duration). This extraction grid was tested with the first ten included studies and modified as needed.

2.2.8. Assessment of risk bias (individual studies)

For this review, previously devised tools for observational studies, namely the Newcastle-Ottawa scale^{41,42} and ROBINS-I⁴³, a more recent tool devised specially for non-randomised studies of interventions, were considered. Considering the benefits and limitations of both these tools and using them as a reference, I developed a tool best suited for this review. Two independent reviewers separately utilised the developed tool to assess the methodological quality of individual studies. This tool considered the bias stemming from selecting participants, measuring variables (exposure, outcome), and controlling confounders such as quality of reporting and total sample size. Six domains for studies reporting prevalence and eight domains for studies reporting incidence were assessed. Each bias item was rated from ‘moderate to high risk of bias’, ‘unclear risk of bias, or low risk of bias’. Discrepancies were resolved through discussion with a third reviewer.

2.2.9. Data synthesis

The degree of heterogeneity across included studies was explored as the first step towards summarising study results. Heterogeneity can be due to clinical or statistical patient differences, study designs or outcomes. Heterogeneity was assessed using the I^2 test to determine which quantitative or narrative data synthesis approaches (quantitative or narrative) should be applied. Meta-analysis would be considered when the I^2 test is acceptable (I^2 statistic $< 50\%$). However, in this review, as heterogeneity was substantial ($>50\%$), I assessed the possibility of conducting analyses by subtypes of ILDs. These subtype analyses could be adjusted for various reasons of heterogeneity, such as methodological quality and disease definition. However, the I^2 statistic after subgroup analysis was greater than 50%, hence unacceptable. The studies included in this chapter have a very high statistical level of heterogeneity ($I^2 > 75\%$) or high heterogeneity after subgroup analysis. Therefore, I chose to perform a narrative synthesis of the evidence.

2.2.9.1. Evidence Synthesis

As mentioned in the above paragraph, I proceeded with data synthesis across ILD subtypes of interest due to high heterogeneity. I used the aetiological classification described in Table 2.1 to synthesise my results.

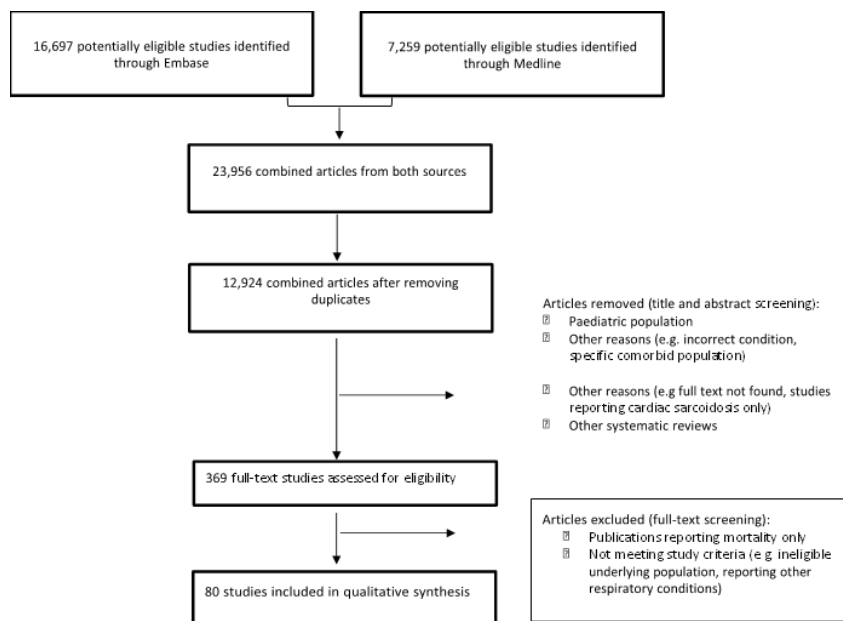
As anticipated, there was a lot of published evidence for IPF. In the literature, many authors have adopted a 'broad' and 'narrow' classification to report IPF results. For example, Raghu et al., across papers, defined patients with an ICD-9 code 516.3 as a broadly defined case of IPF, and those who had this diagnosis code alongside a claim for a surgical lung biopsy, transbronchial lung biopsy, or CT thorax as a narrowly defined case³³. I used a similar approach to present IPF results. Depending on how the authors defined and reported IPF using the published definitions. In literature, categories such as 'general', 'broad', or 'narrow' are used to report incidence and prevalence estimates for IPF. This theme was observed for IPF only. Most IPF studies were drawn from the general population. In contrast, non-IPF ILD studies were conducted in single centres or patients with an underlying condition of interest, with fewer common subgroups drawn from large-scale data or the general population. Therefore, in this chapter, I have reported the results separately for two significant subgroups: IPF and non-IPF ILDs.

2.3. Results

2.3.1. Study characteristics

Overall, 23,956 studies were identified from EMBASE (n=16,697) and MEDLINE (n=7,259) following the electronic systematic search. A total of 12,924 article titles and abstracts were screened after excluding duplicate articles, of which 369 articles were selected for the full-text screening. Of 369 articles screened, 80 articles met the inclusion criteria (Figure 2. 32). A total of 289 articles were excluded at full-text screening for reporting only mortality, other outcomes reported (not incidence and/or prevalence), and for the underlying study population not meeting the study criteria.

Figure 2. 3 PRISMA Flow Diagram to Demonstrate Studies Identified, Screened, and Included



Although 80 unique publications were included, some papers reported on more than one ILD and so the total count of reported estimates is 88. Half of the included publications explored autoimmune-related ILDs (n=44/88) [Table 2.3]. At an individual condition level, the top 3 investigated conditions were RA-ILD (n=14, 15.9%), SSc ILD (n=13, 14.7%), Sjogren's and IPF (n=11, 12.5%) [Table 2. 5]. When examining ILD publications over the past 7 years, I found that the number of studies investigating autoimmune conditions and their association with ILD has vastly increased (from 3 papers in 2015 to 19 papers published in 2021). This sharp increase observed in 2021 Figure 2. 4) is mainly driven by increased interest in autoimmune ILDs. Interestingly, for IPF, the publication trend remained unchanged for the same period. Unsurprisingly, there was a marked decline in the number of published studies in 2020 due to the impact of COVID-19. A paper may have reported an estimate for more than one ILD; I have included the paper for that subtype for each type of ILD reported.

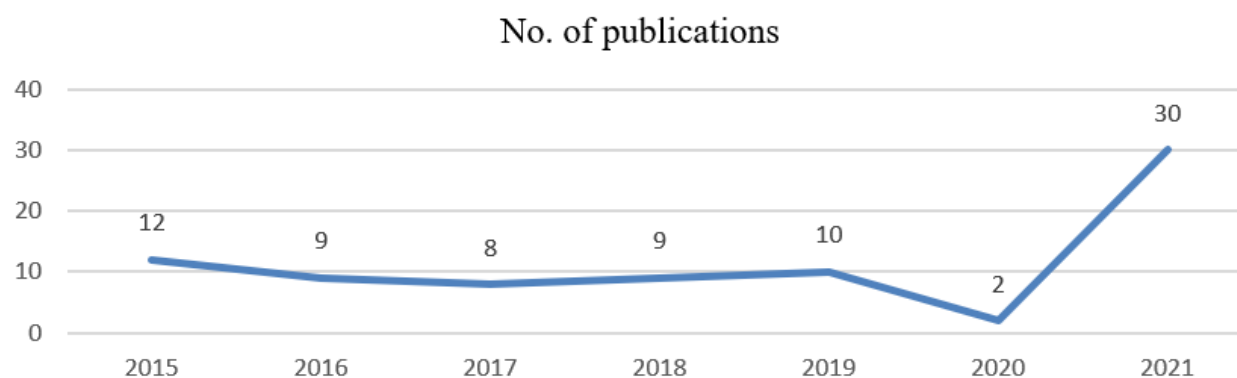
Table 2. 5 Distribution of ILD conditions, sorted by distribution of publications.

ILD subgroup	Individual ILD	Frequency of ILD papers	Distribution of papers
Autoimmune ILD	RA ILD	14	15.9%

Autoimmune ILD	SSc ILD	13	14.8%
Autoimmune ILD	Sjogren's ILD	11	12.5%
IPF	IPF	11	12.5%
Occupational ILD	Silicosis	9	10.2%
Occupational ILD	Asbestosis	5	5.7%
Autoimmune ILD	Myositis ILD	4	4.5%
Occupational ILD	Pneumoconiosis	3	3.4%
Others	Radiation-induced ILD	3	3.4%
Sarcoid	Pulmonary sarcoid	3	3.4%
HP	HP	2	2.3%
IIP	IIP	2	2.3%
PF ILD	PF ILD	2	2.3%
Autoimmune ILD	MCTD ILD	1	1.1%
Autoimmune ILD	CTD ILD	1	1.1%
Occupational ILD	Other occupied	1	1.1%
Others	Unclassifiable ILD	1	1.1%
Others	PLCH	1	1.1%
Others	LAM	1	1.1%

RA ILD, rheumatoid arthritis; SSc, systemic sclerosis; MCTD, mixed connective tissue disorder; CTD, connective tissue disorder; HP, hypersensitivity pneumonitis; IIP, idiopathic interstitial pneumonia; IPF, idiopathic pulmonary fibrosis; PLCH, pulmonary Langerhans cell histiocytosis; LAM, lymphangioleiomyomatosis; PF, progressive fibrosing

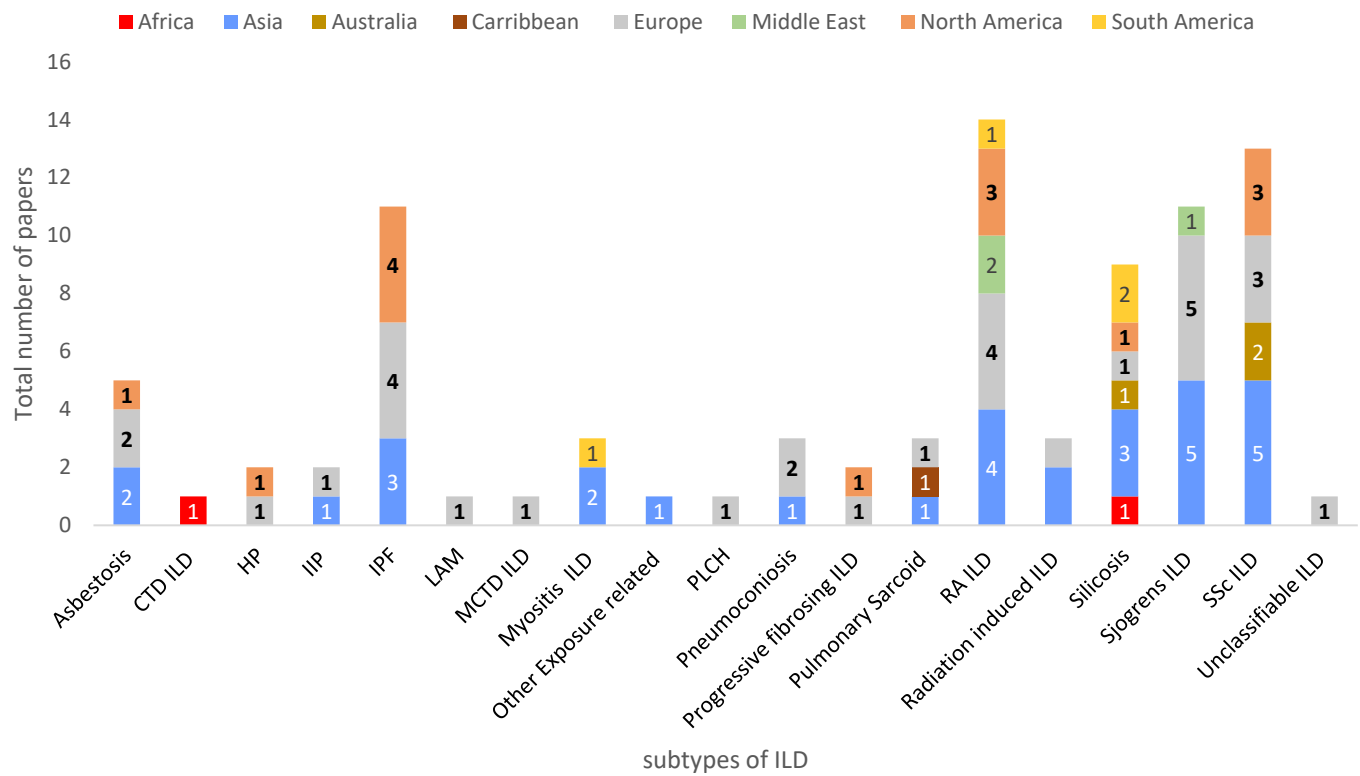
Figure 2. 4 Number of publications for ILD for the study period, 2015 to 2021



Geographically, ILD publications represented all prominent world regions but were mainly from Asia (n = 30, 34.1%) and Europe (n=23, 26.1%). Figure 2.4 shows the geographic distribution of the 80 unique included studies. The distribution also varied by ILD subgroups. For example, environmental and occupational ILDs

were dominant in Asia, whereas autoimmune ILDs were primarily reported in Europe. Table 2. 6 shows each ILD subgroup's top-researched region and country.

Figure 2. 5 Geographical distribution of publications included.



RA ILD, rheumatoid arthritis; SSc, systemic sclerosis; MCTD, mixed connective tissue disorder; CTD, connective tissue disorder; HP, hypersensitivity pneumonitis; IIP, idiopathic interstitial pneumonia; IPF, idiopathic pulmonary fibrosis; PLCH, pulmonary Langerhans cell histiocytosis; LAM, lymphangioleiomyomatosis; PF, progressive fibrosing

Table 2. 6 Top researched region and country for each of the ILD subgroups

ILD subgroups	Dominant region
Occupational or exposure-related ILD	Asia
Autoimmune ILDs	Europe and Asia
Idiopathic interstitial pneumonia	Europe and Asia
Cystic lung disease	Europe*
Pulmonary sarcoidosis	Asia*, Europe and South America
Unclassifiable ILD	North America*

ILD, interstitial lung disease; *Only 1 publication included

2.3.2. Studies Reporting Incidence

The reporting of ILD incidence demonstrated significant variation for several reasons, one being differences in the units used for reporting estimates. Some authors provided incidence rates per 100,000 person-years, while others reported incidence as a proportion. Other discrepancies can largely be attributed to the diverse data sources utilised. As explained in the methods section, different sources and case definitions were employed to establish the incidence of IPF and non-IPF ILDs. This chapter reports the incidence rates separately for IPF and non-IPF ILDs. Most IPF studies drew upon extensive databases such as electronic health records or claims, whereas most non-IPF studies drew upon single or multiple centres.

1.3.2.1 Incidence of IPF

IPF incidence across various regions ranged from 2.4 to 311 per 100,000 person-years and 2.2 to 183.3 per 100,00 persons. This review included eight publications reporting IPF incidence, of which two publications were from North America (1 in Canada and 1 in USA)^{33,44}, three publications were from Europe (1 in UK and 1 in Italy)^{13,45} and three publications were from Asia (South Korea)⁴⁶⁻⁴⁸. The reported incidence in case definitions included various terminologies, such as primary, general, broad, and narrow. Generally, the primary, general, and broad definitions across different studies utilised diagnostic codes associated with IPF. In contrast, the narrow definition specifically focuses on IPF-specific codes.

North America: In the USA, the incidence of IPF was reported using primary, broad, and narrow case definitions, differing mainly in the list of clinical codes used to identify IPF patients. The incidence rates

varied as follows: using the primary definition, the incidence was 5.8 per 100,000 person-years; using a broad list of codes, it was 3.6 to 5.1 per 100,000 person-years; using a narrow list of codes, it was 2.4-2.9 per 100,000 person-years between 2005 and 2010³³. In Canada, the incidence ranged from 9 per 100,000 persons using a conservative list of codes to 18.7 per 100,000 persons using a broader list of codes⁴⁴. Kaul et al., based on veterans' data, reported a much higher incidence: 331 per 100,000 person-years for the broad definition and 210 per 100,000 person-years for the narrow definition, approximately 30 to 70 times higher than other reported estimates⁴⁹. It is important to note that the underlying population for Kaul et al. was not drawn from the general population; hence, the estimates are much higher.

Europe:

In the UK, the incidence of IPF varied based on the case definition used: the broad IPF incidence rate was 8.7 (8.4 - 8.9) per 100,000 persons per year, while the narrow IPF incidence rate was 2.8 (2.7 - 3) per 100,000 persons per year¹³. However, Bellou et al. reported a much higher incidence of 183.3 per 100,000 persons using self-reported biobank data⁵⁰. In Italy, the age- and sex-adjusted incidence rate of IPF was reported for three different case definitions: generic, broad, and narrow. The generic definition yielded an incidence of 5.2 (5.1 - 5.4) per 100,000 persons per year; the broad definition reported an incidence of 3.7(3.6 - 3.9) per 100,000 persons per year; and the narrow definition reported an incidence of 2.3 (2.2 - 2.5) per 100,000 persons per year between 2005 and 2010⁴⁵

Asia: Three studies have reported incidences in South Korea. The estimates ranged from 8 to 48.5 per 100,000 persons⁴⁶⁻⁴⁸. Gjonbrataj et al.⁴⁶ reported the incidence of IPF using different case definitions, which included diagnosis codes, diagnosis codes plus evidence of bronchoalveolar lavage (BAL), diagnosis codes plus BAL, and HRCT (again, from broad, which is only clinical codes to narrow, which includes clinical codes plus evidence of scan or lung biopsy). Table 2.5 provides details on the case descriptions.

Table 2.5 summarises studies reporting the incidence of IPF (sorted by case definitions). Table S1 (appendix) provides detailed characteristics of studies reporting the incidence of IPF.

Table 2. 7 Studies reporting IPF incidence and sorted by case definition (primary general, overall, broad and narrow)

Country,	Incidence estimate	Underlying population (denominator)	Case definitions	Case descriptions
----------	--------------------	-------------------------------------	------------------	-------------------

Author, published year				
USA, Raghu 2016 ³³	5.8 per 100,000 py	Claims database	Primary	Patients with at least one claim for idiopathic fibrosing alveolitis
Italy, Harari 2016 ⁴⁵	5.2 (5.1-5.4) per 100,000 persons	Healthcare administrative database of Lombardy	General	Patients with at least one hospitalisation or at least one outpatient visit with an IPF diagnosis
South Korea, Lee 2016 ⁴⁷	12.9 per 100,000 persons	National Claims Database	Overall	Diagnostic codes
South Korea, Lim 2019 ⁴⁸	35.8 per 100,000 persons	National Claims Database	Overall	Diagnostic codes
USA, Raghu 2016 ³³	3.6 to 5.1 per 100,000 py	Claims database: Optum	Broad	Primary patients with an exclusion of ICD code 515 after the last diagnosis code 516.3.
Canada, Hopkins 2016 ⁴⁴	18.7 per 100,000 persons	National databases	Broad	Patients with IPF diagnosis, excluding cases with a diagnosis for another ILD
USA, Kaul 2021 ⁴⁹	331 per 100,000 py	Claims database: Veterans health	Broad	Patients who had a diagnosis code for IPF
Italy, Harari 2016 ⁴⁵	3.7 (3.6-3.9) per 100,000 persons	Healthcare administrative database of Lombardy	Broad	Patients who satisfied the general definition and had no claims with another ILD
UK, Strongman 2018 ¹³	8.7 (8.4-8.9) per 100,000 py	Electronic health records	Broad	Included Read codes • Idiopathic fibrosing alveolitis • Cryptogenic fibrosing alveolitis • Idiopathic fibrosing alveolitis NOS • UIP • IPF
USA, Raghu 2016 ³³	2.4 to 2.9 per 100,000 py	Claims database: Optum	Narrow	Broad cases are further restricted by requiring a claim for a surgical lung biopsy, TLB or CT thorax scan
Canada, Hopkins 2016 ⁴⁴	9 per 100,000 persons	National databases	Narrow	Patients with IPF diagnosis excluding cases that did not have chest CT, bronchus or lung biopsy, or bronchoscopy record
Italy, Harari 2016 ⁴⁵	2.3 (2.2-2.5) per 100,000 persons	Healthcare administrative database of Lombardy	Narrow	• Patients that satisfied the broad definition • Patients with a claim for SLB, TLB or CT thorax
UK, Strongman 2018 ¹³	2.8 (2.7-3) per 100,000 py	Electronic health records	Narrow	Additional Read codes to the broad definition
USA, Kaul 2021 ⁴⁹	210 per 100,000 py	Veterans Database	Narrow	• Patients who had a diagnosis code for IPF • And no other ILD diagnosis • Patients with SLB or a CT thorax before the last IPF diagnosis

South Korea, Gjonbrataj 2015 ⁴⁶	48·5 per 100,000 persons	Claims database	Definition 1	ICD codes; (BAL) or lung biopsy
South Korea, Gjonbrataj 2015 ⁴⁶	32·2 per 100,000 persons	Claims database	Definition 2	Diagnostic ICD code and HRCT, bronchoalveolar lavage or SLB
South Korea, Gjonbrataj 2015 ⁴⁶	16·2 per 100,000 persons	Claims database	Definition 3	ICD code
South Korea, Gjonbrataj 2015 ⁴⁶	11·4 per 100,000 persons	Claims database	Definition 4	ICD code and HRCT, BAL or SLB; and
South Korea, Gjonbrataj 2015 ⁴⁶	8 per 100,000 persons	Claims database	Definition 5	ICD code based on the 2011 international statement.
UK, Bellou, 2021 ⁵⁰	183·3 per 100,000 persons	Biobank	NR	Self-reported

py, person-years; IPF, Idiopathic pulmonary fibrosis, CT, computed tomography; HRCT, High-resolution computed tomography; TLB, transbronchial lung biopsy

1.3.2.2 Incidence of Non-IPF ILDs

Unlike IPF, I could not do a 1:1 comparison in incidence estimates at the condition level. There is a high discrepancy in how incidence was reported; for example, some papers reported it as a rate (per 100,000 person-years) and some as a proportion of the underlying population, which was considered. This variation was observed within a subgroup and also at the condition level. For autoimmune ILDs, the population with the underlying autoimmune disease were used as the denominator. For example, for RA-ILD, the studies reported new cases of ILD in patients who had evidence of RA. Meanwhile, for occupational or exposure-related ILDs, the underlying population was workers or factories. Table 2. lists the incidence for each subgroup and condition across different countries.

Appendix 2 (Table S2) has the detailed characteristics of non-IPF ILDs.

Table 2. 6 Studies reporting ILD incidence are grouped by ILD subgroups and sorted into ILD subgroups.

ILD subgroup	Country	Unit of Reporting	Incidence
RA-ILD	Asia ⁵¹	% of study RA cohort	37.3
RA-ILD	North America ^{52,53}	Per 100,000	3.8 to 714.0
SSc-ILD; dcSSc	Asia ⁵⁴	Per 100,000 py	58.8
SSc-ILD; dcSSc	Europe ⁵⁵ Carton 2021	Per 100,000 py	3730
SSc-ILD; lcSSc	Asia ⁵⁴	Per 100,000 py	17.3
SSc-ILD; lcSSc	Europe ⁵⁴	Per 100,000 py	2300
Sjogren's syndrome	Europe ⁵⁶	% of Sjogren's cohort	3.4
CTD ILD	Africa ⁵⁷	% of study cohort	2.5
MCTD ILD	Europe ⁵⁸	% of study cohort	1.7
HP	Europe ⁵⁹	Per 100,000 py	0.9
	North America ⁶⁰	Per 100,000	1.3
IIP	Asia ⁴⁷	Per 100,000	34.9
IIP	Europe ⁵⁹	Per 100,000 py	4.4
Pneumoconiosis	Asia ⁶¹	% of study cohort	0.3 to 4.9
Pneumoconiosis	Europe ⁵⁹	Per 100,000 py	0.8
Silicosis ¹	North America ⁶²	Per 100,000	16.6
Asbestosis	Europe ^{63, 64}	% of study cohort	0.03 to 4.9
Asbestosis	North America ⁶⁵	% of study cohort	0.04
LAM	Europe ⁵⁹	Per 100,000 py	0.3
PLCH	Europe ⁵⁹	Per 100,000 py	0.2
Others			
Pulmonary sarcoid	Asia ⁶⁶	Per 100,000 py	0.5
Pulmonary sarcoid	Europe ⁵⁹	Per 100,000 py	4.9
Unclassifiable ILD	Europe ⁵⁹	Per 100,000 py	1.8
Progressive fibrosing ILDs	Europe ⁶⁷	Per 100,000	4.7
Progressive fibrosing ILDs	USA ⁶⁸	Per 100,000	32.5

py, person-years; RA-ILD, Rheumatoid arthritis ILD; SSc, systemic sclerosis ILD; LAM, lymphangioleiomyomatosis; PLCH, pulmonary Langerhans cell histiocytosis; BOOP, bronchiolitis obliterans organizing pneumonia; MCTD, mixed connective tissue disorder; CTD, connective tissue disorder

¹Silicosis was defined as a Medicare beneficiary with any claim that included ICD-9-CM code 502, pneumoconiosis due to other silica or silicates, listed in any position during 1999-2014, with at least one inpatient, skilled nursing or home health agency claim, or at least two outpatient provider claims within 365 days of each other and cases with a chest X-ray or CT scan 30 days before or 30 days after a silicosis claim.

2.3.3. Studies reporting prevalence.

Similar to incidence, studies reporting prevalence employed different methods to establish prevalence. These methods varied mainly between IPF and non-IPF studies. IPF prevalence was mostly established by using secondary use of large datasets, whereas ILD in patients with autoimmune conditions was mainly determined in a tertiary single or multicenter study setting.

2.3.3.1. IPF

Seven studies reporting the prevalence of IPF were included in this review. The prevalence of IPF was commonly reported using 'broad', 'intermediate', and 'narrow' case definitions. Table 2.8 presents IPF prevalence by different case definitions. In the general population, the prevalence of IPF ranged from 7 to 1,650 per 100,000 persons^[33, 49]. These differences were caused by dissimilarities in how IPF was diagnosed and the underlying population studied. Only one study reported IPF prevalence of 8.6% using a multicentre study setting⁶⁹.

North America: Four studies reported IPF prevalence in North America. Prevalence for 'broad' cases ranged from 11³³ to 1,160⁴⁹, and for 'narrow' cases ranged from 7³³ to 725⁴⁹. As noted earlier, Kaul et al. had a different underlying population (veteran health) and not a generalised population. Hence, considering those estimates as outliers, if I observe the range of broad: 11³³ to 42⁴⁴ and narrow: 7³³ to 20⁴⁴ per 100,000 persons, illustrating an approximate 3-to-4-fold difference in the reported IPF prevalence across different case definitions.

Europe: Two studies reported IPF prevalence in Europe. In Italy, IPF prevalence ranged from 13 (narrow) to 36 (primary) per 100,000 persons⁴⁵. Strongman et al. reported 39 per 100,000 persons in the UK for the broad definition¹³.

Asia: 1 study reported that the prevalence of IPF in South Korea using the national claims database was 35 per 100,000 persons⁴⁷.

Overall, across countries, the prevalence of IPF using a broad definition was 11³³ to 42⁴⁴ per 100,000 persons. For a narrow definition, it was 7³³ to 20⁴⁴ per 100,000 persons (in generalised population, excluding Kaul et al.). This shows that even within the same case category of broad or narrow, variation is observed—mainly due to the underlying population (denominator) and how an IPF case was found.

Table 2.7 summarises studies reporting the prevalence of IPF. Table S3 (in the appendix) reports the detailed characteristics of the studies reporting the prevalence of IPF.

Table 2.7 List of studies reporting IPF prevalence, sorted by region.

Case definitions	Prevalence estimate (per 100,000 persons)	Underlying population (denominator)	Case descriptions	Country, author-published year
General / Primary	36	Healthcare administrative database of Lombardy	Patients with at least one hospitalisation or at least one outpatient visits with IPF diagnosis	Italy, Harari 2016 ⁴⁵
General / Primary	13	Claims database: Optum	Patients with at least 1 claim with IPF diagnosis.	USA, Raghu 2016 ³³
Overall IPF	20	Claims database	Patients with at least 1 IPF inpatient claim, or 2 IPF outpatient claims with ICD code with no other ILD claim	USA, Raimundo 2016 ⁷⁰
Overall IPF	35	National database	IPF diagnostic K codes	South Korea, Lee 2016 ⁴⁷
Broad	1,160	Veteran's database	Patients who had a code for IPF with no other ILD	USA, Kaul 2021 ⁴⁹
Broad	11	Claims database: Optum	Patients with code of ICD 516·3, patients excluded if they had a claim with code 515 after the last ICD code 516·3	USA, Raghu 2016 ³³
Broad	42	National health database	Patients with code ICD code J84·1, cases with a diagnosis of another ILD excluded	Canada, Hopkins 2016 ⁴⁴

Broad	22	Healthcare administrative database of Lombardy	Patients with IPF code and no claims for other ILD diagnosis	Italy, Harari 2016 ⁴⁵
Broad	39	Electronic health records	Patients with Read codes: H563·00, H563·12, H563300, H563·13, H563100, H563200 and H563.11	UK, Strongman 2018 ¹³
Narrow	13	Healthcare administrative database of Lombardy	- Patients that satisfied the broad definition - had 1> claim with procedure code for SLB, TLB or CT thorax	Italy, Harari 2016 ⁴⁵
Narrow	725	Veteran's health	- Patients who had a code for IPF - And if they did not have any other code for an alternative ILD - And patients who had procedure code for an SLB or a thorax	USA, Kaul 2021 ⁴⁹
Narrow	7	Claims database: Optum	- Patients with an ICD code of 516·3 and they were excluded if they had a claim with the ICD code 515 - And further restricted by requiring a claim for a SLB, TLB, or CT thorax	USA, Raghu 2016 ³³
Narrow	20	National health databases	Patients with ICD codes J84·1 with no other ILD and excluded cases that did not have chest CT, bronchus or SLB, or bronchoscopy	Canada, Hopkins 2016 ⁴⁴

IPF, Idiopathic pulmonary fibrosis; CT, computed tomography; HRCT, High-resolution computed tomography; ICD, International Classification of Diseases; SLB, surgical lung biopsy; TLB, transbronchial lung biopsy; USA, United States of America; UK, United Kingdom.

2.3.3.1.1. Non-IPF ILDS

Of 62 studies reporting non-IPF ILDs, there were only nine papers (14.5%) that reported non-IPF ILDs in the general population (Table 2.8). Among these studies, only two were outside the USA (1 in France and 1 in South Korea). Non-IPF ILD prevalence in the general population was mainly conducted in the USA, and fewer large healthcare databases were used outside the USA. Two studies reported the prevalence of fibrosing ILDs using the USA claims database and the French national administrative. More detailed information about these studies, including study details, cohort characteristics, and ILD case ascertainment methods, can be found in Table S4 in the appendix.

Table 2.8 Studies reporting non-IPF prevalence, population level prevalence, by ILD condition, and year of publication.

ILD condition	Author, Year	Country	Study population	Unit for reporting	Prevalence
HP	Perez 2018 ⁶⁰	United States	MarketScan Commercial and Medicare Supplemental healthcare claims databases	per 100,000 persons	2.7
HP	Olson 2021 ⁶⁸	USA	IBM MarketScan claims database	per 100,000 persons	1.21
IIP	Lee 2016 ⁴⁷	South Korea	National health insurance database	per 100,000 persons	81.8
MCTD ILD	Olson 2021 ⁶⁸	USA	IBM MarketScan claims database	per 100,000 persons	0.32
Progressive fibrosing ILD	Nasser 2021 ⁶⁷	France	French healthcare database	per 100,000 persons	19.4
MCTD ILD	Olson 2021 ⁶⁸	USA	IBM MarketScan claims database	per 100,000 persons	117.8
RA ILD	Raimundo 2019 ⁵²	USA	Truven market scans commercial claims and encounters database	per 100,000 persons	6
RA ILD	Sparks 2021 ⁵³	USA	Medicare claims data that includes nearly all Americans 65 years of age and some younger individuals with disabilities specific disabilities.	% of the study population	2
RA ILD	Olson 2021 ⁶⁸	USA	IBM MarketScan claims database	per 100,000 persons	2.8
Silicosis	Casey 2019 ⁶²	USA	Medicare claims data that includes nearly all Americans 65 years of age and some younger individuals with disabilities certain disabilities.	Narrow, per 100,000 persons	20.1
SSc ILD	Li 2021 ⁷¹	USA	IBM MarketScan claims database	per 100,000 persons	19
SSc ILD	Pope 2021 ⁷²	Canada	SS Patients were identified using the Canadian Institute for Health Information (CIHI) National Ambulatory Care Reporting System (NACRS) and the Discharge Abstract Database (DAD)	per 100,000 persons	2.3
SSc ILD	Olson 2021 ⁶⁸	USA	IBM MarketScan claims database	per 100,000 persons	2.8

HP, hypersensitivity pneumonitis; IIP, idiopathic interstitial pneumonia; MCTD, mixed connective tissue disorder; RA, rheumatoid arthritis; SSc, systemic sclerosis.

Most studies (53 of 62) established prevalence for non-IPF ILDs in tertiary single centres, multicenter settings, a disease registry, or a factory. These studies primarily reported prevalence as a proportion, with the

denominator representing patients with an underlying condition (autoimmune disorder) or subjects working at a factory exposed to certain agents, such as silica or asbestosis (Table 2.9).

Table 2.9 displays notable variations in the prevalence of SSc ILD, RA ILD, and Sjogren's ILD, not only among different regions but also within the same region. For instance, the prevalence of Sjogren's ILD ranged from 1% (Sweden, 2011) to 87.8% (Saudi Arabia, 2021). Moreover, within China, four papers reported estimates for Sjogren's ILD, which varied from 18.6% (2011) to 78.6% (2014). Several factors, such as the type of study conducted (single or multiple centre studies), the methodology used to identify ILD cases, the severity of the underlying population, and the duration of follow-up for the cases, led to significant differences in the estimates. These aspects were evaluated in the assessment of quality bias.

Table 2.10 shows there was a significant difference in the prevalence estimates of SSc-ILD from 26.1% (Australia, 2015) to 88.1% (India, 2013). Largely, for SSc ILD, studies published in Europe had a lower SSc-ILD prevalence (ranged from 31.2-49.4%), in comparison to the studies based in Asia (ranged from 27.6 to 88.1%). Contributing factors may be a higher proportion of patients who had a CT scan or that ethnicity may also play a role in the clinical features of SSc.

Among the seven papers, the prevalence of Sjogren's ILD ranged from 1.01 to 78.6%. Except for Zhao et al.⁷³, which included patients from multiple centres in China, the others were single-centre studies. Wang et al.⁷⁴ reported the highest prevalence of Sjogren's ILD (78.6%). However, the differentiating factor was that it included only incident Sjogren's cases and had a higher proportion of patients undergoing the CT scan. Six studies reported a prevalence of RA-ILD from RA-specific cohorts. Among these studies, prevalence ranged from 1.8 to 30.01%. Only one study reported prevalence using a claims database of 3.2 per 100,000 persons in the USA⁵².

For exposure-related ILD, silicosis had the maximum number of published papers (n=7). The prevalence of silicosis ranged from 5.6 in Sri Lanka to 37.1% in Brazil (Table 2.10). For all other exposure or occupational ILDs, 1-2 papers were found in the literature, thus making it difficult to compare and get a pooled estimate. Appendix 2 provides more detailed information about these studies, including study details, cohort characteristics, and ILD case ascertainment methods.

Table 2. 9 Studies reporting prevalence for Autoimmune ILDs, cohort-specific prevalence, sorted by autoimmune condition.

ILD condition	Author, Year	Country	Unit for reporting	Prevalence
CTD ILD	Olaosebikan 2021 ⁵⁷	Nigeria	% of study population	7.2
MCTD ILD	Reiseter 2018 ⁵⁸	Norway	% of study population	40
MCTD ILD	Olson 2021 ⁶⁸	USA	per 100,000 persons	0.3
Myositis ILD	Gomez 2021 ⁷⁵	Argentina	% of study population	27.5
Myositis ILD	Huang 2021 ⁷⁶	Taiwan	% of study population	47.4
Myositis ILD	Ishizuka 2016 ⁷⁷	Japan	% of study population	61.3
RA ILD	Duarte 2019 ⁷⁸	UK	% of study population	3.9
RA ILD	Ke 2021 ⁷⁹	China	% of study population	14.8
RA ILD	Kiely 2019 ⁸⁰	UK	% of study population	4.8
RA ILD	Kim 2017 ⁸¹	South Korea	% of study population	1.8
RA ILD	Kronzer 2021 ⁸²	Sweden	% of study population	0.6
RA ILD	McFarlane 2019 ⁸³	USA	% of study population	6.4
RA ILD	Md Yosuf 2017 ⁸⁴	UK	% of study population	8
RA ILD	Paulin 2021 ⁸⁵	Argentina	% of study population	7.5
RA ILD	Raimundo 2019 ⁵²	USA	per 100,000 persons	6
RA ILD	Sakr 2018 ⁸⁶	Egypt	% of study population	0.8
RA ILD	Hammoda 2021 ⁸⁷	Egypt	% of study population	63.7
RA ILD	Shin 2021 ⁸⁸	South Korea	% of study population	0.84
RA ILD	Sparks 2021 ^{*53}	USA	% of study population	2
RA ILD	Zhang 2017 ⁵¹	China	% of study population	5.8
RA ILD	Olson 2021 ⁶⁸	USA	per 100,000 persons	2.8
Sjogrens ILD	Gao 2018 ⁸⁹	China	% of study population	19.3
Sjogrens ILD	Gao 2021 ⁹⁰	China	% of study population	19.1
Sjogrens ILD	Kam 2021 ⁹¹	Singapore	% of study population	1.96
Sjogrens ILD	Kampolis 2018 ⁹²	Greece	% of study population	17.9
Sjogrens ILD	Kvarnstorm 2015 ⁹³	Sweden	% of study population	1
Sjogrens ILD	Manfredi 2017 ⁹⁴	Italy	% of study population	16.9
Sjogrens ILD	Nilsson 2021 ⁹⁵	Sweden	% of study population	38
Sjogrens ILD	Omair 2021 ⁹⁶	Saudi Arabia	% of study population	87.8
Sjogrens ILD	Roca 2017 ⁵⁶	France	% of study population	4.2
Sjogrens ILD	Wang 2018 ⁷⁴	China	% of study population	78.6
Sjogrens ILD	Zhao 2015 ⁷³	China	% of study population	18.6
SSc ILD	Bauer 2020 ⁹⁷	France	% of study population	31.2
SSc ILD	Carton 2021 ⁵⁵	Belgium	% of study population	34
SSc ILD	Fairley 2021 ⁹⁸	Australia	% of study population	26.1
SSc ILD	Jaafar 2021 ⁹⁹	USA	% of study population	53.6
SSc ILD	Janardana 2019 ¹⁰⁰	India	% of study population	88.1
SSc ILD	Li 2021 ⁷¹	USA	per 100,000 persons	19
SSc ILD	Morrisroe 2019 ¹⁰¹	Australia	% of study population	28.3
SSc ILD	Noviani 2019 ¹⁰²	Singapore	% of study population	27.6
SSc ILD	Pope 2021 ⁷²	Canada	per 100,000 persons	2.3
SSc ILD	Sanchez-cano 2018 ¹⁰³	Spain	% of study population	43.3

SSc ILD	Simeon-Aznar 2015 ¹⁰⁴	Spain	% of study population	47.8
SSc ILD	Tomiyama 2016 ¹⁰⁵	Japan	% of study population	47.5
SSc ILD	Wangkaew 2016 ⁵⁴	Thailand	% of study population	69.1
SSc ILD	Olson 2021 ⁶⁸	USA	per 100,000 persons	2.8

Table 2. 10 Studies reporting prevalence for exposure ILDs, cohort specific prevalence, sorted by ILD condition

ILD condition	Author, Year	Country	Unit for reporting	Prevalence
Asbestosis	Lee 2015 ¹⁰⁶	South Korea	% of study population	77.1
Asbestosis	Wickramatillake 2019 ¹⁰⁷	Sri Lanka	% of study population	7
HP	Duchemann 2017 ⁵⁹	France	per 100,000 persons per year	2.3
HP	Perez 2018 ⁶⁰	United States	per 100,000 persons	2.7
HP	Olson 2021 ⁶⁸	USA	per 100,000 persons	1.21
Other exposure related	Choi 2015 ¹⁰⁸	South Korea	% of study population	24
Pneumoconiosis	Duchemann 2017 ⁵⁹	France	per 100,000 persons	3.5
Pneumoconiosis	Simsek 2021 ¹⁰⁹	Turkey	% of study population	38.8
Radiation-induced ILD	Duchemann 2017 ⁵⁹	France	per 100,000 persons	0.6
Silicosis	Casey 2019 ¹⁶²	USA	per 100,000 persons	20.1
Silicosis	Hoy 2021 ¹¹⁰	Australia	% of study population	36
Silicosis	Knight 2015 ¹¹¹	South Africa	% of study population	11.9
				10.5
Silicosis	Souza 2021 ¹¹²	Brazil	% of study population	
Silicosis	Poinen-Rughooputh 2021 ¹¹³	China	% of study population	13.7
Silicosis	Requena-Mullor 2021 ¹¹⁴	Spain	% of study population	20.6
Silicosis	Silanum 2017 ¹¹⁵	Thailand	% of study population	25.4
Silicosis	Siribadda 2016 ¹¹⁶	Srilanka	% of study population	5.6
Silicosis	Souza 2017 ¹¹⁷	Brazil	% of study population	37

2.3.4. Assessment of risk bias for individual studies

The strengths and limitations of the included studies differed based on various factors such as the study setting, data source, ILD diagnosis approach, and research methodology. Six domains for studies reporting prevalence and eight domains for studies reporting incidence were assessed. For prevalence, I assessed the sample's representativeness, sample size, ascertainment of ILD, reporting quality, and exposure type reported. For incidence: Representativeness of a sample, sample size, ascertainment of ILD, quality of descriptive statistics reporting, the demonstration that ILD was not present at the start of the study, follow-up long enough for ILD to occur: adequacy of follow-up and ascertainment of exposure of occupational ILD/HP/autoimmune ILD.

Studies utilising large datasets offered the advantage of a larger sample size and potentially more generalisable results. However, they often had unclear case definitions and limited coding granularity, which could lead to possible misdiagnosis. On the other hand, studies conducted at a single centre or within a specific geographic region, such as a factory, had the advantage of verified clinical ILD diagnoses and the involvement of multidisciplinary teams. Nevertheless, these studies were susceptible to selection bias. This chapter presents separate risk scores for papers reporting incidence and prevalence.

Assessment for studies reporting incidence: As depicted in Table 2.11, among the 40 assessments of ILD incidence, most studies (26 out of 40, or 65%) were classified as medium risk. Two studies were classified as high risk mainly due to insufficient information on ILD diagnosis and poor quality of reporting estimates (e.g., incomplete or lacking descriptive statistics and appropriate measures of dispersion). Additionally, 12 studies (30%) were classified as low-risk.

Regarding the method used for ILD diagnosis, approximately 40% of the papers (16 out of 40) relied on the expertise of a multidisciplinary team to ascertain ILD cases. About 33% of the papers (13 out of 40) used diagnostic codes for ILD assessment. A smaller number of studies solely utilised HRCT (high-resolution computed tomography) as the diagnostic method, and only two studies did not report the specific method of diagnosis. These findings highlight the inconsistencies in determining ILD across different regions and study settings.

Assessment for studies reporting prevalence:

According to Table 2.12, a total of 76 prevalence assessments were conducted. Among these, approximately 18% (14 out of 76) of the studies were classified as high risk, 66% (50 out of 76) as medium risk, and 16% (12 out of 76) as low risk. The studies reporting on idiopathic pulmonary fibrosis (IPF) and autoimmune ILDs contributed to the higher risk category, primarily due to factors such as ILD diagnosis methods, single-centre study designs, or small sample sizes. Conversely, most studies that reported prevalence drawn from large datasets (claims, electronic health records) or disease registries were classified as low risk.

Significant inconsistencies were observed in the ascertainment of ILD across the studies. Approximately 22% (17 out of 76) of the studies did not describe how ILD cases were ascertained. Around 28% (21 out of 76) of the studies relied on diagnosis codes to identify ILD patients, 36% (27 out of 76) used HRCT scans as a diagnostic method, and 15% (11 out of 76) ascertained ILD with the support of a multidisciplinary team.

Similar to the incidence assessments, these findings highlight the disparities in how ILD is determined across different studies.

Table 2. 11 Assessment for studies reporting incidence.

ILD condition	Study author	Study type	Risk of Bias Assessment	Ascertainment of ILD
Asbestosis	Thomsen 2021 ⁶⁴	Retrospective	Low	High
Asbestosis	Szeszenia-Dąbrowska 2015 ⁶³	Retrospective	Medium	High
Asbestosis	DeBono 2021 ⁶⁵	Retrospective	Medium	High
Asbestosis	DeBono 2021 ⁶⁵	Surveillance/ Restrospective	Medium	Low
CTD ILD	Olaosebikan 2021 ⁵⁷	Retrospective	High	None
Cystic Lung disease	Duchemann 2017 ⁵⁹	Cross-sectional	Medium	High
HP	Duchemann 2017 ⁵⁹	Cross-sectional	Medium	High
HP	Perez 2018 ⁶⁰	Retrospective	Medium	Low
IIP	Lee 2016 ⁴⁷	Retrospective	Low	Low
IIP	Duchemann 2017 ⁵⁹	Cross-sectional	Medium	High
IPF	Gjonbrataj 2015 ⁴⁶	Retrospective	Medium	Low
IPF	Harari 2016 ⁴⁵	Retrospective	Medium	Low
IPF	Hopkins 2016 ⁴⁴	Retrospective	Medium	Low
IPF	Lee 2016 ⁴⁷	Retrospective	Low	Low
IPF	Raghu 2016 ³³	Retrospective	Low	Low
IPF	Bellou 2021 ⁵⁰	Prospective	High	None
IPF	Strongman 2018 ¹³	Retrospective	Low	Medium
IPF	Lim 2019 ⁴⁸	Retrospective	Medium	Medium
IPF	Kaul 2021 ⁴⁹	Retrospective	Medium	Low
MTCD ILD	Reiseter 2018 ⁵⁸	Retrospective	Low	Medium
Pneumoconiosis	Cui 2015 ⁶¹	Retrospective	Medium	High
Pneumoconiosis	Duchemann 2017 ⁵⁹	Cross-sectional	Medium	High
Progressive fibrosing ILD	Olson 2021 ¹¹⁸	Retrospective	Medium	Low
Progressive fibrosing ILD	Nasser 2021 ⁶⁷	Retrospective	Low	Medium

Pulmonary Sarcoid	Duchemann 2017 ⁵⁹	Cross-sectional	Medium	High
Pulmonary Sarcoid	Jeon 2020 ⁶⁶	Retrospective	Medium	High
Pulmonary Sarcoid	Coquart 2015 ¹¹⁹	Retrospective	Medium	High
RA ILD	Zhang 2017 ⁵¹	Retrospective	Low	Medium
RA ILD	Raimundo 2019 ⁵²	Retrospective	Low	Low
RA ILD	Sparks 2021 ⁵³	Retrospective	Medium	Low
Silicosis	Casey 2019 ⁶²	Retrospective	Medium	High
Sjogren's Syndrome ILD	Roca 2017 ⁵⁶	Retrospective	Low	Medium
SSc ILD	Wangkaew 2016 ⁵⁴	Prospective	Low	Medium
SSc ILD, overall	Carton 2021 ⁵⁵	Retrospective	Low	Medium
SSc ILD, overall	Li 2021 ⁷¹	Retrospective	Medium	Low
Unclassifiable ILD	Duchemann 2017 ⁵⁹	Cross-sectional	Medium	High

Table 2.12 Assessments for studies reporting prevalence.

ILD condition		Study type	Risk of Bias Assessment	Ascertainment of ILD
Asbestosis	Wickramatillake2019 ¹⁰⁷	Cross-sectional	M	None
Asbestosis	Lee 2015 ¹⁰⁶	Cross-sectional	L	Medium
CTD ILD	Duchemann 2017 ⁵⁹	Cross-sectional	L	High
CTD ILD	Olaosebikan 2021 ⁵⁷	Retrospective	M	Low
HP	Perez 2018 ⁶⁰	Retrospective	M	Low
HP	Duchemann 2017 ⁵⁹	Cross-sectional	L	High
IIP	Lee 2016 ⁴⁷	Retrospective	M	Low
IIP	Duchemann 2017 ⁵⁹	Cross-sectional	L	High
IPF	Harari 2016 ⁴⁵	Retrospective	M	Low
IPF	Hopkins 2016 ⁴⁴	Retrospective	M	Low
IPF	Lee 2016 ⁴⁷	Retrospective	M	Low
IPF	Raghu 2016 ³³	Retrospective	M	Low
IPF	Raimundo 2016 ⁷⁰	Cross-sectional	M	Low
IPF	Strongman 2018 ¹³	Retrospective	M	Low
IPF	Zhang 2022 ⁴⁹	Retrospective	M	Low
IPF	Kaunisto 2015 ⁶⁹	Cross-sectional	L	Medium
LAM	Duchemann 2017 ⁵⁹	Cross-sectional	L	High
MCTD ILD	Reiseter 2018 ⁵⁸	Retrospective	M	Medium
Myositis ILD	Ishizuka 2016 ⁷⁷	Retrospective	M	Medium
Myositis ILD	Huang 2021 ⁷⁶	Retrospective	M	Low
Myositis ILD	Gomez 2021 ⁷⁵	Cross-sectional	H	None
Other exposure related ILD	Choi 2015 ¹⁰⁸	Survey	M	High
PLCH	Duchemann 2017 ⁵⁹	Cross-sectional	L	High
Pneumoconiosis	Duchemann 2017 ⁵⁹	Cross-sectional	M	None
Pneumoconiosis	Simsek 2022 ¹⁰⁹	Cross-sectional	H	None

Progressive fibrosing ILD	Nasser 2021 ⁶⁷	Retrospective	M	Low
Progressive fibrosing ILD	Olson 2021 ¹¹⁸	Retrospective	M	Low
Pulmonary sarcoidosis	Duchemann 2017 ⁵⁹	Cross-sectional	M	High
Pulmonary sarcoidosis	Coquart 2015 ¹¹⁹	Retrospective	H	None
RA ILD	Kiely 2019 ⁸⁰	Prospective	M	Low
RA ILD	Paulin 2021 ⁸⁵	Cross-sectional	M	Medium
RA ILD	Raimundo 2019 ⁵²	Retrospective	M	Low
RA ILD	Duarte 2019 ⁷⁸	Retrospective	M	Medium
RA ILD	Zhang 2017 ⁵¹	Retrospective	M	Medium
RA ILD	Samy 2021 ⁸⁷	Cross-sectional	M	High
RA ILD	Sparks 2021 ⁵³	Retrospective	M	Low
RA ILD	Md Yosuf 2017 ⁸⁴	Retrospective	M	Medium
RA ILD	McFarlane 2019 ⁸³	Cross-sectional	L	High
RA ILD	Kim 2017 ⁸¹	Prospective	H	None
RA ILD	Shin 2021 ⁸⁸	Retrospective	H	None
RA ILD	Sakr 2018 ⁸⁶	Retrospective	H	None
RA ILD	Ke 2021 ⁷⁹	Retrospective	H	None
RA ILD	Kronzer 2021 ⁸²	Case-control	H	Low
Silicosis	Hoy 2021 ¹¹⁰	Population-based health assessment	M	High
Silicosis	Knight 2015 ¹¹¹	Prospective	M	None
Silicosis	Poinen-Rughooputh 2021 ¹¹³	Retrospective	M	Low
Silicosis	Requena-Mullor ¹¹⁴	Cross-sectional	M	Medium
Silicosis	Siribadda 2016 ¹¹⁶	Cross-sectional	M	None
Silicosis	Souza 2017 ¹¹⁷	Cross-sectional	M	None
Silicosis	Casey 2019 ⁶²	Retrospective	H	Low
Silicosis	Silanum 2017 ¹¹⁵	Cross-sectional	H	None
Silicosis	Souza 2021 ¹¹²	Cross-sectional	H	None
Sjogrens ILD	Gao 2018 ⁸⁹	Retrospective	M	Medium
Sjogrens ILD	Gao 2021 ⁹⁰	Retrospective	M	Medium
Sjogrens ILD	Kampolis 2018 ⁹²	Cross-sectional	M	Medium
Sjogrens ILD	Kvarnstorm 2015 ⁹³	Prospective	M	Low
Sjogrens ILD	Manfredi 2017 ⁹⁴	Prospective	M	Medium
Sjogrens ILD	Wang 2018 ⁷⁴	Prospective	M	Medium
Sjogrens ILD	Zhao 2015 ⁷³	Cross-sectional	M	Medium
Sjogrens ILD	Kam 2021 ⁹¹	Retrospective	H	None
Sjogrens ILD	Nilsson 2021 ⁹⁵	Cross-sectional	H	Medium
Sjogrens ILD	Omair 2021 ⁹⁶	Cross-sectional	H	None
Sjogrens ILD	Roca 2017 ⁵⁶	Retrospective	L	Medium
SSc ILD	Janardana 2019 ¹⁰⁰	Retrospective	M	Medium
SSc ILD	Sanchez-cano 2018 ¹⁰³	Cross-sectional	M	Medium
SSc ILD	Bauer 2020 ⁹⁷	Retrospective	M	Medium

SSc ILD	Fairley 2021 ⁹⁸	Prospective	M	Medium
SSc ILD	Jaafar 2021 ⁹⁹	Prospective	M	None
SSc ILD	Tomiyama 2016 ¹⁰⁵	Retrospective	M	Medium
SSc ILD	Wangkaew 2016 ⁵⁴	Prospective	M	Medium
SSc ILD	Li 2021 ⁷¹	Retrospective	M	Low
SSc ILD	Pope 2021 ⁷²	Retrospective	M	Low
SSc ILD	Morrisroe 2019 ¹⁰¹	Prospective	L	Medium
SSc ILD	Noviani 2019 ¹⁰²	Prospective	L	Medium
SSc ILD	Simeon-Aznar 2015 ¹⁰⁴	Retrospective	L	Medium
SSc ILD	Carton 2021 ⁵⁵	Retrospective	L	Medium
Unclassifiable ILD	Duchemann 2017 ⁵⁹	Cross-sectional	L	High

2.4. Discussion

In this study, I comprehensively analysed the evidence regarding the incidence and prevalence of ILD based on studies published between 2015 and 2021. The study period was limited to the past seven years to capture current estimates, considering changes in ILD nomenclature over time. This review included 76 prevalence estimates and 39 incidence estimates for various individual ILD disorders from regions across the globe. In total, 80 unique studies were included that reported 88 estimates, and several studies reported on more than one ILD condition.

When I began this review, I aimed to report the incidence and prevalence of ILD as an umbrella term and then break it down at the subgroup and condition levels. However, during this search, I needed help finding papers that addressed overall ILD. This underscores the heterogeneity of this disease subgroup and the challenges in accurately assessing the burden of this chronic condition.

IPF remains one of the most extensively researched ILDs. During the study period, 11 papers reported on the incidence and prevalence of IPF. The geographic distribution of IPF research was reasonably consistent across different regions. Most studies classified IPF using various case definitions—primary, general, broad, and narrow—each aimed at capturing a spectrum of cases from high sensitivity to high specificity. These definitions sought a comprehensive view of the disease burden by incorporating various diagnostic codes, including 'IPF-like' codes, to account for evolving terminology and coding practices. However, a narrow definition focused specifically on IPF-specific codes. Eight of the reviewed publications reported the incidence of IPF, with many

drawing data from extensive databases such as electronic health records or claims. The incidence rates reported varied widely, ranging from 2.4 to 311 per 100,000 person-years and 2.2 to 183.3 per 100,000 persons^[13,44,49,50]. This wide range is primarily attributed to the study by Kaul et al.,⁴⁹ which reported significantly higher incidence rates due to its focus on veterans, a population at higher risk for IPF.

Similarly, the study by Bellou et al.⁵⁰ used self-reported data, which might not be generalisable. Excluding these studies, the incidence of IPF ranged from 2.4 (Raghu 2016³³) to 8.7 (Strongman 2018¹³), an approximately 4-fold difference. Per 100,000 persons from 2.3 (Harari 2016⁴⁵) to 48.5 (Gjonbrataj 2015⁴⁶), which is approximately 21 times difference, demonstrating the challenge of accurately determining the disease burden due to differences in study methodology (study design and population drawn from) for establishing IPF burden which makes it challenging to derive an accurate disease burden.

The prevalence of IPF also varied across studies, with broadly classified cases ranging from 11 to 42 per 100,000 persons and narrowly classified cases ranging from 7 to 20 per 100,000 persons^{33,44}. These differences were influenced by factors such as the study population and the diagnostic codes' specificity. Despite the extensive research in the field of IPF, the variation could be from different clinical codes used to define IPF in observational data and differences in approaches to ascertaining IPF, which have created inconsistencies in how IPF is studied and reported.

Examining and meta-analysing non-IPF ILDs presents significant challenges due to the diversity and complexity of these conditions. In this review, six non-IPF subgroups encompassing 21 conditions were considered, with the most extensively researched being RA-associated ILD, SSc-associated ILD, Sjogren's ILD, and silicosis. Autoimmune ILDs accounted for about 50% of the estimates, with RA-ILD and SSc-ILD alone contributing approximately 30%. Interest in autoimmune ILDs has surged recently, likely tied to the clinical trials for anti-fibrotic therapies beyond IPF. This is evidenced by a six-fold increase in studies on autoimmune ILDs in 2021 compared to 2015.

For non-IPF ILDs, studies more frequently reported prevalence than incidence, reflecting the difficulties in capturing these conditions using existing data. The studies often focused on specific populations, such as

individuals with rheumatoid arthritis for RA-ILD or workers exposed to silica for silicosis. This introduces selection bias and contributes to inconsistencies in the reported estimates. Geographic distribution showed a dominance of autoimmune ILD studies in Europe and Asia. In contrast, occupational ILD studies were more prevalent in Asia, likely due to the concentration of factories and industries in low- and middle-income countries.

For non-IPF ILDs, studies are typically drawn from single-center cohorts or specific populations of interest, unlike IPF research, which is often based on large, population-based studies. A key distinction lies in the denominator used for calculating disease occurrence, which significantly influences incidence estimates and their comparability across studies. For instance, in studies assessing RA-ILD, the denominator is often the total number of patients diagnosed with RA, with the outcome measuring how many of them subsequently developed ILD. Similarly, in occupational ILDs, such as silicosis, the denominator consists of workers exposed to silica dust, with incidence reflecting the proportion who develop the disease over time. This reliance on specific subpopulations, rather than broader, generalizable databases, introduces a selection bias that can limit the external validity of findings. Differences in population characteristics, risk exposure levels, and study settings can lead to significant variability in reported incidence rates, making it difficult to compare findings for the condition, for example, pneumoconiosis among coal workers in China to coal workers in UK. Additionally, making it difficult to compare burden across ILD conditions. For instance, given the differences in denominator for RA-ILDs (mostly single centers) and IPF (large scale datasets).

Additionally, the choice of incidence measure further complicates interpretation. While some studies report incidence as a proportion (e.g., the percentage of an at-risk population that develops ILD), others use cases per 100,000 individuals, and only a few account for person-time at risk by reporting incidence rates per 100,000 person-years. This variation in denominators and metrics can obscure the true burden of disease. Proportions can overestimate risk when follow-up durations vary, while cases per 100,000 persons may not reflect differences in observation periods or disease latency. The use of person-time metrics helps address these concerns by incorporating the duration of exposure and follow-up, yet such data is often unavailable or inconsistently reported in non-IPF ILD studies.

Furthermore, diagnostic criteria and reporting practices vary widely across studies, contributing to further inconsistencies in incidence estimates. Differences in case definitions, diagnostic methods (e.g., imaging vs. medical records), and physician awareness, all influence reported incidence rates. Geographical and healthcare system variations also play a role, as access to specialist care, infrastructure and diagnostic resources may differ significantly between study populations. These factors collectively make it challenging to derive meaningful comparisons and trends in non-IPF ILDs, underscoring the need for more standardized methodologies and large-scale, population-based studies to improve disease surveillance and epidemiological understanding.

Discrepancies in reported estimates for non-IPF ILDs were observed across different countries and within studies conducted in the same country. These variations were more pronounced for non-IPF ILDs than IPF, driven by factors such as the lack of standardised guidelines for non-IPF ILD assessment, differences in healthcare infrastructure, and variability in diagnostic practices across centres. At the same time, there were attempts to standardise IPF diagnosis through guidelines, no such guidelines exist for non-IPF ILDs, further complicating the assessment and comparison of these diseases.

The introduction of anti-fibrotic therapies for IPF and other progressive ILDs has spurred increased interest in non-IPF fibrosing lung diseases¹²⁰⁻¹²². A recent systematic review highlighted 13-40% of ILDs were estimated to develop a progressive phenotype – thus emphasizing the need for more research in this field¹¹⁸.

This growing interest may explain the rise in studies focusing on autoimmune ILDs in recent years, highlighting the need for more standardised approaches in studying these complex conditions.

A new quality assessment scale was developed to assess the quality of the studies. Depending on the total score, studies were categorized as low (0 to 3), medium (4 to 6/7), or high (6/7 to 9/10) risk of bias. A low risk of bias indicates minimal bias in quality, based on the factors considered in the assessment. Medium risk of bias suggests that there are certain issues, such as how ILD was diagnosed or the sample size, that should be considered. High risk of bias reflects to multiple issues within the study, and such studies should be interpreted with caution, as their results may be misleading or prone to high measurement error.

Like prevalence, incidence scores were also categorized in similar categories. The same interpretation should be applied as for prevalence, considering the risk of bias assessment. The score for incidence considers additional factors, such as person-time, to indicate whether a case is incident. These quality factors are crucial for evaluating the robustness of the estimates reported by a study.

This systematic review has several strengths: it provides an assessment of the incidence and prevalence of various ILD conditions globally; it groups ILD conditions based on their aetiology, allowing for an appraisal of incidence and prevalence at both the ILD condition level and the subgroup level across different countries; it emphasises differences in study settings, such as data sources and clinical and demographic characteristics of the study population, which played a significant role in explaining the discrepancies in reported incidence and prevalence estimates; it uncovers the disproportion in the number of studies reporting prevalence versus incidence, recognising the challenges associated with establishing incidence; it assesses the global landscape, demonstrating differences in prevalence or incidence estimates based on different guidelines and practices, thus accentuating the importance of robust guidelines and standardisation for the ILD community;

Despite these strengths, there are a few limitations: this review cannot confirm the diagnosis of ILD reported in the included studies, although the diagnosis methods adopted by authors have been summarised; studies reporting ILD estimates in cohorts with high comorbidities, such as cancer patients, were excluded to focus on the general population rather than disease burden in specific high-comorbidity groups; and several non-IPF studies, particularly those reporting autoimmune and occupational ILD estimates, were based on single-centre data, which could present a selection bias, risk of bias quality assessment takes this into account.

2.5.Conclusion

In conclusion, this review underscores the considerable uncertainty surrounding the true prevalence of ILDs, with estimates varying significantly—sometimes by factors of 5 to 10 or more, particularly for non-IPF ILDs. IPF, which has been more extensively studied, benefits from greater awareness, enhanced diagnostic tools, and standardised case definitions, leading to a more consistent understanding of its burden. In contrast, the lack of

these factors for non-IPF ILDs has resulted in a poorly characterised disease burden and variable reporting. Given the updates in the field of IPF over the decade, coupled with a lack of recent IPF epidemiology in the UK over the past decade and the insights gained from this review, I have therefore chosen to focus the remainder of this thesis on IPF, where the data is more reliable than other ILD conditions, and the need for updated disease burden assessment is clear.

Chapter 3

Description of Electronic Health Records

Overview

This chapter provides an overview of the data sources used in the subsequent chapters, focusing on the Clinical Practice Research Datalink (CPRD), the leading UK real-world research service for observational studies and interventional research. CPRD has two primary care databases, CPRD GOLD and CPRD Aurum, which contain anonymised EHR data from general practitioners (primary care). This chapter begins by explaining CPRD. Further, I discuss how CPRD works with general practitioners for this data collection process. Understanding how CPRD collects and records data is crucial to ensure an accurate interpretation. I delve into high-level characteristics, such as contributing practices, coverage, and data period, for GOLD and Aurum to elaborate on why I chose to continue my research using Aurum. I explain the structure of Aurum and the variables recorded, which helps me understand how to work effectively with this dataset. Two linked datasets used in my thesis, Hospital Episode Statistics (HES) and Office of National Statistics (ONS) mortality data, are also described. Lastly, I discuss the drawbacks and merits of CPRD.

3.1. What is CPRD and what does it do?

The Clinical Practice Research Datalink (CPRD) is a prominent UK-based real-world research service supporting retrospective and prospective public health studies and interventional research initiatives. Funded by the Medicines and Healthcare Products Regulatory Agency (MHRA) and supported by the National Institute for Health Research (NIHR), CPRD operates as a government-funded, non-profit organisation focused on cost recovery. CPRD is an invaluable resource for public health research, representing over 2,400 general practices and 60 million patient lives¹²³.

Established in 1987, CPRD collects routinely collected records (EHRs) from general practices across the UK. Over the years, CPRD has developed strong relationships with General Practitioners (GP) and Clinical Research Networks (CRNs), building a comprehensive dataset that spans several decades.

CPRD offers a longitudinal view of patient data. Established with a goal beyond mere administrative use, CPRD functions as a clinical database that captures detailed patient information recorded during routine care. While it includes some administrative data, such as patient registration details, its primary focus is on medical records, diagnoses, prescriptions, lifestyle factors, and clinical outcomes. This data supports various research activities, such as i) enabling the study of disease progression, ii) treatment outcomes, iii) healthcare interventions, and iv) drug safety in real-world settings. The ultimate goal is to support robust, data-driven evidence for public health and clinical decision-making.

CPRD data is regularly updated and can be linked to secondary care data (for example, hospital data), disease registries (for example, MINAP), and mortality records (ONS)- thus offering a complete view of the patient journey. The CPRD manages two key databases, CPRD GOLD and CPRD Aurum, which contain anonymised EHRs from GP practices. The choice between these databases or to use both depends on the study research questions and period. In this chapter, I will discuss CPRD Aurum, outlining why I selected it for my research, its key features, how I utilised linked data to address my identified research gaps, and an overview of acquiring access to it ¹²³⁻¹²⁵.

3.2. How is healthcare managed in the UK?

How the health care system works in the UK plays a vital role in describing the usefulness of this data and why I chose to use CPRD databases to address my research objectives. The National Health Service (NHS) is a publicly funded system established in 1948 and is the backbone of the UK's healthcare system. The NHS operates under a single-payer model, providing free healthcare services to all UK residents at the point of use, primarily funded through general taxation¹²⁶. The NHS plays an integral role in delivering healthcare across primary care, where GPs serve as the first point of contact, and secondary care, which includes specialised services and hospitals.

The NHS comprises four distinct entities in England, Wales, Northern Ireland, and Scotland, each with its organisational structure but adhering to a unified framework. In this chapter, we will not dive into details about these entities^{126,127}.

The intent of the data recording directly impacts the research objectives and what can be pursued using these databases. This begins with recognising how healthcare services are delivered across different NHS settings—such as primary versus secondary care—and how the NHS's overall structure influences the data collected. Understanding data nuances provides valuable insights that enhance the quality of the research. For example, IPF treatments such as antifibrotic drugs which are prescribed by specialist hospitals and are therefore outside of GP care may not be fully captured within the CPRD databases or in hospital data; this is important to consider when considering the appropriate research question for this data.

The two main types of care settings are briefly described below.

3.2.1. Primary care

3.2.1.1. What is the role of GP practice in the UK?

Primary care in the UK is the cornerstone of the healthcare system, serving as the first point of contact for individuals seeking medical assistance outside of emergencies. GPs, pharmacists, and dentists offer a broad range of essential healthcare services. GPs are crucial in diagnosing and treating common ailments and managing chronic conditions like diabetes and COPD. They are also responsible for creating care plans and

conducting regular health assessments to monitor and adjust treatments as needed. In addition, GPs are essential to preventive healthcare, providing vaccinations, health screenings, and personalised lifestyle counselling to reduce the risk of chronic diseases such as hypertension and heart disease. The dataset reflects the UK population in terms of socioeconomic measures, age, sex, and ethnicity, providing a reliable foundation for public health research¹²⁸. Around 98% of the UK population is registered with a GP¹²⁵, and they are essential to coordinating and managing individual healthcare needs. Acting as gatekeepers, GPs ensure timely and appropriate care for individuals by referring them to specialists or additional services when necessary. They also coordinate patient care across various healthcare providers, including hospitals and allied health professionals, to maintain continuity of care and comprehensive management of a patient's health needs. Typically, when a patient has a health concern that requires medical attention, their first assessment is scheduled with their GP to determine whether specialist input or further tests are required. Overall, GP practices are integral to the NHS, offering accessible, comprehensive, and continuous care that supports the health and well-being of individuals in communities.

CPRD data can be linked with other datasets, such as HES datasets, to provide comprehensive information on hospital admissions, outpatient visits, and emergency care, thus facilitating the monitoring of patient journeys across different care settings. Additionally, linkages with disease-specific registries, such as cancer registries, improve the understanding of diseases and outcomes in the high-risk population. CPRD is also often integrated with ONS to address survival, causes of death, and mortality-related aims. I used both HES and ONS in my research¹²⁹.

3.2.1.2. How does CPRD work with GP practices?

CPRD works closely with GP practices across the UK to capture routinely collected patient data. This process begins when a GP practice joins CPRD, with practices using one of two primary software systems for recording patient information: Vision or EMIS (Egton Medical Information Systems). For example, if a patient visits their GP with a headache, the GP will assess the situation and may refer the patient to a specialist. To document this visit, the GP uses their software system (Vision or EMIS) to record the assessment. When entering "headache," the GP does not have just one option—there is a range of coding

possibilities from a drop-down menu, with options varying from general to specific types of headaches, as shown in Table 3.1. Based on the assessment, the GP selects the most appropriate code.

There is a difference in the quality of coding between conditions mainly managed in primary care and those like IPF, which tend to be diagnosed and managed in secondary care¹³⁰. For certain chronic conditions, there is incentivisation that a GP gets rewarded (as part of programs like the Quality and Outcomes Framework designed to improve the quality of care and health outcomes for patients, particularly those with long-term chronic conditions) for accurately identifying and monitoring patients with a specific condition, for example, diabetes^{131,132}. Therefore, certain conditions are better recorded than others. CPRD attempts to mitigate quality issues by employing data handling and cleaning steps (described later in this chapter). As with the example of headache, the GP diagnosed it in primary care. However, for chronic conditions like IPF, the diagnosis is by a specialist. So how do GPs record this information? Discharge letters are one of the mechanisms by which information from secondary care is entered into GP records; how efficiently this is translated and recorded depends on the practice. Hence, a researcher must understand how and when certain things get translated and coded.

Another point to consider is the timeliness of when a record shows up in CPRD. There is a difference between when an event occurred in secondary care versus when it was logged into CPRD – again, relying on the efficiency of the practice. Therefore, to ensure I am identifying the 'right' patient population, liaising with clinicians to understand patient treatment and management is necessary, which will facilitate the interpretation of recorded data. When using this data for research, it is essential to understand how GPs enter and code the information, as the accuracy and specificity of these codes significantly impact the quality of the data available for analysis. Thus, validation studies examining the reliability of diagnostic disease particularly those conditions primarily managed in primary care are foundational, a previous systematic review revealed that using simple lists of clinical diagnostic codes from primary care records achieves acceptable positive predictive values (PPVs) for most diseases and conditions¹³³.

When a practice agrees to contribute data to CPRD, the coded part of the patient's health records, including data on deceased patients and those who have left the practice, is shared – only for the patients who have not

opted out to share their information. The data collected includes demographic information (year of birth, sex), records of clinical events (medical diagnoses, signs, and symptoms), referrals to specialists and other secondary care settings, prescriptions issued in primary care, records of immunisations and vaccinations, diagnostic testing, and lifestyle information (body mass index, smoking, and alcohol consumption)^{129,134}.

However, it is important to note that CPRD only captures data from the current GP practice a patient is registered with; if a patient changes practices, past medical history (previous records from a practice) is not automatically transferred, requiring a review appointment for the new GP to document pre-existing conditions.

3.2.2. Secondary care

Secondary care in the UK comprises specialised healthcare services delivered by hospitals and other medical facilities, including specialist consultations, diagnostic tests, surgeries, and inpatient care. These services are typically accessed through referrals from primary care (GP). Unlike primary care, which handles common medical issues and chronic condition management, secondary care involves specialised medical treatments from experts across various disciplines, such as cardiology, oncology, orthopaedics, and neurology.

Specialists diagnose conditions such as ILD. Coordination with primary care is essential for ensuring continuity of care, with secondary and tertiary care providers working closely with GPs to manage referrals, treatment plans, and post-hospital care or rehabilitation services.

For example, when a patient is diagnosed with IPF by a specialist, the specialist and GP must coordinate to ensure holistic and integrated care for the patient. As mentioned above, it is vital to understand how and what information from secondary care is translated to primary care. For instance, in the case of IPF diagnosis, if an ILD specialist has requested a HCRT, in the radiology data, we may see a test was conducted but not the results for it. Therefore, like primary care, secondary and tertiary care is a crucial part of the UK healthcare system that offers specialised treatments and supports patients referred from primary care for more intensive medical interventions, but to be able use the data, it is important to know how it was recorded. Referrals to secondary care or specialists are documented, however, we cannot see what happens at the specialists, but we know the patient had a specialist referral^{126,129}.

3.3. CPRD datasets and which data is suitable for my research?

CPRD offers two primary care databases: GOLD and Aurum. GOLD collects patient-level data using the VISION software, while Aurum uses EMIS¹³⁴. When deciding which database to use for research, it is important to understand their key differences. CPRD regularly collects data from contributing GP practices, which are compiled into database builds—snapshots of the data at specific time points. Early in my research, I assessed these builds to examine the number of contributing practices, patient lives covered, and how well the data represented the UK population.

For example, as of July 2023, CPRD GOLD included data from nearly 21 million patients across 984 GP practices in the UK. In contrast, CPRD Aurum, as of July 2023, contained data for approximately 45 million patients. Aurum has more contributing practices (1,721 in Aurum versus 984 in GOLD) and patient lives, making it a larger dataset than GOLD. Additionally, Aurum covers a broader section of England and continues to grow as more practices switch from VISION to EMIS software. In contrast, the number of practices contributing to GOLD has declined since 2017¹³⁵.

Another important consideration is the difference in data structure and coding systems between the two databases due to the software platforms used. This required careful planning regarding the design of my research studies and patient cohorts. Since I am studying a rare disease like IPF, accessing a larger dataset was crucial for identifying as many relevant patients as possible. Overall, while both databases strive for high data quality and completeness, Aurum benefits from newer technology, ongoing updates, and a larger dataset, which can contribute to more reliable and comprehensive data for research purposes - therefore, I chose Aurum as the more suitable database for my research.

3.4. What information is available in Aurum data?

As described in the earlier section, CPRD Aurum provides a wealth of anonymised patient data collected from general practices in the UK that use the EMIS software system. Aurum utilises SNOMED CT for clinical coding, a more comprehensive and detailed coding system for diagnoses and medical events. This data

includes details of every GP interaction, such as consultations, phone calls, reminders, routine visits and urgent care appointments¹³⁶. This database is updated quarterly.

What is the structure of Aurum?

The information is provided is a set of eight types of data files. These eight main dataset files are shown in Figure 3.1.

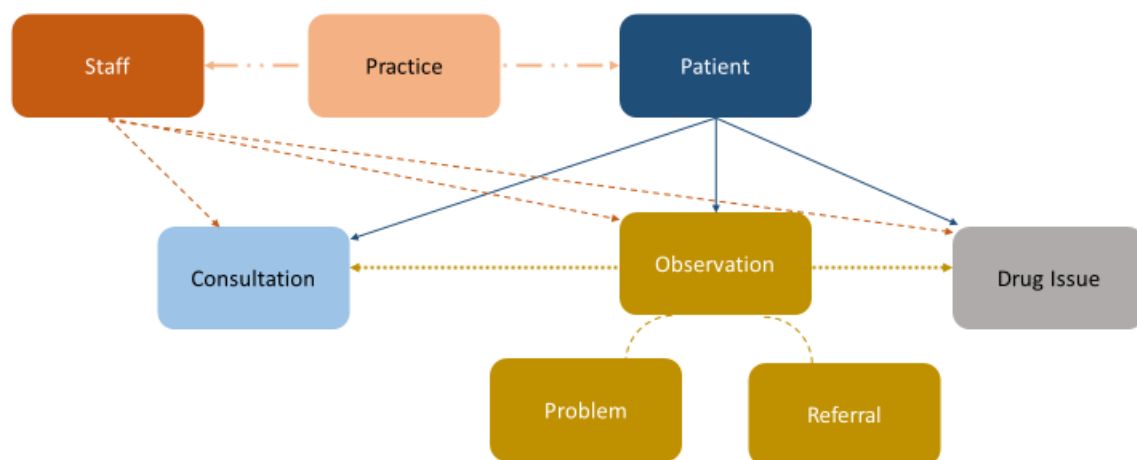


Figure 3. 1 CPRD Aurum dataset structure

Patient: This dataset encompasses patient demographics and registration particulars. This central file contains essential patient demographic data and distinct, de-identified patient identifiers for linking event files. The patient file encompasses details regarding patient enrollment with the general practice, date, and reasons for leaving practice when applicable and recording the date of death if a patient died. Important variables within this file include i) the current registration date, denoting the date of the patient's current general practice registration; ii) the transfer out date, representing the date of departure from a CPRD-affiliated general practice, for example if a patient relocated and joined a new practice that does not contribute to Aurum or due to the patient's passing; iii) the last collection date, which is the date of CPRD's most recent data collection from the general practices and iv) acceptable and regular patient types, acceptable" patients denote those

whose data is deemed suitable for research use, while "regular" patients have a consistent record of interactions with their GP practice, supporting longitudinal analysis.

Practice: This dataset encompasses practice details, including the practice identifier, region, and the last collection date. This practice file can be linked to the patient file using a unique practice number, and geographic location. The last 4-5 digits of the patient's unique pseudonymised identifier are the practice ID. Aurum data at the time of this analysis was limited to GP practices in England only. CPRD and researchers do not have access to specific practice names and locations, ensuring the protection of patient anonymity.

Staff: This dataset includes the job category associated with each staff member within CPRD.

Consultation: This dataset comprises data about the nature of consultations recorded by GPs, specifying types like telephone, home visit, or practice visit. Specific consultations are connected to observations during the consultation using a unique identifier (consid). The consultation file not only encompasses details about the consultation type, entered by the GP, but also allows for internal linkage (within Aurum) to related observations through the consultation identifier. Additionally, it is linked to the staff member who conducted the consultation, which can be found in the staff file.

Observation: This dataset contains medical history information input into the GP system (EMIS Web®) including symptoms, clinical measurements (for example heart rate, blood pressure, body temperature), laboratory tests ordered, and diagnosis). All data in the Aurum dataset is coded, so you cannot directly extract specific values like heart rate from the file. Instead, each value is associated with a Medcode ID and requires a lookup file to determine the units and details of the value listed in the dataset. As shown in Figure 3.1, observations occurring during a consultation can be linked via the consultation identifier. Aurum data are organised in a long format that is several rows per subject, and these multiple observations are associated with a main/ parent observation. For example, several rows of measurements of systolic and diastolic blood pressure will be grouped via a parent observation for blood pressure measurement.

Different datafiles linked to the observation file via obsid:

Referral: This dataset contains referral details recorded via the GP system. Information about the referrals normally to secondary care such as a hospital or specialist is recorded in the GP system. This data in the referral file can be linked to the parent observation file using the identifier, 'obsid'. The referral file contains information for both outbound and inbound patient referrals to or from external care centres.

Problem: This dataset contains information on the patient's medical history that has been defined by the GP as a 'problem'. GPs utilise these 'problems' to handle chronic conditions by facilitating the group of clinical events (such as prescriptions and clinical measurements) based on the primary issue rather than following a chronological order by consultation date. These 'problems' include the problem's significance and expected duration. Information on identifying the problems and its details for it can be found in this data. To obtain the full problem record (including the clinical code for the problem), problems should be linked to the Observation file using the observation identifier (obsid).

Drug: The dataset comprises information about all prescriptions within the GP system, encompassing medications and appliances. This file contains data regarding prescriptions issued by GPs, some connected to observations of specific issues through the Observation file, utilising the observation identifier ('obsid').

3.5. What are the strengths and weaknesses of Aurum data?

3.5.1. Strengths

- **Real-World Data:** Aurum provides one of the largest observational data sets based on routine clinical practice in the UK. This data reflects actual physician behavior and patient pathways, offering valuable insights into disease epidemiology, treatment patterns, and physician behavior in real-world settings.
- **Database Size and Representativeness:** Aurum includes records for millions of patients, covering a significant portion of the population of England [Aurum 2019]. Its large scale ensures representativeness across key demographic variables such as age, sex, and socioeconomic status. The dataset also captures a wide range of information, including lifestyle factors like smoking and alcohol use, providing a comprehensive picture of patients' health and care.

- **Data Quality and Completeness:** CPRD implements a set of data quality and completeness checks for the Aurum dataset. These processes include checking the format and structure of data elements, ensuring all clinical events are linked to individual patients, and assessing the completeness of essential information such as date of birth and registration status. Additionally, CPRD introduces quality markers, like the 'regular' and 'acceptability' flags, which suggest the consistency of key variables. Over 900 routine quality checks are conducted to validate coding accuracy and ensure temporal consistency of events¹²⁸.
- **Data Linkages:** One of Aurum's key strengths is its ability to link with secondary care datasets. As described above, these linkages provide a holistic view of the patient journey.
- **Ongoing Data Collection:** As the newer of the two CPRD databases, Aurum benefits from more registered patients than GOLD. GP practices contribute data on a daily basis, with updates made available for research on a quarterly basis. This ensures that the dataset remains current and reflective of ongoing clinical practices, making it ideal for longitudinal studies and timely research.

3.5.2. Weaknesses

- **Missing Data:** The integrity of data recorded in Aurum depends on the information inputted by GP practices. Any information that is not coded or entered will be unavailable for analysis. This reliance on coding can lead to significant gaps, especially for nuanced health details that may not be routinely documented. Furthermore, Aurum does not include free-text entries due to concerns about data privacy and the risk of re-identification, which limits the richness of the data.
- **Variability in IT Systems and Coding:** Researchers must consider potential variations in coding practices among GP practices and how they may change over time. These discrepancies can affect the consistency and reliability of the data, complicate analyses, and potentially lead to misinterpretations. Standardisation issues may arise, particularly when interpreting data points that have been coded differently across practices.
- **Quality Assurance:** While CPRD performs various checks to ensure that the dataset is high quality, it is important to note that CPRD does not conduct any comprehensive data cleaning. For instance, the organisation may not verify whether recorded values, such as blood measurements, are valid or

logical. This lack of cleaning places the responsibility on researchers to assess the validity and interpretability of the data, which can introduce uncertainty into the research outcomes.

- **Case Ascertainment Challenges:** There may be instances where relevant information is captured but needs to be explicitly coded within the primary care data. For example, while a hospital discharge letter may be scanned and saved as part of the patient's record, the details may not be coded, leading to the loss of critical information. Similarly, variables like BMI or ethnicity might be recorded only in the context of specific health conditions, which can create biases in the data. Researchers must be acutely aware of these limitations when designing studies using Aurum data¹³⁷.
- **Need for Preliminary Data Cleaning:** When working with Aurum data, researchers often need to apply basic data-cleaning steps to ensure the accuracy of their analysis. For example, validating that a patient's date of birth precedes their registration at the GP practice is a crucial check. Such preliminary steps are necessary to enhance the dataset's reliability before conducting any further analysis.

In summary, while CPRD Aurum provides valuable insights and data for research, these weaknesses must be considered. While working with this data, I applied additional cleaning steps such as making sure the patient had a date of birth that predates the registration at the GP practice, if an IPF diagnosis is prior to date of birth or patient registration date at the practice then I did not consider it. Hence, understanding the limitations and challenges of the dataset is essential for designing studies (eligibility criteria) and accurately interpreting their findings.

3.6. What linked data are available, and which did I use for my research?

A subset of participating GP practices has consented to participate in CPRD's data linkage scheme (> 95% of acceptable patients), which enables the integration of primary care data with other healthcare data and selected disease registries. This linkage is managed by NHS Digital (formerly the Health and Social Care Information Centre), a trusted third party (TTP) that ensures patient anonymity throughout the process. GP practices provide de-identified patient data and identifiers such as NHS numbers, postcodes, dates of birth, and sex to the TTP, which then matches these identifiers with data from external sources¹³⁸. The TTP creates a lookup file to link the two datasets without revealing personal identifiers to any individual party (see Figure 3.1).

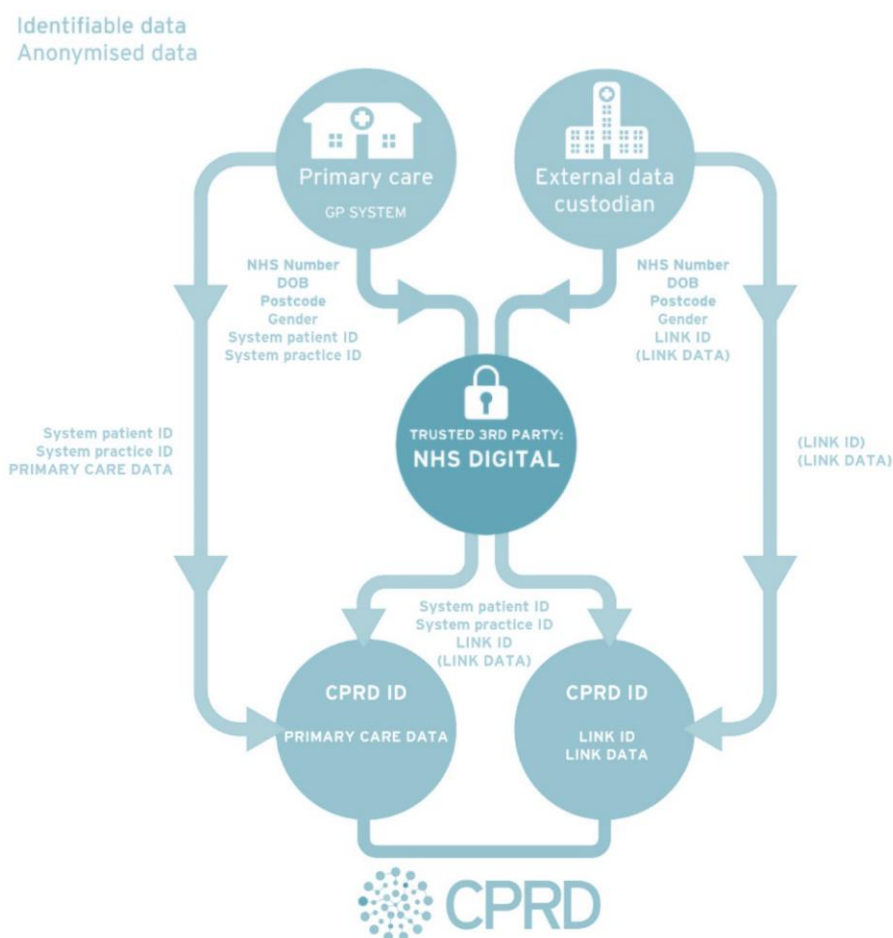


Figure 3. 2 Data linkage process schema ¹²⁹

The linkage is limited to GP practices in England that actively opt into the CPRD scheme. Linkage for GOLD to HES is about 60% and for Aurum to HES is higher at about 80% ¹³⁵. Among the datasets available for linkage, Hospital Episode Statistics (HES) and ONS mortality data are the most frequently used. HES provides detailed clinical and diagnostic information about hospital admissions, procedures, and serious illnesses, while ONS data offers detailed causes of death. These linkages extend the utility of electronic primary healthcare data in several ways. Firstly, linkage allows researchers to combine the longitudinal medical records provided by GPs with secondary or specialist care records to provide a far more detailed, comprehensive picture of an individual's medical history and health trajectory. Additionally, linked data sources can also be used to validate information within CPRD (primary care), or to incorporate additional data that are not part of CPRD or are not typically well recorded within CPRD. An example of linkage in this

analysis is assessing mortality rates for IPF. For this purpose, I used the Aurum data linked to ONS records to determine the mortality rate for IPF from 2008 to 2018. In addition to identifying IPF cases, secondary care data serves as supporting or confirmatory evidence for events observed in primary care. For instance, in the following chapter, I used death records with IPF listed as the underlying cause as a reference standard to verify whether individuals diagnosed with IPF in primary care had the condition.

However, it was important to understand the record linkage methods and each dataset's strengths and limitations to ensure the robust use and appropriate interpretation of linked data.

I have used linked HES and ONS data to validate IPF diagnosis (chapter 4), assess the incidence, prevalence, and mortality of IPF (Chapter 5), and assess the survival of patients with IPF (chapter 6). Thus, to provide some necessary background to the methodological aspects of these studies, a brief description of these two data sources is provided in the following subsections.

3.6.1. Hospital Episodes Statistics

Hospital Episode Statistics (HES) provides a comprehensive collection of records related to all admissions, outpatient appointments, and A&E attendances at NHS hospitals in England, dating back to the late 1980s. Initially, the primary purpose was to capture hospital activity for administrative purposes and to ensure hospitals are appropriately remunerated for the care provided. Over time, this data has been repurposed for a wide range of secondary uses, including healthcare planning, policy development, and research. As CPRD processes the primary data, HES data is processed by the Secondary Uses Service (SUS) at NHS Digital, which transforms raw patient-level data into a format that can be used for research.

HES data are updated monthly. This data is collected for each patient interaction with healthcare services, whether it is an outpatient appointment, inpatient stay, or attendance at an accident and emergency (A&E) department. A patient's distinct phase (s) of care during a hospital stay are captured through several specific data points and processes. An episode is when a patient is under the care of a single consultant or specialty team. An episode begins when a patient is admitted or transferred to the care of a consultant and ends when they are discharged, die, or are transferred to another consultant or department. For each episode, details like the consultant's specialty, admission and discharge dates (or transfer), diagnoses, and procedures are recorded. A spell is a collection of one or more episodes that happen during a continuous inpatient stay. A spell begins when a patient is admitted to the hospital and concludes when they are discharged from the hospital entirely,

regardless of how many different consultants or specialties the patient was transferred between during their stay. Each episode in the spell is recorded separately, capturing the transition of care and any changes in treatment or diagnosis as the patient moves through different stages of their hospital stay. This granular capture of episodes ensures that HES data can track the whole patient journey, providing insights into care pathways, resource utilisation, and clinical outcomes. For instance, if a patient is admitted for surgery, their spell would include multiple episodes—such as pre-surgery care, surgery under a surgical team, and post-surgery recovery under a different consultant. Each would be captured individually but linked as part of the same spell in the HES dataset.

Over the years, the methods of collecting hospital data have evolved with organisational changes in the NHS. Currently, HES data covers records of admitted patient care (APC) from the financial year 1989/90 onwards. The records are updated monthly and are organised as finished consultant episodes (FCEs). Each FCE represents a period of care under a single consultant, and multiple FCEs can occur during a single hospital "spell" or stay, indicating transfers between different teams (e.g., from A&E to a specialised ward).

Clinical Information: Each FCE can include up to 20 clinical diagnoses, coded using the ICD-10 system (International Classification of Diseases, 10th revision). The diagnosis listed in the first position is the primary diagnosis—the main reason for hospital admission. Secondary and subsequent diagnoses often represent comorbidities or underlying conditions, and they are recorded to ensure complete patient health status documentation. For instance, chronic diseases like COPD are often recorded as comorbidities even when they are not the primary reason for admission, as these conditions are associated with higher reimbursement rates for the hospital.

Demographic Information: Records also capture essential demographic details, such as age, sex, and ethnicity, allowing researchers and policymakers to analyse healthcare patterns across different population segments¹³⁹⁻

¹⁴¹.

Administrative Information: Besides clinical and demographic data, HES records include administrative details, such as waiting times, admission dates, discharge dates, and the admission method (e.g., emergency or elective).

Data on Outpatient and A&E Attendances

In addition to APC, HES includes records of A&E attendances and outpatient appointments¹³⁹. However, these datasets, known as AE (Accident & Emergency) and OP (Outpatient), need to be more detailed than APC.

- Outpatient data (OP) only provides information about the consultant's speciality for a given patient's visit, with limited clinical detail.
- A&E data (AE) offers more granularity than outpatient data but is constrained by the limited list of 56 possible diagnoses available for selection during an A&E visit.
- The history of these datasets varies: Outpatient data is available from 2003 onwards, while A&E data has been collected since 2007.

The Richness and Utility of HES Data

HES data contains information on hospital events, diagnoses, procedures, and hospitalisation details. All events are coded using **ICD-10** codes, allowing standardisation of diagnoses across hospitals. The data is structured by **hospitalisations**, where each hospitalisation may consist of multiple **care episodes**, reflecting different treatment or care phases under various consultants.

For example, during a single hospitalisation, a patient might be transferred from the A&E department to an acute care ward and later to a specialised unit. Each transfer generates a new **FCE** within the same hospital "spell."

3.6.2. Office of National Statistics (ONS)

Table 3. 3 List of linked datasets and the ones used in my research.

List of Linked datasets	Used in my research
Hospital Episode Statistics Admitted Patient Care (HES APC)	Yes

Hospital Episode Statistics Outpatient (HES OP)	Yes
Hospital Episode Statistics Accident and Emergency (HES A&E)	Yes
Hospital Episode Statistics Diagnostic Imaging Dataset (HES DID)	Yes
Office of National Statistics (ONS) Death Registration	Yes
National Cancer Registration and Analysis Service (NCRAS) data from Public Health England (PHE) including:	
Cancer registration data	
Cancer Patient Experience Survey (CPES) data	
Systemic Anti-Cancer Treatment (SACT) data	
National Radiotherapy Dataset (RTDS)	
Mental Health Dataset (MHDS) data	
Measures of relative deprivation and rural-urban classification at Lower Layer Super Output Area (LSOA) level for practices and patients	

3.7. What are the steps I followed to get access to CPRD data?

To extract data, the following steps are required:

3.7.1. ISAC Approval and Data Access

The process began with submitting a research proposal to CPRD's Independent Scientific Advisory Committee (ISAC) for approval. Once approved, I gained access to download CPRD Aurum data from CPRD's customer website. This prerequisite ensures that the research is scientifically and ethically sound.

Table 3. 4 List my applications to CPRD to request Aurum data for my analyses.

Protocol ID	Study Title
-------------	-------------

20_000068	An investigation of the reliability and granularity of the recording of interstitial lung diseases in primary and secondary care datasets
20_000108	To investigate the prevalence, incidence and mortality of Interstitial Lung Disease (ILD) in the UK for the study period 2009 to 2019
23_003093	Changes in survival and causes of death in people with idiopathic pulmonary fibrosis in England, 2010 to 2022

3.7.2. Two-Step Data Download Process

The data download is a two-step process: Define and Extract.

Step 1: Define

The first step is to define the cohort by specifying the characteristics of the group of interest (i.e., the source population). For my research, the focus was on individuals diagnosed with IPF.

To define this group, I prepared a code list of diagnostic codes (SNOMED concept codes) that identified patients with IPF. CPRD Aurum uses SNOMED CT codes and their corresponding CPRD equivalents (medcodes) to prepare essential code lists. In the next section, I have described how the code list was constructed.

Exclusion Criteria and Study Period

At this stage, patients can be excluded based on predefined criteria such as age, sex, or data quality flags (e.g., 'acceptable' records). For instance, I applied basic criteria such as a minimum age of 18 and excluded those with missing date of birth information. This was useful as IPF is a relatively rare condition, so I captured as many relevant cases as possible without overloading the data with extraneous entries. If a more common condition is being studied, such as COPD, a study period can be defined, including patients with at least one event of interest (e.g., an IPF diagnosis) during the specified timeframe. This helps streamline the dataset to only the relevant period for analysis.

Output from the Define Step

Once the defining criteria are established, CPRD produces a list of patient identifiers (patios) who meet the specified inclusion criteria. Next, I identified which Aurum patients were eligible for HES data linkage, narrowing the dataset to patients whose records existed in both datasets.

All these steps take place on the CPRD portal. I can access the data under the multi-study license with Imperial College London.

Step 2: Extract

After defining the cohort, the second step is to extract and download all relevant records for the patients identified in the defined step. Importantly, this involves extracting all available records for these patients, not just the specific events of interest, as this allows for a comprehensive view of patient history. This includes all clinical encounters, diagnoses, prescriptions, and lifestyle factors recorded in primary care. For my thesis, I was looking for all records for patients with a record for IPF.

Additional Steps for Hospital Data (HES APC)

My study also required hospital data from the HES, so I incorporated another step to integrate data from this source.

HES Linkage Process

CPRD handles the HES linkage process since researchers do not have direct access to HES APC as they do with Aurum. After defining the patient cohort in Aurum, I requested that CPRD provide a list of patient IDs with IPF records in HES data. This ensured I captured patients diagnosed with IPF in secondary care, if not within primary care.

All requests for HES data are handled via the erap portal within the approved study remit.

De-duplication of Patient IDs

To avoid redundancy, I removed duplicate patient identifiers across the Aurum and HES datasets, ensuring I had a unique list of patients for further analysis.

Final Data Extraction

After cleaning and de-duplication, I submitted a request to extract all relevant data from both Aurum and HES for the unique list of patients. To capture a comprehensive clinical history, the extraction included all patient

records across the two datasets, not limited to specific events or periods. This included primary care interactions, hospital admissions, and other medical encounters.

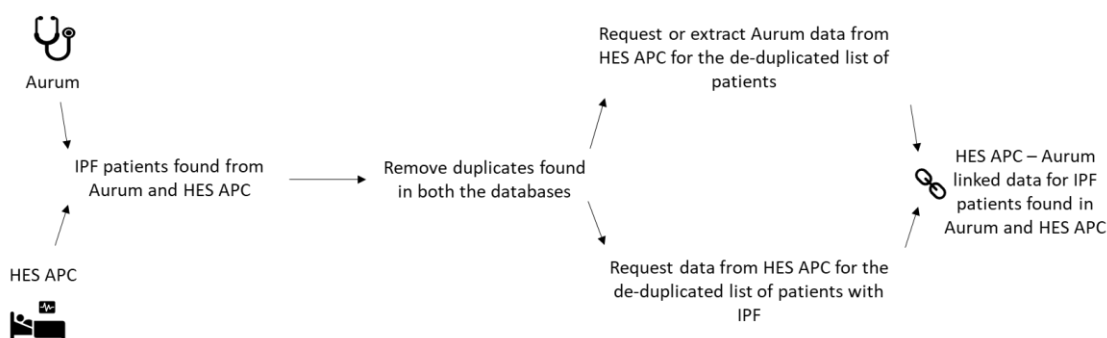
Data Preparation and Processing

Once the data was downloaded, I needed to perform routine data manipulation, such as restructuring and cleaning the dataset, before it could be used for statistical analysis. This step involved removing irrelevant data, handling missing values, and ensuring the data was in the correct format for further processing.

An overview of this process is shown in Figure 3.3.

Figure 3.3. Steps followed to find IPF patients in NHS universe

I followed the steps to get linked data to find IPF patients in the NHS universe.



3.7.3. Constructing code list

I followed an established methodology to create a comprehensive code list (<https://github.com/NHLI-Respiratory-Epi/SNOMED-CT-codelists>). Below is a detailed step-by-step approach to how I developed an IPF code list.

- Identify Search Terms

I began by identifying all relevant terms and synonyms associated with IPF. This step was crucial for ensuring comprehensive coverage of the topic to include new and dated terms. For instance, given the study period

from 2010 onwards, I listed dated terms like ‘cryptogenic fibrosing alveolitis’, generic terms such as ‘pulmonary fibrosis’, and disease-specific terms such as ‘idiopathic pulmonary fibrosis’.

- Create Search Terms

I constructed "search terms" that helped locate each identified term. Wildcard characters (*) accounted for variations in spelling, capitalisation, hyphenation, and other grammatical differences. For example, to search for idiopathic pulmonary fibrosis, I used the following search terms: “idio. pulmonary fibrosis,” idiopathic pul*,” and “pulmonary fib*.” In the literature, I have observed that IPF is defined as ‘broad’ and ‘narrow – so I created similar terms to include broad and narrow IPF terms.

- Search the Medical Terminology Dictionary

As the next step, I imported a medical terminology dictionary with SNOMED CT term descriptions and their Description ID and Concept ID. Using the search terms developed above (in step 2). After completing this search, I retained the SNOMED CT terms that matched at least one of your search terms.

- Manual Screening of Codelist to Remove Undesired Terms

I manually reviewed the generated code list, removing any undesired codes based on their Description ID (or miscoded if using CPRD Aurum).

- Use SNOMED CT Concept ID to Find Additional Synonymous Terms

Using the SNOMED CT Concept ID, which represents a single clinical concept that may have multiple associated terms. For example, the concept of "fibrosis" encompasses synonyms like "fibrosis of lung".

Using the Concept IDs from my current code list, I searched for further synonymous terms to expand my code list.

- Review by experts

These codes were independently reviewed and rated by the clinicians with expert knowledge of ILD as “yes” for the codes which were strongly indicative of an IPF diagnosis, “maybe” for codes that conceivably might suggest IPF (to allow for variability in coding practices) and “no” for highly unlikely that a person has IPF. Codes rated as “no” were rejected and not used further in this analysis. This categorisation is aligned with how IPF is reported in the literature.

Chapter 4

Validation of the recording of IPF using primary and secondary care data in England

Overview

Routinely collected healthcare data are invaluable for epidemiological research, although validation studies are sparse for diseases predominantly managed in secondary and tertiary care, such as idiopathic pulmonary fibrosis (IPF). Using the UK's Clinical Practice Research Datalink (CPRD) Aurum dataset, which integrates primary care records with hospital admissions and death records, I evaluated the positive predictive value (PPV) of IPF diagnostic algorithms. These algorithms, developed based on IPF guidelines and literature using SNOMED-CT or ICD-10 codes alone or combined with additional clinical information, were assessed against death records as the gold standard. This study observed changes in IPF coding practices over the 2008-2018 study period. Among 17,559 individuals with at least one IPF record across the datasets, algorithm PPVs ranged from 64.4% (95% CI: 63.3–65.3) for a broad code set to 74.9% (95% CI: 72.8–76.9) for a narrow, highly-specific code set. Adding confirmatory evidence, such as CT scan results, improved the PPV of narrow code-based algorithms to 79.2% (95% CI: 76.4–81.8), albeit at a reduced sensitivity below 10%.

Similarly, incorporating evidence of hospitalisation enhanced PPV (78.4% vs. 64.4%) while increasing sensitivity (53.5% vs. 38.1%). The study underscores a shift towards more specific IPF codes over time. These findings support using a focused IPF code set for higher diagnostic validity, though the trade-off between increased accuracy and reduced sample size warrants consideration. For broader studies, an algorithm combining a comprehensive IPF code set with evidence of hospitalisation offers a balanced approach. This chapter is published – *Validation of the recording of idiopathic pulmonary fibrosis in routinely collected electronic healthcare records in England*: <https://doi.org/10.1186/s12890-023-02550-0>.

4.1. Background

IPF is the most common and aggressive type of pulmonary fibrosis and the most established ILD condition. It is recognised as the most prevalent and aggressive form of pulmonary fibrosis globally²³, garnering substantial research attention due to rising incidence and prevalence rates reported not only in the UK but also internationally. In Chapter 1 and Chapter 2 of my thesis, I highlighted the wide-ranging prevalence (7-42 per 100,000 persons) and incidence (2.4-8.7 per 100,000 person-years) estimates of IPF in the UK. I elaborated on the complexities in meta-analysing various ILDs. Even though IPF is well-researched compared to other ILDs, there is a wide range of inconsistencies in reporting practices at global and local scales.

So far in my research, I have covered the clinical viewpoint of IPF in Chapter 1 and in detail the burden of ILD, particularly IPF in Chapter 1 and Chapter 2. IPF has garnered substantial research interest due to its progressive and aggressive nature, the approval of antifibrotic treatments, and the increasing evidence of its rising prevalence and incidence rates, both in the UK and worldwide. Despite continued efforts, the reliability of previously published estimates of IPF incidence and prevalence in the UK—primarily based on primary care data^{13,142}—remains under scrutiny. Additionally, these studies are now outdated (up to 2012) and were conducted before the introduction of antifibrotic therapies and other critical updates in the field of IPF.

Electronic health records in the UK serve as a rich resource for studying IPF, as they offer a large sample size and have longitudinal data on patients' medical history. Datasets such as CPRD Aurum provide a detailed view of patients' journeys¹²⁵, offering more comprehensive insights compared to commercial databases in the US, like claims databases. Despite the wealth of data available, there is a critical need for a validation study to address several key issues related to identifying and diagnosing IPF within these datasets. Over the past two decades, numerous studies have utilised various electronic health record databases to investigate IPF.

However, a significant challenge remains: no standardised definition or consistent list of diagnostic codes is used to identify IPF in large databases. The diagnostic codes used to identify IPF patients vary from study to study, leading to potential inconsistencies in data and findings.

Moreover, to the best of my knowledge, a study has yet to be conducted in the past ten years to update the disease burden of IPF in the UK despite evolving diagnostic criteria and increasing recognition of the

condition. As discussed in my first chapter, the diagnosis of IPF has undergone significant changes, mainly influenced by the guidelines from the ATS and the ERS. These guidelines have been crucial in refining the terminology and classification of ILDs, especially IPF. The latest 2022 ATS/ERS/JRS/ALAT Clinical Practice Guideline¹¹ underscores the importance of a multidisciplinary approach to diagnosing IPF, with continued reliance on HRCT as a critical diagnostic tool. The diagnosis of IPF now involves a combination of clinical judgment by a multidisciplinary team, patient history, physical examination, and HRCT assessment. These advancements have significantly impacted the diagnosis and management of IPF, highlighting the need for accurate and updated data to reflect these changes.

Given the complexities of diagnosing IPF and the evolving nature of the field, it is challenging to capture these nuances when using retrospective data to identify IPF patients. Retrospective evaluations may misclassify patients, mainly relying on commonly used definitions like usual interstitial pneumonia (UIP), which may not fully align with the most current understanding of IPF. The best methods for identifying IPF patients and improving the accuracy of such identifications in EHRs are not well-defined, particularly in the UK. A single validation study published in 2000 demonstrated a high positive predictive value (PPV) of 95% for diagnoses of cryptogenic fibrosing alveolitis (an earlier term for IPF) within the UK primary care database¹⁴³. However, this study needs to be updated, as it predates the transition to medical coding systems that are based on SNOMED CT terminology and codes. There is a need for new research that reflects the current diagnostic criteria and coding practices. However, several studies in the United States have focused on validating IPF codes based on the International Classification of Diseases (ICD) codes to assess the accuracy and reliability of using existing data to identify IPF patients^{33,144,145}.

In the UK, the introduction of NICE guidelines on IPF diagnosis and the availability of antifibrotics have likely increased disease awareness and incentivised physicians to accurately diagnose and manage IPF^{25,146-148}. However, the impact of these advancements on GPs coding practices remains unclear. There is limited information on the correctness and completeness of IPF recordings in primary care settings, whether the appropriate diagnostic codes are consistently used, and whether coding practices have changed over time. While routinely collected healthcare data are a valuable resource for epidemiological research, their usefulness depends on the accuracy and consistency of condition recording. Therefore, a validation study is essential to

evaluate the current practices, ensure the reliability of IPF diagnoses in EHRs, and ultimately improve the quality of research and patient care in this evolving field. Understanding the coding practices in primary care settings is crucial, particularly whether GPs predominantly use generic pulmonary fibrosis codes or codes specifically identifying IPF cases. These considerations raise important questions regarding the recording of IPF diagnosis in primary care settings, the accuracy of recording, the timeliness of case documentation, and discrepancies between the date of diagnosis and the initial appearance of diagnostic codes in databases, all of which influence the definition and identification of incident cases. Moreover, changes in coding practices over time, owing to disease name revisions and reclassifications over recent decades, are likely to affect the estimation and interpretation of trends in IPF prevalence, incidence, and mortality. Therefore, in this chapter, I focused on assessing the quality and reliability of recording IPF in primary (CPRD Aurum) and secondary care data (Hospital Episode Statistics, HES) using Office of National Statistics (ONS) mortality data as the gold standard in the UK. Specifically, I set out to determine whether reliance on a list of clinical diagnostic codes alone was sufficient to find an IPF case in the primary care data or whether I needed further evidence from radiology or hospital admission data to improve confidence in IPF case finding. And lastly, I describe any discernible changes in GP coding for IPF over time.

4.2. Methods

4.2.1. Data sources

For this study, the Aurum – HES APC linked dataset served as the primary source of data. This dataset was linked to several HES datasets, including HES Admitted Patient Care (APC), HES Outpatient (OP), HES Accident and Emergency (A&E), and HES Diagnostic Imaging Dataset (DID) along with mortality data from the Office for National Statistics (ONS), renowned for its reliability in recording deaths in the UK.

In Chapter 3 (Data sources), I describe the structure, information, and variables provided in the Aurum database. For this analysis, I explored the use of the following datasets: Aurum, HES APC, HES OP, HES A&E, HES DID, and ONS. However, there were several key points to consider when using these data. In Table 4.1, I have summarised the features and limitations of each concerning my analysis. High-level details

on dataset specifics for this analysis can be found in Table 4.1; a comprehensive of these data is provided in Chapter 3: Data Sources.

Table 4. 7 Key Considerations for using the primary and secondary data sources

Data sources	Characteristics	Limitations
Aurum	Symptoms and clinical diagnoses, demographic characteristics, vaccination history, lifestyle and behaviors (e.g., smoking), lab test results, prescriptions, referrals to secondary care (SNOMED codes)	Aurum data lacks the severity of the condition and is unable to check whether the IPF diagnosis was delayed.
ONS	Date, place and cause of death, including underlying and contributory causes of death (ICD-10) This analysis used it as a gold standard; that is, if someone had IPF as the cause of death, this was conserved to be true.	ONS data may not capture individual-level variations or nuances. This can make it challenging to explore detailed trends or the impact of specific interventions on mortality. Misclassification or inaccuracies in the recorded cause of death can affect the data's reliability.
HES APC	Details of all admissions to NHS hospitals in England (including acute, mental, and primary care trust hospitals), including clinical diagnoses (ICD-10) and procedures performed (OPCS)	In-hospital treatments are not captured. Reliance on diagnoses captured at later positions typically reflects coexisting conditions or comorbidities and not the actual reason for admission.
HES OP	Date of outpatient event	A significant limitation of HES OP data is the absence of ICD-10 codes, which means I can identify that an outpatient visit occurred but cannot determine the specific reason for it. While I can see the consultant's specialty, such as respiratory or cardiology, the detailed information about the nature of the consultation is missing.
HES A&E	Date of emergency room visit	ICD-10 codes are not generally provided. Similar to HES OP, HES A&E data often lacks detailed diagnostic information. The data may include the reason for the A&E visit, but it typically lacks the specificity found in inpatient records, such as detailed ICD-10 codes. A&E data primarily captures acute events and emergencies, so it may not provide a comprehensive view of a patient's overall health, particularly for chronic conditions that are managed outside of emergency settings.

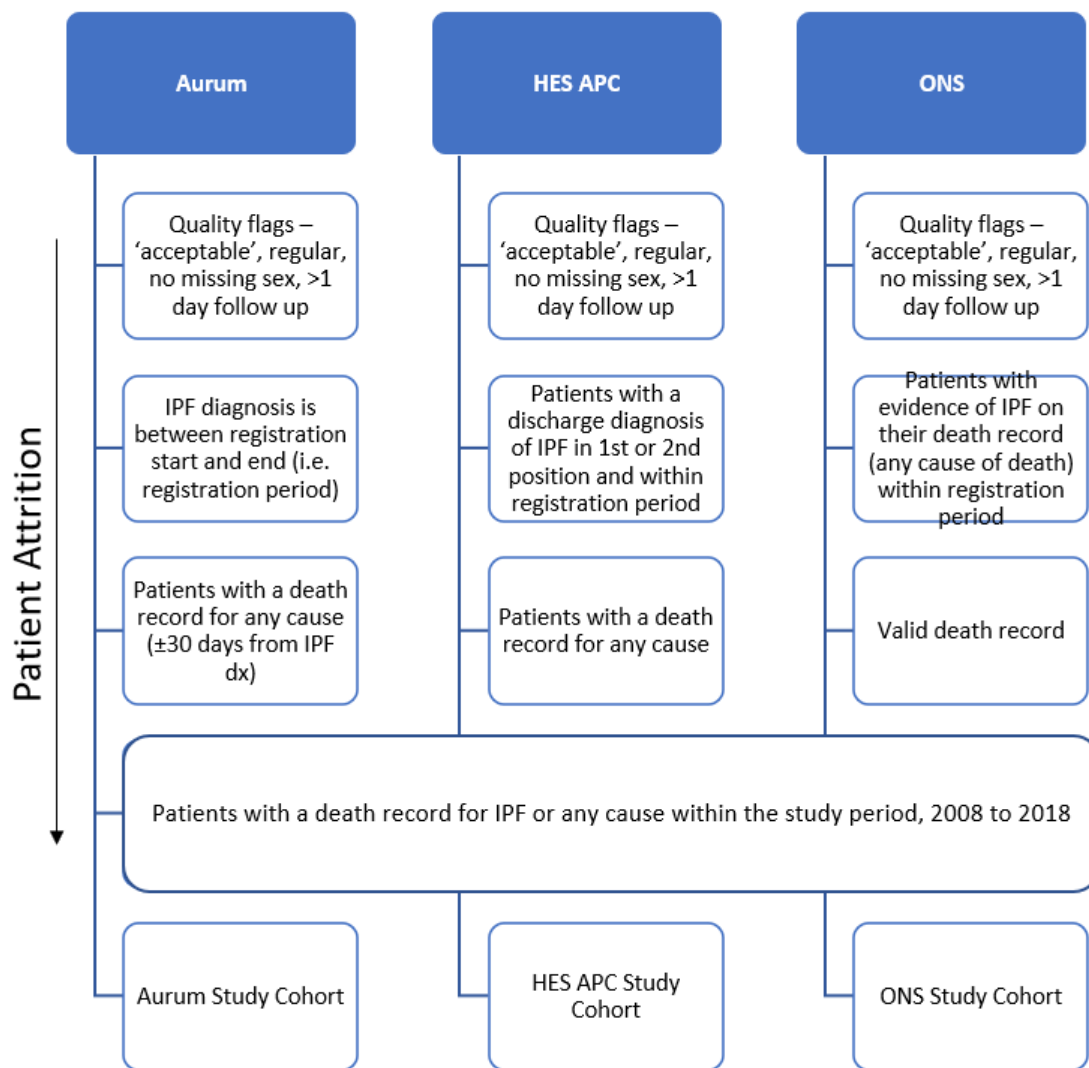
HES DID	Date and type of imaging	HES DID primarily capture information related to diagnostic imaging procedures, such as the type of scan performed and the date. However, it often lacks detailed clinical context, such as the reason for the imaging, specific diagnoses, or the outcome of the procedure. Without data on patient outcomes following imaging limits the usefulness of understanding the impact of the imaging on patient care.
---------	--------------------------	---

ONS, office of national statistics; DID, digital imaging dataset; OP; outpatient; A&E, accident and emergency

4.2.2. Study design and population

This study employs a historical cohort design utilising linked data from Aurum and HES datasets. The study period spans from 1 January 2008 to 31 December 2018, aligning with data availability, Aurum data up to December 2019, and HES data until December 2018, ensuring coverage across the two sources. The end date was 31 December 2018. Eligibility criteria were stringent: participants must have been aged 18 years or older at their initial IPF diagnosis, with linked Aurum data meeting quality standards of 'acceptable' and 'regular'. Additionally, eligibility required at least one record of IPF diagnosis in Aurum, hospital admission in HES APC with IPF noted as a primary or secondary diagnosis, or IPF recorded in ONS mortality data during the study period. Follow-up periods were established from the latest of the IPF diagnosis date, 1 January 2008, start of current registration, or 18th birthday, continuing until the earliest of 31 December 2018, date of death, last date of data collection at the practice, or patient transfer out date. Figure 4.1 illustrates the application of these eligibility criteria across Aurum, HES-APC, and ONS datasets, ensuring a robust selection of study participants and a comprehensive capture of relevant clinical and mortality data.

Figure 4. 2 Eligibility criteria for Aurum-HES APC – ONS linked datasets



4.2.3. Codelists

In Chapter 3 (section 3.7.3), I have described in detail the concept and importance of the code list. I also elaborated on the process of deriving a thorough list of codes. I followed those steps to generate an IPF codelist. This search encompassed old and newer IPF terms such as 'idiopathic pulmonary fibrosis', 'pulmonary fibrosis', 'cryptogenic alveolitis', 'idiopathic interstitial pneumonia', 'interstitial pneumonia', and 'Hamman Rich syndrome'. The text search terms were developed in consultation with ILD experts (JKQ and PMG). The resulting list integrated previously documented codes from literature (Strongman 2018) alongside discoveries from the text search. Two clinicians (JKQ and PMG) independently reviewed and categorised

these codes into three ratings: "yes" for codes strongly indicative of IPF, "maybe" for potentially IPF-related ("IPF-like") codes accounting for variability in coding practices, and "no" for codes deemed highly unlikely to represent IPF. Diagnostic clinical codes were classified in the literature as 'broad' and 'narrow' code lists^{13,33,144,149,150}. The broad list encompassed generic and less specific codes suggestive of any type of pulmonary fibrosis, not exclusive to IPF (but IPF-like), while the narrow list comprised highly specific codes strongly indicative of IPF diagnosis. Like published literature, I categorised my codes as 'broad' and 'narrow'. JKQ and PMG independently assigned codes to either the 'narrow' or 'broad' categories, with any discrepancies resolved through consensus among the research group. The narrow category represented a subset of the broader list, and Table 4.2 provides a detailed breakdown of all medcode IDs included in both categories. In total, 27 medcode IDs were classified as broad list and nine medcode IDs were classified as narrow. For **HES APC and ONS** data, IPF was found using the ICD-10 code J84.1, J84.8, or J84.9 to define IPF hospitalisations (HES APC) and/ or deaths (ONS). These codes were consistent with what was found in the literature^{33,144,150}.

Table 4. 8 List of Medcode ids included in the broad and narrow categories.

Medcodeid (Aurum code identifier)	Read code	SnomedCT concept id (not unique identifier)	Term	Narrow	Broad
909731000006117		909731000006101	[rfc] pulmonary fibrosis	0	1
196654011	H563.12	700250006	Cryptogenic fibrosing alveolitis	1	1
5054601000006119		233703007	Diffuse parenchymal lung disease	0	1
621121000006112	H563100	196125002	Diffuse pulmonary fibrosis	0	1
4782331000006111		196125002	DIPF - Diffuse interstitial pulmonary fibrosis	0	1
4782341000006118		196125002	Fibrosing alveolitis	0	1
3336391000006117		51615001	Fibrosis of lung	0	1
815341000006115	H563.11	236302005	Hamman - Rich syndrome	0	1
885481000006119	H563.99	45157009	idiopath. fibrosing alveolitis	1	1
12757781000006113		45157009	idiopathic fibrosing alveolitis	1	1
75302014	H563.00	700250006	Idiopathic fibrosing alveolitis	1	1
301709019	H563z00	700250006	Idiopathic fibrosing alveolitis NOS	1	1
12716191000006115		45157009	idiopathic interstitial pneumonia	0	1

1786131000006116	H563.13	700250006	Idiopathic pulmonary fibrosis	1	1
5054591000006110		233703007	ILD - Interstitial lung disease	0	1
107470010	H56y100	64667001	Interstitial pneumonia	0	1
3551861000006113		64667001	Interstitial pneumonitis	0	1
254002014	23E5.11	162974009	O/E - fibrosis of the lung	0	1
254003016	23E5.00	162974009	O/E - fibrosis of lung present	0	1
4555561000006111		162974009	On examination - fibrosis of the lung is present	0	1
5492481000006115		266368002	Post-inflammatory pulmonary fibrosis	0	1
350255018	H48..00	233749003	Progressive massive fibrosis	0	1
85960018	H563200	51615001	Pulmonary fibrosis	0	1
885491000006116	H563199	51615001	Pulmonary fibrosis	0	1
4782361000006119		196125002	UIP - Usual interstitial pneumonitis	1	1
7532701000006115		700250006	Usual interstitial pneumonia	1	1
2475606014	H563300	196125002	Usual interstitial pneumonitis	1	1

4.2.4. Sensitivity analyses

I conducted two sensitivity analyses: i) to decide on how to ‘accurately’ find patients in hospital records and ii) death records in ONS data as a gold standard.

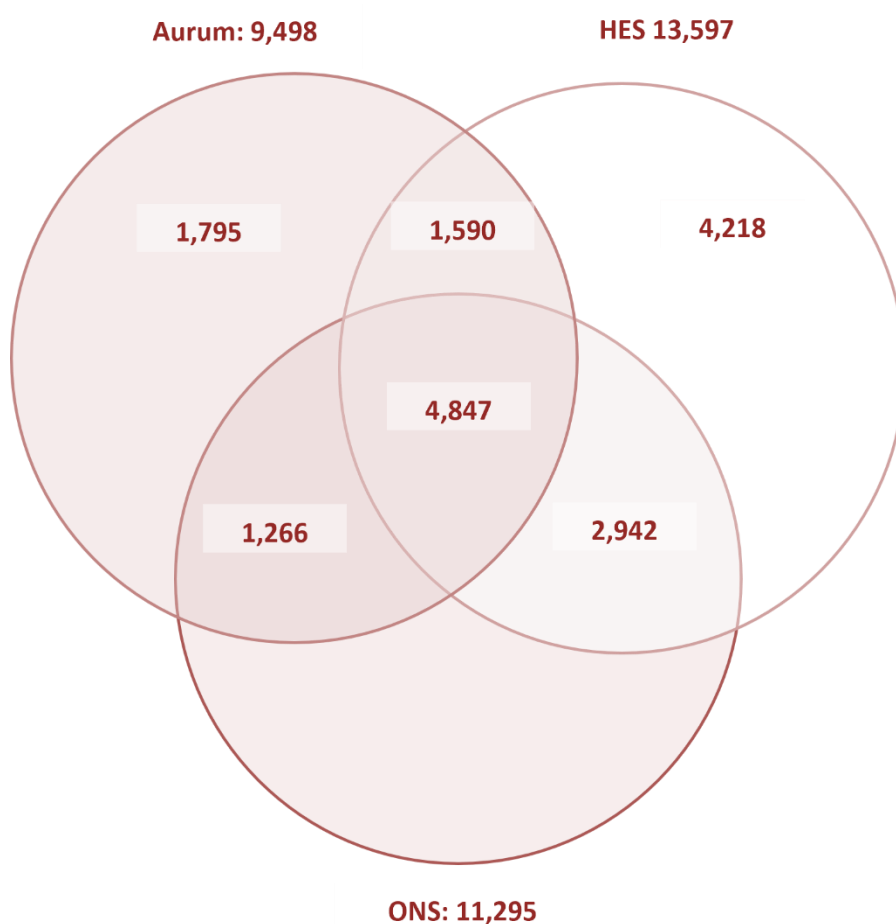
4.2.4.1. Evidence of IPF on hospital discharge

The goal here was to identify individuals hospitalised specifically due to their IPF (such as an acute exacerbation) rather than those admitted for another reason who happen to have IPF as a comorbidity.

In HES data, an episode is the period of care for an admitted patient overseen by a given healthcare professional, and a spell is defined as a continuous period spent as a patient within a trust, which may include one or more episodes. For this sensitivity analysis, I assessed the coding of ICD-10 of IPF in any position in an episode of hospital spell. The frequency of using IPF codes beyond position three was reduced. Hence, I performed further checks limited to the top three diagnosis positions. I incorporated patients with ICD-10 codes for IPF (J84.1, J84.8, or J84.9) in any of the top three diagnostic positions. The cohort size was 13,957 patients when ICD-10 codes were in the top three positions. However, in the top two positions, the cohort size was 10,714. However, when I assessed the overlap of 13,957 and 10,714 in ONS, the proportion of overlap

was 55.8% and 62.1%, respectively (Figure 4.2). Therefore, to ensure the selection of appropriate cases, that is, cases in the hospital for IPF (or IPF-related reasons) versus IPF listed as comorbidity, I opted to restrict the discharge diagnosis to include only those listed in the first and second positions within HES-APC. The rationale is that earlier positions are more likely to indicate the primary reason for hospitalisation, while later positions typically reflect coexisting conditions or comorbidities that might influence care.

Figure 4. 2 . Sensitivity analysis: effect of extending the HES APC cohort to include those who had IPF listed in the third diagnostic position on the concordance in the recording of IPF between ONS, Aurum and HES APC (restricted to patients who died during the study period 2008–2018)

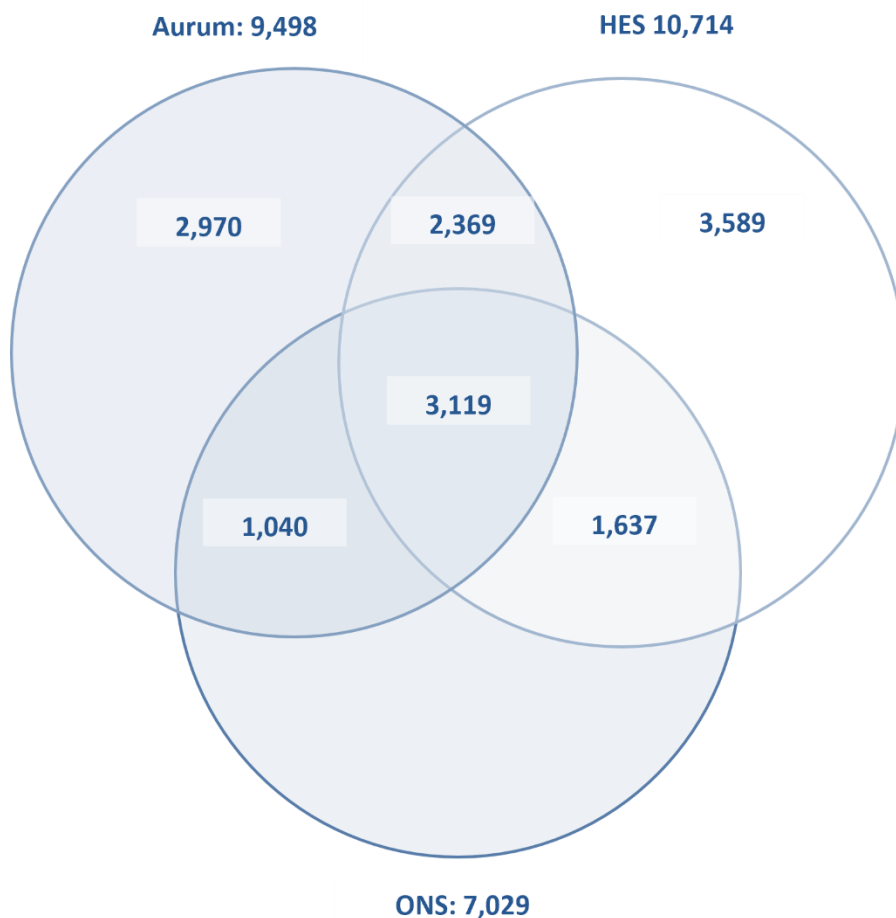


4.2.4.2. Restricting IPF as an underlying cause of death is the gold standard

For this validation study, I chose to use the death record in ONS for IPF as the gold standard. In the absence of a GP survey and access to GP notes, the reason for death in ONS will serve as a reference. When I focused on individuals in the ONS dataset with IPF listed as the underlying cause of death ($n = 7,029$), it resulted in a

reduction in the total number of patients across multiple datasets, as illustrated in Figure 4.3. Specifically, the percentage of Aurum patients also documented in ONS decreased from 64.7% (6,113 patients) to 43.7% (4,159 patients) out of the total Aurum population (n = 9,498). Similarly, among patients admitted to hospital with a diagnosis of IPF who subsequently died from IPF, the percentage in HES-APC decreased from 65.5% (6,702 patients) to 44.4% (4,756 patients) when IPF was specifically identified as the underlying cause of death. Figure 4.3 visually represents the alignment across all three datasets, which decreased to 19.5% of the total study population (n = 15,957), equating to 3,119 patients. This analysis indicates that a greater proportion of cases captured in Aurum and HES-APC align with confirmed IPF cases in ONS when IPF is identified anywhere on the death certificate rather than solely as the underlying cause. Therefore, I adopted the former definition as the gold standard in this study.

Figure 4. 3 Sensitivity analysis: effect of restricting the ONS cohort to those who had IPF coded as the underlying cause of death in the concordance in the recording of IPF between ONS (the underlying cause of death as IPF), Aurum and HES APC (restricted to patients who died during the study period 2008–2018)



4.2.5. Diagnostic algorithms

The case-finding algorithms were built upon the code lists developed for identifying Aurum, HES APC, and ONS patients. In section 4.2.1, I discussed the information captured in the HES datasets (HES OP, HES DID, and HES A&E) and their shortcomings. Keeping those in mind, the data from HES OP, DID, and A&E were not directly utilised for identifying patients with IPF diagnosis but rather employed to enhance and refine the identification from primary datasets (Aurum, HES APC, or ONS) in refining the case-finding algorithms. For example, I had difficulty finding IPF patients from the HES OP dataset, as the ICD-10 codes are not coded well in this dataset. Still, I used HES OP as a proxy for confirmatory evidence that someone likely attended an outpatient centre with symptoms consistent for an exacerbation. Similarly, for HES DID, exploratory work showed a small proportion (~20%) had a scan conducted at any point in their medical history. HES OP, DID, and A&E's main purpose was to get confirmatory evidence to support the IPF diagnosis from HES APC and/or Aurum.

In total, I developed ten distinct case-finding algorithms, each based on different combinations of either "broad" or "narrow" clinical code sets or evidence of IPF-related hospitalisation (recorded in HES APC), along with additional criteria such as evidence of thoracic HRCT, absence of other known causes of ILD (such as connective tissue disease, sarcoidosis, environmental exposures, drug toxicity, and other non-IPF ILDs), outpatient visits (HES OP), or visits to A&E. These algorithms ranged from the most specific to the most sensitive constructs, designed to find IPF cases comprehensively. I relied on the literature to inform the choice of my diagnostic algorithms (DA); previous work on how IPF is captured in the UK and USA formed the basis for how I began these definitions^{13,33,144,150}. The DAs were defined as follows:

- DA1: Any read code in the 'broad' IPF category in Aurum
- DA2: Any read code in the 'broad' IPF category in Aurum plus evidence of HRCT of the thorax at any time in HES DID, and no other known causes of ILD recorded in Aurum (such as connective tissue disease, sarcoidosis, domestic and occupational environmental exposures, drug toxicity and other non-IPF ILDs).
- DA3: Any read code in the 'narrow' IPF category in Aurum

- DA4: Any read code in the 'narrow' IPF category in Aurum plus evidence of HRCT of the thorax at any time in HES DID, and no other known causes of ILD recorded in Aurum (such as connective tissue disease, sarcoidosis, domestic and occupational environmental exposures, drug toxicity and other non-IPF ILDs).
- DA5: Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.
- DA6: Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell plus evidence of HRCT of thorax at any time
- DA7: Any read code in the 'broad' IPF category in Aurum plus primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.
- DA8: Any read code in the 'narrow' IPF category in Aurum plus primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.
- DA9: Any read code in the 'broad' IPF category in Aurum plus primary plus two A&E visits associated with IPF.
- DA10: Any read code in the 'broad' IPF category in Aurum plus primary plus two OP visits associated with IPF.

Taking into consideration the recent IPF diagnostic guidelines, I assessed whether it was feasible to include pulmonary function tests, lung transplantation, and/ or multidisciplinary review in the diagnostic definitions but the initial feasibility showed less than 10% of the IPF cases had any record of either of these three parameters at any time in the Aurum database. Therefore, I did not include these parameters while building the algorithms.

4.2.6. Statistical analysis plan

4.2.6.1. Descriptive analysis

In this study, I analysed coding practices in primary care (Aurum) over the period from 2008 to 2018 to understand any temporal changes in the frequency of IPF codes (broad list). Additionally, I examined the demographic and clinical characteristics of patients identified with IPF in Aurum, HES, or ONS datasets (those who died with underlying reasons as IPF prior to the study end), focusing on comorbidities such as

COPD, asthma, lung cancer, ischemic heart disease, heart failure, myocardial infarction, and gastroesophageal reflux. This list of comorbidities was developed based on the literature^{13,33,144,150} and discussions with experts, JKQ, and PMG. For each comorbidity, I assessed the presence of relevant diagnostic codes in their medical history before IPF diagnosis. Furthermore, I evaluated the smoking status among these patients, assigning the status closest to the IPF diagnosis date. All code lists for the comorbidities are available in Appendix to Chapter 4.

4.2.6.2. Concordance of IPF across Aurum, HES APC and ONS

For the patients who met this study criteria (section 4.2.2) across the three data sources, I evaluated the overlap among them. To assess the concordance with ONS data, the Aurum and HES-APC cohorts were limited to those patients who died within this study period. I checked the concordance for the following terms: i) patients in Aurum and HES-APC, ii) Aurum and ONS, iii) HES-APC and ONS, and iv) Aurum, HES-APC, and ONS.

4.2.6.3. Diagnostic algorithms

Lists of medcode IDs (broad and narrow) and ICD-10 codes for IPF were derived as explained in section 4.2.3. A combination of code lists and additional forms of evidence, such as the presence of CT or evidence of IPF in both primary and secondary care, were used to make up the algorithms. Of the ten algorithms described in section 4.2.3, the first four definitions were founded on the list of a broad or narrow IPF diagnosis code in primary care (Aurum). The other two definitions were derived using ICD-10 codes in the hospital database (HES -APC). The last four definitions were based on both presence in primary and secondary care.

Validation of Diagnostic Algorithm

The accuracy of finding an IPF patient using each of the predefined DAs was measured by PPV. PPV was computed as the proportion of ‘true positives’ (individuals with IPF) in each algorithm as determined by the gold standard (deaths in ONS with IPF at any position). PPV was calculated as the number of IPF cases correctly found by the algorithm (i.e., confirmed by a record in ONS), i.e. 'A' divided by the total number of IPF cases found by that algorithm 'A+B'. As the main aim was to find an algorithm that ‘accurately’ finds IPF patients, I computed sensitivity for each. Sensitivity was computed as the number of “true” IPF cases correctly found by the algorithm ‘A’ divided by the total number of ONS IPF deaths ‘A+C’. Death is the gold standard;

this analysis was limited to those patients who died. Please see Figure 4.2. to see how PPV and sensitivity were computed for this analysis.

Figure 4. 4 Method for positive predictive value and sensitivity

		Gold standard diagnosis (ONS)		Total
		Disease	No disease	
Diagnostic algorithm	Disease	A [true cases correctly identified by the DA]	B [non-cases wrongly coded as cases by DA]	A+B
	No disease	C [true cases not identified by the DA]	D [true non-cases correctly identified by DA]	C+D
Total		A+C	B+D	

DA, diagnostic algorithm

All the programming was executed in STATA, and confidence intervals were reported for all statistics.

4.3. Results

4.3.1. Coding practices in primary care over time

The analysis of IPF coding in primary care (Aurum) revealed a notable trend: a steady increase in the use of broad IPF codes (a broad list that is indicative of IPF includes 27 medcode IDs) starting from 2006, peaking around 2012, and stabilising thereafter. Figure 4.5. plots the frequency of use of IPF codes among ILD patients found in Aurum. It demonstrates the gradual increase in the use of IPF medcode IDs during the study period, and Table 4.3 shows the frequency of use by IPF terms in Aurum data for the same period. The predominant IPF terms accounting for more than 80% of GP-coded IPF visits throughout the study period were: 'Idiopathic pulmonary fibrosis', 'Pulmonary fibrosis', 'fibrosis of lung', and 'Diffuse pulmonary fibrosis'. A further examination of the coding practices during the study period shows an uptake in the use of specific IPF terms such as 'idiopathic pulmonary fibrosis' and 'pulmonary fibrosis' and a gradual decline in the use of broader terms such as 'diffuse pulmonary fibrosis'. Figure 4.6 shows the trend of usage of the 'broad' and 'narrow' list of codes between 2008 and 2018; as anticipated, there is a change in coding behaviour, an

increase in the use of IPF-specific terms, and a decrease in the use of generic (less-specific) IPF terms around 2012 – tied to the approval of antifibrotics and release of guidelines.

Figure 4. 5 Frequency of IPF medcode ids over time (Aurum data)

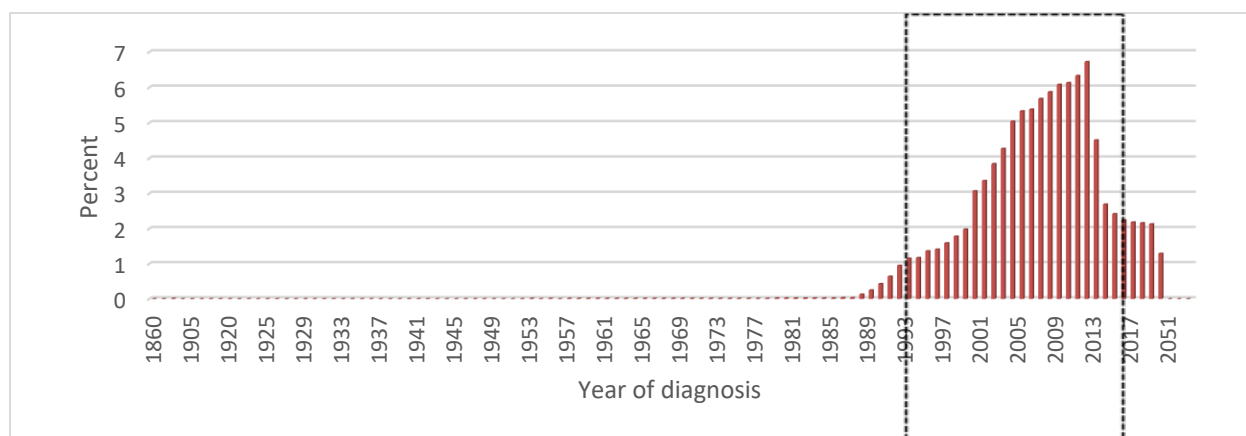
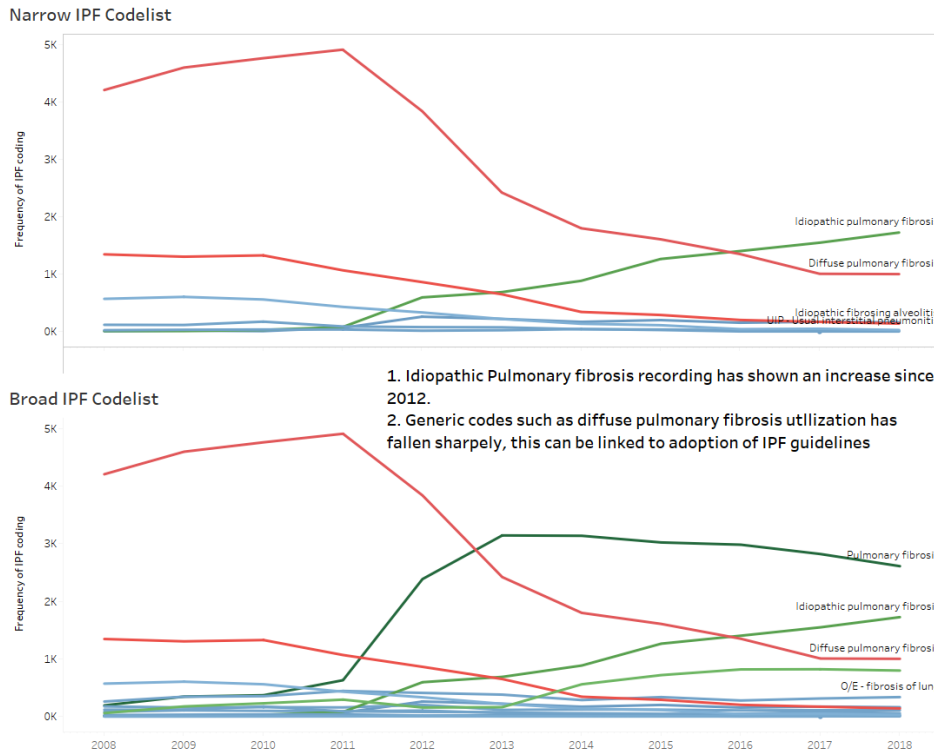


Table 4. 9 Frequency of use of IPF clinical codes in UK primary care (CPRD Aurum data set), 2008–2018, ascending order

Snomed Concept ID	Description	Frequency of codes (n)	Frequency of codes (%)
236302005	Hamman - Rich syndrome	10	0.0%
233749003	Progressive massive fibrosis	77	0.1%
45157009	idiopath. fibrosing alveolitis	251	0.3%
64667001	Interstitial pneumonia	1535	1.8%
162974009	O/E - fibrosis of lung ; O/E - fibrosis of lung present	4570	5.3%
909731000006101	[rfc] pulmonary fibrosis	4789	5.6%
700250006	Idiopathic pulmonary fibrosis	19826	23.1%
51615001	Pulmonary fibrosis: fibrosis of the lung	21720	25.3%
196125002	Diffuse interstitial pulmonary fibrosis, fibrosing alveolitis	32951	38.4%

Figure 4.6 Trends in the use of ‘broad’ and ‘narrow’ IPF clinical codes in CPRD Aurum primary care data, 2008–2018



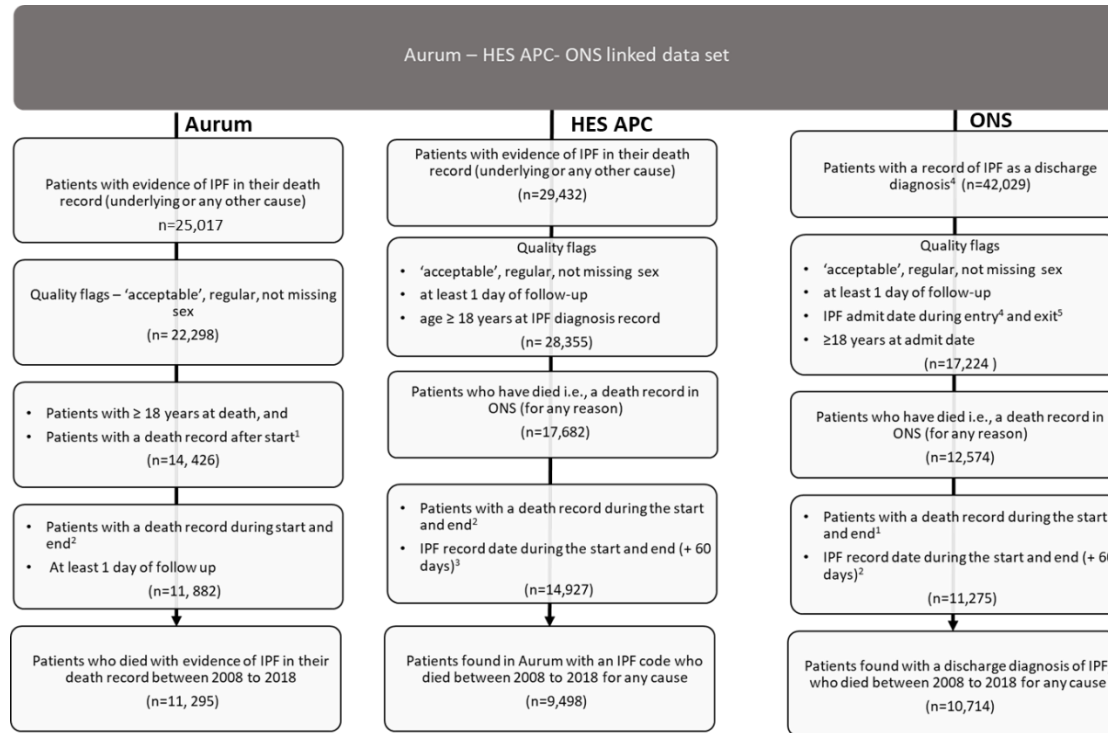
A narrow, specific list of read codes strongly indicative of IPF; broad, non-specific, generic codes denoting IPF.

4.3.2. Study population

For the study period, a cohort of 17,559 adults was identified across Aurum, HES-APC, and ONS datasets, meeting the criteria for IPF diagnosis in at least one of these sources and this study eligibility. Specifically, Aurum documented 13,627 adults with clinical codes indicating IPF according to the broad definition, of whom 9,498 (69.7%) died during the study period. 14,719 adults had evidence of hospital admissions for IPF, with 10,714 (72.7%) experiencing mortality. Furthermore, 11,295 adults had IPF listed as a cause of death (any) on their death certificates within the study period. Figure 4.7 illustrates the patient attrition across these three data sources, highlighting the cohort's composition and distribution. Detailed clinical and demographic characteristics for these cohorts are provided in Table 4.4, offering insights into the study cohort's profile across different datasets. The median age across the three cohorts was 77 years. The ONS cohort had a higher proportion of individuals aged 65 and above (90%, $n=10715$) compared to the HES APC (85.3%, $n=9145$) and Aurum cohorts (86.2%, $n=8191$), where the proportions were similar (85.3% in HES APC and 86.2% in Aurum). A greater proportion of patients in the Aurum (66.2%, $n=6288$) and HES APC (65.6%, $n=7030$) cohorts had a history of smoking. The HES APC cohort also had a higher proportion of patients with

respiratory comorbidities, such as COPD and asthma (within the past two years), compared to the other two cohorts. However, the proportion of patients with cardiovascular conditions, including ischemic stroke, heart failure, and myocardial infarction, was similar across all three cohorts.

Figure 4.7 Patient Attrition in the three datasets: Aurum, HES-APC and ONS



HES, hospital episode statistics; APC, admitted patient care; ONS, office of national statistics

Table 4. 10 Clinical and demographic characteristics of base cohorts: Aurum, HES-APC and ONS

Characteristic	ONS (% of total n unless otherwise stated)	Aurum (% of total n unless otherwise stated)	HES APC (% of total n unless otherwise stated)
Total cohort	11,295 (100%)	9,948 (100%)	10,714 (100%)
<i>Demographic</i>			
Median (IQR) age in years at IPF diagnosis in the study period	77 (10.3)	77 (13)	77 (14)
≥ 65 years	90.0% (n = 10,176)	86.2% (n = 8,191)	85.3% (n = 9,145)
Male	63.0% (n = 7,123)	63.2% (n = 6,011)	60.7% (n = 6,512)
Smoking history (ever/current)	48.9% (n = 5,529)	66.2%(n = 6,288)	65.6% (n = 7,030)
<i>Comorbidities</i>			
COPD	18.2% (n = 2,060)	19.0% (n = 1,806)	20.9% (n = 2,240)
Asthma (in past 2 years)	6.93% (n = 784)	10.1% (n = 959)	10.3% (n = 1,111)

Lung cancer	3.0% (<i>n</i> = 348)	2.1% (<i>n</i> = 206)	2.8% (<i>n</i> = 305)
Ischemic stroke	13.4% (<i>n</i> = 1,513)	12.0% (<i>n</i> = 1,142)	13.2% (<i>n</i> = 1,418)
Heart failure	17.9% (<i>n</i> = 2,024)	17.1% (<i>n</i> = 1,629)	17.6% (<i>n</i> = 1,889)
Ischemic heart disease	13.4% (<i>n</i> = 1,523)	13.6% (<i>n</i> = 1,300)	13.4% (<i>n</i> = 1,441)
Myocardial infarction	13.1% (<i>n</i> = 1,481)	13.3% (<i>n</i> = 1,267)	13.2% (<i>n</i> = 1,414)
Gastroesophageal reflux	17.8% (<i>n</i> = 2,012)	17.6% (<i>n</i> = 1,677)	17.6% (<i>n</i> = 1,892)

4.3.3. Concordance of Aurum and HES-APC

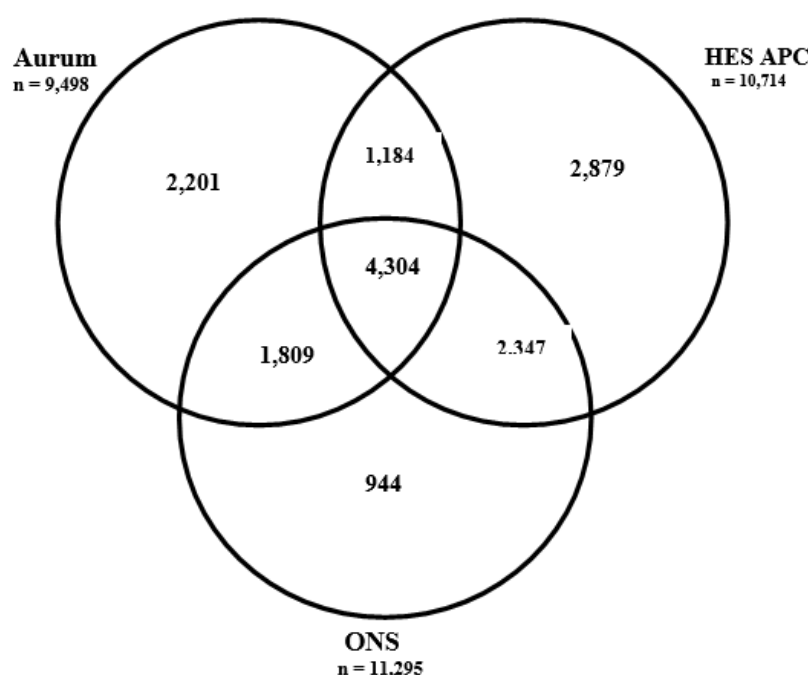
In this analysis, less than 20% (*n*=9621, 18.2%) of individuals in the study population had evidence of IPF in the Aurum and HES APC dataset. In HES APC, primary and secondary diagnoses were considered. A high number of patients were found in HES APC (*n*=28,355) with evidence of IPF as compared to Aurum (*n*=22,298).

4.3.4. Concordance between Aurum, HES-APC and ONS

In this analysis, a quarter of individuals in the study population (24.5%, *n* = 4,304) had evidence of IPF in all three of the linked concurrent data sets (Figure 4.8). Among the cohort identified in ONS with IPF listed on their death certificate (11,295 adults), 54.1% (*n* = 6,113) also had clinical codes denoting IPF in Aurum before or shortly after death (or within 60 days after their death), while 58.9% (*n* = 6,651) had records of hospital admissions for IPF before death, slightly more than those found in Aurum. Approximately 75% (*n* = 8,460) had IPF documented in either Aurum or HES-APC. Conversely, only 38.1% (*n* = 4,304) of those in ONS had IPF documented in both Aurum and HES-APC. Interestingly, among Aurum patients who died with IPF (*n* = 9,498), 35.6% (*n* = 3,385) did not have IPF mentioned on their death certificate. Common underlying causes of death in this group included pneumonia, bronchopneumonia, and unspecified malignant neoplasm of the bronchus or lung (ICD-10 codes J18.9, J18.0, C34.9, respectively). Other frequently reported causes were COPD (J44.9), “other general symptoms and signs” (R68.8), and “respiratory failure, not elsewhere classified” (J96.9). Similarly, for patients hospitalised with IPF but without IPF listed on their death certificate (*n* = 4,063), the leading causes of death were pneumonia, bronchopneumonia, and unspecified malignant

neoplasm of the bronchus or lung. Additional frequently reported causes were myocardial infarction, heart failure, and sepsis.

Figure 4. 8 Concordance in the recording of IPF across ONS, Aurum and HES-APC. ONS, study eligible population with any cause of death as IPF in their death record during the study period; Aurum, patients with one clinical code denoting IPF in Aurum (using a broad set of codes) and had died for any reason during the study period; HES-APC, patients admitted to the hospital for IPF at some point before their death for any reason (during the study period)



HES, hospital episode statistics; APC, admitted patient care; ONS, office of national statistics.

4.3.5. Validation of diagnostic algorithms

The PPVs and sensitivities for the DAs presented in the methods section are shown in Table 4.5. Previous publications, primarily from the United States, have considered visits to A&E or outpatient departments associated with IPF as a criterion to improve patient identification accuracy. However, the HES A&E and OP database lacks specific information about the reason for a patient's visit, making it difficult to ascertain the purpose of their attendance. Given this constraint, I focused my analysis on the first eight DAs, excluding those with A&E and OP visits (Table 4.5).

Table 4. 11 Positive Predictive Value of the Code-Based IPF Algorithms, sorted based on high PPV and sensitivity.

Diagnostic Algorithm (DA)	Data source (Aurum and/ or HES)	Description	No. of patients found (N)	No. of patients meeting the gold std.	PPV (%)	Sensitivity (%)
DA1	Aurum	Any read code in the 'broad' IPF category in Aurum	9,498	6,113	64.3%	54.1%
DA2	Aurum	Any read code in the 'broad' IPF category in Aurum plus evidence of HRCT of the thorax at any time in HES DID and no other known causes of ILD recorded in Aurum (such as connective tissue disease, sarcoidosis, domestic and occupational environmental exposures, drug toxicity and other non-IPF ILDs).	4,721	3,187	67.5%	28.2%
DA3	Aurum	Any read code in the 'narrow' IPF category in Aurum	1,828	1,370	74.9%	12.1%
DA4	Aurum	Any read code in the 'narrow' IPF category in Aurum plus evidence of HRCT of the thorax at any time in HES DID and no other known causes of ILD recorded in Aurum (such as connective tissue disease, sarcoidosis, domestic and occupational environmental exposures, drug toxicity and other non-IPF ILDs).	891	706	79.2%	6.3%
DA5	HES	Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.	10,714	6,651	62.0%	58.8%
DA6	HES	Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell plus evidence of HRCT of thorax at any time	2,985	2,379	79.7%	21.0%
DA7	Aurum and HES	Any read code in the 'broad' IPF category in Aurum plus Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.	5,488	4,304	78.4%	38.1%
DA8	Aurum and HES	Any read code in the 'narrow' IPF category in Aurum plus Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.	1,190	974	81.8%	8.6%

HES, hospital episode statistics; ICD, international classification of disease; HRCT, high-resolution computed tomography; DID, diagnostic imaging dataset

Aurum-based diagnostic algorithms: Among the diagnostic algorithms based on a broad list of IPF codes (DA1, DA2), there was marginal variation in PPVs. PPV for DA1 was 64.3% (95% CI, 63.3–65.3) refining

this criterion, that is, adding evidence of HRCT and excluding any known causes of ILD, did not materially alter the PPV (67.5% [95% CI 66.1% – 68.6%]) but reduced the sensitivity from 53.5% (95% CI, 52.9–54.1) % to 28.2% (95% CI, 27.3–29%).

The narrow-code-based algorithms (DA3, DA4) had notably higher PPVs compared to the broad-code-based algorithms. The PPV for DA3 was 74.9% (72.8–76.9) and in contrast to DA2, refining the ‘narrow’ criteria by adding HRCT improved the PPV for narrow-based algorithms to 79.2% (76.4–81.8). However, the sensitivities for narrow-based algorithms were considerably lower than for broad-based algorithms. (6.3% to 12.1%).

HES-APC-based diagnostic algorithms: Patients found in HES-APC (DA5) had a PPV of 62.0% (61.1–63.0); this was improved considerably to 79.7% (78.2–81.1) by refining the criteria (Table 4.5). As seen previously, refining the criteria and increase in PPV was inversely accompanied by a substantial reduction in sensitivity from 57% (56.4–57.7) to 21% (20.3 – 21.8)

Aurum-plus-HES APC matched Code-Based Diagnostic Algorithms: The algorithms (DA7, DA8) were based on patients found in Aurum and HES-APC had overall a higher PPV. PPVs were 78.4% (77.3 – 79.5) and 81.8% (79.5 – 84) for DA7 and DA8, respectively. The sensitivity for DA7 was 38.1% (37.2-39.0) higher than the narrow-based algorithms and for DA8 was 8.6% (8.1 – 9.1) similar to narrow based algorithms.

Table 4.4. shows the median age at IPF diagnosis was 77 years across the three base cohorts (Aurum, HES-APC, and ONS) and most patients in the cohort were 60 years and above (>80% of the cohort). Hence, in these algorithms increasing the age criteria for example from 18 years to 60 years would not influence the PPV – given the baseline demographics.

I compared cohort profiles for four algorithms selecting those that represented the spectrum; DA1 (high sensitivity but modest PPV), DA3 (higher PPV but low sensitivity), and intermediates that represent more of a compromise between PPV and sensitivity - DA7 and DA5. The idea was to compare algorithms that range from a high PPV (with a modest sensitivity) to a high sensitivity (but modest PPV).

4.3.6. Demographic and Clinical Characteristics

Table 4.6 describes the demographic and clinical characteristics of the patients found using algorithms DA1, DA3, DA5, and DA7. The four patient cohorts are broadly comparable both in terms of the demographic characteristics and the comorbidity profile, except for a slightly lower prevalence of comorbid COPD (and, to a lesser extent, asthma and heart failure) in the Aurum narrow cohort. In comparison to values reported in the literature, median survival times were relatively low, ranging from 1.8 years (DA6) to 2.6 years (DA3). HES only cohort (DA5) had the lowest median survival of 1.8 years (IQR, 3.5).

Table 4. 12 Demographic and clinical characteristics for the patients found using algorithms DA1, DA3, DA6 and DA7

Characteristic	DA1 Aurum broad % of the total n unless otherwise noted (N= 9,498)	DA3 Aurum narrow % of the total n unless otherwise noted (N=3,449)	DA5 HES Only % of the total n unless otherwise noted (N= 2,069)	DA7 Aurum broad plus HES % of the total n unless otherwise noted (N= 5,488)
Demographic				
Median (IQR) age in years	77 (13)	76(14)	76(14)	76 (14)
≥ 65 years	86.2% (n = 8,191)	82.9% (n = 1,516)	84.4% (n = 2,522)	83.7% (n = 4,593)
Male	63.2% (n = 6,011)	63.8% (n = 661)	64.4% (1,922)	62.5% (n = 3,434)
Median (IQR) survival in years	2.0 (3.7)	2.6 (4.8)	1.8 (3.5)	2.2 (3.9)
Smoking history (ever/current)	66.2%(n = 6,288)	62.4%(n = 1,1420)	68.5(2,046)	66.4% (n = 3,642)
Comorbidities				
COPD	19.0% (n = 1,806)	14.7% (n = 269)	17.8% (n = 532)	16.1% (n = 886)
Asthma (in past 2 years)	10.1% (n = 959)	7.9% (n = 146)	8.9% (n = 268)	9.4% (n = 518)
Lung cancer	2.1% (n = 206)	1.5% (n = 28)	2.3% (n = 70)	1.6% (n = 85)
Ischemic stroke	12.0% (n = 1,142)	10.1% (n = 185)	12.4% (n = 370)	11.8% (n = 645)
Heart failure (HF)	17.1% (n = 1,629)	13.6% (n = 249)	16.3% (n = 489)	16.4% (n = 899)
Ischemic heart disease	13.6% (n = 1,300)	11.8% (n = 217)	14.4% (n = 430)	13.3% (n = 735)
Myocardial infarction (MI)	13.3% (n = 1,267)	11.2% (n = 206)	12.4% (n = 370)	12.7% (n = 693)
Gastroesophageal reflux (GERD)	17.6% (n = 1,677)	15.8% (n = 289)	20% (n = 597)	18.2% (n = 1,002)

4.4. Discussion

The primary aim of this study was to investigate the accuracy and consistency of IPF recordings across UK primary and secondary care records. Over the study period from 2008 to 2018, this study revealed significant changes in coding practices, coinciding with the introduction and adoption of revised IPF guidelines⁷⁻¹², as well as the authorisation of antifibrotic medications^{28,29}. This analysis revealed a shift towards the use of IPF-specific diagnostic codes, marking a change from outdated terms like cryptogenic fibrosing alveolitis and generic codes such as diffuse pulmonary fibrosis to more specific IPF terms. This transition suggests efforts to standardise and improve the precision of IPF diagnoses within healthcare records.

The diagnostic algorithms developed in this analysis delve into coding practices and their implications for research. Findings in this chapter highlighted distinct variations in PPVs depending on whether the algorithm was based on primary or secondary care data. Notably, algorithms reliant solely on secondary care data exhibited the lowest PPVs (DA6 = 62.4%, 95% CI: 61.1–63.0), whereas those combining primary care clinical codes with evidence of hospitalisation for IPF showed the highest PPV (DA8 = 81.8%, 95% CI: 79.5–84.0). However, this improvement in diagnostic accuracy was tempered by a decline in the number of identified patients, underlining the trade-off between accuracy and sample size in algorithm selection. Overall, broad-based algorithms exhibited a higher sensitivity in detecting IPF cases but at the expense of PPV, whereas narrow-focused criteria improved PPV while reducing sensitivity.

Moreover, while this study demonstrated comparable demographic and comorbidity profiles between IPF cohorts defined by different lists of diagnostic codes, those identified through secondary care data only tended to exhibit a slightly lower median survival time, indicative of potentially more advanced disease stages within this cohort.

Further granularity in algorithm specificity—such as incorporating age criteria, excluding alternative causes of ILD, and requiring evidence from thoracic imaging—yielded marginal enhancements in PPV but significantly impacted sensitivity. This suggests that the presence of a specific IPF diagnostic code alone is sufficient to accurately identify patients with IPF from Aurum – this wasn’t unexpected. Previous validation studies have

shown that a chronic condition is relatively well captured in CPRD. For example, to be diagnosed with IPF, the assumption is that the individual would have seen a specialist and undergone the required tests. This means that by the time they were reviewed in primary care, the diagnosis was confirmed. This parallels findings from other validation studies^{151,152}, including one for COPD, which similarly showed that incorporating spirometry evidence and/or COPD medications only marginally improved the accuracy of diagnostic code-based algorithms (89.4% vs. 87.5% vs. 86.5%)¹⁵².

When comparing the number of cases found in primary and secondary data sources, I found that secondary data (HES APC) identified a higher number of IPF cases using ICD-10 codes compared to Aurum. This is surprising but can be explained with a few reasons: i) the codes used in Aurum were more restrictive to IPF and 'IPF-like' terms, ii) when someone is admitted to hospital, the hospital record has the reason for admission (primary diagnosis) but in addition captures other conditions or comorbidities (up to 20 diagnosis fields) that are relevant to the patient's care during the episode, which could influence the treatment or management. In my analysis, I limited the cohort to the first three diagnosis positions within a hospital spell (that is, everything between hospital admission and discharge). As a result, I may have included patients who were admitted to the hospital for reasons unrelated to IPF, and I would not know when or where their IPF diagnosis was made. The cases identified solely through HES APC had higher sensitivity but a lower PPV. In contrast, IPF cases identified through Aurum's narrow definition showed much lower sensitivity but a higher PPV, which aligns with the typical diagnostic pathway—beginning in specialist settings and continuing with primary care management. An ideal situation would be the ability to link the IPF registry to primary care. But in the absence of that, when designing a study, it is crucial to consider these differences and choose the appropriate data based on the research objectives.

This study has several strengths and limitations. The first strength of this study was the use of primary and secondary linked data, which allows for the capture of IPF records in any healthcare setting. It is a large population-based representative dataset, thus providing a unique opportunity to find cases in different healthcare settings. Second, the study period, 2008 to 2018, allowed us to observe the impact of IPF guidelines and how they have changed coding practices over time. Despite the strengths, some limitations of

this study were: i) This study relied on the physician's accuracy in recording the correct code at diagnosis; ii) I did not have access to GP notes or questionnaires; therefore, ONS was used as a gold standard to validate IPF diagnosis in Aurum. ONS is limited to patients with a death record only; it has missing data, and not all patients with IPF will necessarily be recorded as an IPF death in ONS. In certain cases, IPF can be the contributory factor and not the main reason for death. I have seen that a large proportion of people who had a GP-recorded IPF diagnosis but no mention of IPF on their certificate died of a respiratory reason (e.g. pneumonia), which suggests that this group of patients did, in all likelihood, have IPF and that these do represent IPF-related deaths despite the absence of IPF recorded on their death certificate. Nonetheless, in the absence of traditional methods of Validation, such as GP questionnaires, which are prohibitively expensive and burdensome to conduct, ONS served as a well-suited option. With this in mind, one can speculate the gold standard likely underestimates the true IPF population. If that is the case, the PPVs (particularly for Aurum) might also be lower than expected.

This is the first study to validate the completeness and quality of recording of IPF in primary care, especially in Aurum and linked data sources in England. This study attempts to call attention to the new set of codes to identify genuine IPF patients in Aurum. This study developed diagnostic code-based algorithms considering previous quality assessments conducted with this same data^{142,143,152-155} and using IPF guidelines⁷⁻¹² as a reference. The diagnosis of IPF recorded in the Aurum database is relatively accurate. Depending on the research question, Aurum can be used solely with the right set of clinical codes. To warrant an appropriate case detection method is employed, researchers should cautiously consider the following: study design, use of an appropriate and recent list of clinical codes that reflect GP coding practices and use of linked data (such as HES APC) to improve the capture of events. The landscape of IPF diagnosis has evolved significantly with advancements in guidelines, diagnostic methods, and therapeutic options over the past decade. This evolution is reflected in the uptake of IPF-specific coding practices across healthcare settings, enhancing the accuracy and consistency of IPF diagnosis. Future studies could benefit from access to more comprehensive clinical records or alternative validation methodologies.

Chapter 5

Incidence, prevalence, and mortality of idiopathic pulmonary fibrosis in England from 2008 to 2018

Overview

This study aims to build upon previous work (Strongman et al), which only extends to 2012 and has not been updated since, particularly concerning anti-fibrotic treatments. Like Strongman and colleagues, I explored how coding choices impact epidemiological estimates (broad and narrow) and analysed trends over time. Additionally, I expanded the analysis by incorporating additional diagnostic algorithms, which we have validated to the best of our ability, and we utilised a different CPRD dataset (Aurum versus GOLD).

Using Aurum-HES APC linked data, I estimated the incidence and prevalence of IPF using previously validated diagnostic-code-based algorithms. IPF mortality was assessed using the registered number of deaths with the underlying cause recorded as IPF (from ONS) for the same period. This study's findings indicate that the incidence of IPF increased over time using Aurum-based definitions: a 100% increase for Aurum narrow (from 3 to 6.1 per 100,000 person-years) and a 25% increase for Aurum broad (from 22.4 to 28.6 per 100,000 person-years). However, HES-based definitions showed a decrease in incidence over the same period, falling between the extremes of the Aurum-based definitions. IPF mortality in 2018 was 7.9 per 100,000 person-years, representing a 53% increase between 2008 and 2018. Scaling Aurum-based incidence estimates to England's 2018 population suggests 8,000-9,000 new cases of IPF per year, which is higher than previous estimates (5,000-6,000). This increased burden of new IPF cases each year has significant implications for future health service planning and resource allocation. This chapter has been published – ‘Incidence, prevalence and mortality of idiopathic pulmonary fibrosis in England from 2008 to 2018: a cohort study’, <https://doi.org/10.1136/thorax-2023-220887>.

5.1. INTRODUCTION

In Chapter 2, I have discussed variations in the incidence and prevalence of IPF reporting on a global landscape. The discrepancy in estimates was observed at the country level and within a country, depending on the data source and the method used to find IPF, for example, some studies were drawn on electronic health records, while some were based on registries or single centers. As IPF research has been growing over the years, I noticed more studies based on EHR or health insurance claims databases.

From Chapter 2, I learned how different methods were applied to diagnose IPF. In Chapter 4, I conducted a validation study to examine the recording of IPF in primary and secondary care databases. To build on what I have learned about IPF so far, the next step was to update the epidemiology of IPF in the U.K, particularly after the introduction of guidelines on diagnosing and managing IPF⁷⁻¹², and approval of antifibrotic treatments^{29,156} which I speculate has further raised awareness of IPF. To my knowledge, there has been no recent study updating the incidence and prevalence of IPF in the U.K. The previous study conducted by Strongman et al. reported the epidemiology of IPF using primary care data from 2000 to 2012. The authors reported the incidence of IPF ranged from 2.85 to 8.65 per 100,000 person-years, and prevalence ranged from 19.94 to 38.82 per 100,000 patients¹³. However, this study predates the recent developments in this field and uses the GOLD database (details on various CPRD databases are covered in Chapter 3).

Given the growing interest in IPF and the absence of recent data to support health service planning in the U.K., this study aimed to update and extend previous research by estimating the incidence and prevalence of IPF in England from 2008 to 2018 using Aurum primary care data. Additionally, trends in IPF-specific mortality were investigated using ONS data, with the study leveraging validated diagnostic algorithms for IPF case findings developed in earlier work.

5.2. METHODS

5.2.1. *Data sources*

For this study, I used the Aurum and HES-APC linked datasets. These datasets are described in Chapter 3. We used the same data build from Chapter 4 (Validation study). In Chapter 4, section 4.2.1 Data sources, I have described the data sources used, and the same data sources were used for this analysis.

5.2.2. Study population

The study population was drawn from subjects who were aged at least 18 years at the date of diagnosis of IPF, the practice the subject was registered to had passed CPRD's quality check, there was no missing sex information, and had at least 1 day of follow-up after they entered into this study. Study entry was the most recent of the following dates: study start date (1 January 2008), the patient's eighteenth birthday, the date a patient joined the practice, or the date of first diagnosis with IPF. Study exit was first of the following dates: study end date (31 December 2018), the date they left the practice, had died or the last date that the practice contributed data to Aurum. Follow-up time was computed from study entry to study exit. Lung transplantation was not considered for censoring because of the very small number of cases, fewer than 120 performed annually¹⁵⁷.

In addition to these primary criteria, the additional criteria for the study population were to meet the criteria for one of the four case-defining algorithms for IPF. The four separate cohorts were defined using previously validated diagnostic-code-based algorithms; the derivation and validation of these case definitions are discussed in depth in Chapter 4. The four algorithms considered for further assessment spanned the extremes of IPF case finding in the databases; two algorithms achieved the highest PPV in my previous validation study (DA2 and DA4), and two had a lower PPV but higher sensitivity (DA1 and DA3). The diagnostic algorithms used to define the four cohorts for this analysis were:

- DA1: Any clinical code in the 'broad' IPF category in Aurum
- DA2: Any clinical code in the 'narrow' IPF category in Aurum
- DA3: Primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.
- DA4: Any clinical code in the 'broad' IPF category in Aurum plus primary or secondary hospital discharge diagnosis of IPF (ICD 10 code J84.1, J84.8 or J.84.9) in any episode within a hospital spell.

5.2.2.1. Study Population for Incidence and Prevalence

In this study, incidence and prevalence were established for the four study cohorts, as mentioned in the above section. To ensure capture of incident cases, I additionally applied the criterion that patients had to have had at

least 1 year of registration with their current G.P. before a diagnosis of IPF (washout period with no diagnosis of IPF in that one-year period). Subjects were considered for prevalence analysis if the date of diagnosis for IPF was before 2008 and was alive and registered with a linked Aurum G.P. practice on 1 January 2008.

5.2.2.2. Study Population for Mortality of IPF

For assessing the mortality related to IPF, I used ONS data to find persons who had IPF recorded as their underlying cause of death. The number of IPF deaths was found by using any of the ICD-10 codes, J84.1, J84.8 and J84.9 (<https://github.com/NHLI-Respiratory-Epi/Validation-of-the-recording-of-Idiopathic-Pulmonary-Fibrosis-in-routinely-collected-electronic-health>) listed as the underlying cause of death in their ONS record. In addition to the primary criteria, the date of death had to be after the patient was registered with a practice and before the practice's last data collection date. The number of IPF deaths was obtained for each calendar year between the study period 2008 to 2018.

5.2.3. Data analysis

5.2.3.1. Incidence and Prevalence

Crude **incidence** was calculated by dividing the total number of eligible persons in that year with newly diagnosed IPF by the total person-years at risk. For example, for 2010, to compute incidence for Aurum narrow, the numerator included newly diagnosed IPF patients meeting the criteria for Aurum narrow in 2010 obtained from the study cohort (Aurum -HES APC linked) as defined in section 5.2.2. The denominator was obtained by calculating the total time contributed by eligible patients in the Aurum database (this would be people registered at linked G.P. practice contributing to that build (these denominator files are supplied by CPRD for each build). Patient time was censored on the study exit, that is, the earliest of 31 December 2018, the date of death, the last day of data collection at the practice, or the patient's transfer out date. Incidence was reported as the number of cases per 100,000 person-years. Crude incidence by year of diagnosis (2008 to 2018) was determined for the four case definitions defined in section 5.2.2: DA1 – DA4.

Incidence rate ratios were estimated using the Poisson regression model (reference year was 2008). In the Poisson regression modelling, I controlled for age and sex. Multiplicative interaction terms were used to test for effect modification by age and sex. I chose to categorise them into two age groups (below and above

70 years at the date of diagnosis for IPF). This decision was informed by the age structure of my study cohorts (median age at IPF diagnosis ranged from 76 to 77 years).

Period prevalence was calculated by dividing the total number of eligible patients in that year (newly or previously diagnosed) by the total number of currently registered patients active in the mid-year (1 July). For example, to calculate the prevalence for Aurum narrow in the year 2010, I used the total number of new and prevalent cases meeting the definition for Aurum narrow in 2010 obtained from the study cohort (Aurum – HES APC linked database), divided by the total number of eligible patients, that is, people alive and actively contributing data at the mid-year point (1 July). The denominator file for incidence is the same as described in the above paragraph. Prevalence was expressed as the number of IPF cases per 100,000 persons. Prevalence was determined for the four definitions defined in section 5.2.2 (DA1 to DA4) and reported for each year from 2008 to 2018. Similar to incidence, the prevalence was stratified prevalence by age group (< 70 years and aged ≥ 70 years) and by sex. The change in IPF prevalence over the study period for each definition was evaluated using regression models to estimate odds ratios (O.R.s).

5.2.3.2. Cause-specific Mortality

IPF-specific mortality was calculated by dividing the total number of IPF deaths in that year by the total person-years at risk by the eligible population (using the same denominator file used for incidence and prevalence). Mortality was expressed as per 100,000 person-years. Age- and sex-stratum-specific mortality rates for each calendar year were calculated. I used direct standardisation to calculate age-standardised mortality rates, and mortality was standardised to the 2018 population of England (using ONS data). Multiplicative interaction terms were used to test for effect modification by sex and age. Poisson regression modelling was used to estimate annual mortality rate ratios (MRR, controlled for sex and age) – using 2008 as the baseline year. IPF-specific mortality was assessed in total population at risk over the study period, this assessment was not within the incident cases found in these analyses.

5.3. RESULTS

5.3.1. Incidence of IPF

5.3.1.1. Aurum-based definitions

During the study period (2008–2018), IPF incidence estimates ranged from 4.52 to 27.2 per 100,000 person-years, depending on the algorithm used to define a case. The lowest crude incidence rate estimate corresponded to the "Aurum (narrow)" case definition, while the highest was for the "Aurum (broad)" definition.

For the Aurum (narrow) definition, the crude incidence rate for the study period was 4.5 per 100,000 person-years. Over the 10-year study period, there was a gradual increase in the incidence rate, from 3.0 per 100,000 person-years in 2008 to 6.2 per 100,000 person-years in 2018 (Table 5. 8). Compared to 2008, a 100% increase in the incidence of IPF was observed in 2018 (IRR 2.06, 95% CI 1.6–2.5) (Table 5. 8).

For the Aurum (broad) definition, the overall incidence rate was 27.7 per 100,000 person-years. The incidence rate increased throughout the study period, from 22.4 per 100,000 person-years in 2008 to 28.6 per 100,000 person-years in 2018 (Table 5. 9). There were small peaks observed in 2012 and 2015. Overall, there was an increase in incidence in 2018 compared to 2008.

Both Aurum-only-based case definitions showed a notable rise in crude incidence between 2011 and 2012.

*Table 5. 8 Incidence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for **Aurum narrow** cohort definition*

	<i>Number of incident cases</i>	<i>Person years (denominator)</i>	<i>Crude Incidence Rate per 100,000 person-years</i>	<i>Incidence Rate ratios (95% CI)</i>
<i>Year</i>				
2008	139	4,633,499	3.0 (2.52-3.54)	Ref
2009	130	4,729,643	2.7 (2.30-3.26)	0.9 (0.72-1.16)
2010	132	4,812,161	2.7 (2.30-3.25)	0.9 (0.72-1.16)
2011	116	4,892,054	2.3 (1.96-2.84)	0.7 (0.62-1.01)
2012	272	4,977,189	5.4	1.8

			(4.83-6.15)	(1.49-2.24)
2013	229	4,973,702	4.6 (4.03-5.24)	1.5 (1.25-1.90)
2014	230	4,983,217	4.6 (4.04-5.25)	1.5 (1.25-1.91)
2015	293	5,093,740	5.7 (5.11-6.45)	1.9 (1.57-2.36)
2016	317	5,223,448	6.0 (5.42-6.77)	2.0 (1.66-2.48)
2017	298	5,338,231	5.5 (4.97-6.25)	1.8 (1.52-2.28)
2018	335	5,430,009	6.1 (5.53-6.87)	2.0 (1.69-2.51)

Table 5. 9 Incidence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum broad cohort definition

	<i>Number of incident cases (numerator)</i>	<i>Person years (denominator)</i>	<i>Crude Incidence Rate per 100,000 person-years</i>	<i>Incidence Rate ratios (95% CI)</i>
<i>Year</i>				
2008	1,039	4,633,499	22.4 (21.09-23.84)	Ref
2009	1,182	4,729,643	25.0 (23.6-26.47)	1.1 (1.03-1.21)
2010	1,196	4,812,161	24.8 (23.48-26.32)	1.1 (1.02-1.21)
2011	1,258	4,892,054	25.7 (24.33-27.19)	1.1 (1.05-1.24)
2012	1,547	4,977,189	31.1 (29.57-32.69)	1.4 (1.26-1.48)
2013	1,477	4,973,702	29.7 (28.22-31.27)	1.3 (1.2-1.41)
2014	1,427	4,983,217	28.6 (27.19-30.19)	1.3 (1.16-1.36)
2015	1,547	5,093,740	30.4 (28.9-31.95)	1.3 (1.23-1.44)
2016	1,533	5,223,448	29.3 (27.92-30.88)	1.23 (1.19-1.39)
2017	1,550	5,338,231	29.1 (27.63-30.55)	1.3 (1.17-1.37)
2018	1,549	5,430,009	28.6 (27.15-30.01)	1.3 (1.15-1.35)

5.3.1.2. HES APC-based definition

The "HES only" definition yielded an average crude incidence rate of 21.9 per 100,000 person-years over the study period, which was slightly lower than the 27.8 per 100,000 person-years observed for the Aurum (broad) definition. This definition showed an incidence rate (per 100,000 person-years) which rose from 21.9 in 2008 to 22.4 in 2014, and then dipped down from 2014 onwards, from 22.4 to 17.4 in 2018. This dip alludes to a decline in the number of cases found in hospitals after 2014.

The decline in hospitalised cases suggests that, in the HES-only cohort where I am focusing solely on hospitalised patients, we are looking at the count of individuals admitted with IPF for the first time in each year. Based on this data, I believe we can infer that fewer patients have been hospitalised for IPF since 2014.

*Table 5. 10 Incidence rate Poisson regression modelling of idiopathic pulmonary fibrosis for **HES-APC** cohort definition*

	<i>Number of incident cases (numerator)</i>	<i>Person years (denominator)</i>	<i>Crude Incidence Rate per 100,000 person-years</i>	<i>Incidence Rate ratios (95% CI)</i>
<i>Year</i>				
2008	1016	4633065	21.9 (20.6-23.32)	<i>Ref</i>
2009	1041	4729093.5	22.0 (20.7-23.39)	1.0(0.93-1.1)
2010	1029	4811507.5	21.4 (20.1-22.73)	0.9(0.9-1.07)
2011	1092	4891337	22.3 (21.02-23.69)	1.0(0.95-1.13)
2012	1167	4976449	23.5 (22.12-24.84)	1.0(1-1.18)
2013	1188	4972975.5	23.9 (22.55-25.29)	1.1(1.03-1.21)
2014	1207	4982517	24.2 (22.88-25.63)	1.1(1.03-1.22)
2015	1183	5093117.5	23.2 (21.92-24.59)	1.1(0.99-1.17)
2016	1176	5222935.5	22.5 (21.25-23.84)	1.0(0.95-1.12)
2017	1028	5337937.5	19.3 (18.1-20.47)	0.8(0.8-0.95)
2018	944	5430103	17.4 (16.29-18.53)	0.8(0.71-0.85)

5.3.1.3. Aurum plus HES APC definition

For the combined Aurum and HES definition, the crude incidence rate for the study period was 5.4 per 100,000 person-years. Using the "Aurum (broad) plus HES" case-finding algorithm, the crude annual incidence rate remained relatively constant during the first half of the study period, peaking at 12.5 per 100,000 person-years in 2012 before declining to 4.9 per 100,000 person-years in 2018. In 2012, the incidence rate peaked at 6.8 per 100,000 person-years, and post-2012, there was a decline in the rate. In 2018, there was a nearly 50% decline in incidence (as compared to 2008) (IRR 0.5, 95% CI 1.6–2.5) (Table 5. 11).

Table 5. 11 Incidence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum plus HES APC cohort definition

	Number of incident cases (numerator)	Person years (denominator)	Crude Incidence Rate per 100,000 person-years	Incidence Rate ratios (95% CI)
Year				
2008	486	8,664,321	5.6 (5.1-6.1)	Ref
2009	545	8,789,264	6.2 (5.6-6.7)	1.1 (0.9-1.2)
2010	541	8,890,141	6.1 (5.5-6.6)	1.1 (0.96-1.23)
2011	531	8,972,461	5.9 (5.4-6.4)	1.1 (0.93-1.19)
2012	622	9,081,988	6.8 (6.3-7.4)	1.2 (1.08- 1.37)
2013	573	9,029,539	6.3 (5.8-6.8)	1.1 (1.00-1.27)
2014	538	9,010,140	5.9 (5.4-6.5)	1.1 (0.93-1.19)
2015	535	9,165,557	5.8 (5.3-6.3)	1.0 (0.91-1.16)
2016	439	9,379,464	4.7 (4.2-5.1)	0.8(0.71-0.92)
2017	361	9,608,641	3.7 (3.3-4.1)	0.6 (0.56-0.74)
2018	272	9,811,609	2.8 (2.4-3.1)	0.45 (0.40-0.54)
Sex				
Men	3,400	50,061,481	6.8 (6.55-7.02)	Ref
Women	2,043	50,341,648	4.1 (3.88-4.23)	0.5 (0.46-0.52)

< 70 years	1,268	84,756,083	1.5 (1.41-1.58)	Ref
≥ 70 years	4,175	15,647,046	26.7 (25.88-27.50)	18.9 (17.70-20.08)

5.3.2. The incidence rate for the four cohorts was stratified by sex and age.

- For Aurum-based definitions

When stratified by sex, men had a consistently higher incidence rate than women for both the aurum-based cohorts (Figure 5. 10, Figure 5. 11). Men aged 70 years and above had the highest incidence rate. There was a 2-fold increase in the incidence rate in men aged 70 years in 2018 compared with 2008 (Figure 5. 12, Figure 5. 13). Similar age and sex patterns were evident for both Aurum narrow and broad definitions.

- For HES only

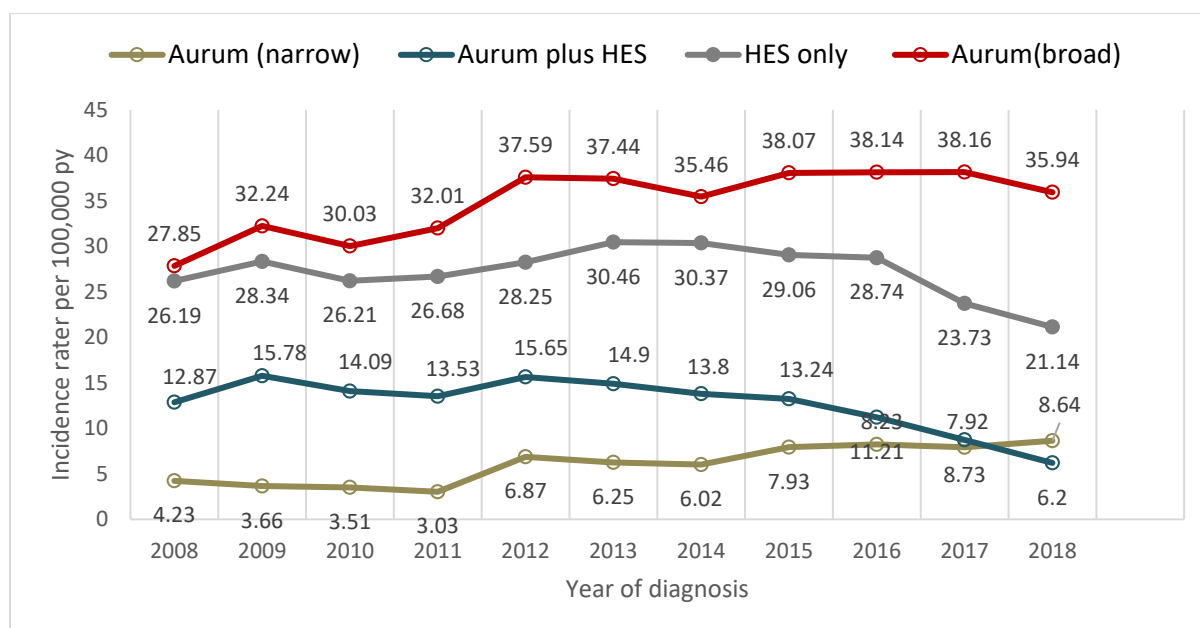
Similar to the Aurum definitions, the overall incidence rate for HES-based definitions was higher in men than women (Figure 5. 10, Figure 5. 11) and in those aged ≥ 70 years (Figure 5. 12, Figure 5. 13), particularly men (aged 70 years and above), whose crude incidence rate was consistently higher across the study period. It was interesting to see that in older subgroups (≥ 70 years), the Aurum broad and HES-only rates are similar; this is logical as older patients are more likely to be hospitalised.

- For Aurum plus HES APC

When stratified by sex, the overall incidence rate was higher in men (6.8 per 100,00 person-years). The age group ≥ 70 years had the highest incidence (26.7 per 100,000 person-years), particularly men aged 70 years and above (37.5 per 100,000 person-years, (Figure 5. 12, Figure 5. 13). The decrease in incidence post-2012 was more pronounced in men (44.1%) than in women (23.9%) (Figure 5. 10, Figure 5. 11).

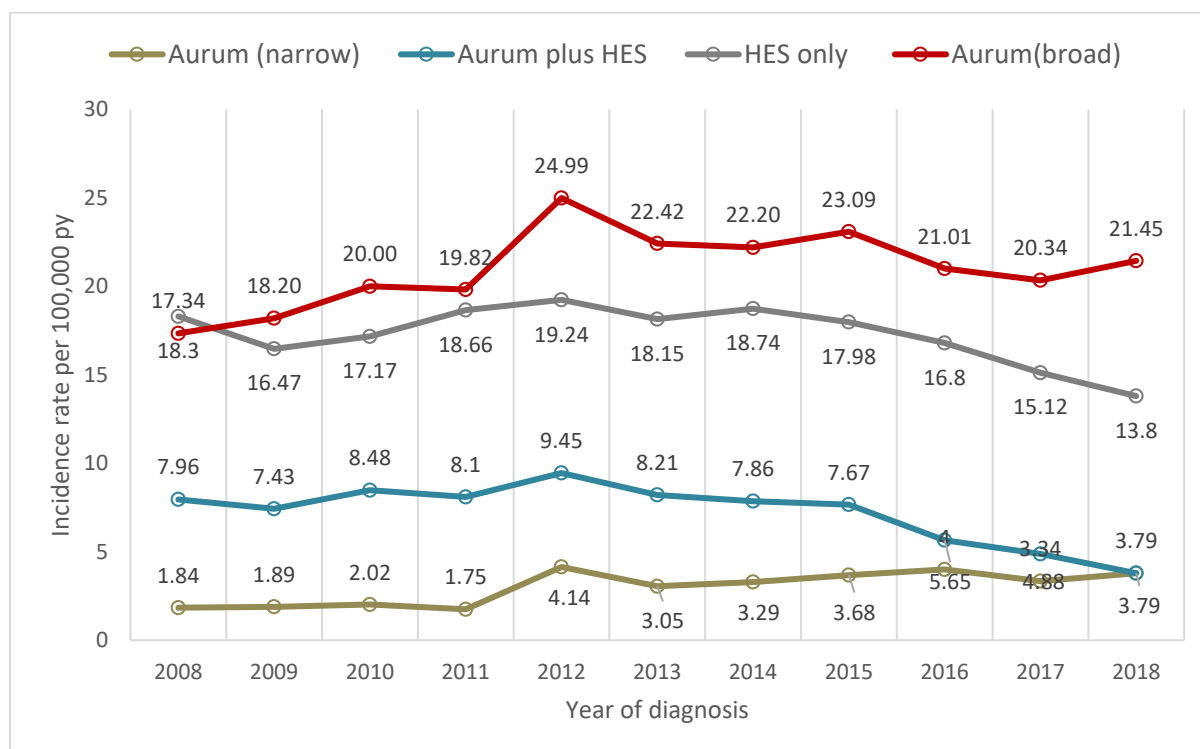
For all definitions, there was evidence of a statistical interaction to suggest higher incidence in men and older age groups (p-value>0.05).

Figure 5. 10 Incidence rate per 100,000 person-years in England, from 2008 to 2018 for men(all ages)



The Aurum (narrow)" definition uses a set of SNOMED CT codes which were considered strongly indicative of IPF; the "Aurum (broad)" definition was based on a larger set of more generic, less specific SNOMED CT codes denoting a probable diagnosis of pulmonary fibrosis, HES only definition uses evidence of hospitalisation with IPF using ICD 10 codes; Aurum plus HES definition detected using a broad list of codes in Aurum and also evidence of hospitalisation for IPF in HES.

Figure 5. 11 Incidence rate per 100,000 person-years in England, from 2008 to 2018 for women (all ages)



The Aurum (narrow)" definition uses a set of SNOMED CT codes which were considered strongly indicative of IPF; the "Aurum (broad)" definition was based on a larger set of more generic, less specific SNOMED CT codes denoting a probable diagnosis of pulmonary fibrosis, HES only definition uses evidence of

hospitalisation with IPF using ICD 10 codes; Aurum plus HES definition detected using a broad list of codes in Aurum and also evidence of hospitalisation for IPF in HES.

Figure 5. 12 Incidence rate per 100,000 person-years in England, from 2008 to 2018 for men aged ≥ 70 years

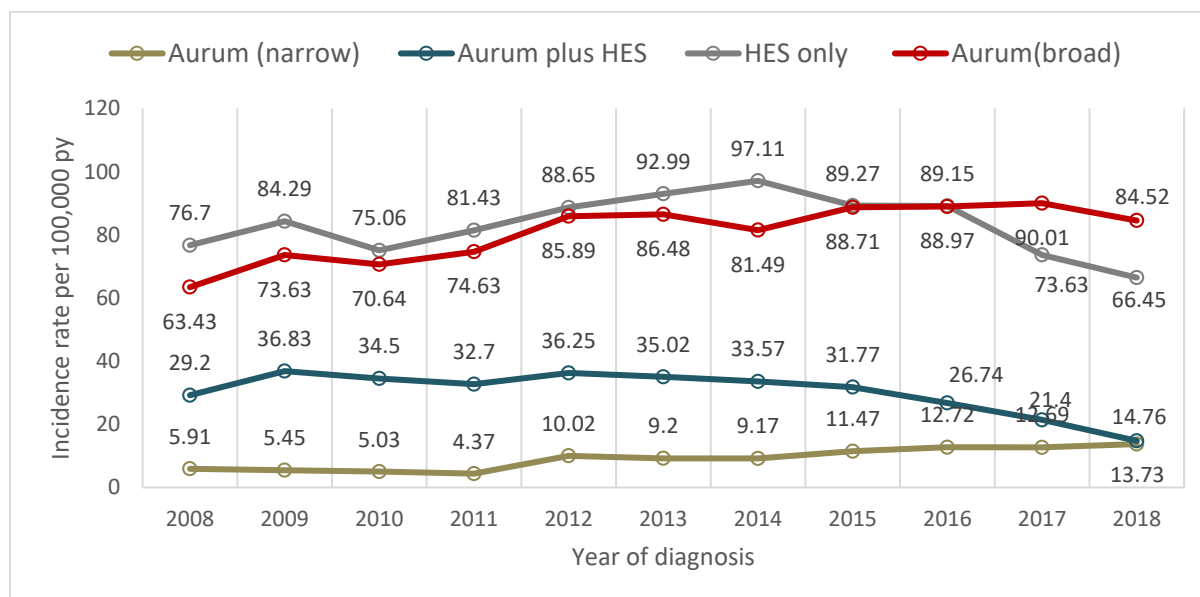
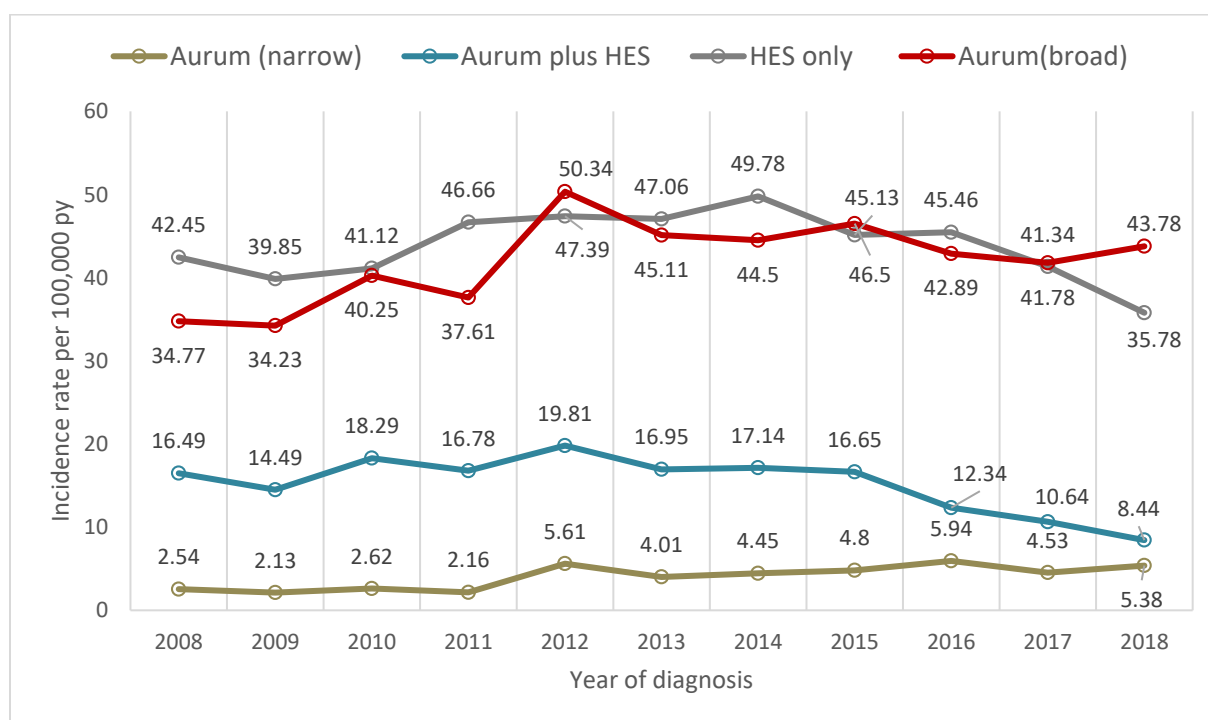


Figure 5. 13 Incidence rate per 100,000 person-years in England, from 2008 to 2018 for women aged ≥ 70 years



5.3.3. Prevalence of IPF

5.3.3.1. Aurum-based definitions

The Aurum (narrow) definition resulted in the lowest prevalence estimate at 16.0 per 100,000 persons [Table 5. 12], while the Aurum (broad) definition yielded the highest at 86.4 per 100,000 persons [Table 5. 13], demonstrating considerable variation, about a factor of five, depending on the definition used. Using the Aurum (narrow) definition, the prevalence remained relatively stable at around 17 cases per 100,000 persons throughout the study period, with no statistically significant change observed (OR=0.99; 95% CI, 0.93–1.06)[Table 5. 12]. However, this overall trend masked an initial decline between 2008 and 2011, followed by a steady increase. In contrast, the Aurum (broad) definition showed a much higher and steadily increasing prevalence, rising from 73.4 per 100,000 persons in 2008 to 91.8 per 100,000 persons in 2018. This represented a 25% increase in prevalence over the study period (OR=1.2, 95% CI 1.21-1.29, p<0.001) [Table 5. 13].

Table 5. 12 Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for Aurum narrow cohort definition

	Number of prevalent cases (numerator)	Persons (denominator)	Crude Prevalence rate per 100,000 persons	Prevalence odds ratios (95% CI)
Year				
2008	1,513	8,721,598	17.3 (16.48-18.24)	Ref
2009	1,429	8,833,792	16.1 (15.35-17.04)	0.9 (0.87-1.00)
2010	1,340	8,929,093	15.0 (14.21-15.83)	0.8 (0.80-0.93)
2011	1,256	9,002,025	13.9 (13.19-14.75)	0.8 (0.75-0.87)
2012	1,367	9,128,385	14.9 (14.19-15.79)	0.8 (0.80-0.93)
2013	1,410	9,033,293	15.6 (14.80-16.45)	0.9 (0.84-0.97)
2014	1,426	9,038,352	15.7 (14.97-16.62)	0.9 (0.85-0.98)
2015	1,509	9,205,258	16.3 (15.58-17.24)	0.9 (0.88-1.01)
2016	1,572	9,420,406	16.6 (15.87-17.53)	0.9 (0.90-1.03)
2017	1,617	9,656,397	16.7	0.9

			(15.94-17.58)	(0.90-1.04)
2018	1,698	9,863,903	17.2 (16.41-18.05)	0.9 (0.93-1.06)

*Table 5. 13 Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for **Aurum broad** cohort definition*

	Number of prevalent cases (numerator)	Persons (denominator)	Crude Prevalence rate per 100,000 persons	Prevalence odds ratios (95% CI)
Year				
2008	6,399	8,721,598	73.4 (71.58-75.19)	Ref
2009	6,684	8,833,792	75.7 (73.86-77.5)	1.0(1-1.07)
2010	6,983	8,929,093	78.2 (76.38-80.06)	1.0(1.03-1.1)
2011	7,297	9,002,025	81.1 (79.21-82.94)	1.1(1.07-1.14)
2012	7,824	9,128,385	85.7 (83.82-87.63)	1.1(1.13-1.21)
2013	8,155	9,033,293	90.3 (88.33-92.26)	1.2(1.19-1.27)
2014	8,331	9,038,352	92.2 (90.21-94.17)	1.3(1.22-1.3)
2015	8,625	9,205,258	93.7 (91.73-95.69)	1.3(1.24-1.32)
2016	8,819	9,420,406	93.6 (91.67-95.59)	1.3(1.24-1.32)
2017	8,974	9,656,397	92.9 (91.02-94.88)	1.3(1.23-1.31)
2018	9,059	9,863,903	91.8 (89.96-93.75)	1.3(1.21-1.29)

5.3.3.2. HES APC-based definition

Mirroring the patterns observed with incidence, when hospital admissions data is used to define IPF cases, the prevalence remained stable for the first half of the study period but declined in the second half, most notably after 2013. This definition is associated with an overall decline in prevalence between 2008 and 2018 [Table 5. 14]).

Table 5. 14 Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for **HES-only** cohort definition

	Number of prevalent cases (numerator)	Persons (denominator)	Crude Prevalence rate per 100,000 persons	Prevalence odds ratios (95% CI)
Year				
2008	2993	8721598	34.3 (33.1-35.57)	Ref
2009	3175	8833792	35.9 (34.7-37.21)	1.0(1-1.1)
2010	3324	8929093	37.2 (35.97-38.51)	1.1(1.03-1.14)
2011	3437	9002025	38.2 (36.91-39.48)	1.1(1.06-1.17)
2012	3571	9128385	39.1 (37.85-40.42)	1.1(1.09-1.2)
2013	3637	9033293	40.3 (38.96-41.59)	1.2(1.12-1.23)
2014	3614	9038352	39.9 (38.69-41.31)	1.2(1.11-1.22)
2015	3585	9205258	38.9 (37.68-40.24)	1.1(1.08-1.19)
2016	3484	9420406	36.9 (35.77-38.23)	1.1(1.03-1.13)
2017	3171	9656397	32.8 (31.71-34)	0.9(0.91-1.01)
2018	2772	9863903	28.1(27.07-29.17)	0.8(0.78-0.86)

5.3.3.3. Aurum plus HES APC definition

For this definition, a similar trend to incidence was observed. From 2008 to 2015, the prevalence is stable (~30 per 100,000 persons), but from 2015 onwards, we see a decline, and by 2018, the prevalence is almost half of 2008 (OR=0.54, 95% CI 0.50-0.57, $p<0.001$) [Table 5. 15].

Table 5. 15 Prevalence rates and Poisson regression modelling of idiopathic pulmonary fibrosis for **Aurum plus HES APC** cohort definition

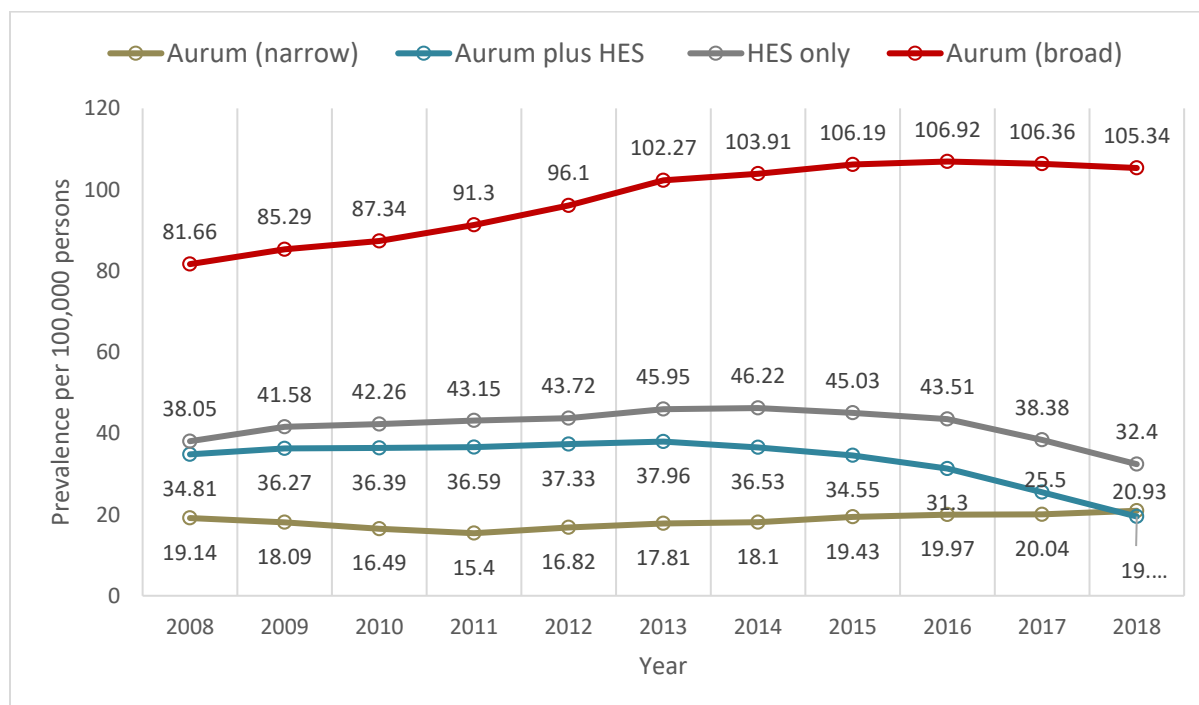
	Number of prevalent cases (numerator)	Persons (denominator)	Crude Prevalence rate per 100,000 persons	Prevalence odds ratios (95% CI)
Year				
2008	2657	8721598	30.4 (29.32-31.65)	Ref

2009	2731	8833792	30.9 (29.77-32.1)	1.0 (0.96-1.07)
2010	2785	8929093	31.1 (30.04-32.37)	1.0 (0.97-1.08)
2011	2798	9002025	31.0 (29.94-32.26)	1.0 (0.97-1.08)
2012	2894	9128385	31.7 (30.56-32.88)	1.0 (0.99-1.11)
2013	2856	9033293	31.6 (30.47-32.8)	1.0 (0.99-1.09)
2014	2765	9038352	30.5 (29.46-31.75)	1 (0.96-1.06)
2015	2639	9205258	28.6 (27.59-29.78)	0.9 (0.89-0.99)
2016	2413	9420406	25.6 (24.6-26.66)	0.8 (0.80- 0.89)
2017	2048	9656397	21.2 (20.3-22.15)	0.7 (0.66- 0.74)
2018	1608	9863903	16.3 (15.51-17.12)	0.5 (0.50-0.57)

5.3.4. The prevalence rate for the four cohorts was stratified by sex and age.

Prevalence was consistently higher in males than females and in the over 70s than in the under 70s irrespective of the choice of case definition.

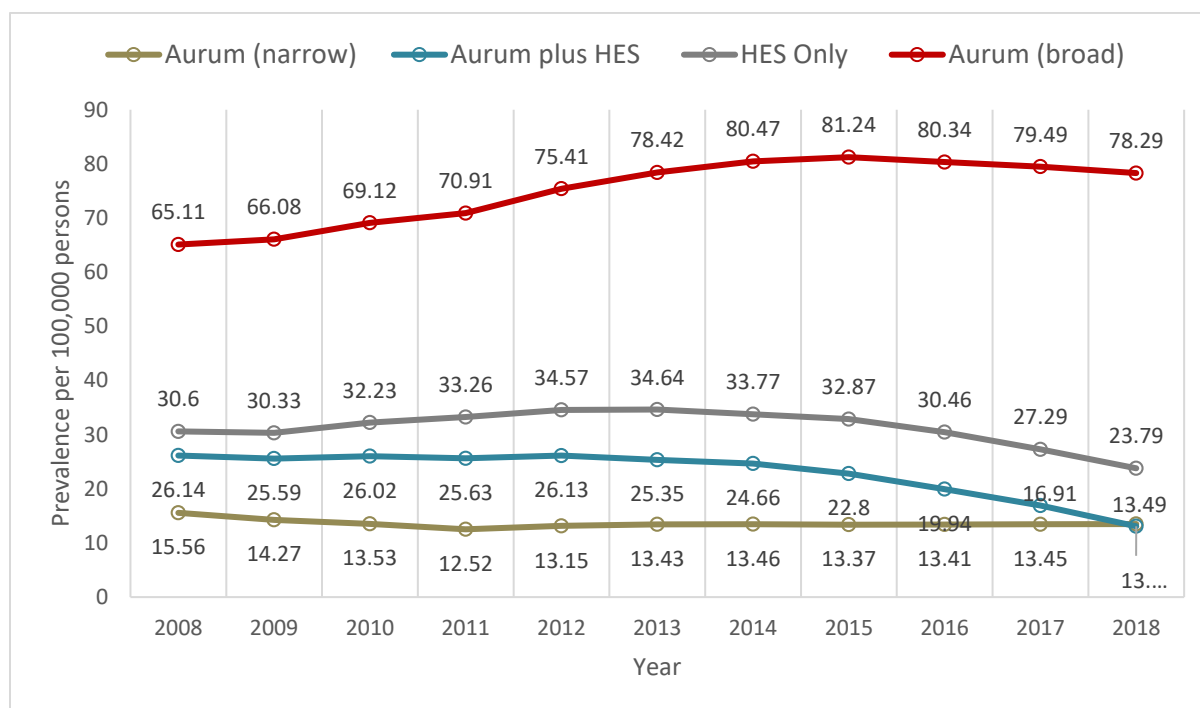
Figure 5. 14 Prevalence per 100,000 persons- in England, from 2008 to 2018 for men (all ages)



The Aurum (narrow)" definition uses a set of SNOMED CT codes which were considered strongly indicative of IPF; the "Aurum (broad)" definition was based on a larger set of more generic, less specific SNOMED CT

codes denoting a probable diagnosis of pulmonary fibrosis, HES only definition uses evidence of hospitalisation with IPF using ICD 10 codes; Aurum plus HES definition detected using a broad list of codes in Aurum and also evidence of hospitalisation for IPF in HES.

Figure 5. 15 Prevalence per 100,000 persons- in England, from 2008 to 2018 for women (all ages)



The Aurum (narrow)" definition uses a set of SNOMED CT codes which were considered strongly indicative of IPF; the "Aurum (broad)" definition was based on a larger set of more generic, less specific SNOMED CT codes denoting a probable diagnosis of pulmonary fibrosis, HES only definition uses evidence of hospitalisation with IPF using ICD 10 codes; Aurum plus HES definition detected using a broad list of codes in Aurum and also evidence of hospitalisation for IPF in HES

Figure 5. 16 Prevalence per 100,000 persons- in England, from 2008 to 2018 for men aged ≥ 70 years

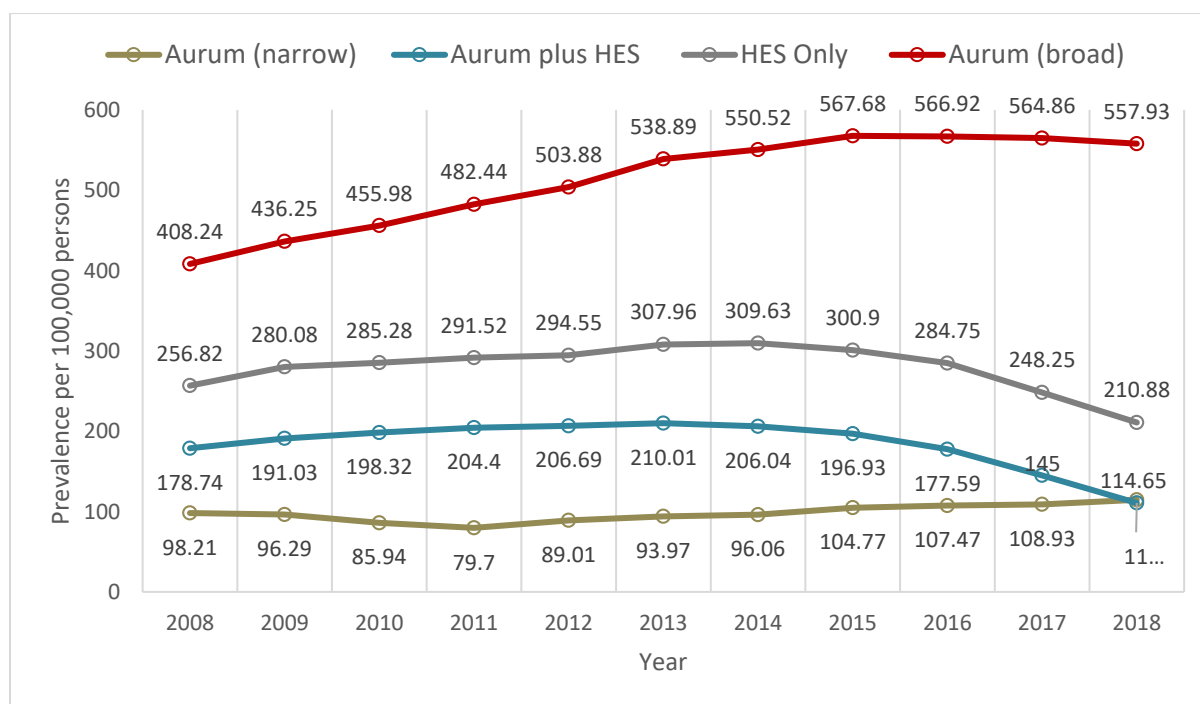
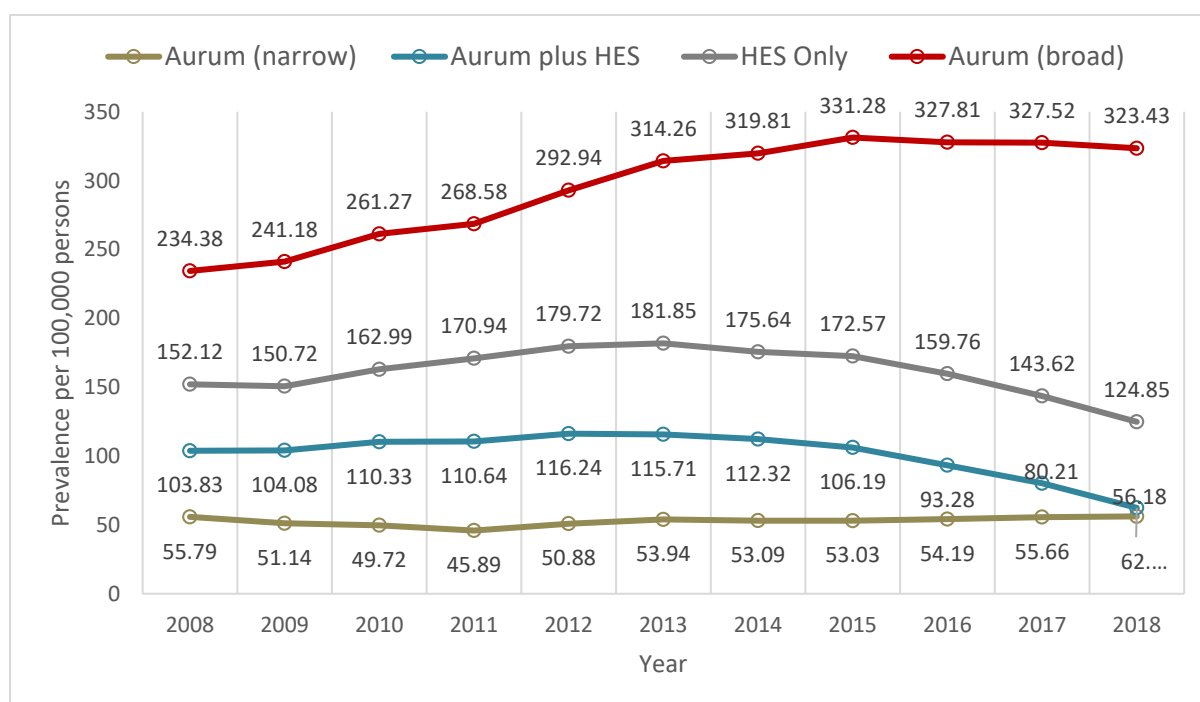


Figure 5. 17 Prevalence per 100,000 persons- in England, from 2008 to 2018 for women aged ≥ 70 years



5.3.5. Mortality rates of IPF

The crude mortality rate for the study period was 7.01 per 100,000 person-years, showing an increase up to 2016 before stabilising in 2017 and 2018 [Table 5. 16]. Men had a higher mortality rate than women,

with 8.65 versus 2.68 per 100,000 person-years, respectively. When standardised to the 2018 England population, the overall mortality rate showed a similar trend to the crude mortality rate, increasing from 5.22 per 100,000 person-years in 2008 to 7.93 per 100,000 person-years in 2018 [Table 5. 16]. The mortality rate ratio (MRR) using 2008 as a reference indicates a roughly 1.5-fold increase in mortality over the study period. As noted above, until 2016 the mortality rate showed an increase, the MRR was 1.64 (64% increase in 2016 as compared to 2008). However, in 2018, the MRR was 1.53, overall, still an increase compared to 2008, but slightly lower than in 2016. Statistically significant interactions suggest higher mortality in men and individuals aged 70 years and older (p-value <0.05). Figure 5. 18 illustrates the adjusted and crude mortality rates per 100,000 persons over the study period.

Figure 5. 18 Adjusted and Crude Mortality rates per 100,000 person-years

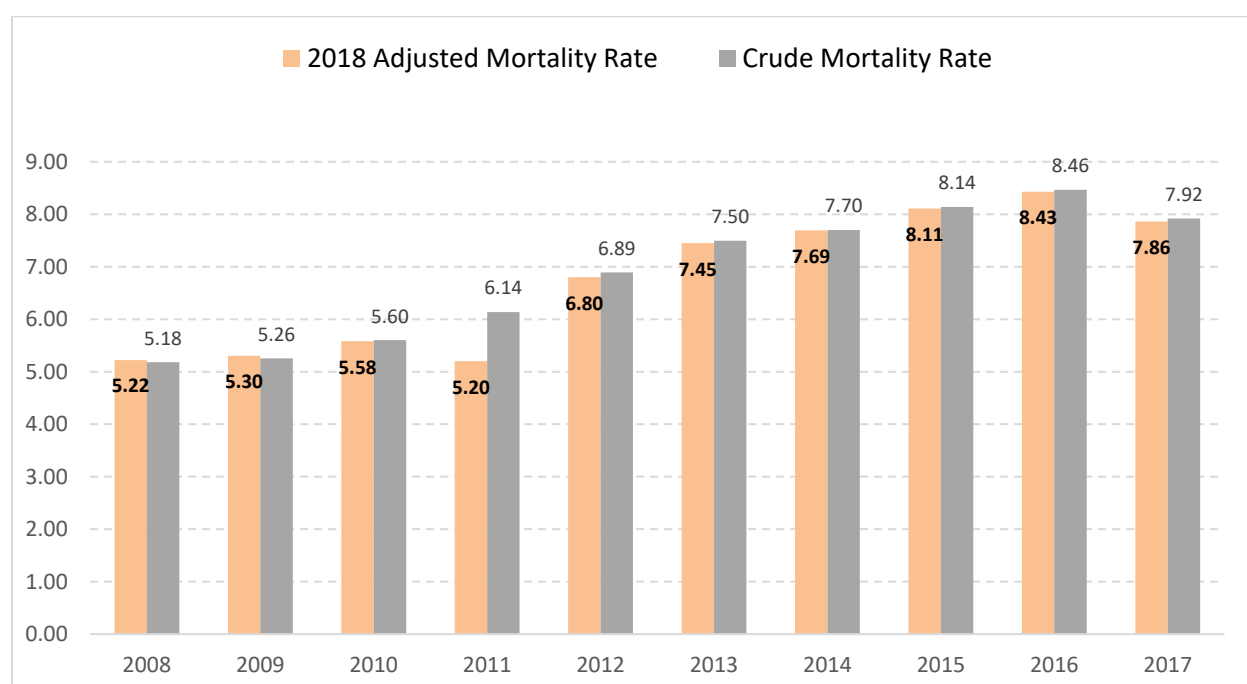


Table 5. 16 Age standardised mortality rates and Poisson regression modelling of deaths from idiopathic pulmonary fibrosis

	Observed Number of Deaths (numerator)	Observed person-years (denominator)	Crude mortality rates per 100,000 person-years	Mortality rates standardised to 2018 population per 100,000 person-years	Mutually adjusted rate ratios (95% CI)
Year					

2008	449	8665627	5.1	5.2	Ref
2009	462	8790555	5.2	5.3	1.0(0.89-1.15)
2010	498	8891392	5.6	5.5	1.1(0.94-1.22)
2011	557	8973684	6.2	5.2	1.2(1.04-1.33)
2012	626	9083174	6.8	6.8	1.3(1.15-1.47)
2013	677	9030670	7.5	7.4	1.4(1.28-1.63)
2014	694	9011196	7.7	7.6	1.5(1.33-1.69)
2015	746	9166544	8.1	8.1	1.5(1.41-1.78)
2016	794	9380399	8.4	8.4	1.6(1.46-1.84)
2017	761	9609504	7.9	7.8	1.5(1.36-1.72)
2018	778	9812383	7.9	7.9	1.5(1.37-1.72)
Sex					
Men	4355	50346855	8.6	4.2	Ref
Women	2687	100415128.5	2.6	2.6	0.44 (0.41-0.46)

5.4. DISCUSSION

To my knowledge, this is the most recent study providing an update on the epidemiological burden of IPF in England following the approval of antifibrotic treatments and the release of the 2011 ATS/ERS IPF diagnostic guidelines^{9,29,156}. In this study, I used four previously validated algorithms to define four study cohorts to establish the incidence and prevalence of IPF: Aurum narrow, Aurum plus HES, Aurum broad, and HES only.

Depending on whether primary or secondary care databases were used to find IPF patients, the crude incidence rate per 100,000 person-years ranged from 4.5 to 27.2 per 100,000 person-years, approximately 6 times higher estimates over the 10 years. The incidence trend varied by study cohort over the decade, with Aurum narrow showing an increase and Aurum plus HES showing a decrease.

Using Aurum-based definitions, incidence doubled for the Aurum narrow definition (3 to 6.1 per 100,000 person-years) and increased by a quarter for the Aurum broad definition (22.4 to 28.6 per 100,000 person-years) over the study period. For both these definitions, incidence peaked in 2012 and stabilised thereafter. As noted above, the difference in estimates between narrow and broad aurum definitions is around 5 to 6-fold. These findings suggest increased use of IPF-specific codes – aligned to the findings observed in Chapter 4. This change in physician behavior suggests increased physician awareness likely related to conformity in the way IPF is labelled and classified, such as the introduction of ATS/ERS diagnostic criteria in 2011, and the approval of antifibrotics for treatment between 2011 and 2014 ^{8-12,29,156,158}.

Conversely, IPF cases found using hospital admission data (HES only and Aurum plus HES) showed approximately a 20% (HES only) to 50% (aurum plus HES) decrease in incidence over the study period. The drop in incidence is towards the later years of the study period. For HES only, at the start of the study period, (2008 – 2015), there was a slight increase (21.9 in 2008 to 23.2 in 2015), and post-2015 there was a gradual decrease in the new cases found in the hospital. A similar trend was observed for the definition that included Aurum and HES (5.6 in 2008 to 6.4 in 2014), with a gradual increase in new cases with a peak in 2012 (6.85). From 2014 onwards there was a drop in new cases (5.9 in 2014 to 2.7 in 2018). This could suggest several possibilities: fewer patients are being admitted for IPF-related issues as they are being effectively managed in primary care. It may also indicate that early diagnosis leads to earlier treatment and better overall management. Additionally, the rise in palliative care may be contributing to a reduced number of hospital visits for these patients.

To put the findings into perspective, several updates in the IPF field since 2008 have influenced trends in physician coding behaviour. Over time, there has been a notable increase in the use of standardised codes for

IPF, accompanied by a decline in the use of generic or outdated codes. This shift likely reflects the impact of updated guidelines and the availability of treatments translating to improved disease awareness. Additionally, greater awareness has contributed to earlier diagnoses, which, in turn, may have led to better management of chronic IPF patients. As a result, patients may experience fewer hospital visits, and this trend could explain the observed decrease in the number of new IPF cases diagnosed in hospital settings.

Prevalence (per 100,000 persons) trends were also dependent on the case definition. For the Aurum narrow definition, prevalence remained relatively stable over the 10 years (17.3 in 2008 to 17.2 in 2018), while the Aurum broad definition showed a steady increase by 25% (73.34 in 2008 to 91.84 in 2018). For the HES only definition, prevalence showed a similar trend to incidence, which increased from 2008 (34.3) to 2013 (40.3) but declined after 2013 (from 40.3 in 2008 to 28.1 in 2018). For the definition which included Aurum and HES, the prevalence halved over the study period. For the first half, the prevalence is stable (around 31), but post-2013, there is a decrease from 31.6 in 2013 to 16.3 in 2018. (Consistent with the literature, this study found that incidence and prevalence were higher in men and individuals aged ≥ 70 year¹³).

To better understand the varying trends in IPF prevalence across different definitions, it's essential to consider the clinical and demographic differences between patient populations identified using these criteria. For instance, a decrease in prevalence when using HES definitions suggests that patients identified through hospital data tend to have more severe diseases compared to those identified through primary care data, such as the Aurum definitions. This observation points to the fact that individuals captured in hospital settings are often sicker and may present with more advanced disease or exacerbations, potentially due to later-stage disease, which naturally may impact survival (as shown in Chapter 4, Table 4.6). On the other hand, early diagnosis in primary care (Aurum) could contribute to improved health outcomes. Early identification of IPF allows for timely interventions, such as anti-fibrotic therapies and better disease management, which may help slow disease progression and reduce the need for hospital admissions. This explains why prevalence estimates using Aurum definitions may be higher if patients are being diagnosed earlier, managed more effectively, and living longer. Overall, the variation in prevalence estimates across different definitions highlights the impact of where and how patients are diagnosed and managed (same as seen in incidence).

Overall, keeping in mind the incidence and prevalence findings across different definitions, HES, tends to capture more advanced cases with poorer outcomes, while primary care-based definitions, such as Aurum, may reflect earlier diagnoses and better-managed conditions, which may lead to longer survival. This emphasises the importance of considering different case definitions and clinical settings when defining IPF and interpreting findings.

Comparison to literature

To date, two large database studies have reported the incidence of IPF in the U.K. Strongman et al.¹³ used the CPRD GOLD dataset for the study period 2000 to 2012, categorising clinical codes to define IPF as 'broad' and 'narrow.' The Aurum narrow definition used in this study is comparable to the 'narrow' definition in Strongman et al. Authors reported an incident rate per 100,000 patient-years for the narrow case definition of 2.8 (95% CI 2.7–3.0), while here, I estimated 4.5 (95% CI 4.3 to 4.7) for Aurum narrow. For the overlapping years (2008 to 2012), this study's incidence rate per 100,000 person-years ranged from 3.0 to 5.4, compared to Strongman et al.'s 2.14 to 2.18. Key differences between the studies can be explained by the following reasons i) this study's calculation of the incidence rate was post-major advances in IPF, and ii) differences in the list of clinical codes used to define 'narrow,' with my list being more restrictive. Despite this, I found a higher incidence rate, suggesting a change in coding practices over time coupled with an increased use of standardised IPF codes.

Navaratnam et al.²⁴ reported an incidence rate of 8.04 (95% CI, 7.12-9.08) in 2008 using the Health Improvement Network (THIN) dataset, whereas, in my analysis, a lower rate (2.8–4.5) was observed. The comparison is limited to one year due to the different study periods. Differences in underlying data (THIN vs. Aurum) and clinical code lists used to define IPF contribute to discrepancies (this analysis had a more restrictive and IPF-specific list for Aurum narrow).

Recent literature reviews reported a global IPF incidence range of 3–9 per 100,000 person-years, noting heterogeneity in reporting and IPF ascertainment methods¹⁵⁹. Most studies used 'broad' and 'narrow'

definitions, suggesting that the actual incidence lies between the two. In the US, Raghu et al.¹⁵⁰, examining Medicare patients (aged ≥ 65 years), found similar results to Aurum's narrow findings here.

This study's findings are aligned with previous findings that show wide variations in estimated incidence rates and trends depending on the IPF definition. Here I highlight the impact of different diagnostic codes (generic or focused) and data sources (Aurum or HES APC) on incidence and prevalence estimates. For example, using a broad list of generic codes in Aurum likely overestimates due to nonspecific IPF codes, while relying on hospital admissions likely underestimates due to a drop in IPF-related admissions. This study reported an increase in incidence for narrow codes over the study period, coinciding with increased use of IPF-focused codes in Aurum.

The mortality rate was determined using registered death records, with the crude mortality rate ranging from 5.1 to 7.9 per 100,000 person-years from 2008 to 2018. There is an increase in IPF-specific mortality by 50% over the study period (MRR=1.5, 95% CI 1.37–1.72). However, in the later years, there has been no change or marginal reduction in the mortality rate. To keep in mind, cause specific – mortality assessment was not conducted within the incident cohort. The aim here was to determine IPF-specific mortality within the population at risk during the study. Some reasons for the reduced number of deaths due to IPF in later years may include the expansion of palliative care services, although palliative care focuses on improving quality of life by managing symptoms and reducing hospitalisations rather than directly reducing death rates. However, it may be as a result of better outcomes and, in some cases, longer survival by preventing complications and managing the conditions. The rollout of antifibrotic drugs, pirfenidone and nintedanib, approved for use in the U.K. in 2013 and 2016, respectively, alongside the phasing out of the previous standard of care, may have also contributed to this trend (which the 2012 PANTHER trial revealed was associated with accelerated mortality and hospitalisation)¹⁶⁰.

Comparison to literature

Navaratnam et al.¹⁴² also reported a mortality rate of 5.1 per 100,000 person-years from 2005 to 2008, similar to my 2008 rate (5.2 per 100,000 person-years). In an updated article, they reported a crude mortality rate of

7.7 per 100,000 persons from 2011 to 2015, comparable to my rate for the same period (7.29 per 100,000 person-years)²⁴. A comparison of mortality rates with Strongman et al. was not feasible due to methodological differences (Strongman et al. assessed mortality in patients who had IPF versus in this chapter, I determined how many patients died of IPF irrespective of whether they had IPF diagnosis in their medical history). A recent study by Gonnelli et al.¹⁶¹, aimed to provide estimates of IPF mortality across Europe from 2013 to 2018 using death certification – a similar approach to this study for the U.K. Gonnelli and colleagues reported similar findings to this study, the age and sex-standardised IPF mortality rate in the U.K. was 8.9 in 2013 (this study: 7.5) to 9 in 2018 (this study: 7.9). Similar to what I observed, there was a rise in the IPF mortality rate from 2013 to 2016, which then stabilised.

This study demonstrates that estimates of IPF incidence and prevalence derived from routinely collected healthcare data are highly sensitive to the choice of clinical codes used to identify cases. In the U.K. context, case-finding algorithms that use evidence of hospitalisation for IPF are less suited for estimating incidence and prevalence (using more recent data). Based on my findings and estimates reported in the literature to date, I approximate the burden of IPF in the U.K. is likely halfway between the two Aurum-based estimates, suggesting an incidence of around 15-16 per 100,000 person-years. To understand how many newly diagnosed cases each year, I scaled up to England's population¹⁶², which suggests 8,000 to 9,000 new cases of IPF diagnosed each year. Similarly, to assess the burden of IPF and the total number of people living with this chronic condition, I scaled up prevalence (around 35 per 100,000 persons). This suggests 24,000-25,000 people currently living with a diagnosis of IPF in England. When looking at the number of people dying each year of this condition, I estimate that approximately 3,500 adults die each year due to IPF.

Several strengths of this study include using primary and secondary linked data. The strength of this data is detailed in Chapter 3. Another merit of this study is the long study period of 11 years, which allowed for an observation of estimates of disease occurrence, mortality, and temporal changes. However, it is important to note this chapter relies on how physicians recorded this chronic condition. The estimates reported are sensitive to any changes in this disease recording. In summary, this study provides a comprehensive update on the epidemiological burden of IPF in England. What we see so far is that increased awareness of IPF has led to

better detection and documentation in primary care settings. However, how IPF is defined and which codes are used play a crucial role in estimating disease burden.

Chapter 6

Changes in survival and all-cause mortality in people with Idiopathic Pulmonary Fibrosis in England, 2010 to 2021

Overview

Despite trial evidence suggesting a benefit from antifibrotic therapies, it remains uncertain whether survival rates for IPF patients in the U.K. have improved since their introduction in the mid-2010s. This study used four previously validated cohorts to calculate median survival times for newly diagnosed IPF patients from 2010 to 2018. Additionally, age- and sex-adjusted hazard ratios for death (relative to 2010) and age- and sex-standardised all-cause mortality rates were estimated for the period from 2010 to 2021. Results indicated no significant difference in median survival between those newly diagnosed in 2010 and 2018. Age- and sex-adjusted Cox models with all-cause mortality as the outcome also did not show improved survival, while age- and sex-standardised mortality rates increased over the study period up to 2019 with a marginal decline or flattening for the last two years. These trends were consistent regardless of IPF case definitions. The conclusion was that between 2010 and 2021, survival time among IPF patients remained unchanged, while the all-cause mortality rate increased, indicating the need for longer follow-up to fully evaluate the impact of antifibrotics on this population. This chapter is being drafted into a manuscript for publication and was presented as a poster at the European Respiratory Society Conference 2024.

6.1. Background

Historically, treatment options for IPF were limited, and data before the availability of antifibrotic treatments showed a median survival of just 3-4 years post-diagnosis^{13,149,150,163-165}. However, several important developments in the U.K. have aimed to improve patient outcomes. These advancements include the development of specialist care and the expansion of palliative care services in the U.K. Reimbursement of antifibrotic therapies (pirfenidone and nintedanib, along with the publication of NICE guidelines for diagnosing and managing IPF, have contributed significantly to patient care^{9,10,25,31,122,146,166-169}.

Clinical trials have demonstrated nintedanib and pirfenidone's efficacy in reducing the decline in lung function in IPF patients¹⁷⁰. A few major trials, such as INPULSIS, CAPACITY, and ASCEND, have shown that in patients with mild to moderate FVC impairment, antifibrotics reduced the rate of decline in FVC by about 50% over one year^{28,29}. Additional data has shown a sustained effect in reducing the decline in lung function beyond year 1¹⁷¹. While these drugs have been shown to improve outcomes for IPF patients, they are only able to slow down disease progression but not halt, much less reverse fibrosis (cure the disease).

What have the clinical trials demonstrated to date? There is growing evidence that nintedanib and pirfenidone reduce the risk of acute deterioration in lung function and improve life expectancy by reducing the rate at which IPF progresses. Antifibrotics have not been shown in large clinical trials to provide significant relief of cough, dyspnea, or quality of life that is impaired due to IPF. It is unclear whether antifibrotic therapies have a meaningful effect on symptoms. However, evidence shows worsening of FVC is associated with a greater worsening in health-related quality of life; therefore, reducing the rate at which FVC declines using antifibrotic therapies may reduce the rate at which functional capacity, symptoms and quality of life worsen^{172,173}. These individual clinical trials were not powered to show significant effects on acute exacerbations and mortality. RCTs demonstrated that nintedanib and pirfenidone reduce lung function deterioration in IPF patients, with significant effect across subjects with both preserved FVC^{28,29,156,174} and more advanced disease. However, these RCTs were not powered to evaluate survival benefits. Despite the introduction of these therapies, IPF continues to be a leading cause of death among ILDs, with primary mortality causes including

respiratory failure and acute exacerbations. The mortality rate associated with an IPF diagnosis is high and is comparable to that of cardiovascular diseases and certain cancers¹⁷⁵. Meta-analyses of trials, registry-based studies, and large database studies, have emerged to assess whether antifibrotic therapy provides a survival benefit¹⁷⁶⁻¹⁷⁸. These publications indicate that antifibrotics reduce the risk of death and enhance life expectancy. However, these recent studies are limited by methodological constraints¹⁷⁹. All-cause mortality in the trial population was low [pirfenidone: hazard ratio [H.R.] 0.52 [95% CI 0.31–0.87; p=0.0107; nintedanib [H.R.] 0.70 (0.46–1.08), p = 0.0954]^{177,179} and to what extent the results of this highly selected population followed using a robust research methodology are generalisable to real-world patients remains unclear¹⁵⁸. Methodological challenges and a smaller sample size limit these findings. Accounting for the challenges with RCTs and observational studies (immortal time bias), there is need to further explore trends in IPF patient mortality and survival, a follow-up study with data beyond 2023 with robust methodology and extended follow-up time is required to fully assess survival or mortality changes.

In this chapter of my thesis, I aim to analyse changes in survival and mortality in people with IPF in the U.K. over the last decade, focusing on the period following the introduction of antifibrotics. Building on earlier work where I explored the robustness of IPF recording in primary care databases and examined incidence, prevalence, and IPF-related mortality, this final analysis looks specifically at trends in survival and all-cause mortality. To ensure a thorough investigation, I used four defined cohorts (previously validated) to assess whether the trends are influenced by how IPF cases are identified. This approach helps determine if case definition variations affect the observed outcomes, providing deeper insights into the impact of evolving treatment strategies and care on patient survival and mortality.

The analysis is crucial in determining the potential impact of evolving treatment strategies and care pathways on patient outcomes.

6.2. Methods

6.2.1. Data sources

For this study, I used the Aurum and HES-APC linked datasets. I also used the most recent Aurum data build available at the time of the analyses – July 2023¹⁸⁰. HES data was available until March 2021. Chapter 3: Data Sources provides a detailed description of the data sources employed in this study.

6.2.2. Study Population

I used the same cohort definitions validated in Chapter 4 and used in Chapter 5. Similar to Chapter 5, I focused this on four separate cohorts defined using previously validated diagnostic-code-based algorithms: Aurum broad, Aurum narrow, HES only, and Aurum plus HES.

Additional primary criteria for the cohorts were: all men and women at least 40 years of age, registered with a G.P. in England, had data available in the Aurum database, and patients were alive at any time during the study period. The study period ranged from 1 January 2010 to 30 June 2023. The cohort entry period ranged from 1 January 2010 to 31 March 2021 (the end date depended on the availability of HES data at the time of the study).

I considered two cohorts for this study; incident and prevalent cohorts. The prevalent cohort included all old and new cases of IPF; that is, it included a valid IPF diagnosis before this study's start date. An individual was eligible for the incident cohort if the date of the first valid record of the IPF diagnosis code (that meets this study's criteria) occurred during the study period. I also applied a criterion that patients had to have had at least one year of registration with their current G.P. before a diagnosis of IPF (washout period) with no diagnosis of IPF in that one year prior. As defined in Chapter 5, the study entry was the most recent of the following dates: the study start date (1 January 2010), the patient's eighteenth birthday, the date a patient joined the practice or the date of first diagnosis with IPF. Study exit was first of the following dates: study end date (last date of HES data), the date they left the practice, had died, or the last date that the practice contributed data to Aurum. Follow-up time was computed from study entry to study exit. Lung transplantation was not considered for censoring because of the small number of cases, fewer than 120 performed annually¹⁵⁷.

6.2.3. Outcomes

The outcome is death – this was derived using the date of death from CPRD aurum data. The date of death in the ONS data is generally considered more accurate, but the CPRD death date is perfectly fine. Herrett et al. (2015) evaluated the agreement between death dates recorded in CPRD and those in ONS mortality data¹³⁴.

The study found that CPRD data on death dates were highly accurate compared to the ONS data. Another study by Padmanabhan et al.¹³⁸ also examined the quality of mortality data in CPRD by comparing it with ONS data, reinforcing the reliability of death date recording in both datasets.

Survival and all-cause mortality were assessed for the four study cohorts of interest. Aurum-HES APC linked data has recorded the date of death. This variable was used to derive overall survival from diagnosis and all-cause mortality. In addition, I assessed year-on-year survival and mortality for the study cohorts.

6.2.4. Comorbidities

As described in Chapter 4, I assessed the prevalence of various demographic and clinical risk factors (age, sex, smoking history, COPD, asthma, lung cancer, ischemic stroke, heart failure, ischemic heart disease, myocardial infarction, and gastroesophageal reflux) as recorded in primary care. All code lists are available in the appendix to Chapter 4. These comorbidities were assessed at the time of index diagnosis, which is the date of IPF diagnosis.

6.3. Statistical Methods

At first, I examined the baseline characteristics of patients in each cohort, focusing on age, sex, comorbidities of interest, and year of diagnosis.

6.3.1. All-cause mortality and Year-on-year mortality

I computed crude all-cause mortality rates for each year from 1 January 2010 to 21 December 2020 (end of HES data was 31 March 2021) stratified by age groups (18 to 44 years, 45 to 54 years, 55 to 64 years, 65 to 74 years, 75 to 85 years and, 85 years and above) and sex in our IPF cohorts. All-cause mortality was calculated as the number of deaths in each IPF cohort by the total population at midyear (1 July) for that year. Mortality was expressed as per 100,000 persons. Age- and sex-stratum-specific mortality rates for each calendar year were calculated. I calculated age- and sex-standardised mortality rates for England (using the ONS data as standard population). I used the Cox Proportional Hazard model to assess the hazard of death for each year of diagnosis, adjusting for sex and age.

6.3.2. Median and year-on-year survival

Using the Kaplan-Meier survival analysis, I estimated the overall median survival time (with 95% CI). I used only incident cases for this to evaluate for a difference in survival time over the years. Survival was assessed by year of diagnosis in the four IPF cohorts. For instance, for patients who met the Aurum narrow criteria, I computed survival times for those diagnosed between 2010 and 2018. I limited the follow-up time to the latest 2018 for sufficient follow-up time (at least 4 years).

Life table analysis was carried out to estimate cumulative mortality (with 95% C.I.s) at 3 years, 3-5 years, 5-7 years, and > 7 years after incident diagnosis. These estimates represent the proportion of individuals who survived for at least 3,5-, and 7-years post-diagnosis.

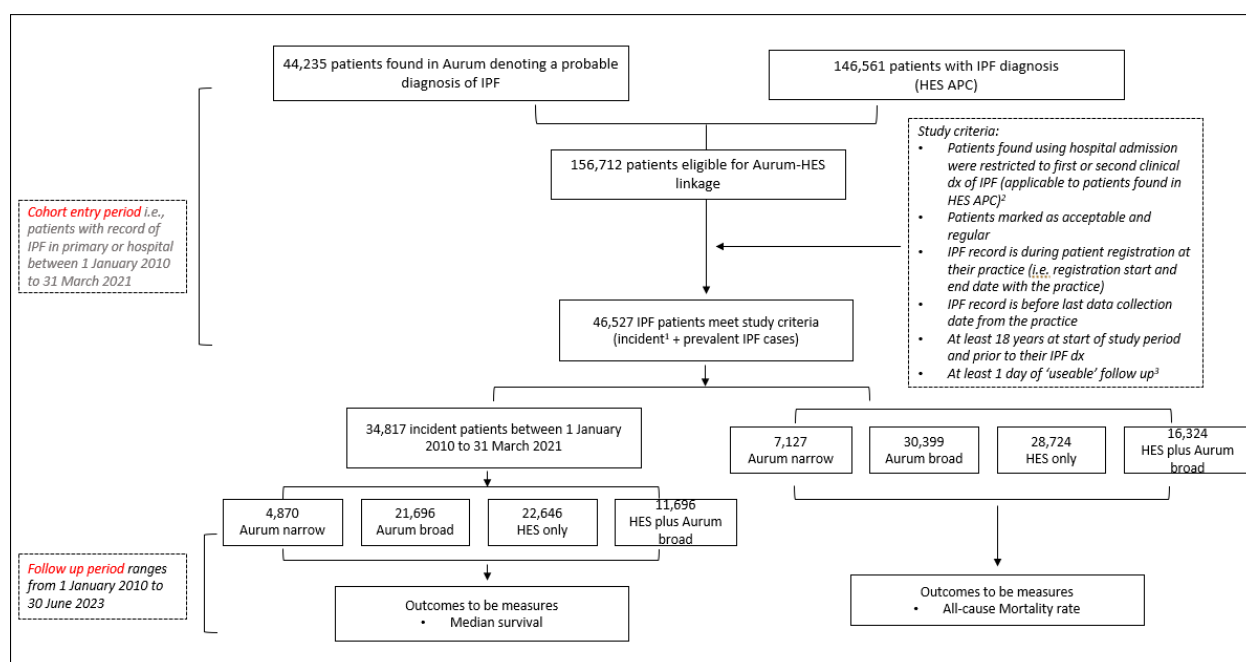
All analyses were performed separately for the four study cohorts.

6.4. Results

6.4.1. Cohort Characteristics

A total of 46,527 patients met the study criteria. Of these 34,817 (74%) were incident cases between 2010 and 2021. The Aurum broad cohort (prevalent, N=30,399) had the highest patient count, unlike the Aurum narrow cohort (prevalent, N=7,127), the most conservative definition, with the least number of patients. HES-based cohorts were in between the Aurum broad and narrow definitions (HES only, N=28,724; HES plus Aurum, N=16,324) (Figure 6.1).

Figure 6. 6 Flow chart of patient inclusion



Cohort definition numbers do not sum up to total numbers as patients were eligible for multiple definitions. IPF, idiopathic pulmonary fibrosis; HES, hospital episode statistics.

Table 6.1 shows the demographic and clinical characteristics of these study cohorts. The median age (75-76 years) was comparable for all four cohorts. Broadly, the findings were similar to Chapter 4 – this was expected as the cohort was designed similarly but with a newer data build. A majority (>80%) of the patients for all four cohorts were in the age group 65 years and above. Cohorts have predominantly male populations, ranging from 58% to 65%. The study population, including patients from HES APC (HES only and HES plus Aurum) appeared marginally sicker, with a higher proportion of patients presenting with underlying conditions and the shortest median follow-up length (2.1 years, IQR: 4.1). The Aurum broad definition had a slightly higher proportion of patients with COPD (12.4% in Aurum narrow versus 17.3% in Aurum broad) and heart failure (11.1% in Aurum narrow versus 14.3% in Aurum broad) compared to Aurum narrow. A majority of the patients (> 64%) had a smoking record in the past 2 years.

Table 6. 6 Clinical and demographic characteristics for the four prevalent study cohorts

	Aurum narrow N (%)	Aurum broad N (%)	HES only N (%)	HES plus Aurum broad N (%)
No. of patients	7127	30,399	28,724	16,324
No. of deaths	4594 (64.4)	18,665 (61.4)	19,026 (66.2)	11,379 (69.7)

Length of follow-up (years, median, IQR)	3.4 (4.7)	3.1 (4.5)	2.1 (4.1)	2.9 (4.5)
Female	2476 (34.7)	11,884 (39.1)	12,109 (42.2)	6206 (38.0)
Age at IPF diagnosis				
Mean (sd)	73.4 (10.5)	73.9 (11.2)	73.1 (12.9)	73.9 (10.9)
Median (IQR)	75(14)	76(14)	75(16)	75(14)
Age groups at IPF diagnosis				
18-44 years	107 (1.5)	552 (1.8)	984 (3.4)	241 (1.5)
45 -54 years	279 (3.9)	1249 (4.1)	1563 (5.4)	623 (3.8)
55-64 years	930 (13.0)	3776 (12.4)	3684 (12.8)	2024 (12.4)
65-74 years	2129 (29.8)	8396 (27.6)	7396 (25.8)	4675 (28.6)
75 -84 years	2729 (38.23)	11437 (37.6)	9807 (34.1)	6228 (38.1)
85 years and above	953 (13.4)	4989 (16.4)	5290 (18.4)	2533 (15.5)
Comorbidities				
COPD	882(12.4)	5267(17.3)	5442(18.9)	2467(15.1)
Asthma (in past 2 years)	633(8.9)	3300(10.8)	3328(11.6)	1676(10.3)
Heart failure	794(11.1)	4342(14.3)	4080(14.2)	2302(14.1)
Ischemic heart disease	887(12.5)	4054(13.3)	3557(12.4)	2192(13.4)
Myocardial infarction	555(7.8)	2684(8.8)	2473(8.6)	1431(8.8)
Gastroesophageal reflux	1260 (17.7)	5380(17.7)	4756(16.6)	2906(17.8)
Smoking in the past 2 years	4630(64.9)	20,431 (67.2)	19,045 (66.3)	10,903(66.8)

6.4.2. All-cause mortality rates

The total deaths were 4,594, 18,665, 19026, and 11,379 for Aurum narrow, Aurum broad, HES only, and HES plus Aurum cohorts, respectively. A higher proportion of deaths was observed in cohorts derived using HES case definitions, with more than 65% mortality. In the HES plus Aurum cohort, the proportion of deaths was ~70%. These are crude numbers of deaths not adjusting for follow-up time.

The average crude mortality rates per 100,000 persons ranged from 3.02 (95% CI [2.7-3.3]) for the Aurum narrow cohort to 13.4 per 100,000 persons (95% CI [12.7-14.1]) for the HES-only cohort. Table 6.2 presents annual mortality, that is, the number of deaths each year per 100,000 persons from 2010 to 2021. Across all study cohorts, there was a gradual increase in the crude mortality rate over time. However, this rate stabilised

over the last 2 years (2019 and 2020) of the study period, showing marginal or no growth in rate. As the HES APC data was available up to 31 March 2021 for HES-derived cohorts, the decline in mortality from 2020 to 2021 should be interpreted cautiously. For Aurum-based cohorts, the rate was flattened.

Table 6. 7 Crude mortality rate per 100,000 persons for four study cohorts

Year of Death	Aurum Narrow	Aurum Broad	HES only	HES plus Aurum
2010	2.2(1.93-2.53)	9.2(8.7-9.92)	10.0(9.47-10.73)	5.7(5.32-6.28)
2011	2.0(1.75-2.32)	10.5(9.88-11.16)	11.2(10.58-11.91)	6.1(5.68-6.66)
2012	2.2(1.95-2.54)	11.5(10.89-12.23)	12.8(12.12-13.52)	7.4(6.9-7.97)
2013	2.5(2.29-2.93)	12.5(11.86-13.25)	13.3(12.63-14.07)	7.9(7.42-8.53)
2014	2.6(2.35-2.99)	12.8(12.2-13.61)	14.2(13.53-15.02)	8.1(7.59-8.71)
2015	3.0(2.69-3.37)	13.8(13.18-14.62)	14.7(14.03-15.52)	8.4(7.94-9.07)
2016	3.3(3.02-3.72)	13.0(12.37-13.76)	14.6(13.89-15.36)	8.4(7.88-9)
2017	3.1(2.87-3.55)	13.9(13.26-14.68)	14.5(13.81-15.26)	8.5(8-9.12)
2018	3.6(3.32-4.05)	13.4(12.79-14.17)	14.8(14.18-15.63)	8.7(8.2-9.31)
2019	3.7(3.37-4.09)	13.7(13.03-14.42)	15.1(14.41-15.87)	8.7(8.24-9.36)
2020	3.8(3.48-4.21)	13.9 (13.22-14.61)	14.5(13.86-15.28)	8.7(8.18-9.29)
2021	3.8(3.47-4.2)	13.0(12.4-13.76)	11.3(10.64-11.9)	7.34(6.88-7.9)

Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum; HES APC cohort patients detected using hospitalisations for IPF in HES APC; Aurum and HES cohort, patients detected using a narrow list of codes in Aurum.

When standardised for age and sex using the ONS population, the all-cause mortality rates for all cohorts increased. Both the crude and age—and sex-standardised mortality rates (ASMR) increased over the study period, a trend that was similar across the four IPF cohorts. Table 6.3 shows the ASMR per 100,000 persons for the study cohorts. ASMR is higher than crude, which indicates adjusting for the elderly population (for all cohorts).

For the Aurum narrow cohort, ASMR increased from 4.3 per 100,000 persons in 2010 to 8.0 per 100,000 persons in 2021 (an increase of 86.0% overall or an average of 7.8% per year); ASMR estimates based on an Aurum broad case definition were higher and increased from 18.0 per 100,000 persons to 27 per 100,000 persons over the same period (a 62% increase overall or an average of 5.7 % per year). For HES only, ASMR increased from 19.33 per 100,000 persons in 2010 to 34.21 per 100,000 persons in 2019, but for the years 2020 and 2021, there has been a decrease compared to 2019. But for the 10 years, there is an increase from 19.33 to 30.15 per 100,000 persons. Similarly, for HES plus Aurum, the definition increased from 11.23 per

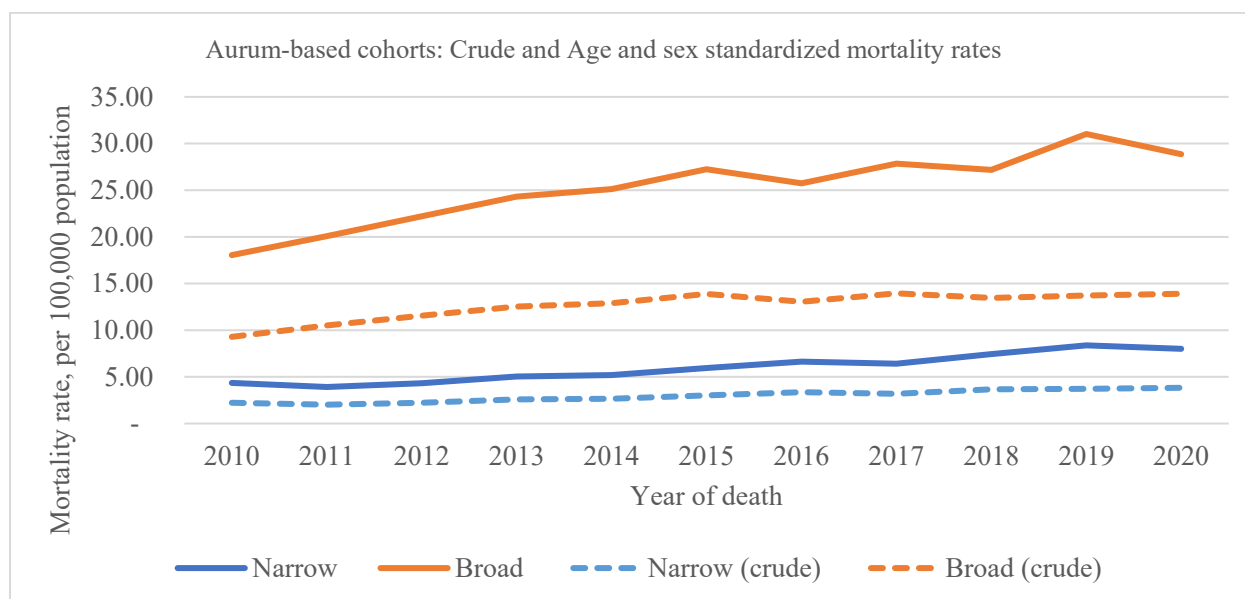
100,000 persons in 2010 to 19.56 per 100,000 persons in 2020 (a 74.1% increase overall). However, similar to the HES cohort, there is a decrease in ASMR for 2020 and 2021 compared to 2019. Crude and ASMRs increased with age and were higher in men than in women. Figure 6.2 and Figure 6.3 show the crude and ASMR for cohorts derived using Aurum and HES data.

Table 6. 8 Age and sex-standardised mortality rate per 100,000 persons for the four study cohorts

Year of death	Aurum Narrow	Aurum Broad	HES only	HES plus Aurum
2010	4.3	18.0	19.3	11.2
2011	3.9	20.0	21.3	11.3
2012	4.3	22.2	24.5	14.9
2013	5.0	24.3	25.6	15.6
2014	5.1	25.1	27.7	15.7
2015	5.9	27.2	28.8	16.9
2016	6.6	25.7	28.7	16.3
2017	6.4	27.8	28.9	17.9
2018	7.4	27.1	29.9	17.4
2019	8.3	31.0	34.2	19.6
2020	8.0	28.8	30.1	18.1
2021	7.7	27.7	23.6	15.3

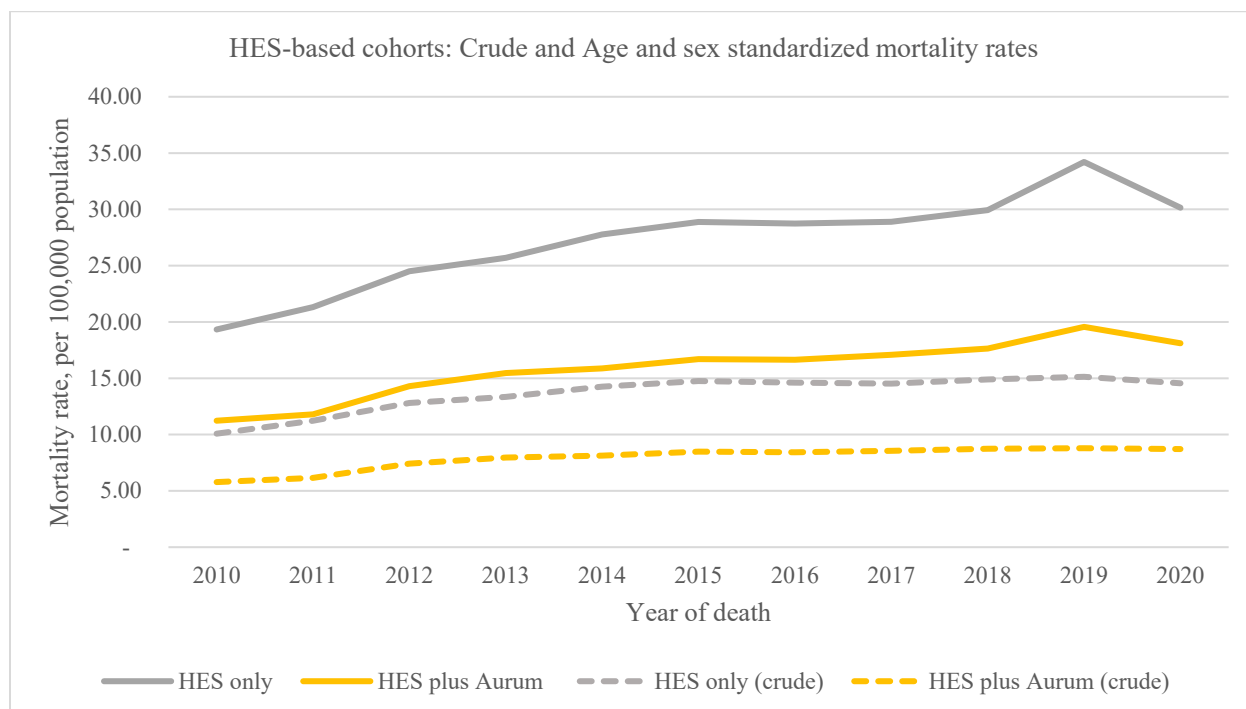
Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum; HES APC cohort patients detected using hospitalisations for IPF in HES APC; Aurum and HES cohort, patients detected using a narrow list of codes in Aurum.

Figure 6. 7 Crude and ASMR for Aurum narrow and Aurum broad cohorts



Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum.

Figure 6. 8 Crude and ASMR for HES only and HES plus Aurum cohorts



HES APC cohort patients were detected using hospitalisations for IPF in HES APC; Aurum and HES cohort patients were detected using a narrow list of codes in Aurum.

6.4.3. Survival

Among the incident cases, there were 4870, 21696, 22,646, and 11696 deaths in the Aurum narrow, Aurum broad, HES only, and HES plus Aurum cohorts, respectively. The proportion of deaths ranged from 56.9% in the Aurum broad cohort to 66.5% in the Aurum plus HES cohort. Table 6.4. illustrates the median survival (95% CI) for these cohorts from 2010 to 2018, with a follow-up cut-off in 2018 for adequate tracking. The Aurum-based cohorts generally exhibited higher median survival than the HES-only and Aurum plus HES cohorts. However, there was no clear upward trend in survival over the years for any cohort and overlapping confidence intervals.

On the contrary, the Aurum narrow cohort showed a decline in median survival over time. Aurum broad cohort saw a minor increase of 0.5 years in median survival from 2010 to 2018 but no consistent year-on-year improvement. Similarly, the HES-based cohorts did not exhibit a clear trend in survival changes. Due to insufficient follow-up time after the introduction of antifibrotics and the absence of a definitive trend, it is challenging to conclude any improvement in patient survival over time. Additionally, evaluating survival rates

before and after introducing antifibrotic treatments (pre-2010-2014 and post-2015-2018) revealed no meaningful improvement in median survival among the four cohorts.

Table 6. 9 Median survival in years (95% CI) among all incident patients for the four study cohorts

Year of diagnosis	Aurum Narrow	Aurum Broad	HES only	HES plus Aurum
2010	4.6(3.7-5.4)	3.7(3.4-4.0)	2.6(2.2-2.8)	3.4 (3.1-3.7)
2011	4.3(3.6-5.1)	3.6(3.2-3.9)	2.4(2.2-2.7)	3(2.8-3.4)
2012	3.9(3.4-4.5)	3.9(3.6-4.2)	2.6(2.3-2.8)	3.3(2.9-3.7)
2013	4.04(3.6-4.7)	4.03(3.7-4.3)	2.6(2.3-2.8)	3.4(3.1-3.7)
2014	4.3(3.7-5.0)	3.8(3.4-4.0)	2.6(2.3-2.7)	3.2(2.8-3.5)
2015	3.9(3.6-4.4)	3.9(3.6-4.2)	2.9(2.3-3.1)	3.3(3.1-3.6)
2016	4.2(3.6-4.6)	4.1(3.7-4.5)	2.8(2.5-3.1)	3.4(3.1-3.7)
2017	3.9(3.6-4.6)	4.3(3.9-4.5)	2.9(2.6-3.1)	3.3(3.0-3.6)
2018	3.7(3.4-4.3)	4.2(3.8-4.4)	2.9(2.6-3.1)	3.1(2.8-3.3)

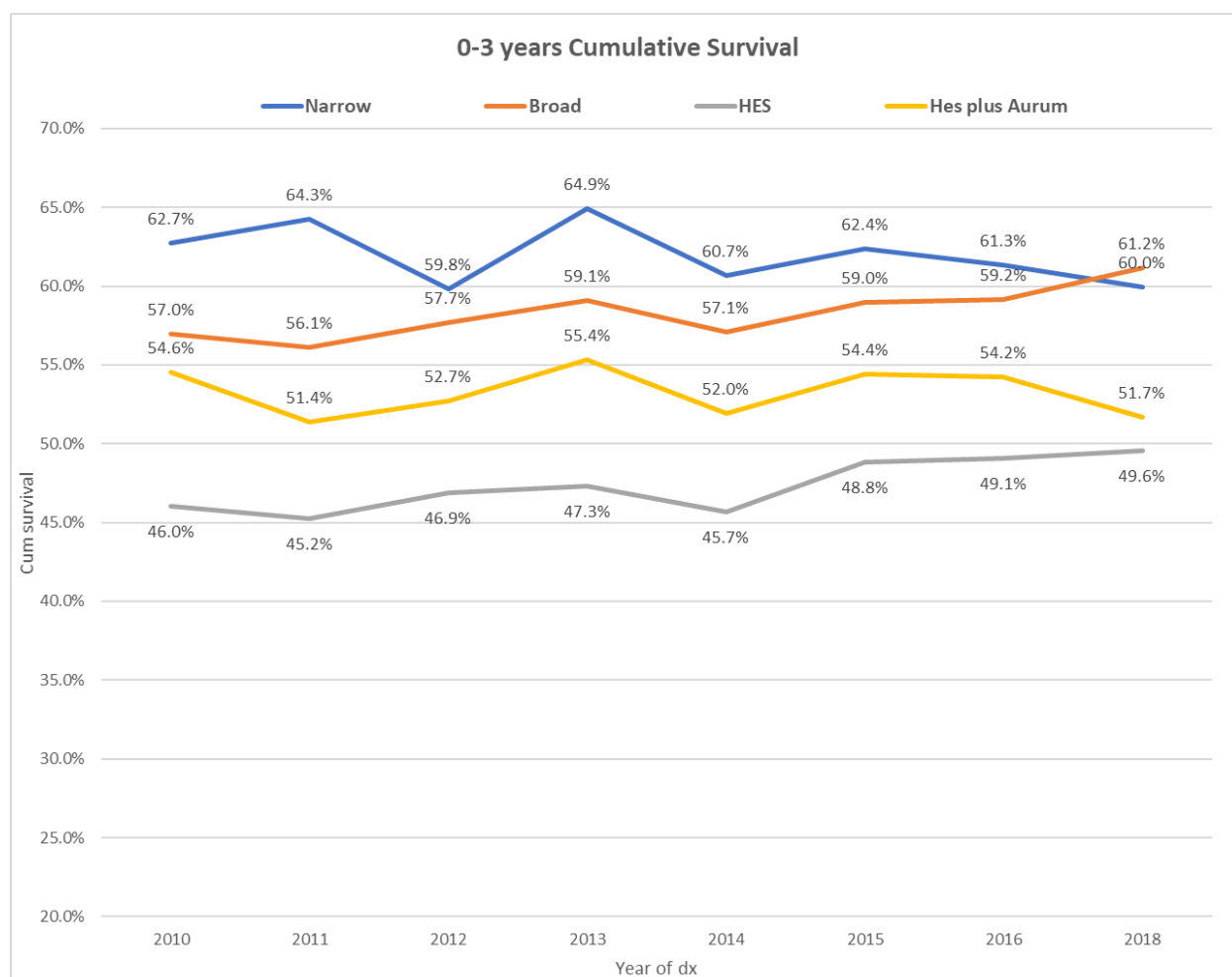
Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum; HES APC cohort patients detected using hospitalisations for IPF in HES APC; Aurum and HES cohort, patients detected using a narrow list of codes in Aurum.

Cumulative survival for the incident cases was assessed by year of diagnosis for the four study cohorts. In general, no distinct improvement in 3-year survival was observed across the four cohorts (Figure 6.4). The 5-year survival showed a decline in only Aurum-based cohorts but no improvement for HES-based cohorts. For the Aurum narrow cohort, more than 60% of patients were alive at 3 years, about 45-50% at 5 years, and less than 40% at 7 years post-IPF diagnosis. When analysed by year of diagnosis (2010 to 2018), the 3-year survival rate remained relatively stable. In contrast, the 5-year survival rate declined, with 47% of patients diagnosed in 2010 still alive at 5 years compared to 39% of those diagnosed in 2018. In the Aurum broad cohort, over half (56% to 60%) of patients were alive at 3 years and over 40% at 5 years, with no clear trend of improvement over time. For the HES-only cohort, 3-year survival ranged from ~52% to ~55%, while for HES plus Aurum, it remained below 50% for all years from 2010 to 2018. A decrease in the 3-year survival over the study period was observed in the HES-only cohort, whereas an improvement was seen from 2015 onwards in the HES plus Aurum cohort. However, follow-up time was insufficient to establish a clear trend.

Although 7-year survival was calculated for the cohorts (Figure 6.6), insufficient follow-up rendered these numbers potentially misleading and not meaningful. No statistically significant survival differences were

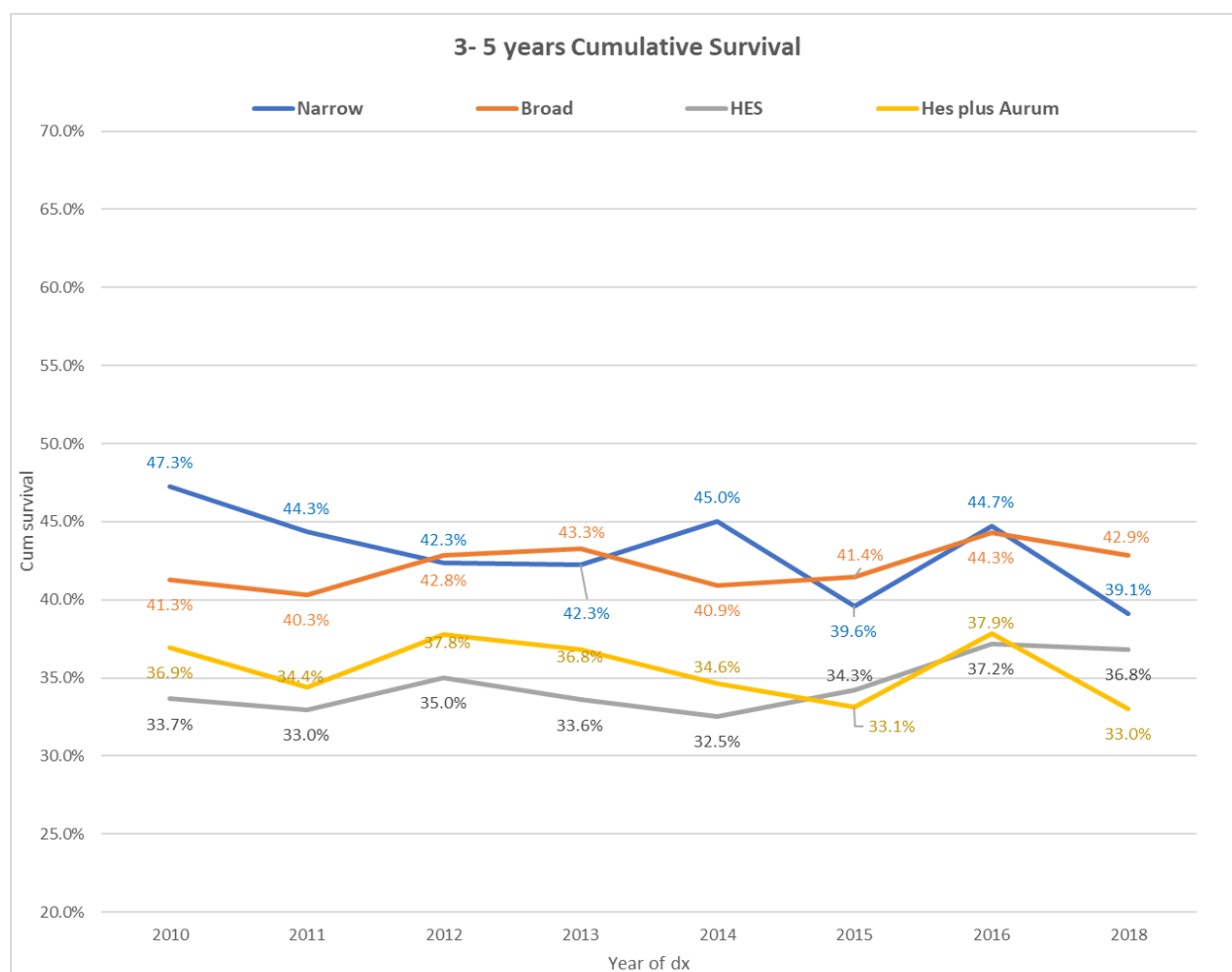
found between the study cohorts ($p > 0.05$), nor was there a significant difference in survival function by year of diagnosis among the cohorts.

Figure 6. 9 Cumulative survival for incident cases for 0 to 3 years for the four study cohorts



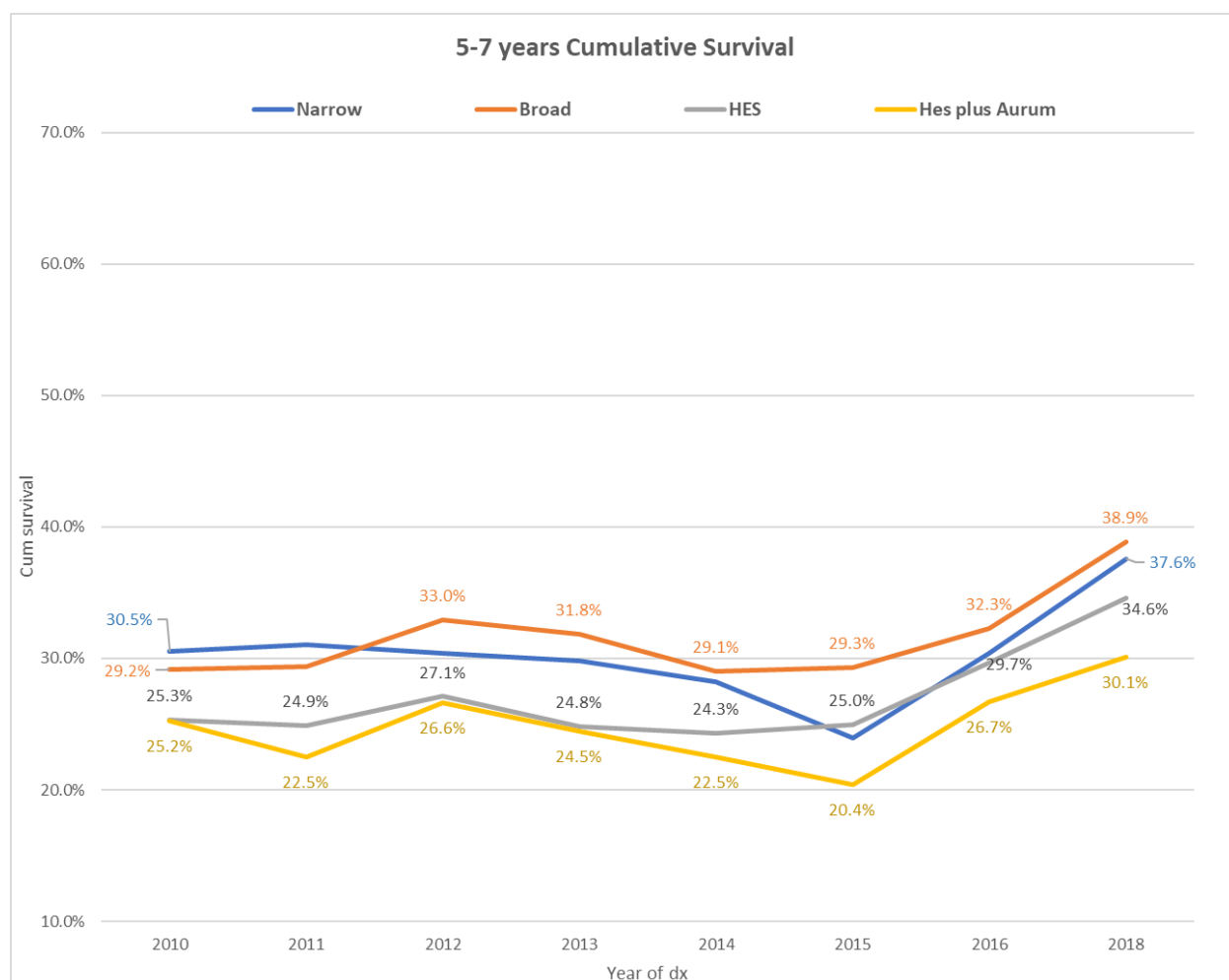
Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum; HES APC cohort patients detected using hospitalisations for IPF in HES APC; Aurum and HES cohort, patients detected using a narrow list of codes in Aurum.

Figure 6. 10. Cumulative survival for 3 to 5 years in incident cases for the four study cohorts



Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum; HES APC cohort patients detected using hospitalisations for IPF in HES APC; Aurum and HES cohort, patients detected using a narrow list of codes in Aurum.

Figure 6. 11. Cumulative survival for 5 to 7 years in incident cases for the four study cohorts



Aurum narrow cohort, patients detected using a broad list of codes in Aurum and also found in HES APC; Aurum broad cohort, patients detected using a broad list of codes in Aurum; HES APC cohort patients detected using hospitalisations for IPF in HES APC; Aurum and HES cohort, patients detected using a narrow list of codes in Aurum.

The Cox proportional hazard model adjusted for demographic variables did not demonstrate that patients diagnosed in 2010 had a higher mortality rate than those diagnosed in subsequent years, indicating no clear evidence of improvement over time (Table 6.5). Although some improvement was observed for the Aurum broad and HES-only cohorts after 2015, the hazard ratio was not statistically significant. Across all four study cohorts, females consistently had a lower mortality hazard than males. Additionally, an increase in age was associated with a higher hazard ratio, with the highest risk observed in patients aged 85 years and above (Table 6.5).

Table 6. 10 Relative effects of calendar year of diagnosis, sex, and age group on the risk of mortality for the four study cohorts

Year of diagnosis	Aurum narrow	Aurum broad	HES only	HES plus Aurum broad
2010 (reference year)	1	1	1	1
2011	0.9 (0.7-1.2)	1.0(0.9-1.1)	1.0(0.9-1.1)	1.1 (0.9-1.2)
2012	1.0(0.8-1.2)	0.9(0.8-1.0)	0.9(0.8-1.0)	0.9(0.8-1.1)
2013	0.9(0.8-1.2)	0.9 (0.9-0.9)	0.9(0.9-1.0)	0.9(0.8-1.1)
2014	1.0(0.8-1.2)	1 (0.9-1.0)	1.0(0.9-1.0)	1.0(0.9-1.2)
2015	1.1(0.9-1.3)	0.9 (0.9-0.9)	0.9(0.8-1.0)	1.0(0.9-1.2)
2016	0.9(0.7-1.1)	0.8(0.8-0.97)	0.8(0.8-0.96)	0.9(0.8-1.1)
2017	0.9(0.8-1.2)	0.8(0.8-0.96)	0.9(0.8-0.98)	0.9(0.8-1.1)
2018	1.1(0.8-.2)	0.8(0.8-0.99)	0.9(0.8-0.98)	1.0(0.9-1.1)
2019	1.1(0.9-1.4)	0.8(0.8-0.99)	0.9(0.8-1.0)	1.0(0.9-1.1)
2020	1.1(0.9-1.5)	0.9 (0.89-0.99)	0.9(0.9-1.1)	1.1(1.0-1.3)
2021	1.2(0.8-1.9)	0.8 (0.80-0.85)	0.9(0.7-1.1)	1.0(0.7-1.3)
Female	0.7 (0.7-0.8)	0.7 (0.7-0.8)	0.7 (0.7-0.8)	0.7(0.7-0.8)
Age (categories) years at IPF diagnosis				
18-45	1	1	1	1
45-54	1.6 (0.6-3.8)	1.9(1.3-2.7)	2.1(1.6-2.6)	1.4(0.9-2.2)
55-64	3(1.3-6.7)	3.4(2.4-4.8)	4.0(3.2-4.9)	2.5(1.8-3.7)
65-74	3.5(1.6-7.9)	4.9(3.5-6.8)	6.2(5.1-7.6)	3.1(2.2-4.6)
75-84	4.9(2.2-10.9)	7.2(5.2-10.1)	9.5(7.8-11.7)	4.5(3.1-6.6)
85 years and above	7.7(3.4-17.4)	11.9(8.5-16.7)	15.8(12.9-19.4)	7.4(5.1-10.8)

6.5. Discussion

The primary aim of this study was to assess whether survival and all-cause mortality have improved in patients diagnosed with IPF, particularly after the introduction of antifibrotic therapies. Across the four study cohorts, there was a consistent rise in crude as well as age- and sex-adjusted all-cause mortality rates over the study period. The data showed a gradual increase in all-cause mortality until 2019, after which there was an indication that this rise was slowing or flattening (particularly in the last two years of follow-up). However, insufficient follow-up data makes it difficult to draw a definitive conclusion. The modest decline seen particularly in the HES-based cohorts suggests a possible reduction in mortality, although the short follow-up period of only two years prevents any firm conclusions. For the HES-only cohort, there was a notable decrease in the ASMR rate per 100,000 persons, from 34.2 per 100,000 in 2019 to 23.2 in 2021. The Aurum-based cohorts exhibited more of a flattening pattern rather than a decline. The HES-only cohort consistently had the highest mortality rates across the study period, likely due to a sicker patient population with a higher proportion of individuals with multiple comorbidities and shorter survival times (median of 2.1 years versus 3-3.4 years in other cohorts). This suggests that patients diagnosed in the hospital setting may have experienced delayed diagnoses, with many showing signs of advanced disease by the time they sought hospital care. The incidence analysis in the previous chapter shows that new diagnoses in hospitals have decreased over time, which hints at improved care within the primary care setting. Thus, this substantiates the idea that patients in the HES cohort represent a more advanced disease stage. Across all four cohorts consistently over the study period, a higher ASMR compared to the crude mortality rate confirmed that age and sex substantially impact IPF disease burden. The unadjusted (crude) rate underestimates the true burden in populations with older and at-risk (older men) demographics.

Another aim of this study was to determine whether survival improved for patients diagnosed after the advent of antifibrotic treatments, for example, whether a newly diagnosed patient in 2010 had a shorter survival than someone newly diagnosed in 2015 (post-antifibrotic availability). However, in this analysis, no improvement in survival was observed.

Our findings differ from those reported by Cottin et al., who analysed an untreated cohort from French National Health System claims data (2015-2016), showing a 3-year cumulative survival of 26-31%, which is notably lower than the 40-60% survival rates observed in our study, depending on the cohort definition¹⁶³. These differences likely stem from variations in the study populations. Our study included primary and secondary care patients, including those on antifibrotic treatment. In contrast, Cottin's study focused on untreated, hospitalised patients (secondary care), who may have been at a more advanced stage of the disease and did not benefit from antifibrotic therapies. This could explain the poorer outcomes observed in the French cohort.

Additionally, Hutchinson et al.¹⁵⁹ reported the highest mortality rates for IPF in the E.U. were in the U.K., showing a strong increase over time. Gonnelli et al. using EUROSTAT data, reported the overall standardised mortality rate for the U.K. ranged between 8.9 and 9.3 per 100,000 person-years, which aligns more closely with our study's a priori narrow estimates¹⁶¹.

Within the U.K., two large epidemiological studies conducted by Strongman et al. and Navaratnam et al. provide additional U.K. context. Strongman et al. reported a median survival of 2.7 to 3 years using 'IPF-like' clinical codes from a primary care database (2000-2012), with no significant improvement in survival over time¹³. Navaratnam et al. noted an increase in age-standardised mortality rates from 1.66 per 100,000 person-years in 1979 to 8.29 per 100,000 in 2019, with an annual rise in IPF deaths of approximately 5%²⁴.

However, it is important to note that these studies predominantly analysed data from the pre-antifibrotic era, which limits direct comparisons with our findings, as our study spans both the pre- and post-antifibrotic periods. Furthermore, Navaratnam et al. assessed IPF-specific mortality while we evaluated all-cause mortality among IPF patients, making direct comparison less feasible. Comparing the risk of all-cause mortality in our study to the literature presents challenges, as there are no exact 1:1 comparisons for the U.K. In the U.S., two recent studies of Medicare beneficiaries demonstrated a reduction in all-cause mortality and a survival benefit for antifibrotic-treated patients compared to untreated patients^{174,181,182}.

Similarly, Iommi et al.¹⁸³ reported comparable findings using an Italian healthcare database. The key difference between these studies is that we did not compare treated versus untreated patients. Instead, we included all incident IPF patients within our study period. We assessed survival trends from the pre-

antifibrotic era through the approval and availability of antifibrotic treatments. This broader inclusion may explain why our findings did not show a marked survival improvement as seen in studies focused on antifibrotic treatment benefits.

The lack of improvement in survival and mortality in IPF patients, as observed in this study, can be defended to several interrelated factors, including the variable uptake of antifibrotic treatments¹⁸⁴, the restrictive lung function eligibility criteria for these therapies¹⁴⁶, and the “watch and wait” approach sometimes employed by physicians. While antifibrotics such as pirfenidone and nintedanib have been shown to slow disease progression, their overall impact on survival in real-world settings remains limited, in part due to challenges in treatment access and initiation. When we think about the low uptake of antifibrotic therapies, one would assume a contrary behaviour given the severity of the condition. However, there are several factors to consider: i) Up until May 2023, the **National Institute for Health and Care Excellence (NICE)** guidelines¹⁴⁶, recommended antifibrotic therapy only for patients with an FVC between **50% and 80%** of the predicted value. This means that patients with higher lung function (FVC > 80%) have not qualified for immediate treatment, leading physicians to adopt a "wait and watch" approach, particularly for those with mild or stable disease. This delay in treatment may allow the disease to progress before any intervention is initiated, which could negatively impact long-term survival outcomes. Another reason could be the side effects of antifibrotic therapies, such as nausea, diarrhoea, and liver toxicity, which are significant. Given that these treatments are not curative, physicians may prefer to defer initiating therapy in slowly progressing patients to maintain their quality of life for as long as possible. Diagnosing IPF can present challenges and there remains diagnostic uncertainty when it comes to early diagnosis. Early symptoms of IPF, such as a persistent cough or shortness of breath, can be nonspecific and overlap with other respiratory conditions. In addition, there are no reliable biomarkers to predict disease progression in IPF. When there is uncertainty in diagnosis and the inability to definitively assess which patients will experience rapid progression, physicians may adopt a more cautious approach and monitor the patient until signs of disease progression are evident rather than immediately start long-term treatment. The insufficient follow-up time post-antifibrotic availability may also explain the lack of observed improvements in survival and mortality. Since antifibrotic drugs were approved in the U.K. in **2013** (pirfenidone) and **2016** (nintedanib), widespread adoption and integration into clinical practice have taken time. The data in this study may not yet fully capture the long-term effects of these

treatments on survival, as many patients likely only began receiving antifibrotics after substantial disease progression, limiting the potential for improved outcomes. While antifibrotic therapies have been shown to slow the decline in lung function, their impact on **long-term survival** remains unclear. Clinical trials have primarily been powered to demonstrate improvements in lung function rather than overall survival, and the real-world benefit of these therapies on mortality may take years to become apparent. Moreover, the **British Thoracic Society registry** highlighted that many patients were not receiving antifibrotics earlier in their disease course¹⁸⁵. By the time patients begin treatment, they may already have advanced disease, limiting the potential for survival improvements. The lack of improvement in survival and mortality among IPF patients in this study can be attributed to the delayed uptake of antifibrotic therapies, restrictive treatment eligibility criteria, and the "wait and watch" approach often employed by physicians for patients with milder disease. While antifibrotic drugs hold promise in slowing disease progression, several barriers—such as side effects, diagnostic uncertainty, and limited access—have hindered their potential impact on survival. As a result, increasing physician awareness of the benefits of antifibrotic treatment across all stages of IPF, along with addressing diagnostic and treatment access issues, will be crucial in improving outcomes for IPF patients in the future.

The primary limitation of this study is the inability to capture antifibrotic treatment administered in hospital settings. Due to this lack of information, I am unable to accurately assess whether there was an improvement in those patients who received treatment. Additionally, as noted in the above paragraph, there is insufficient follow-up time to fully evaluate long-term outcomes, particularly considering the updated NICE guidelines that came into effect in 2023, whose impact I could not examine. Furthermore, changes in IPF diagnostic criteria over time may have led to some incident cases being misclassified as prevalent cases. The limitations of the data used are covered in Chapter 3.

In conclusion, this analysis should be repeated in the future when we have a longer follow-up time, which allows the changes in the NICE guidelines to be reflected in practice and expansion of the work to capture patients on antifibrotic therapy as that will allow a more nuanced assessment of the survival benefit.

Chapter 7

General Discussion: Summary of Main Findings

This final discussion chapter will synthesise the key insights from each chapter into a cohesive discussion that addresses the thesis's overarching aims. While detailed discussions of specific results can be found at the end of each relevant chapter, where findings are contextualised within the existing literature, and the strengths and limitations are explored, this chapter will provide a broader perspective. Additionally, it will highlight areas for future research that could enhance our understanding of the burden of this complex disease, which was beyond the scope of my thesis.

7.1. Global Landscape of Incidence and Prevalence of ILDs

Re: Chapter 2 (Systematic review) and Aim 1

When I started my research, I aimed to understand the disease landscape of ILDs. Given how complex and heterogeneous this disease pool is, as the first step, I assessed how well the various conditions within ILD were recorded and whether the disease burden for various ILDs was established.

As the initial phase of my thesis, I conducted a systematic literature review to explore the epidemiology of ILDs and their subtypes. This review served as a crucial foundation for evaluating the existing research on ILDs, providing insights into the scope and depth of the literature on ILDs. The aim was to assess the relevance and extent of published research, which would help identify gaps in the current knowledge base of this disease burden.

The literature on the epidemiology of ILDs is quite fragmented. Much of the research focuses on individual conditions like IPF or RA-ILD or broader categories based on aetiology, such as drug-induced ILDs. However, there is relatively little on ILDs as a whole, which is understandable given that the term encompasses over 150 distinct conditions—and added complexity of evolving classifications for ILDs as we learned more about the underlying mechanisms. This review uncovered the significant diversity and heterogeneity in how various ILDs, especially non-IPF ILDs, were studied and reported. Notably, data on the incidence of non-IPF conditions were particularly sparse, with most studies conducted in small, highly selected populations. While the research on non-IPF ILDs is expanding, most studies so far have focused on prevalence—how many patients currently live with the condition—rather than the incidence or the number of new diagnoses. This underscores the challenges in accurately estimating the incidence of these conditions. This review highlighted the considerable variation in study design and methods and an inconsistent approach to diagnosing various ILDs. Despite IPF being one of the most published IPFs, there was a lack of standardisation in how IPF diagnosis was confirmed.

In certain countries, variations were observed in how IPF was ascertained, whereas in others, in how IPF data was collected and thus reported. In a resource-limited country, the clinical process of arriving at a diagnosis varies depending on the availability of technology for diagnosing rare conditions (e.g. HRCT scan). Also, data collection and reporting accuracy are affected in areas with limited infrastructure or resources. In a country like the UK, health systems are generally better at keeping accounts of more common diseases that GPs are incentivised to treat or manage (for example, diabetes and high blood pressure). For rare conditions such as IPF, there is a heavy reliance on specialists, to confirm the diagnosis. Another weighty factor for discrepancies is the lack of standardised diagnostic guidelines. While there have been longstanding efforts to standardise the diagnosis of IPF, the broader field of non-IPF ILDs is still evolving. As medical evidence and understanding develop, definitions of these diseases can change, complicating the estimation of their actual burden and making it difficult to compare rates over time. For example, certain ILDs may go unrecognised initially, leading to underestimating early cases. However, once a new diagnostic criteria or nomenclature is adopted, a sudden increase in the recorded cases may occur—raising the question of whether this reflects an actual increase or simply better recognition and reporting.

A combination of various factors discussed so far - lack of infrastructure, incentives, disease awareness, particularly for non-IPF ILDs influences inconsistencies in data collection at both global and national levels. This fragmented data, combined with the lack of centralised databases and incomplete medical records, impacts the quality and availability of information, making it challenging to accurately estimate ILD prevalence and incidence.

I have seen in the literature that IPF is the most extensively researched ILD, which is expected given its progressive and severe nature. IPF is generally easier to diagnose compared to many other ILDs. Several factors contribute to this: i) the introduction of diagnostic guidelines based on a combination of clinical, radiological, and histopathological features, which makes the diagnostic pathway more evident for physicians; ii) more research into IPF biomarkers and genetic markers compared to other ILDs. Although there is no single biomarker for IPF, the advances in this area provide more tools for diagnosis, iii) greater awareness of IPF, and more educational resources are available for diagnosing it. As it is better recorded than other ILDs,

studies on IPF can be drawn on routinely collected data from medical records, such as claims and electronic health records, to estimate incidence and prevalence in the general population. This option is preferred over single-centre or registry-based fragmented data, which is the option for most non-IPF ILDs.

Nevertheless, despite the use of large databases, inconsistencies in how IPF is 'accurately' defined still need to be addressed. A recurring theme in the literature is the classification of IPF into narrow and broad categories, reflecting discrepancies in diagnostic practices and recording of IPF. The broad classification includes a more comprehensive range of 'IPF-like' conditions, such as pulmonary and diffuse pulmonary fibrosis. At the same time, narrow categories are limited to codes specifically representing IPF. These differences in classification result in significant discrepancies in incidence estimates, with studies reporting a 4 to 21-fold variation. Prevalence estimates, while less variable than incidence, also show discrepancies of 2 to 4-fold across studies – prevalence is less sensitive to changes in diagnostic practices or short-term reporting than incidence. They are primarily driven by the specificity of diagnosis codes and the availability of diagnostic resources, such as HRCT scans, which are essential for confirming IPF. In regions with limited resources, such as rural areas in middle or low-income countries, the reliance on less definitive diagnostic tools further complicates the accurate identification of IPF.

Before the approval of antifibrotic therapies in 2015, treatment options for IPF were limited to symptom management. This could have impacted the motivation to record the condition accurately. However, with the introduction of antifibrotics and updated guidelines for IPF diagnosis, there is now greater awareness. To be eligible for antifibrotics, a patient requires a confirmed diagnosis of IPF, which translates to a better and cleaner capture of the condition. Therefore, this should lead to more consistency in how IPF is documented in routinely collected data.

In reviewing the literature, I encountered several studies that explored fibrosing ILDs beyond IPF,. A subset of patients with non-IPF ILDs can develop a progressive fibrosing phenotype or progressive pulmonary fibrosis, where the disease mirrors IPF, leading to deteriorating lung function, increased mortality, and reduced quality of life. Recognising this progression has shifted focus toward early detection and management of fibrosis across the ILD spectrum. The success of antifibrotic therapies such as pirfenidone and nintedanib in

IPF has prompted interest in their potential benefits for other types of fibrosing ILDs. Clinical trials have shown that these treatments also slow disease progression in progressive non-IPF ILDs³¹, sparking greater research into comprehensive treatment strategies¹⁸⁶. This expanded focus has the potential to lead to better diagnostic tools, more effective treatments, and improved outcomes for a larger group of patients affected by these progressive and debilitating diseases.

Given the timing of these advancements, it is crucial to revisit and update previous reviews to include more recent studies that reflect these changes and include data not limited to IPF but in other fibrosing ILDs.

7.2. Defining IPF In Electronic Health Records

Re: Chapter 4 (Validation) and Aim 2

The validation study was essential for several reasons. Routinely collected data, such as electronic health records (EHRs) from primary care and hospital administrative records, have become invaluable for studying chronic conditions like IPF. These data sources are increasingly relied upon for regulatory decision-making because traditional clinical trials are costly, often lack generalizability, and require long follow-up periods to capture significant outcomes, such as the impact of antifibrotics on patient survival.

For my thesis, using primary and secondary care data was appropriate for estimating the disease burden in England. However, my main limitation is my level of confidence in reliably defining a case.

IPF has garnered significant research interest, especially after the approval of antifibrotic treatments. With the incentive to treat these patients who have poor survival rates, there is a heightened need for accurate diagnosis to ensure that suitable patients receive these treatments at the right time. This heightened awareness, combined with increased interest in research, raises foundational questions: How many patients have the condition? Who will benefit most from the treatment? What are the survival rates, and what is the impact of delayed diagnosis? To address these critical questions, reliable data are necessary, and we must look beyond clinical trials or prospective data collection to promptly address some of these important questions. The first step of estimating disease burden at the population level is understanding how these patients are recorded in routine practice. In the field of ILD, particularly IPF, there have been several changes, with guidelines updated based on research and evidence. However, whether and how these updates have affected how these patients are diagnosed and eventually recorded is unclear.

Johnston and colleagues determined the accuracy of coding cryptogenic fibrosing alveolitis in hospital data; they concluded their findings were not adequate to determine how to find an IPF case in the data accurately, but for any planning purposes to estimate the patient population, this would be adequate ¹⁵³. Another study was conducted by Hubbard and colleagues, who validated the IPF diagnosis in primary care using hospital letters and discharge summaries and found the proportion of true positives to be as high as 95% - construing it is unlikely that patients will have a diagnosis of IPF in primary care without referral to a specialist ¹⁶⁴. As far

as I know, no recent work (post-2000) has been done on how IPF is recorded in the UK. Previously published estimates of IPF incidence and prevalence in the UK, primarily based on primary care data, are outdated (up to 2012) and do not reflect significant advancements described above in this field. Based on my research, one of the major data gaps is the lack of updated IPF epidemiological search. It is foundational work to determine where to find these patients 'accurately'— is it primary or secondary care? Hence, understanding how and where these patients are diagnosed and captured in the UK.

Before I started with this study, a few points to consider were: i) **Reliability of Recordings:** IPF is typically diagnosed outside of primary care, and so relying solely on primary care data might lead to either under- or overestimation of IPF cases; ii) **Granularity of Data:** It was unclear how accurately these patients would be captured in the database. Specifically, whether it would be straightforward to identify patients using IPF-specific clinical codes or if a broader term, like interstitial pneumonia, would be needed; iii) **Impact of Nomenclature Updates:** As discussed in Chapter 1, there have been several updates to ILD nomenclature over time. It is unknown how these changes have affected the patient recording. These factors are linked to physician behaviour, which impacts the case recording and, in turn, disease burden over time.

The definition of an IPF case and how to define an IPF cohort varied across studies; for example, in certain studies based on US claims data^{33,70,150}, a record of IPF along with A&E visits for IPF exacerbation served as confirmed cases. However, what differed across studies was the confirmatory evidence. IPF diagnosis involves a myriad of assessments, including a multidisciplinary clinical discussion, patient history, physical examination, and HRCT assessment. In the UK, a diagnosis of IPF is confirmed by specialists, but its day-to-day management often falls to GPs. It is the responsibility of the GP to accurately (as much as possible) translate information from specialist to routine data. How the GPs record this information speaks to the reliability of the coding as well as the granularity of it. Thus, understanding coding practices in primary care and physician behaviour is important to truly understand how a GP would code a patient diagnosed with IPF - whether GPs predominantly use generic pulmonary fibrosis or specific IPF codes. For example, a GP may record “*pulmonary fibrosis*” over specific “*idiopathic pulmonary fibrosis*”. Over time, there have been several updates to IPF terms; hence, in this analysis, it was interesting to see how, with time, we see a change in

physician coding, uptake of more IPF-specific codes, and decline in the use of obsolete terms (cryptogenic fibrosing alveolitis) as well as generic codes (diffuse pulmonary fibrosis). In the literature, IPF cohorts were commonly defined using a broad and narrow definition, with the idea of capturing different aspects of the disease and balancing sensitivity and specificity. I followed this theme in my research to capture IPF and IPF-like patients. Considering the changes in IPF terminology and increasing disease awareness, within this study period, I noticed an uptake of a narrow, IPF-specific list of codes.

This validation study was conducted to gain insights into using clinical diagnostic codes for identifying individuals with IPF from primary care records in the UK. I learned that selecting the "right" set of codes is a crucial strategy for most chronic conditions ^{151,152}. It is important to consider the choice of codes – that should be tailored to the specific research objectives: for studies requiring high validity, a narrower set of IPF-specific codes is appropriate (but comes at the expense of reduced sample size), while studies aiming for larger sample sizes such as incidence or prevalence estimation, may benefit from a broader set of codes indicative of IPF or IPF-like conditions (but comes at the expense of a lower PPV).

This study offers various strategies for defining an IPF cohort and provides the PPV associated with each approach. When designing a study, researchers should carefully consider the limitations of each definition and select the most suitable one based on their research goals. The aim was not to establish a single ideal definition without limitations but rather to understand how this chronic condition is recorded and identify the best way to capture IPF patients—depending on the specific objectives of the research.

This study is the first to validate the completeness and quality of IPF recording in routinely collected health data in the UK (in recent times). IPF may likely be underreported or misclassified due to factors such as a lack of clinical details, evolving diagnostic coding practices, or inconsistencies in how multiple causes of death are recorded. For example, death certificates may often capture complications like pneumonia or respiratory failure – as shown in previous studies using ONS as the gold standard ^{133,151}. Without access to physician's notes, I used ONS data as my gold standard – despite its limitations, it is a good dataset for assessing mortality trends. Conducting validation studies is crucial, as there is growing use and reliance on large databases to address important research questions. This ensures the reliability of recording and case finding in UK

electronic health records, as these data are used to estimate disease prevalence and inform health service planning and resource allocation. Specifically, this work demonstrates that using a simple set of clinical codes to define a cohort of IPF patients in primary care data is feasible for many observational research purposes. However, researchers should tailor their choice of codes to their research objectives and consider using linked hospital data to improve diagnostic accuracy.

7.3. Aim 3: To establish the incidence and prevalence of IPF in England from 2008 to 2018 and investigate IPF-related mortality.

Re: Chapter 5 and Aim 3

This study demonstrated that IPF incidence and prevalence estimates are highly sensitive to how the disease is defined. This finding is unsurprising given the relative rarity of IPF, especially when compared to more common and established conditions like lung cancer. The diagnostic pathway for IPF is further complicated by the nonspecific nature of symptoms, the complex and often lengthy diagnostic process, and evolving diagnostic criteria and nomenclature changes over time.

When estimating the burden of IPF using primary care records, incidence and prevalence estimates varied substantially depending on the set of clinical codes used to define the disease. For instance, our incidence estimates ranged from 4.5 per 100,000 person-years using specific codes to 27.2 per 100,000 person-years when a broader set of codes was applied. The more specific codes, while conservative, likely underestimated the disease burden, while the broader codes potentially overestimated it. Based on these findings, I hypothesise that the true incidence of IPF lies somewhere halfway between these two extremes, with the most accurate estimate being around 15-16 per 100,000 person-years. This would suggest approximately 8,000 to 9,000 new cases of IPF annually in the UK, updating the previously widely cited figure of 5,000 to 6,000 new cases per year.

Another key finding was that hospitalisation data is less suited for estimating IPF incidence and prevalence. Over time, there has been a decline in hospital visits for IPF, which I speculate may be due to several reasons, such as increased awareness, which may be leading to an earlier diagnosis; another reason can be the

availability of antifibrotics, which provides an incentive for diagnosis of the condition, and thus translate to improved disease management (such as palliative care and primary care management). These factors are likely to contribute towards better care. That is, if these patients are managed well within the realm of primary care, then they show up less in the hospital for acute management (for example, exacerbation). However, this is speculative and needs to be investigated further once longer follow-up data is available.

While this study indicated an increase in the number of people diagnosed with IPF over the study period, it remains unclear whether this reflects a true rise in disease incidence or simply better detection and reporting due to advances in diagnostic guidelines and awareness. Additionally, the study found a 50% increase in IPF-specific mortality during the period analysed. Although IPF mortality has increased, the rate of increase has slowed in more recent years, suggesting possible early impacts of improved treatment and management, phasing out of the previous standard of care, which was associated with increased mortality – but at this point, this is purely speculative, and I was not able to confirm this as part of this study.

Importantly, as awareness and diagnosis of IPF continue to improve, patients will likely be captured more accurately within primary care data, reducing the need for reliance on hospital or secondary care data for certain research purposes. This study emphasises the importance of improving and standardising coding practices, which will play a critical role in the accuracy of future research on IPF (and eventually other fibrosing ILDs).

What is the implication if the burden of IPF is truly rising? From the context of health resource planning, it is important to know if incidence and prevalence are on the rise as it is crucial for allocating the necessary resources to manage the disease effectively. Studies have shown managing an IPF patient is costly, increase in cases would have significant economic implications, especially in a government-funded system such as in the UK^{68,187,188,189} access to antifibrotics has been a challenge (as discussed in the earlier chapters). If these treatments have improved quality of life (for example, management of symptoms) and survival, it is imperative to reconsider the eligibility criteria, potentially making antifibrotics available earlier in the disease course rather than limiting them to advanced-stage cases – this is reflected in the recent update of the NICE guidelines. This analysis should be repeated in the future to allow the changes in the NICE guidelines to be reflected in practice and expansion of the work to capture patients on antifibrotic therapy as that will allow a more nuanced assessment of the survival benefit¹⁴⁶.

In conclusion, this study provides a clearer understanding of the burden of IPF in England recently, suggesting that the true incidence may be higher than previously estimated. However, the findings also highlight the complexities of diagnosing and tracking this disease and the ongoing challenges in improving survival outcomes despite the availability of new treatments.

7.4. Aim 4: To investigate changes in survival and causes of death in people diagnosed with IPF over the last 10 years in England.

A crucial question is whether patients are being managed more effectively and if early diagnosis, along with the availability of treatments, leads to better survival outcomes. Unfortunately, this remains uncertain¹⁹⁰.

While randomised controlled trials (RCTs) have demonstrated some survival benefits for IPF patients and broader fibrosis cohorts, these studies are often limited in scope¹⁹¹. These are not necessarily generalisable to a real-world population and often lack the statistical power needed to assess survival benefits fully. Therefore, more real-world data are needed to understand the true impact of current treatments on survival.

In Chapter 5, I analysed death records where IPF was listed as the underlying cause. Interestingly, many of these patients did not have an IPF diagnosis in their medical history, raising questions about how well this disease is documented. What was missing from this analysis was the survival and mortality among patients who had been diagnosed with IPF. To truly understand the mortality associated with IPF – it is important to corroborate information on who died of IPF and what the survival and mortality in patients who were diagnosed with IPF. Building on that, my most recent research phase focused on evaluating survival for patients identified through primary care databases, hospital records, or both. Regardless of how the cohort was defined, I found no improvements in survival among IPF patients over time. The patients diagnosed through hospital records had a higher incidence of comorbidities and greater mortality rates, likely reflecting a sicker population or delayed diagnosis – aligned with my previous findings. Despite covering a study period that spanned the pre- and post-antifibrotic approval years, no significant improvements in survival were observed.

The stagnation in survival outcomes for IPF patients can be explained by several factors rooted in treatment access, disease complexity, and patient characteristics. These factors include:

- i. **Limited Uptake of Antifibrotic Treatments Due to Eligibility Criteria:** Despite the approval of antifibrotic drugs like pirfenidone and nintedanib, many IPF patients do not receive these treatments due

to restrictive eligibility criteria. According to the British Thoracic Society registry¹⁸⁵, only 21-22% of IPF patients in the UK were prescribed antifibrotics at diagnosis. Eligibility is primarily based on FVC thresholds set by the NICE. A 2019 survey indicated that more than half of patients were ineligible for antifibrotics because their FVC exceeded the NICE-approved limit of 80% (ref). This lung function restriction for access to nintedanib was subsequently removed in 2023. Additionally, some clinical centres lack the resources to prescribe these medications in a timely fashion, limiting access.

- ii. **Conservative 'Watch and Wait' Approach:** A review across five European countries revealed that 54% of IPF patients did not receive antifibrotic treatment despite drug availability¹⁸⁴. This approach can delay intervention until the disease has progressed, diminishing the effectiveness of antifibrotics. It looks like a "watch and wait" approach is adopted rather than immediate treatment for patients with mild or stable disease.
- iii. **Severe Side Effects of Antifibrotic Drugs:** The side effects of antifibrotic medications, pirfenidone and nintedanib, are wide-ranging from mild to severe. Effects include gastrointestinal issues, liver function, skin reactions (rash or photosensitivity) and loss of appetite; some of the effects are more cumbersome to manage but overall, these can significantly impact quality of life (refs). As a result, even those eligible for treatment may choose not to continue it, thus limiting the potential benefits of these therapies.
- iv. **Delayed Diagnosis:** The nonspecific early symptoms of IPF, such as a persistent dry cough or shortness of breath, are often misdiagnosed as asthma or chronic obstructive pulmonary disease (COPD). As a result, many patients are diagnosed at a later stage, when lung damage is already extensive. Despite improvements in diagnostic tools like high-resolution CT scans, early detection remains a significant challenge. Furthermore, early-stage biomarkers are still under development, leaving clinicians to rely on clinical presentation and imaging, which often occurs after considerable lung function has been lost.
- v. **Comorbidities and Age:** In this research, we have seen median age at IPF diagnosis higher than 70 years and these patients often have other health conditions such as heart disease, diabetes, or pulmonary hypertension, which complicate treatment and worsen survival outcomes. Many of these comorbid conditions, like heart failure or pneumonia, can be more fatal than the fibrosis itself. Managing these comorbidities alongside IPF is difficult, and the ageing population tends to be less resilient to aggressive treatments, leading to higher mortality rates.

- vi. **High Rates of Treatment Discontinuation:** For some of the above-mentioned reasons, such as side effects or a perceived lack of improvement, there may be a higher discontinuation rate, which undermines their impact on disease progression and survival. Furthermore, patients experiencing medication side effects frequently undergo treatment interruptions once again limiting cumulative exposure to treatment.
- vii. **Insufficient Long-Term Data on Antifibrotic Efficacy:** Antifibrotic drugs have been available since 2013 and 2015. However, owing to slow uptake, there is insufficient long-term data to evaluate their impact on survival fully. Longer-term data collection and analysis are necessary to assess the true benefit of these treatments. The study period, from 2010 to 2023, included years before and shortly after introducing these drugs, meaning there has not been enough follow-up time to evaluate their full potential in improving long-term survival rates.

While there is growing awareness of IPF, treatment access issues and diagnostic uncertainty need to be addressed. Until treatment eligibility and early diagnosis improve, significant advancements in survival outcomes are unlikely.

When we think about the complex nature of the disease and its debilitating effect on a person's quality of life, the goal of IPF treatment should not be solely linked to survival benefits or focusing on extending survival by a few months. Patient preferences for quality-of-life improvements are essential in IPF treatment strategies. Therefore, extending the focus beyond mortality and survival benefits to include quality of life is crucial in managing IPF. Future advancements in genetic research, biomarker development, and lung tissue repair may hold promise.

7.5 Clinical Implications: Patient Care

In the context of patient care for these comorbid, chronic patients it is crucial to consider outcomes beyond survival and mortality rates. From the patient's perspective, factors such as quality of life and effective symptom management are equally important indicators of care quality.

While research has significantly advanced in the IPF field over the past decade, the impact of these developments on the overall quality and type of care received by IPF patients remains an area for reflection. Notable improvements include the expansion of palliative care and pulmonary rehabilitation services for IPF patients. These services are now widely acknowledged as key components of an integrated care package that addresses both patient and caregiver needs.

Originally designed for individuals with COPD, pulmonary rehabilitation has increasingly demonstrated benefits for IPF patients. Likewise, integrated palliative care services provide comprehensive support that extends beyond symptom management to encompass the emotional, social, and spiritual needs of individuals with terminal illnesses. Studies have shown that such holistic approaches contribute to improving the quality of life for people with IPF. Despite these positive strides, gaps remain in ensuring equitable access to these beneficial services. A recent study Morgan and colleagues¹⁹² showed the referral rates to these non-pharmacological therapies is low (12%) and even a smaller proportion complete the programme (< 2%). The referral and the completion rates for pulmonary rehabilitation and palliative care for IPF patients are often lower than those for other respiratory conditions such as lung cancer and COPD, limiting their impact on improving patient-centred outcomes. This highlights gaps in the patient care.

Another challenge is the uptake of antifibrotic treatments. Although these medications are available, their use is limited due to concerns about side effects, which can deter patients from continuing treatment. This hesitation is especially significant given that IPF patients often manage multiple comorbidities. Adding

additional medication side effects further burdens patients, particularly when the treatments have not yet demonstrated substantial survival benefits.

In conclusion, while advancements in IPF research have enhanced knowledge and improved care strategies, there remains a need to address gaps in service provision and treatment uptake. By focusing on patient centric approach such as quality-of-life and ensuring better access to supportive care services, healthcare providers can offer a more comprehensive and patient-centred approach to IPF care.

REFERENCES

- 1 Medicine, S. S. o. *Understanding ILD*, <<https://med.stanford.edu/ild/patient-resources/understanding-ild>> (2024).
- 2 Guler, S. A. & Corte, T. J. Interstitial Lung Disease in 2020: A History of Progress. *Clin Chest Med* **42**, 229-239 (2021). <https://doi.org/10.1016/j.ccm.2021.03.001>
- 3 Wells, A. U. *et al.* What's in a name? That which we call IPF, by any other name would act the same. *Eur Respir J* **51** (2018). <https://doi.org/10.1183/13993003.00692-2018>
- 4 Brownell, R. *et al.* The use of pretest probability increases the value of high-resolution CT in diagnosing usual interstitial pneumonia. *Thorax* **72**, 424-429 (2017). <https://doi.org/10.1136/thoraxjnl-2016-209671>
- 5 Caban, J. J., Yao, J., Bagci, U. & Mollura, D. J. Monitoring pulmonary fibrosis by fusing clinical, physiological, and computed tomography features. *Annu Int Conf IEEE Eng Med Biol Soc* **2011**, 6216-6219 (2011). <https://doi.org/10.1109/IEMBS.2011.6091535>
- 6 Raghu, G. *et al.* The accuracy of the clinical diagnosis of new-onset idiopathic pulmonary fibrosis and other interstitial lung disease: A prospective study. *Chest* **116**, 1168-1174 (1999). <https://doi.org/10.1378/chest.116.5.1168>
- 7 American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS). *Am J Respir Crit Care Med* **161**, 646-664 (2000). <https://doi.org/10.1164/ajrccm.161.2.ats3-00>
- 8 Raghu, G. Idiopathic pulmonary fibrosis: guidelines for diagnosis and clinical management have advanced from consensus-based in 2000 to evidence-based in 2011. *Eur Respir J* **37**, 743-746 (2011). <https://doi.org/10.1183/09031936.00017711>
- 9 Raghu, G. *et al.* An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* **183**, 788-824 (2011). <https://doi.org/10.1164/rccm.2009-040GL>
- 10 Raghu, G. *et al.* Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med* **198**, e44-e68 (2018). <https://doi.org/10.1164/rccm.201807-1255ST>
- 11 Raghu, G. *et al.* Idiopathic Pulmonary Fibrosis (an Update) and Progressive Pulmonary Fibrosis in Adults: An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med* **205**, e18-e47 (2022). <https://doi.org/10.1164/rccm.202202-0399ST>
- 12 Shih, A. R. *et al.* Practical application and validation of the 2018 ATS/ERS/JRS/ALAT and Fleischner Society guidelines for the diagnosis of idiopathic pulmonary fibrosis. *Respir Res* **22**, 124 (2021). <https://doi.org/10.1186/s12931-021-01670-7>
- 13 Strongman, H., Kausar, I. & Maher, T. M. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. *Adv Ther* **35**, 724-736 (2018). <https://doi.org/10.1007/s12325-018-0693-1>
- 14 Vancheri, C., Failla, M., Crimi, N. & Raghu, G. Idiopathic pulmonary fibrosis: a disease with similarities and links to cancer biology. *Eur Respir J* **35**, 496-504 (2010). <https://doi.org/10.1183/09031936.00077309>
- 15 Ye, Y. *et al.* Prevalence, incidence, and survival analysis of interstitial lung diseases in Hong Kong: a 16-year population-based cohort study. *Lancet Reg Health West Pac* **42**, 100871 (2024). <https://doi.org/10.1016/j.lanwpc.2023.100871>
- 16 Kaul, B., Cottin, V., Collard, H. R. & Valenzuela, C. Variability in Global Prevalence of Interstitial Lung Disease. *Front Med (Lausanne)* **8**, 751181 (2021). <https://doi.org/10.3389/fmed.2021.751181>
- 17 Molina-Molina, M. *et al.* Importance of early diagnosis and treatment in idiopathic pulmonary fibrosis. *Expert Rev Respir Med* **12**, 537-539 (2018). <https://doi.org/10.1080/17476348.2018.1472580>
- 18 Salciccioli, J. D. *et al.* Interstitial lung disease incidence and mortality in the UK and the European Union: an observational study, 2001-2017. *ERJ Open Res* **8** (2022). <https://doi.org/10.1183/23120541.00058-2022>
- 19 Patterson, K. C. & Streck, M. E. Pulmonary fibrosis in sarcoidosis. Clinical features and outcomes. *Ann Am Thorac Soc* **10**, 362-370 (2013). <https://doi.org/10.1513/AnnalsATS.201303-069FR>
- 20 Wells, A. U. Managing diagnostic procedures in idiopathic pulmonary fibrosis. *Eur Respir Rev* **22**, 158-162 (2013). <https://doi.org/10.1183/09059180.00001213>
- 21 Nathan, S. D. & du Bois, R. M. Idiopathic pulmonary fibrosis trials: recommendations for the jury. *Eur Respir J* **38**, 1002-1004 (2011). <https://doi.org/10.1183/09031936.00068611>
- 22 Kreuter, M. *et al.* Health related quality of life in patients with idiopathic pulmonary fibrosis in clinical practice: insights-IPF registry. *Respir Res* **18**, 139 (2017). <https://doi.org/10.1186/s12931-017-0621-y>
- 23 Adegunsoye, A. *et al.* Evaluation of Pulmonary Fibrosis Outcomes by Race and Ethnicity in US Adults. *JAMA Netw Open* **6**, e232427 (2023). <https://doi.org/10.1001/jamanetworkopen.2023.2427>
- 24 Navaratnam, V. & Hubbard, R. B. The Mortality Burden of Idiopathic Pulmonary Fibrosis in the United Kingdom. *Am J Respir Crit Care Med* **200**, 256-258 (2019). <https://doi.org/10.1164/rccm.201902-0467LE>

- Excellence, N. I. f. H. a. C. *Nintedanib for treating idiopathic pulmonary fibrosis* <<https://www.nice.org.uk/guidance/ta379>> (2016).
- Demedts, M. & Costabel, U. ATS/ERS international multidisciplinary consensus classification of the idiopathic interstitial pneumonias. *Eur Respir J* **19**, 794-796 (2002). <https://doi.org/10.1183/09031936.02.00492002>
- Jacob, J. *et al.* Predicting outcomes in rheumatoid arthritis related interstitial lung disease. *Eur Respir J* **53** (2019). <https://doi.org/10.1183/13993003.00869-2018>
- King, T. E., Jr. *et al.* A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* **370**, 2083-2092 (2014). <https://doi.org/10.1056/NEJMoa1402582>
- Richeldi, L. *et al.* Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* **370**, 2071-2082 (2014). <https://doi.org/10.1056/NEJMoa1402584>
- Flaherty, K. R. *et al.* Design of the PF-ILD trial: a double-blind, randomised, placebo-controlled phase III trial of nintedanib in patients with progressive fibrosing interstitial lung disease. *BMJ Open Respir Res* **4**, e000212 (2017). <https://doi.org/10.1136/bmjresp-2017-000212>
- Flaherty, K. R. *et al.* Nintedanib in Progressive Fibrosing Interstitial Lung Diseases. *N Engl J Med* **381**, 1718-1727 (2019). <https://doi.org/10.1056/NEJMoa1908681>
- Ma, X. *et al.* Global and regional burden of interstitial lung disease and pulmonary sarcoidosis from 1990 to 2019: results from the Global Burden of Disease study 2019. *Thorax* **77**, 596-605 (2022). <https://doi.org/10.1136/thoraxjnl-2020-216732>
- Raghu, G., Chen, S. Y., Hou, Q., Yeh, W. S. & Collard, H. R. Incidence and prevalence of idiopathic pulmonary fibrosis in US adults 18-64 years old. *Eur Respir J* **48**, 179-186 (2016). <https://doi.org/10.1183/13993003.01653-2015>
- Olson, A. L., Brown, K. K. & Swigris, J. J. Understanding and optimizing health-related quality of life and physical functional capacity in idiopathic pulmonary fibrosis. *Patient Relat Outcome Meas* **7**, 29-35 (2016). <https://doi.org/10.2147/PROM.S74857>
- Roubille, C. & Haraoui, B. Interstitial lung diseases induced or exacerbated by DMARDS and biologic agents in rheumatoid arthritis: a systematic literature review. *Semin Arthritis Rheum* **43**, 613-626 (2014). <https://doi.org/10.1016/j.semarthrit.2013.09.005>
- Deidda, A. *et al.* Interstitial lung disease induced by fluoxetine: Systematic review of literature and analysis of Vigibase, Eudravigilance and a national pharmacovigilance database. *Pharmacol Res* **120**, 294-301 (2017). <https://doi.org/10.1016/j.phrs.2017.04.010>
- Salliot, C. & van der Heijde, D. Long-term safety of methotrexate monotherapy in patients with rheumatoid arthritis: a systematic literature research. *Ann Rheum Dis* **68**, 1100-1104 (2009). <https://doi.org/10.1136/ard.2008.093690>
- Skeoch, S. *et al.* Drug-Induced Interstitial Lung Disease: A Systematic Review. *J Clin Med* **7** (2018). <https://doi.org/10.3390/jcm7100356>
- Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021). <https://doi.org/10.1136/bmj.n71>
- Nussbaumer-Streit, B. *et al.* Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. *J Clin Epidemiol* **118**, 42-54 (2020). <https://doi.org/10.1016/j.jclinepi.2019.10.011>
- Lo, C. K., Mertz, D. & Loeb, M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. *BMC Med Res Methodol* **14**, 45 (2014). <https://doi.org/10.1186/1471-2288-14-45>
- Moskalewicz, A. & Oremus, M. No clear choice between Newcastle-Ottawa Scale and Appraisal Tool for Cross-Sectional Studies to assess methodological quality in cross-sectional studies of health-related quality of life and breast cancer. *J Clin Epidemiol* **120**, 94-103 (2020). <https://doi.org/10.1016/j.jclinepi.2019.12.013>
- Sterne, J. A. *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4919 (2016). <https://doi.org/10.1136/bmj.i4919>
- Hopkins, R. B., Burke, N., Fell, C., Dion, G. & Kolb, M. Epidemiology and survival of idiopathic pulmonary fibrosis from national data in Canada. *Eur Respir J* **48**, 187-195 (2016). <https://doi.org/10.1183/13993003.01504-2015>
- Harari, S., Madotto, F., Caminati, A., Conti, S. & Cesana, G. Epidemiology of Idiopathic Pulmonary Fibrosis in Northern Italy. *PLoS One* **11**, e0147072 (2016). <https://doi.org/10.1371/journal.pone.0147072>
- Gjonbrataj, J. *et al.* Incidence of idiopathic pulmonary fibrosis in Korea based on the 2011 ATS/ERS/JRS/ALAT statement. *Int J Tuberc Lung Dis* **19**, 742-746 (2015). <https://doi.org/10.5588/ijtld.14.0650>
- Lee, H. E. *et al.* Incidence and prevalence of idiopathic interstitial pneumonia and idiopathic pulmonary fibrosis in Korea. *Int J Tuberc Lung Dis* **20**, 978-984 (2016). <https://doi.org/10.5588/ijtld.16.0003>
- Lim, S. S. *et al.* The cumulative incidence and trends of rare diseases in South Korea: a nationwide study of the administrative data from the National Health Insurance Service database from 2011-2015. *Orphanet J Rare Dis* **14**, 49 (2019). <https://doi.org/10.1186/s13023-019-1032-6>

49 Kaul, B. *et al.* Epidemiology of Idiopathic Pulmonary Fibrosis among U.S. Veterans, 2010-2019. *Ann Am Thorac Soc* **19**, 196-203 (2022). <https://doi.org:10.1513/AnnalsATS.202103-295OC>

50 Bellou, V., Belbasis, L. & Evangelou, E. Tobacco Smoking and Risk for Pulmonary Fibrosis: A Prospective Cohort Study From the UK Biobank. *Chest* **160**, 983-993 (2021). <https://doi.org:10.1016/j.chest.2021.04.035>

51 Zhang, Y., Li, H., Wu, N., Dong, X. & Zheng, Y. Retrospective study of the clinical characteristics and risk factors of rheumatoid arthritis-associated interstitial lung disease. *Clin Rheumatol* **36**, 817-823 (2017). <https://doi.org:10.1007/s10067-017-3561-5>

52 Raimundo, K. *et al.* Rheumatoid Arthritis-Interstitial Lung Disease in the United States: Prevalence, Incidence, and Healthcare Costs and Mortality. *J Rheumatol* **46**, 360-369 (2019). <https://doi.org:10.3899/jrheum.171315>

53 Sparks, J. A. *et al.* Prevalence, incidence and cause-specific mortality of rheumatoid arthritis-associated interstitial lung disease among older rheumatoid arthritis patients. *Rheumatology (Oxford)* **60**, 3689-3698 (2021). <https://doi.org:10.1093/rheumatology/keaa836>

54 Wangkaew, S., Euathrongchit, J., Wattanawittawas, P., Kasitanon, N. & Louthrenoo, W. Incidence and predictors of interstitial lung disease (ILD) in Thai patients with early systemic sclerosis: Inception cohort study. *Mod Rheumatol* **26**, 588-593 (2016). <https://doi.org:10.3109/14397595.2015.1115455>

55 Vandecasteele, E. *et al.* Incidence, prevalence and long-term progression of Goh algorithm rated interstitial lung disease in systemic sclerosis in two independent cohorts in flanders: A retrospective cohort study. *Semin Arthritis Rheum* **51**, 969-976 (2021). <https://doi.org:10.1016/j.semarthrit.2021.07.018>

56 Roca, F. *et al.* Interstitial lung disease in primary Sjogren's syndrome. *Autoimmun Rev* **16**, 48-54 (2017). <https://doi.org:10.1016/j.autrev.2016.09.017>

57 Olaosebikan, H. *et al.* Connective tissue disease--associated interstitial lung disease: an underreported cause of interstitial lung disease in Sub-Saharan Africa. *Clin Rheumatol* **40**, 3455-3460 (2021). <https://doi.org:10.1007/s10067-020-05336-5>

58 Reiser, S. *et al.* Progression and mortality of interstitial lung disease in mixed connective tissue disease: a long-term observational nationwide cohort study. *Rheumatology (Oxford)* **57**, 255-262 (2018). <https://doi.org:10.1093/rheumatology/kex077>

59 Duchemann, B. *et al.* Prevalence and incidence of interstitial lung diseases in a multi-ethnic county of Greater Paris. *Eur Respir J* **50** (2017). <https://doi.org:10.1183/13993003.02419-2016>

60 Fernandez Perez, E. R. *et al.* Epidemiology of Hypersensitivity Pneumonitis among an Insured Population in the United States: A Claims-based Cohort Analysis. *Ann Am Thorac Soc* **15**, 460-469 (2018). <https://doi.org:10.1513/AnnalsATS.201704-288OC>

61 Cui, K. *et al.* Comparison of the Cumulative Incidence Rates of Coal Workers' Pneumoconiosis between 1970 and 2013 among Four State-Owned Colliery Groups in China. *Int J Environ Res Public Health* **12**, 7444-7456 (2015). <https://doi.org:10.3390/ijerph120707444>

62 Casey, M. L. & Mazurek, J. M. Silicosis prevalence and incidence among Medicare beneficiaries. *Am J Ind Med* **62**, 183-191 (2019). <https://doi.org:10.1002/ajim.22944>

63 Szeszenia-Dabrowska, N., Swiatkowska, B., Sobala, W., Szubert, Z. & Wilczynska, U. Asbestos related diseases among workers of asbestos processing plants in relation to type of production and asbestos use. *Med Pr* **66**, 1-9 (2015).

64 Thomsen, R. W. *et al.* Risk of asbestosis, mesothelioma, other lung disease or death among motor vehicle mechanics: a 45-year Danish cohort study. *Thorax* **77**, 477-485 (2022). <https://doi.org:10.1136/thoraxjnl-2020-215041>

65 DeBono, N. L. *et al.* Incidence of mesothelioma and asbestosis by occupation in a diverse workforce. *Am J Ind Med* **64**, 476-487 (2021). <https://doi.org:10.1002/ajim.23245>

66 Jeon, M. H. *et al.* The incidence, comorbidity and mortality of sarcoidosis in Korea, 2008-2015: a nationwide population-based study. *Sarcoidosis Vasc Diffuse Lung Dis* **37**, 24-26 (2020). <https://doi.org:10.36141/svdld.v37i1.7660>

67 Nasser, M. *et al.* Estimates of epidemiology, mortality and disease burden associated with progressive fibrosing interstitial lung disease in France (the PROGRESS study). *Respir Res* **22**, 162 (2021). <https://doi.org:10.1186/s12931-021-01749-1>

68 Olson, A. L. *et al.* Prevalence and Incidence of Chronic Fibrosing Interstitial Lung Diseases with a Progressive Phenotype in the United States Estimated in a Large Claims Database Analysis. *Adv Ther* **38**, 4100-4114 (2021). <https://doi.org:10.1007/s12325-021-01786-8>

69 Kaunisto, J. *et al.* Re-evaluation of diagnostic parameters is crucial for obtaining accurate data on idiopathic pulmonary fibrosis. *BMC Pulm Med* **15**, 92 (2015). <https://doi.org:10.1186/s12890-015-0074-3>

70 Raimundo, K. *et al.* Clinical and economic burden of idiopathic pulmonary fibrosis: a retrospective cohort study. *BMC Pulm Med* **16**, 2 (2016). <https://doi.org:10.1186/s12890-015-0165-1>

71 Li, Q. *et al.* Disease frequency, patient characteristics, comorbidity outcomes and immunosuppressive therapy in systemic sclerosis and systemic sclerosis-associated interstitial lung disease: a US cohort study. *Rheumatology (Oxford)* **60**, 1915-1925 (2021). <https://doi.org:10.1093/rheumatology/keaa547>

- Pope, J. E. *et al.* Systemic Sclerosis and Associated Interstitial Lung Disease in Ontario, Canada: An Examination of Prevalence and Survival Over 10 Years. *J Rheumatol* **48**, 1427-1434 (2021). <https://doi.org/10.3899/jrheum.201049>
- Zhao, Y. *et al.* Primary Sjogren syndrome in Han Chinese: clinical and immunological characteristics of 483 patients. *Medicine (Baltimore)* **94**, e667 (2015). <https://doi.org/10.1097/MD.0000000000000667>
- Wang, Y., Hou, Z., Qiu, M. & Ye, Q. Risk factors for primary Sjogren syndrome-associated interstitial lung disease. *J Thorac Dis* **10**, 2108-2117 (2018). <https://doi.org/10.21037/jtd.2018.03.120>
- Gomez, G. N. *et al.* Myositis-specific antibodies and clinical characteristics in patients with autoimmune inflammatory myopathies: reported by the Argentine Registry of Inflammatory Myopathies of the Argentine Society of Rheumatology. *Clin Rheumatol* **40**, 4473-4483 (2021). <https://doi.org/10.1007/s10067-021-05797-2>
- Huang, H. L., Lin, W. C., Lin, P. Y., Weng, M. Y. & Sun, Y. T. The significance of myositis autoantibodies in idiopathic inflammatory myopathy concomitant with interstitial lung disease. *Neurol Sci* **42**, 2855-2864 (2021). <https://doi.org/10.1007/s10072-020-04911-7>
- Ishizuka, M. *et al.* Long-term follow-up of 124 patients with polymyositis and dermatomyositis: Statistical analysis of prognostic factors. *Mod Rheumatol* **26**, 115-120 (2016). <https://doi.org/10.3109/14397595.2015.1054081>
- Duarte, A. C., Porter, J. C. & Leandro, M. J. The lung in a cohort of rheumatoid arthritis patients-an overview of different types of involvement and treatment. *Rheumatology (Oxford)* **58**, 2031-2038 (2019). <https://doi.org/10.1093/rheumatology/kez177>
- Ke, Y. *et al.* Features and Outcomes of Elderly Rheumatoid Arthritis: Does the Age of Onset Matter? A Comparative Study From a Single Center in China. *Rheumatol Ther* **8**, 243-254 (2021). <https://doi.org/10.1007/s40744-020-00267-8>
- Kiely, P. *et al.* Is incident rheumatoid arthritis interstitial lung disease associated with methotrexate treatment? Results from a multivariate analysis in the ERAS and ERAN inception cohorts. *BMJ Open* **9**, e028466 (2019). <https://doi.org/10.1136/bmjopen-2018-028466>
- Kim, D. *et al.* Impact of interstitial lung disease on mortality of patients with rheumatoid arthritis. *Rheumatol Int* **37**, 1735-1745 (2017). <https://doi.org/10.1007/s00296-017-3781-7>
- Kronzer, V. L. *et al.* Respiratory Diseases as Risk Factors for Seropositive and Seronegative Rheumatoid Arthritis and in Relation to Smoking. *Arthritis Rheumatol* **73**, 61-68 (2021). <https://doi.org/10.1002/art.41491>
- McFarlane, I. M. *et al.* Assessment of interstitial lung disease among black rheumatoid arthritis patients. *Clin Rheumatol* **38**, 3413-3424 (2019). <https://doi.org/10.1007/s10067-019-04760-6>
- Md Yusof, M. Y. *et al.* Effect of rituximab on the progression of rheumatoid arthritis-related interstitial lung disease: 10 years' experience at a single centre. *Rheumatology (Oxford)* **56**, 1348-1357 (2017). <https://doi.org/10.1093/rheumatology/kex072>
- Paulin, F. *et al.* Lung involvement prevalence in patients with early rheumatoid arthritis without known pulmonary disease: a multicentric cross sectional study. *Adv Rheumatol* **61**, 52 (2021). <https://doi.org/10.1186/s42358-021-00209-0>
- Sakr, B. R. *et al.* Rheumatoid arthritis: A single-center Egyptian experience. *Immunol Invest* **47**, 293-302 (2018). <https://doi.org/10.1080/08820139.2018.1425700>
- Hammada, R. M., Salah-ElDin, H., El-Gharbawy, N. H. & Ragab, D. Krebs von den Lungen-6, a promising marker in evaluating the severity of interstitial lung disease in Egyptian rheumatoid arthritis patients. *Egypt J Immunol* **28**, 241-249 (2021).
- Shin, S. *et al.* Sex differences in clinical characteristics and their influence on clinical outcomes in an observational cohort of patients with rheumatoid arthritis. *Joint Bone Spine* **88**, 105124 (2021). <https://doi.org/10.1016/j.jbspin.2020.105124>
- Gao, H. *et al.* Prevalence, risk factors, and prognosis of interstitial lung disease in a large cohort of Chinese primary Sjogren syndrome patients: A case-control study. *Medicine (Baltimore)* **97**, e11003 (2018). <https://doi.org/10.1097/MD.00000000000011003>
- Gao, H. *et al.* Characteristics and mortality in primary Sjogren syndrome-related interstitial lung disease. *Medicine (Baltimore)* **100**, e26777 (2021). <https://doi.org/10.1097/MD.00000000000026777>
- Kam, J. K., Charan, N., Leong, R. W., Loh, Z. W. & Thong, B. Y. Clinical features and outcomes from the Singapore Sjogren's syndrome study. *Lupus* **30**, 248-255 (2021). <https://doi.org/10.1177/0961203320976932>
- Kampolis, C. F. *et al.* Prevalence and spectrum of symptomatic pulmonary involvement in primary Sjogren's syndrome. *Clin Exp Rheumatol* **36 Suppl 112**, 94-101 (2018).
- Kvarnstrom, M., Ottosson, V., Nordmark, B. & Wahren-Herlenius, M. Incident cases of primary Sjogren's syndrome during a 5-year period in Stockholm County: a descriptive study of the patients and their characteristics. *Scand J Rheumatol* **44**, 135-142 (2015). <https://doi.org/10.3109/03009742.2014.931457>
- Manfredi, A. *et al.* Erratum to: Prevalence and characterization of non-sicca onset primary Sjogren syndrome with interstitial lung involvement. *Clin Rheumatol* **36**, 1931 (2017). <https://doi.org/10.1007/s10067-017-3635-4>

- 95 Nilsson, A. M. *et al.* Mixed Airway and Pulmonary Parenchymal Disease in Patients With Primary Sjogren Syndrome: A 6-year Follow-up. *J Rheumatol* **48**, 232-240 (2021). <https://doi.org/10.3899/jrheum.200247>
- 96 Omair, M. A. *et al.* Disease phenotype and diagnostic delay in Saudi patients with primary Sjogren's syndrome: An exploratory cross-sectional study. *Saudi Med J* **42**, 405-410 (2021). <https://doi.org/10.15537/smj.2021.42.4.20200767>
- 97 Bauer, E. *et al.* Relationship between calcifications and structural lesions on hand radiography and axial calcifications on CT-scan: A retrospective study in systemic sclerosis. *Medicine (Baltimore)* **99**, e22443 (2020). <https://doi.org/10.1097/MD.00000000000022443>
- 98 Fairley, J. L. *et al.* Clinical Features of Systemic Sclerosis-Mixed Connective Tissue Disease and Systemic Sclerosis Overlap Syndromes. *Arthritis Care Res (Hoboken)* **73**, 732-741 (2021). <https://doi.org/10.1002/acr.24167>
- 99 Jaafar, S. *et al.* Clinical characteristics, visceral involvement, and mortality in at-risk or early diffuse systemic sclerosis: a longitudinal analysis of an observational prospective multicenter US cohort. *Arthritis Res Ther* **23**, 170 (2021). <https://doi.org/10.1186/s13075-021-02548-1>
- 100 Janardana, R. *et al.* Unique clinical and autoantibody profile of a large Asian Indian cohort of scleroderma-do South Asians have a more aggressive disease? *Clin Rheumatol* **38**, 3179-3187 (2019). <https://doi.org/10.1007/s10067-019-04659-2>
- 101 Morrisroe, K. *et al.* Incidence, Risk Factors, and Outcomes of Cancer in Systemic Sclerosis. *Arthritis Care Res (Hoboken)* **72**, 1625-1635 (2020). <https://doi.org/10.1002/acr.24076>
- 102 Noviani, M. *et al.* Mortality and hospitalization outcomes of interstitial lung disease and pulmonary hypertension in the Singapore systemic sclerosis cohort. *Semin Arthritis Rheum* **50**, 473-479 (2020). <https://doi.org/10.1016/j.semarthrit.2019.11.005>
- 103 Sanchez-Cano, D. *et al.* Interstitial lung disease in systemic sclerosis: data from the spanish scleroderma study group. *Rheumatol Int* **38**, 363-374 (2018). <https://doi.org/10.1007/s00296-017-3916-x>
- 104 Simeon-Aznar, C. P. *et al.* Registry of the Spanish Network for Systemic Sclerosis: Survival, Prognostic Factors, and Causes of Death. *Medicine (Baltimore)* **94**, e1728 (2015). <https://doi.org/10.1097/MD.0000000000001728>
- 105 Tomiyama, F. *et al.* High Prevalence of Acute Exacerbation of Interstitial Lung Disease in Japanese Patients with Systemic Sclerosis. *Tohoku J Exp Med* **239**, 297-305 (2016). <https://doi.org/10.1620/tjem.239.297>
- 106 Lee, E. K., Kim, J. S., Kim, Y. & Park, J. S. CT Findings in People Who Were Environmentally Exposed to Asbestos in Korea. *J Korean Med Sci* **30**, 1896-1901 (2015). <https://doi.org/10.3346/jkms.2015.30.12.1896>
- 107 Wickramatillake, B. A., Fernando, M. A. & Frank, A. L. Prevalence of Asbestos-Related Disease Among Workers in Sri Lanka. *Ann Glob Health* **85** (2019). <https://doi.org/10.5334/aogh.2575>
- 108 Choi, S. *et al.* Interstitial lung disorders in the indium workers of Korea: an update study for the relationship with biological exposure indices. *Am J Ind Med* **58**, 61-68 (2015). <https://doi.org/10.1002/ajim.22402>
- 109 Simsek, C., Sari, G., Uzmezoglu, B. A. & Simsek, B. The relationship between respiratory health and hard metal dust exposure: A cross-sectional study. *Arch Environ Occup Health* **77**, 227-233 (2022). <https://doi.org/10.1080/19338244.2020.1870911>
- 110 Hoy, R. F. *et al.* Identification of early-stage silicosis through health screening of stone benchtop industry workers in Victoria, Australia. *Occup Environ Med* **78**, 296-302 (2021). <https://doi.org/10.1136/oemed-2020-106897>
- 111 Knight, D. *et al.* Trends in silicosis prevalence and the healthy worker effect among gold miners in South Africa: a prevalence study with follow up of employment status. *BMC Public Health* **15**, 1258 (2015). <https://doi.org/10.1186/s12889-015-2566-8>
- 112 Souza, T. P., van Tongeren, M. & Monteiro, I. Respiratory health and silicosis in artisanal mine workers in southern Brazil. *Am J Ind Med* **64**, 511-518 (2021). <https://doi.org/10.1002/ajim.23242>
- 113 Poinen-Rughooputh, S. *et al.* Sex-Related Differences in the Risk of Silicosis Among Chinese Pottery Workers: A Cohort Study. *J Occup Environ Med* **63**, 74-79 (2021). <https://doi.org/10.1097/JOM.0000000000002068>
- 114 Requena-Mullor, M. *et al.* Association between Crystalline Silica Dust Exposure and Silicosis Development in Artificial Stone Workers. *Int J Environ Res Public Health* **18** (2021). <https://doi.org/10.3390/ijerph18115625>
- 115 Rechaipichitkul, K. S. N. C. W. Prevalence of Silicosis in Stone Carving Workers Being Exposed to Inorganic Dust at Sikhiu District Nakhonratchasima Province, Thailand; Preliminary Results. *Journal of the Medical Association of Thailand* **100**, 598 (2017).
- 116 Siribaddana, A. D. *et al.* A study on silicosis among employees of a silica processing factory in the Central Province of Sri Lanka. *Ceylon Med J* **61**, 6-10 (2016). <https://doi.org/10.4038/cmj.v61i1.8252>
- 117 Souza, T. P. *et al.* Silicosis prevalence and risk factors in semi-precious stone mining in Brazil. *Am J Ind Med* **60**, 529-536 (2017). <https://doi.org/10.1002/ajim.22719>
- 118 Olson, A. *et al.* Estimation of the Prevalence of Progressive Fibrosing Interstitial Lung Diseases: Systematic Literature Review and Data from a Physician Survey. *Adv Ther* **38**, 854-867 (2021). <https://doi.org/10.1007/s12325-020-01578-6>

Coquart, N., Cadelis, G., Tressieres, B. & Cordel, N. Epidemiology of sarcoidosis in Afro-Caribbean people: a 7-year retrospective study in Guadeloupe. *Int J Dermatol* **54**, 188-192 (2015). <https://doi.org/10.1111/ijd.12633>

Highland, K. B. *et al.* Efficacy and safety of nintedanib in patients with systemic sclerosis-associated interstitial lung disease treated with mycophenolate: a subgroup analysis of the SENSICIS trial. *Lancet Respir Med* **9**, 96-106 (2021). [https://doi.org/10.1016/S2213-2600\(20\)30330-1](https://doi.org/10.1016/S2213-2600(20)30330-1)

Wells, A. U. *et al.* Nintedanib in patients with progressive fibrosing interstitial lung diseases-subgroup analyses by interstitial lung disease diagnosis in the INBUILD trial: a randomised, double-blind, placebo-controlled, parallel-group trial. *Lancet Respir Med* **8**, 453-460 (2020). [https://doi.org/10.1016/S2213-2600\(20\)30036-9](https://doi.org/10.1016/S2213-2600(20)30036-9)

Behr, J. *et al.* Pirfenidone in patients with progressive fibrotic interstitial lung diseases other than idiopathic pulmonary fibrosis (RELIEF): a double-blind, randomised, placebo-controlled, phase 2b trial. *Lancet Respir Med* **9**, 476-486 (2021). [https://doi.org/10.1016/S2213-2600\(20\)30554-3](https://doi.org/10.1016/S2213-2600(20)30554-3)

CPRD. *Introudction to CPRD* (2023).

CPRD. *Using linked data*, <<https://cprd.com/using-linked-data>> (2023).

Wolf, A. *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* **48**, 1740-1740g (2019). <https://doi.org/10.1093/ije/dyz034>

Grosios, K., Gahan, P. B. & Burbidge, J. Overview of healthcare in the UK. *Epma j* **1**, 529-534 (2010). <https://doi.org/10.1007/s13167-010-0050-1>

Anderson, M. *et al.* United Kingdom: Health System Review. *Health Syst Transit* **24**, 1-194 (2022).

Mahadevan, P. *et al.* Completeness and representativeness of small area socioeconomic data linked with the UK Clinical Practice Research Datalink (CPRD). *J Epidemiol Community Health* **76**, 880-886 (2022). <https://doi.org/10.1136/jech-2022-219200>

CPRD. *Using CPRD primary care data*, <<https://cprd.com/using-cprd-primary-care-data>> (2023).

Jick, S. S. *et al.* Quality and completeness of diagnoses recorded in the new CPRD Aurum Database: evaluation of pulmonary embolism. *Pharmacoepidemiol Drug Saf* **29**, 1134-1140 (2020). <https://doi.org/10.1002/pds.4996>

Persson, R. *et al.* Quality and Completeness of Myocardial Infarction Recording in Clinical Practice Research Datalink Aurum. *Clin Epidemiol* **13**, 745-753 (2021). <https://doi.org/10.2147/CLEP.S319245>

Persson, R. *et al.* CPRD Aurum database: Assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiol Drug Saf* **29**, 1456-1464 (2020). <https://doi.org/10.1002/pds.5135>

Herrett, E., Thomas, S. L., Schoonen, W. M., Smeeth, L. & Hall, A. J. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* **69**, 4-14 (2010). <https://doi.org/10.1111/j.1365-2125.2009.03537.x>

Herrett, E. *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* **44**, 827-836 (2015). <https://doi.org/10.1093/ije/dyv098>

Yu, D. *et al.* Trends in the Annual Consultation Incidence and Prevalence of Low Back Pain and Osteoarthritis in England from 2000 to 2019: Comparative Estimates from Two Clinical Practice Databases. *Clin Epidemiol* **14**, 179-189 (2022). <https://doi.org/10.2147/CLEP.S337323>

(2023), C. P. R. D. (ed CPRD) (March 2023).

Shiekh, S. I. *et al.* Completeness, agreement, and representativeness of ethnicity recording in the United Kingdom's Clinical Practice Research Datalink (CPRD) and linked Hospital Episode Statistics (HES). *Popul Health Metr* **21**, 3 (2023). <https://doi.org/10.1186/s12963-023-00302-0>

Padmanabhan, S. *et al.* Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol* **34**, 91-99 (2019). <https://doi.org/10.1007/s10654-018-0442-4>

Boyd A., C. R., Johnson L., Simmonds S., Syddall H., Westbury L., Cooper C., & J, M. Understanding Hospital Episode Statistics (HES). (UCL Institute of Education 2017).

Digital, N. *Hospital Episode Statistics (HES)*, <<https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics>> (2024).

Herbert, A., Wijlaars, L., Zylbersztejn, A., Cromwell, D. & Hardelid, P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* **46**, 1093-1093i (2017). <https://doi.org/10.1093/ije/dyx015>

Navaratnam, V. *et al.* The rising incidence of idiopathic pulmonary fibrosis in the U.K. *Thorax* **66**, 462-467 (2011). <https://doi.org/10.1136/thx.2010.148031>

Hubbard, R., Venn, A., Lewis, S. & Britton, J. Lung cancer and cryptogenic fibrosing alveolitis. A population-based cohort study. *Am J Respir Crit Care Med* **161**, 5-8 (2000). <https://doi.org/10.1164/ajrccm.161.1.9906062>

Esposito, D. B. *et al.* Idiopathic Pulmonary Fibrosis in United States Automated Claims. Incidence, Prevalence, and Algorithm Validation. *Am J Respir Crit Care Med* **192**, 1200-1207 (2015). <https://doi.org/10.1164/rccm.201504-0818OC>

Ley, B. *et al.* Code-based Diagnostic Algorithms for Idiopathic Pulmonary Fibrosis. Case Validation and Improvement. *Ann Am Thorac Soc* **14**, 880-887 (2017). <https://doi.org/10.1513/AnnalsATS.201610-764OC>

146 Excellence, N. I. f. H. a. C. *Nintedanib for treating idiopathic pulmonary fibrosis when forced vital capacity is*
147 *above 80% predicted*, <<https://www.nice.org.uk/guidance/ta864>> (2023).
148 in *Idiopathic pulmonary fibrosis in adults: diagnosis and management* National Institute for Health and Care
149 Excellence: Guidelines (2017).
150 Laurensen, S., Sidhu, R., Goodall, M. & Adler, A. I. NICE guidance on nintedanib for treating idiopathic
151 pulmonary fibrosis. *Lancet Respir Med* **4**, 176-177 (2016). [https://doi.org/10.1016/S2213-2600\(16\)00022-9](https://doi.org/10.1016/S2213-2600(16)00022-9)
152 Gribbin, J. *et al.* Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Thorax* **61**,
153 980-985 (2006). <https://doi.org/10.1136/thx.2006.062836>
154 Raghu, G. *et al.* Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: incidence,
155 prevalence, and survival, 2001-11. *Lancet Respir Med* **2**, 566-572 (2014). [https://doi.org/10.1016/S2213-2600\(14\)70101-8](https://doi.org/10.1016/S2213-2600(14)70101-8)
156 Morgan, A., Sinnott, S. J., Smeeth, L., Minassian, C. & Quint, J. Concordance in the recording of stroke across
157 UK primary and secondary care datasets: a population-based cohort study. *BJGP Open* **5** (2021).
158 <https://doi.org/10.3399/BJGPO.2020.0117>
159 Quint, J. K. *et al.* Validation of chronic obstructive pulmonary disease recording in the Clinical Practice
160 Research Datalink (CPRD-GOLD). *BMJ Open* **4**, e005540 (2014). <https://doi.org/10.1136/bmjopen-2014-005540>
161 Johnston, I. D., Bleasdale, C., Hind, C. R. & Woodcock, A. A. Accuracy of diagnostic coding of hospital
162 admissions for cryptogenic fibrosing alveolitis. *Thorax* **46**, 589-591 (1991). <https://doi.org/10.1136/thx.46.8.589>
163 Khan, N. F., Harrison, S. E. & Rose, P. W. Validity of diagnostic coding within the General Practice Research
164 Database: a systematic review. *Br J Gen Pract* **60**, e128-136 (2010). <https://doi.org/10.3399/bjgp10X483562>
165 Tulloch, J. S. *et al.* GP coding behaviour for non-specific clinical presentations: a pilot study. *BJGP Open* **4**
166 (2020). <https://doi.org/10.3399/bjgpopen20X101050>
167 Noble, P. W. *et al.* Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised
168 trials. *Lancet* **377**, 1760-1769 (2011). [https://doi.org/10.1016/S0140-6736\(11\)60405-4](https://doi.org/10.1016/S0140-6736(11)60405-4)
169 Transplant, N. B. a. Annual Report on
170 Cardiothoracic Organ Transplantation 2021/2022, NHS Blood and Transplant. (2022).
171 Richeldi, L. *et al.* Nintedanib in patients with idiopathic pulmonary fibrosis: Combined evidence from the
172 TOMORROW and INPULSIS(R) trials. *Respir Med* **113**, 74-79 (2016).
173 <https://doi.org/10.1016/j.rmed.2016.02.001>
174 Hutchinson, J., Fogarty, A., Hubbard, R. & McKeever, T. Global incidence and mortality of idiopathic
175 pulmonary fibrosis: a systematic review. *Eur Respir J* **46**, 795-806 (2015).
176 <https://doi.org/10.1183/09031936.00185114>
177 Idiopathic Pulmonary Fibrosis Clinical Research, N. *et al.* Prednisone, azathioprine, and N-acetylcysteine for
178 pulmonary fibrosis. *N Engl J Med* **366**, 1968-1977 (2012). <https://doi.org/10.1056/NEJMoa1113354>
179 Gonnelli, F., Bonifazi, M. & Hubbard, R. Mortality trends in idiopathic pulmonary fibrosis in Europe between
180 2013 and 2018. *Eur Respir J* **64** (2024). <https://doi.org/10.1183/13993003.02080-2023>
181 Statistics, O. o. N. (2023).
182 Cottin, V. *et al.* Mortality and Respiratory-Related Hospitalizations in Idiopathic Pulmonary Fibrosis Not
183 Treated With Antifibrotics. *Front Med (Lausanne)* **8**, 802989 (2021). <https://doi.org/10.3389/fmed.2021.802989>
184 Hubbard, R., Johnston, I. & Britton, J. Survival in patients with cryptogenic fibrosing alveolitis: a population-
185 based cohort study. *Chest* **113**, 396-400 (1998). <https://doi.org/10.1378/chest.113.2.396>
186 Mapel, D. W. *et al.* Idiopathic pulmonary fibrosis: survival in population based and hospital based cohorts.
187 *Thorax* **53**, 469-476 (1998). <https://doi.org/10.1136/thx.53.6.469>
188 Excellence, N. I. f. H. a. C. *Idiopathic pulmonary fibrosis in adults: diagnosis and management*
189 <<https://www.nice.org.uk/guidance/cg163>> (2013).
190 Excellence., N. I. f. H. a. C. *Pirfenidone for treating idiopathic pulmonary fibrosis*
191 <<https://www.nice.org.uk/guidance/ta504>> (2021).
192 Knuppel, L. *et al.* A Novel Antifibrotic Mechanism of Nintedanib and Pirfenidone. Inhibition of Collagen Fibril
193 Assembly. *Am J Respir Cell Mol Biol* **57**, 77-90 (2017). <https://doi.org/10.1165/rcmb.2016-0217OC>
194 Maher, T. M. *et al.* Pirfenidone in patients with unclassifiable progressive fibrosing interstitial lung disease:
195 design of a double-blind, randomised, placebo-controlled phase II trial. *BMJ Open Respir Res* **5**, e000289
196 (2018). <https://doi.org/10.1136/bmjresp-2018-000289>
197 Maher, T. M. & Streck, M. E. Antifibrotic therapy for idiopathic pulmonary fibrosis: time to treat. *Respir Res* **20**,
198 205 (2019). <https://doi.org/10.1186/s12931-019-1161-4>
199 Crestani, B. *et al.* Long-term safety and tolerability of nintedanib in patients with idiopathic pulmonary fibrosis:
200 results from the open-label extension study, INPULSIS-ON. *Lancet Respir Med* **7**, 60-68 (2019).
201 [https://doi.org/10.1016/S2213-2600\(18\)30339-4](https://doi.org/10.1016/S2213-2600(18)30339-4)
202 Bahmer, T. *et al.* Clinical Correlates of Reduced Physical Activity in Idiopathic Pulmonary Fibrosis.
203 *Respiration* **91**, 497-502 (2016). <https://doi.org/10.1159/000446607>

- 173 Nathan, S. D. *et al.* Validation of test performance characteristics and minimal clinically important difference of the 6-minute walk test in patients with idiopathic pulmonary fibrosis. *Respir Med* **109**, 914-922 (2015). <https://doi.org/10.1016/j.rmed.2015.04.008>
- 174 Dempsey, T. M. *et al.* Clinical Effectiveness of Antifibrotic Medications for Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* **200**, 168-174 (2019). <https://doi.org/10.1164/rccm.201902-0456OC>
- 175 Thong, L., McElduff, E. J. & Henry, M. T. Trials and Treatments: An Update on Pharmacotherapy for Idiopathic Pulmonary Fibrosis. *Life (Basel)* **13** (2023). <https://doi.org/10.3390/life13020486>
- 176 Canestaro, W. J., Forrester, S. H., Raghu, G., Ho, L. & Devine, B. E. Drug Treatment of Idiopathic Pulmonary Fibrosis: Systematic Review and Network Meta-Analysis. *Chest* **149**, 756-766 (2016). <https://doi.org/10.1016/j.chest.2015.11.013>
- 177 Nathan, S. D. *et al.* Effect of pirfenidone on mortality: pooled analyses and meta-analyses of clinical trials in idiopathic pulmonary fibrosis. *Lancet Respir Med* **5**, 33-41 (2017). [https://doi.org/10.1016/S2213-2600\(16\)30326-5](https://doi.org/10.1016/S2213-2600(16)30326-5)
- 178 Rogliani, P., Calzetta, L., Cavalli, F., Matera, M. G. & Cazzola, M. Pirfenidone, nintedanib and N-acetylcysteine for the treatment of idiopathic pulmonary fibrosis: A systematic review and meta-analysis. *Pulm Pharmacol Ther* **40**, 95-103 (2016). <https://doi.org/10.1016/j.pupt.2016.07.009>
- 179 Di Martino, E., Provenzani, A., Vitulo, P. & Polidori, P. Systematic Review and Meta-analysis of Pirfenidone, Nintedanib, and Pamrevlumab for the Treatment of Idiopathic Pulmonary Fibrosis. *Ann Pharmacother* **55**, 723-731 (2021). <https://doi.org/10.1177/1060028020964451>
- 180 Datalink, C. P. R. *CPRD Aurum July 2023* (2023).
- 181 Dempsey, I. J. *et al.* Relationship Between Physical Therapy Characteristics, Surgical Procedure, and Clinical Outcomes in Patients After ACL Reconstruction. *J Sport Rehabil* **28**, 171-179 (2019). <https://doi.org/10.1123/jsr.2017-0176>
- 182 Mooney, J. *et al.* Antifibrotic therapies reduce mortality and hospitalization among Medicare beneficiaries with idiopathic pulmonary fibrosis. *J Manag Care Spec Pharm* **27**, 1724-1733 (2021). <https://doi.org/10.18553/jmcp.2021.27.12.1724>
- 183 Iommi, M. *et al.* Prognosis and Survival in Idiopathic Pulmonary Fibrosis in the Era of Antifibrotic Therapy in Italy: Evidence from a Longitudinal Population Study Based on Healthcare Utilization Databases. *Int J Environ Res Public Health* **19** (2022). <https://doi.org/10.3390/ijerph192416689>
- 184 Maher, T. M. *et al.* Unmet needs in the treatment of idiopathic pulmonary fibrosis-insights from patient chart review in five European countries. *BMC Pulm Med* **17**, 124 (2017). <https://doi.org/10.1186/s12890-017-0468-5>
- 185 Spencer, L. G., Loughenbury, M., Chaudhuri, N., Spiteri, M. & Parfrey, H. Idiopathic pulmonary fibrosis in the UK: analysis of the British Thoracic Society electronic registry between 2013 and 2019. *ERJ Open Res* **7** (2021). <https://doi.org/10.1183/23120541.00187-2020>
- 186 Egashira, R. & Raghu, G. Acute exacerbation of fibrotic interstitial lung disease beyond idiopathic pulmonary fibrosis: time to intervene. *Eur Respir J* **61** (2023). <https://doi.org/10.1183/13993003.00459-2023>
- 187 Farrand, E. *et al.* Impact of Idiopathic Pulmonary Fibrosis on Longitudinal Health-care Utilization in a Community-Based Cohort of Patients. *Chest* **159**, 219-227 (2021). <https://doi.org/10.1016/j.chest.2020.07.035>
- 188 Swigris, J. J. & Fairclough, D. Patient-reported outcomes in idiopathic pulmonary fibrosis research. *Chest* **142**, 291-297 (2012). <https://doi.org/10.1378/chest.11-2602>
- 189 Olson, A. L. *et al.* Healthcare Resource Utilization and Related Costs in Chronic Fibrosing Interstitial Lung Diseases with a Progressive Phenotype: A US Claims Database Analysis. *Adv Ther* **39**, 1794-1809 (2022). <https://doi.org/10.1007/s12325-022-02066-9>
- 190 Davis, S., Rafia, R., Carroll, C., Hamilton, J. & Essat, M. Pirfenidone for Treating Idiopathic Pulmonary Fibrosis: An Evidence Review Group Perspective of a NICE Single Technology Appraisal. *Pharmacoeconomics* **37**, 763-775 (2019). <https://doi.org/10.1007/s40273-018-0727-1>
- 191 Raghu, G. Idiopathic pulmonary fibrosis: lessons from clinical trials over the past 25 years. *Eur Respir J* **50** (2017). <https://doi.org/10.1183/13993003.01209-2017>

9.1 Appendix 1. Supplementary material to Chapter 2

Table S1 List of studies and detailed characteristics of studies reporting the **incidence** of IPF, Sorted by Publication Year

Author, year	Study design	Country	Study population	ILD assessment	Sample size	No. of IPF cases by ILD definition	Mean age; Female (%)	Reported Incidence
Gjonbrataj 2015	Rtr	South Korea	HIRA	Definition 1: code J84, ICD-10 code; (BAL) or lung biopsy.	36,765,374	21,805	NR; 38	48.5 per 100,000
				Definition 2: code J84 and HRCT, bronchoalveolar lavage or lung biopsy		13,680	NR; 25	32.2 per 100,000
				Definition 3: code J84.1, ICD-10 code;		5,881	NR; 12	16.2 per 100,000
				Definition 4: code J84.1 and HRCT, BAL or lung biopsy; and		4,219	NR: 8.5	11.4 per 100,000
				Definition 5: code J84.1A, based on the 2011 international statement.		615	NR; 0.9	1.7 per 100,000
Lee 2016	Rtr	South Korea	HIRA	Patients with IPF were defined as those with a disease code of K-J84.18.	51,026,868	2011,2012;657	NR	13.1 and 12.9 per 100,000
Harari 2016	Rtr	Italy	Healthcare administrative database of Lombardy	IPF was defined as: <u>GCD</u> : Defined as IPF cases all individuals with at least one hospitalization with diagnosis of IPF or at least one outpatient visits with diagnosis of IPF (ICD-9-CM code 516.3) during the period from 1st January 2000 to 31st December 2010.	Approx. 10 million	2,951	Majority >65 years; 43	5.24 (5.065.44) per 100,000 py
				<u>BCD</u> : Defined IPF cases those patients that satisfied the GCD and had no claims (inpatient or outpatient) with a diagnosis code for any other type of ILDs on or after the date of the last claim with IPF diagnosis		2,093	Majority >65 years; 40	3.74 (3.583.90) per 100,000 py
				<u>NCD</u> : Defined IPF cases those patients that satisfied the BCD and had one or more claim with a procedure code for surgical lung biopsy, transbronchial lung biopsy or computed tomography of the thorax, on or before the date of the last claim with a diagnosis code for IPF.		1,309	Majority >65 years; 41	2.33(2.202.46) per 100,000 py
Hopkins 2016	Rtr	Canada	National databases:		NR	6,390	NR	18.7 per 100,000

			DAD and NACRS	<u>Broad</u> definition: by excluding cases with an ICD-10 CA code for another ILD after the J84.1 code, like other studies.				
Author, year	Study design	Country	Study population	ILD assessment	Sample size	No. of IPF cases by ILD definition	Mean age; Female (%)	Reported Incidence
				<u>Narrow</u> definition further excluded cases that did not have chest CT, bronchus or lung biopsy, or bronchoscopy prior to establishing a J84.1 code.		3,057	NR	9 per 100,000
Raghu 2016	Rtr	USA	Optum Clinformatics	<u>Primary</u> cohort: IPF patients were first identified if they had at least one claim with ICD-9-CM diagnosis code 516.3 (idiopathic fibrosing alveolitis, which is used by providers to identify IPF) during the study period. Two subgroups of the primary cohort were further identified by applying additional criteria to form more restrictive cohorts of prevalent IPF patients:	9,375,020	4,206	53;51	5.8 per 100,000
				<u>Broad</u> case definition excluded patients if they had a claim with the diagnosis code 515 after the last diagnosis code 516.3.		2,670	53;52	3.6-5.1 per 100,000
				<u>Narrow</u> case definition further restricted the broad case subgroup patients by requiring a claim for a surgical lung biopsy, transbronchial lung biopsy, or computed tomography scan of the thorax prior to the last diagnosis code 516.3.		1,685	54;53	2.4-2.9 per 100,000
Strongman 2018	Rtr	UK	CPRD	<u>Narrow</u> IPF-CS case definition were H563.00 (Idiopathic fibrosing alveolitis), H563.12 (Cryptogenic fibrosing alveolitis), H563z00 (Idiopathic fibrosing alveolitis NOS), H563300 (Usual interstitial pneumonitis), and H563.13 (Idiopathic pulmonary fibrosis).	9,748,108	1,491	NR	2.85(2.7-3) per 100,000 py
				<u>Broad</u> IPF-CS definition included the following three additional Read codes: H563100 (Diffuse pulmonary fibrosis), H563200 (Pulmonary		4,527		8.7 (8.4-8.9) per 100,000 py

				fibrosis), and H563.11 (Hamman-Rich syndrome).				
Lim 2019	Rtr	South Korea	Korean National Health Insurance Service (NHIS) database for patients registered with the co-payment assistance policy for rare and incurable diseases.	ICD codes	2011: 50,908,646, 2012: 51,169,141, 2013: 51,448,491, 2014: 51,757,146 2015: 52,034,424	NR	NR	2011: 27.85 2012: 28.36 2013: 31.28 2014: 32.21 2015: 35.82 Cumulative 2011-2015: 31.03 per 100,000
Bellou 2021	Prosp	UK	Biobank, self-reported (all unrelated individuals of White ethnicity participating in the UK Biobank)	<u>Broad</u> included all individuals with an ICD, Tenth Revision, diagnosis of J84.1 and J84.9 or an ICD, Ninth Revision, diagnosis of 516.3.	437,453 individuals	1,234	63,41	282.1 per 100,000
				<u>Narrow</u> : individuals with pulmonary fibrosis excluding other medical conditions that cause pulmonary fibrosis or interstitial lung disease		802	NR	183.3 per 100,000
Zhang 2021	Rtr	USA	Veterans Health Administration (VHA) electronic health record	<u>Broad</u> : Patients who had a diagnosis code for IPF (ICD-9-CM code 515, 516.3, 516.31 or ICD10-CM code J84.111, J84.112, J84.89, J84.9, J84.10, J84.17) recorded during study period • Patients were considered to have IPF if they did not have any other diagnosis code for an alternative interstitial lung disease	10.7 million	139,116	71;7	331 per 100,000
				<u>Narrow</u> : Patients who had a diagnosis code for IPF (ICD-9-CM code 515, 516.3, 516.31 or ICD10-CM code J84.111, J84.112, J84.89, J84.9, J84.10, J84.17) recorded during study period Patients were considered to have IPF if they did not have any other diagnosis code for an alternative interstitial lung disease Patients who had procedure code for a lung biopsy or a CT scan of the thorax before the last IPF diagnosis		82,557	707	210 per 100,000 py

Table S2 List of studies and detailed characteristics of studies reporting the **incidence** of non-IPF, Sorted by Publication Year

ILD	Author, year	Country	Study design	Study population description	ILD assessment	Study population (N)	ILD subtype	No. of ILD cases	Mean age; Female (%)	Incidence
IIP	Lee 2016	Korea	Rtr	HIRA database	Diagnosis codes; Patients with IIP were defined as those with two or more medical visits or at least one hospitalisation with an IIP diagnostic code	51,026,868	IIP	17,826	NR; 50	34.93 per 100,00 persons
IIP	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: - Public and private hospitals - Community Pulmonologists and General Practitioners	- A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. - ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/A LAT diagnostic criteria for IPF. - ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. - For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.	1,194,601	IIP	52	56;50	4.4 per 100,000 person per year
IIP							NSIP	10		0.8 per 100,000 person per year
							DIP	3		0.3 per 100,000 person per year
							OP	1		0.1 per 100,000 person per year
SSc	Wangkaew 2016	Thailand	Prosp	Maharaj Nakorn Chiang Mai Hospital, Chiang Mai University	All HRCTs were obtained with one of two MDCT platforms. All images were reviewed by an experienced thoracic radiologist blinded to clinical and laboratory data. HRCTs were categorized by the pattern of the lung parenchyma findings that represented ILD,	89	dcSSc	NR	53.4;58	58.8 per 100,000 py

SSc	Wangkaew 2016	Thailand	Prosp	Maharaj Nakorn Chiang Mai Hospital, Chiang Mai University	All HRCTs were obtained with one of two MDCT platforms All images were reviewed by an experienced thoracic radiologist blinded to clinical and laboratory data. HRCTs were categorized by the pattern of the lung parenchyma findings that represented ILD,	24	lcSSc	NR	53.4;58	17.3 per 100,000 py
SSc	Carton 2021	Belgium	Rtr	Consecutive SSc patients, included in two Flemish cohort	•Each HRCT was centrally analysed by an experienced investigator who was blinded. Patients were classified into three subgroups (“no ILD”, limited ILD or extensive ILD), according to the simplified flow diagram described by Goh et al	722	SSc, overall	39		2,570 (1,800-3,560) (8%)
SSc	Carton 2021	Belgium	Rtr	Consecutive SSc patients, included in two Flemish cohort	•Each HRCT was centrally analysed by an experienced investigator who was blinded. Patients were classified into three subgroups (“no ILD”, limited ILD or extensive ILD), according to the simplified flow diagram described by Goh et al	281	lcSSc	26		2,300 (1,500-3,360) (9%)
SSc	Li 2021	USA	Rtr	IBM MarketScan claims database	ICD codes	78,964,708	Overall, SSc	8,252	58,82	4.3 (4.2 - 4.4)
RA ILD	Zhang 2017	China	Rtr	Patients who were diagnosed with RA at the Capital Medical University, Beijing Chaoyang Hospital, during the study period	Not reported, ILD cases were reported as seen in the medical history	550	205	57.6;63	Percentage of RA study cohort	37.3
RA ILD	Raimundo 2019	USA	Rtr	Truven Health MarketScan Commercial and Medicare Supplemental health insurance databases	Patients meeting ≥ 1 of the following criteria were identified as potential RAILD cases: ≥ 2 claims with an International Classification of Diseases-9-Clinical Modification (ICD-9-CM) diagnosis code for pulmonary disease (fibrosis: ICD-9-CM 515, 516.3, 516.31; rheumatic lung disease: ICD-9-CM 714.81) on different days, with the second claim occurring within 12 months of the first plus ≥ 2 claims with a diagnosis of RA (ICD-9-CM 714.xx, excluding 714.3x and 714.4x) on different days in the 12 months before or	over 150 million lives	NR	NR	2004, Per 100,00 persons	2.7 (95% CI 2.5-2.9)

					after the first claim for pulmonary disease; ≥ 2 claims with a diagnosis of rheumatic lung disease (ICD-9-CM 714.81) on different days, with the second claim occurring within 12 months of the first based on a previously published algorithm.					
RA ILD	Raimundo 2019	USA	Rtr	Truven Health MarketScan Commercial and Medicare Supplemental health insurance databases	Patients meeting ≥ 1 of the following criteria were identified as potential RAILD cases: ≥ 2 claims with an International Classification of Diseases-9-Clinical Modification (ICD-9-CM) diagnosis code for pulmonary disease (fibrosis: ICD-9-CM 515, 516.3, 516.31; rheumatic lung disease: ICD-9-CM 714.81) on different days, with the second claim occurring within 12 months of the first plus ≥ 2 claims with a diagnosis of RA (ICD-9-CM 714.xx, excluding 714.3x and 714.4x) on different days in the 12 months before or after the first claim for pulmonary disease; ≥ 2 claims with a diagnosis of rheumatic lung disease (ICD-9-CM 714.81) on different days, with the second claim occurring within 12 months of the first based on a previously published algorithm.	over 150 million lives	NR	NR	2013, Per 100,00 persons (general population)	3.8 (95% CI 3.5-4.0)
RA ILD	Sparks 2021**	USA	Rtr	Medicare claims data that includes nearly all Americans 65 years of age and some younger individuals with disabilities certain disabilities.	Among patients ever identified with RA, we identified ILD by the presence of two or more ICD-9 or ICD-10 codes for ILD in any position by a rheumatologist or pulmonologist separated by 7–365 days, as previously validated (PPV 72.4%) RA-ILD date was defined as the second billing code for ILD that fulfilled the algorithm. Incident RA-ILD was defined as occurring after the RA index date.	499,481	13,372	NR	Per 100,000 person-years among RA patients	714 (700-730) per 100,000 persons (2.7%)

Sjogrens ILD	Roca 2017	France	Rtr	Medical records from 2 participating institutional centers	ILD was investigated by PFTs, and HRCT scan of the lungs; no patient underwent lung biopsy.	263	9	63;86	p-SS	3.4 (%)
MCTD	Reiseter 2018	Norway	Rtr	Nationwide MCTD cohort from depts of Rheumatology	CT scans and PFTs	119	2	91;38		1.7%
CTD	Olaosebikan 2021	Nigeria	CS	CTD patients managed at the Rheumatology Unit of Lagos State University Teaching Hospital	Clinical information including socio-demographics, signs and symptoms of ILD, pattern(s) of CTD, serology, pulmonary function, and radiology findings as well as treatment outcomes were retrieved from patients' records	318	8	39;90		2.5%
HP	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: - Public and private hospitals - Community Pulmonologists and General Practitioners	MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.	1,194,601	11		HP	0.9 per 100,000
HP	Perez 2018	United States	Rtr	Truven Health Market Scan Commercial and Medicare	Novel claims-based coding algorithms to identify HP, CHP, FHP cases The case validation procedure involved medical record review performed using the electronic medical records An expert chest radiologist reviewed CT scans, blinded to subjects' clinical information and the primary radiologist's interpretation, to determine whether pulmonary fibrosis was present and whether the	158,270,350	3405		HP	1.28-1.9 (1.7 – 2.2)

					CT met criteria for consistent or possible HP or was inconsistent with HP. Last, a clinical-radiological consensus diagnosis of HP or not HP was applied.					
HP	Perez 2018	United States	Rtr	Truven Health Market Scan Commercial and Medicare	Novel claims-based coding algorithms to identify HP, CHP, FHP cases The case validation procedure involved medical record review performed using the electronic medical records An expert chest radiologist reviewed CT scans, blinded to subjects' clinical information and the primary radiologist's interpretation, to determine whether pulmonary fibrosis was present and whether the CT met criteria for consistent or possible HP or was inconsistent with HP. Last, a clinical-radiological consensus diagnosis of HP or not HP was applied.	158,270,350			Chronic HP	0.63 (95% 0.5– 0.7) – 1.08 (95% CI, 0.9– 1.3)
HP	Perez 2018	United States	Rtr	Truven Health Market Scan Commercial and Medicare	Novel claims-based coding algorithms to identify HP, CHP, FHP cases The case validation procedure involved medical record review performed using the electronic medical records An expert chest radiologist reviewed CT scans, blinded to subjects' clinical information and the primary radiologist's interpretation, to determine whether pulmonary fibrosis was present and whether the CT met criteria for consistent or possible HP or was inconsistent with HP. Last, a clinical-radiological consensus diagnosis of HP or not HP was applied.	158,270,350			Fibrosing HP	0.29 (95% CI, 0.2–0.4) - 0.43 (95% CI, 0.4– 0.5)
Cystic lung disease	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: • Public and	MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable.	1,194,601	4		LAM	0.3 per 100,000 persons per year

				private hospitals • Community Pulmonologists and General Practitioners	ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.					
Cystic lung disease	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: • Public and private hospitals • Community Pulmonologists and General Practitioners	MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.	1,194,601	2		PLCH	0.2 per 100,000 persons per year
Unclassifiable ILDs	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: • Public and private hospitals • Community Pulmonologists and General Practitioners	MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis	1,194,601	22			1.8 per 100,000 persons per year

					was taken from the completed questionnaires or the ICDs-10 codes.					
Silicosis	Casey 2019	USA	Rtr	Medicare beneficiaries	Three case definitions were used to identify beneficiaries with silicosis: - Broad: Any claim that included ICD-9-CM code 502, pneumoconiosis due to other silica or silicates, listed in any position during 1999-2014 - Intermediate: Among broad definition cases those who used i) at least one inpatient, skilled nursing, or home health agency claim or ii) at least two outpatient provider claims within 365 days of each other. Narrow: Among intermediate cases, a narrow definition applied included, that cases have a chest X-ray or CT scan 30 days before or 30 days after a silicosis claim.	4,99,23,987	19,696	65-84 years: 80;19	broad	32.1 per 100,000
Silicosis	Casey 2019	USA	Rtr	Medicare beneficiaries	Three case definitions were used to identify beneficiaries with silicosis: - Broad: Any claim that included ICD-9-CM code 502, pneumoconiosis due to other silica or silicates, listed in any position during 1999-2014 - Intermediate: Among broad definition cases those who used i) at least one inpatient, skilled nursing, or home health agency claim or ii) at least two outpatient provider claims within 365 days of each other. Narrow: Among intermediate cases, a narrow definition applied included, that cases have a chest X-ray or CT scan 30 days before or 30 days after a silicosis claim.	4,99,23,987	11,469	65-84 years: 80;19	Interim	19.1 per 100,000
Silicosis	Casey 2019	USA	Rtr	Medicare beneficiaries	Three case definitions were used to identify beneficiaries with silicosis: Broad: Any claim that included ICD-9-CM code 502, pneumoconiosis due to other silica or silicates, listed in any position during 1999-2014	4,99,23,987	10,026	65-84 years: 67;19	Narrow	16.6 per 100,000

					any position during 1999-2014 Intermediate: Among broad definition cases those who used i) at least one inpatient, skilled nursing, or home health agency claim or ii) at least two outpatient provider claims within 365 days of each other. Narrow: Among intermediate cases, a narrow definition applied included, that cases have a chest X-ray or CT scan 30 days before or 30 days after a silicosis claim.					
Silicosis	Casey 2019	USA	Rtr	Medicare beneficiaries	Three case definitions were used to identify beneficiaries with silicosis: - Broad: Any claim that included ICD-9-CM code 502, pneumoconiosis due to other silica or silicates, listed in any position during 1999-2014 - Intermediate: Among broad definition cases those who used i) at least one inpatient, skilled nursing, or home health agency claim or ii) at least two outpatient provider claims within 365 days of each other. Narrow: Among intermediate cases, a narrow definition applied included, that cases have a chest X-ray or CT scan 30 days before or 30 days after a silicosis claim.	4,99,23,987			Avg. annual	2.4 – 4.6 cases per 100,000
Asbestosis	Szeszenia-Dąbrowska 2015	Poland	Rtr	Based on data obtained from former asbestos company records and the Central Register of Occupational Diseases, this study included 18 large state-owned asbestos processing enterprises	Chest radiography and detection of the changes in the X-ray chest image are crucial in diagnosing asbestosis. The diagnosis of asbestosis relies on detecting the presence of X-ray changes as small irregular s-, t-, u-type opacities with density starting from 1/0. It is in accordance with the radiological classification of pneumoconiosis developed by ILO.	43,650	2,160	NR		49.8 per 100,0 workers (4.9%)
Asbestosis	DeBono 2021	Canada	Longitudinal	Workers employed in over 900 industry groups as part of	□ Cases were defined as those who had any of the following: (1) two or more physician	1.76 million	737	66;44		0.042%

				ongoing surveillance program	visits with a diagnosis of ICD-9 code 501 (asbestosis), (2) two or more visits to an ambulatory care center with a main diagnosis of ICD-10 code J61 (pneumoconiosis due to asbestos and other mineral fibers), or (3) one or more hospital discharges with a main diagnosis of ICD-10 code J61.					
Asbestosis	Thomsen 2021	Denmark	Longitudinal	This study took advantage of the new nationwide DOC*X Project research database, which has been approved by the Danish Data Protection Agency, self-reported	- On 1 January 1980, we ascertained all hospital contacts up to and including the year 2015 for asbestosis (ICD- 8 code 515.2; ICD- 10 code J61) - Pathological confirmation remains the gold standard for diagnosis, 19 35 but lung biopsy is not routinely recommended in Denmark	138,276 (Vehicle mechanics)	47 (Number at risk: 138,226 Person-years at risk: 2,751,423)	NR		0.03%
Pneumoconiosis	Cui 2015	China	Rtr	Workers registered in the selected 4 colliery groups in North China	- Qualified experts who were all members of the Pneumoconiosis Diagnosis Committee independently read the chest radiographs of CWP patients and other investigated workers - If there was difference among the experts on the diagnosis, the diagnosis judgment principle was that the minority should be subordinate to the majority	NR	2,873	Exposure duration Datong: 19.8 (7.2) years Kailuan: 24.9 (7.1) years Fuxin: 23.5 (5.1) years Tiefa: 20.5 (6.2) years	Avg: 4.06% Cum (1980-1970): 6.6-14.1% Avg: 4.9% Cum (1980-1970): 2.9-8.5% Avg: 1.4% Cum (1980-1970): 1.6-77.3% Avg: 0.3% Cum (1980-1970): 0.4-3.6%	
Pneumoconiosis	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: - Public and private hospitals - Community Pulmonologists and General Practitioners	- A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. - ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF.	1,194,601	9	NR		0.8 per 100,000 per year

					• ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.					
Pulmonary Sarcoid	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: • Public and private hospitals • Community Pulmonologists and General Practitioners	• A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. • ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. • ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.	1,194,601	58			4.9 per 100,000 persons per year
Pulmonary Sarcoid	Jeon 2020	South Korea	Rtr	NHI linked to RID	A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the	402,345,679,01	1,955			0.48

					completed questionnaires or the ICDs-10 codes.					
Olson 2021	Olson 2021**	USA	Rtr	IBM MarketScan claims database	ICD codes					
Nasser 2021	Nasser 2021	France	Rtr	French healthcare database	PF-ILD case definitions was based on three algorithms modified from Olson et al 2020 using diagnosing codes	538,472	14,119			4.0 to 4.7

**Adjusted for age and sex

*No papers reporting Incidence of myositis were found

Table S3 List of studies and detailed characteristics of studies reporting the **prevalence** of non-IPF, Sorted by Publication Year

ILD	Author, year	Country	Study design	Study population description	ILD assessment	Study population (N)	ILD subtype	No. of ILD cases	Mean age; Female (%)	Prevalence
IIP	Lee 2016	Korea	Rtr	HIRA database	Diagnosis codes; Patients with IIP were defined as those with two or more medical visits or at least one hospitalisation with an IIP diagnostic code	51,026,868	IIP	41,770	NR; 50	81.86 per 100,000 persons
	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: - Public and private hospitals - Community Pulmonologists and General Practitioners	- A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. - ILD aetiology was determined when a diagnosis could be	1,194,601	IIP	145	56;50	12.14 per 100,000 persons
							NSIP	20		1.7 per 100,000 persons
							DIP	10		0.8 per 100,000 persons
							OP	9		0.8 per 100,000 persons

					attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. • ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. • For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.		RB-ILD	2		0.2 per 100,000 persons
SSc	Simeon-Aznar 2015	Spain	Rtr	RESCLE registry	Pulmonary involvement is defined by the presence of ILD or PH. ILD was established if any of the following criteria were identified: • restrictive pulmonary pattern with FVC below 80% of expected value on PFTs, • pulmonary interstitial pattern evidenced by chest radiograph HRCT scan, alveolitis confirmed by bronchoalveolar	879	Overall SSc	421	44.4; 85	47.8%

					lavage (defined as neutrophilia of $\geq 3\%$, eosinophilia of $\geq 2\%$, or lymphocytosis $\geq 15\%$).					
	Wangkaew 2016	Thailand	Prosp	Maharaj Nakorn Chiang Mai Hospital	All HRCTs images were reviewed by an experienced thoracic radiologist We used HRCT to categorize pattern of the lung parenchyma findings that represented ILD, which were unexplainable by any other causes.	113	Overall SSc	80	53.4; 58	69.1%
						89	dcSSc,	70	54.1; 51	78.7%
						24	LcSSc,	11	50.6; 88	45.8%
	Tomiyama 2016	Japan	Rtr	SSc who visited hospital department	ILD were evaluated by radiologists using high-resolution computed tomography (HRCT). Follow-up CT scans were performed when patients experienced a worsening of respiratory symptoms	139	Overall SSc	66	49.2; 82	47.5%
						116	lcSSc	48	NR	41.4%
						23	dcSSc	18	NR	78.3%
	Sanchez-cano 2018	Spain	CS	RESCLE is a nation-wide, multi-center registry	ILD: evidence consistent with pulmonary fibrosis present on chest radiograph or HRCT. Additionally, FVC value was considered to categorise as ILD severity, we took a 70% value as cutoff, between mild	1,374	Overall SSc	595	57; 86	43.3%

				and moderate–severe ILD.					
					1,374	lcSSc	316	60;87	53.1%
					1,374	dcSSc	240	52;85	40.3%
					1,374	ssSSc	39	58;80	6.6%
Janardana 2019	India	Rtr	Department of Clinical Immunology & Rheumatology, Christian Medical College	The diagnosis of ILD was made HRCT of the chest reported by radiologists at our institute in majority of the cases. Only for occasional patients, the diagnosis was based on clinical examination findings and CXR	327	Overall SSc	288	35.1, 8	88.1%
Morrisroe 2019	Australia	Prosp	SSc patients enrolled in the Australian Scleroderma Cohort Study (ASCS),	ILD was defined by characteristic changes on HRCT.	1,727	Overall SSc	489	46.6; 86	28.3%
Noviani 2019	Singapore	Prosp	Three tertiary rheumatology centres in Singapore	ILD was diagnosed by HRCT and the severity of ILD was determined by PFT. HRCT was performed in patients with clinical symptoms (dyspnea or cough) or signs suggestive of ILD (end-inspiratory bibasilar crepitations), chest radiographs suspicious of ILD or restrictive pattern on PFT with FVC or TLC <80%	490	Overall SSc	135	49; 83	27.6%

				predicted or DLCO <75% predicted. Significant ILD was defined by FVC <70% predicted and categorised under the ILD group					
Bauer 2020	France	Rtr	Patients treated at Nancy University Hospital	Pulmonary involvement was determined with functional measurement in terms of obstructive and restrictive defects and decreased diffusing capacity of the lungs for carbon monoxide DLCO (<80%) With a specific examination of the CT scan by an experienced radiologist	77	Overall SSc	24	NR	31.2%
					58	lcSSc	14		24.1%
					19	dcSSc	10		52.6%
Jaafar 2021	USA	Prosp	Multicentre national Prospective Registry of Early Systemic Sclerosis (PRESS) registry	Not clearly reported. Based on the results, ILD was diagnosed using HRCT and significant functional progression of ILD was defined as an absolute FVC decline of $\geq 10\%$ as compared to baseline FVC during the whole course of the study.	239	Overall SSc	128	51,70	53.6%
Li 2021	USA	Rtr	IBM MarketScan claims database	ICD codes	78,964,708	Overall , SSc	15,005	NR	19 per 100,000 persons

	Author, year	Country	Study design	Study population description	ILD assessment	Study population (N)	Type of SSc	No. of ILD cases	Mean age; Female (%)	Prevalence
	Pope 2021	Canada	Rtr	SSc Patients were identified using the CIHI, NACRS, DAD.	ICD 10 codes	NR	Overall, SSc For year: (2017/18)	257	Age groups: 18-29: 0.7 30-50: 1.51-64: 4.2 65+: 3.2; 3.6	2.3 per 100,000 persons (12% SSc cases with ILD)
	Fairley 2021	Australia	Prosp	Consecutive patients with SSc who prospectively enrolled in the Australian Scleroderma Cohort Study (ASCS)	- HRCT were used to diagnose ILD - These were usually performed in response to clinical examination findings (chest crepitations) or abnormal respiratory function testing	1,505	SSc only	393	NR	26.1%
	Carton 2021	Belgium	Rtr	Consecutive SSc patients, included in two Flemish cohort	- Each HRCT was centrally analysed by an experienced investigator or who was blinded. - Patients were classified into three subgroups ("no ILD", limited ILD or extensive ILD), according to the simplified flow diagram described by Goh et al	722	Overall, SSc	243	NR	34%
RA	Md Yosuf 2017	UK	Prosp	Patients with moderate to severe RA who were treated with RTX at participating site	PFT collected at 6–12 months pre-RTX, at the time of treatment with RTX, 6–12	700		56	64; 98	8 (% of cohort)

					months post-RTX and at the most recent follow-up. • HRCT scans were acquired (when clinically indicated) in patients with worsening dyspnoea and/or deterioration in lung function using a standardised method. The scans were scored independently by two radiologists, both blinded to lung function information and the sequence of scans. In order to account for missing PFT data of those with severe ILD who were unable to perform the test, data from the HRCT and survival status were incorporated into the overall lung response.					
Kim 2017	South Korea	Prosp	RA Patients from the KOREan Observational study Network for Arthritis cohort	• ILD was diagnosed based on the results of CXR or CT scans. • The CXRs and CT scans were read by expert radiologists at each hospital and the chest imaging records were collected from each hospital.	3,555		64	63.2; 70	1.8 (95% CI 1.4–2.3) (% of cohort)	

				The ILD group consisted of all patients whose chest image reports contained key words such as “pulmonary fibrosis”, “interstitial fibrosis”, “interstitial pneumonia”, “interstitial lung disease” and “ILD”, USA in addition to a descriptions of patterns of the lung disease such as “usual interstitial pneumonia” and “nonspecific interstitial pneumonia”.					
Zhang 2017	China	Rtr	Patients who were diagnosed with RA at the Capital Medical University, Beijing Chaoyang Hospital, during the study period	Not reported, ILD cases were reported as seen in the medical history	550		32	57.6; 63	5.8 (% of cohort)
Sakr 2018	Egypt	Rtr	RA patients attending the Rheumatology outpatient clinic, Kasr Alainy Hospital, Cairo University	ILD presence was extracted from patient’s medical history, details of ILD assessment are not reported.	3,219		27	NR	0.8 (% of cohort)
Duarte 2019	UK	Rtr	RA patients that attended the Rheumatology Department of University College London Hospital	- Types of lung involvement were based on HRCT, with the date of the exam being considered the date of lung disease diagnosis. - Subsequent HRCT results were also recorded.	1,129		45	63.2; 67	3.98 (% of cohort)

				PFTs including gas transfer (transfer factor for carbon monoxide), were collected for patients with lung involvement					
McFarlane 2019	USA	CS	RA inpatient discharges during study period, that took place at NY hospitals	A chest radiologist, pulmonary specialist, and a rheumatologist i.e. MDD reviewed the clinical course, management, pulmonary function tests, and chest CT images for each patient during the multidisciplinary reviews.	503		32	64;88	6.4 (% of cohort)
Kiely 2019	UK	Prosp	The study used data from ERAS (1986–2001) and ERAN (2002–2012), two multicentre early RA inception cohorts.	The diagnosis of ILD at each centre was according to standard practice, with confirmatory evidence from standard investigations including pulmonary function tests, chest radiographs and HRCT scans. ILD was deemed to be present if the terms pulmonary fibrosis or ILD were listed on the CRF or the death certificate using ICD codes.	1114		53	NR	4.8 (% of cohort)
Raimundo 2019	USA	Rtr	Truven Health MarketScan Commercial and Medicare Supplemental	Patients meeting ≥ 1 of the following criteria were identified as potential RA-ILD cases: ☐ ≥ 2 claims with an International Classification of Diseases-9-Clinical	over 150 million lives		648	66;67	3.2 (95% CI 3.0–3.4) (% of cohort)

					Modification (ICD-9-CM) diagnosis code for pulmonary disease (fibrosis: ICD-9-CM 515, 516.3, 516.31; rheumatic lung disease: ICD-9-CM 714.81) on					
Myositis ILS	Ishizuka 2016	Japan	Rtr	Enrolled patients with PM/DM who were referred to our department	The presence of ILD was evaluated by computed tomography (CT).	124		76	53.5; 65	61.3
	Ishizuka 2016					46 (PM)		25	NR	54.3
	Ishizuka 2016					78 (DM)		51	NR	65.4
	Gomez 2021	Argentina	CS	Argentine registry of IMM that includes patients older than 18 years who have been diagnosed with myositis ILD	HRCT or biopsy	360		99	NR	27.5
	Huang 2021	Taiwan	Rtr	Patients who underwent complete myositis autoantibody serology examination at university hospital	NR	97		46	56;80	47.4
SS	Kvarnstorm 2015	Sweden	Prosp	All incident patients referred to the dept. of Rheumatology at the Karolinska University Hos.	☐ ILD assessment not reported	199	2	55;94	p-SS	1
	Zhao 2015	China	CS	Patients from 16 Chinese medical centers nationwide	☐ Pulmonary involvement was indicated by persistent cough and/or dyspnea with chronic diffuse interstitial infiltrates on X-rays, altered patterns on	317	59	49;94	p-SS	18.6

				pulmonary function tests, and/or evidence of lung alveolitis or fibrosis in computed tomography (CT) scans.					
Manfredi 2017	Italy	Prosp	All consecutive patients diagnosed as pSS during the study period	• ILD assessment was based on HRCT and PFTs. • HRCT was categorized as definite, possible, or inconsistent with usual interstitial pneumonia (UIP) pattern according to the 2011 ATS/ERS/JRS/ALAT Guidelines for Diagnosis and Management of IPF	77	13	67;77	p-SS	16.9
Roca 2017	France	Rtr	Medical records from 2 participating institutional centers	☐ ILD was investigated by PFTs, and HRCT scan of the lungs; no patient underwent lung biopsy.	263	11	63;86	p-SS	4.2
Gao 2018	China	Rtr	Hospitalized patients admitted to Peking	☐ Diagnosis of pSS-ILD was made according to HRCT abnormalities and/or impaired PFTs, with or without pulmonary signs	853	165	61;92	p-SS	19.3
					488	126	61;92	s-SS	25.8
Kampolis 2018	Greece	CS	Department of Pathophysiology (Medical School, University of Athens)	☐ HRCT scan of the lungs was performed in all patients at the day of pulmonary function testing. Abnormal HRCT findings were classified into four major categories: reticular (septal thickening, honeycombing),	Overall, 384; Asymptomatic patients who	7	63;95	p-SS	17.9

					nodular (multiple large or small nodules), high attenuation					
Other Autoimmune condition	Reiseter 2018	Norway	Rtr	Nationwide MCTD cohort from depts of Rheumatology	CT scans and PFTs	119	Prevalent: 47 Incident: 2	91;38	40%	1.7%
	Olaosebikan 2021	Nigeria	CS	CTD patients managed at the Rheumatology Unit of Lagos State University Teaching Hospital	Clinical information including socio-demographics, signs and symptoms of ILD, pattern(s) of CTD, serology, pulmonary function, and radiology findings as well as treatment outcomes were retrieved from patients' records	318	Prevalent: 23 Incident: 8	39;90	7.2%	2.5%
Hypersensitivity pneumonia	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: - Public and private hospitals - Community Pulmonologists and General Practitioners	☐ ☐ ☐ ☐	MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on	1,194,601	HP; 28	Per 100,000 persons per year	2.3

						available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.				
Perez 2018	United States	Rtr	Truven Health Market Scan Commercial and Medicare		This research developed novel claims-based coding algorithms to identify HP, CHP, FHP cases, these algorithms validity and reliability were assessed with clinical data from National Jewish Health. The case validation procedure involved medical record review performed using the electronic medical records. An expert chest radiologist reviewed CT scans, blinded to subjects' clinical information and the primary radiologist's interpretation, to determine whether pulmonary fibrosis was present and whether the CT met criteria for consistent or possible HP or was inconsistent with HP. Last, a clinical-	158,270,350	HP; 7,498	Per 100,000 persons	1.67 (95% CI, 1.5 - 1.8) - 2.71 (95% CI, 2.6 -2.9)	

					radiological consensus diagnosis of HP or not HP was applied.				
						4,093	FLD; 157	% of the HP coho rt	3.8
						4,093	Bagasso sis; 11	% of the HP coho rt	0.3
						4,093	Bird fancier' s; 89	% of the HP coho rt	2.2
						4,093	Suberos is; 71	% of the HP coho rt	1.7
						4,093	Malk workers ; 21	% of the HP coho rt	0.5
						4,093	Mushro m workers ; 55	% of the HP coho rt	1.3
						4,093	Maple bark stripper s; 26	% of the HP coho rt	0.6
						4,093	Ventilat ion pneumo nitis; 201	% of the HP coho rt	4.9
						4,093	Other Spec; 482	% of the HP coho rt	11.8
						4,093	Unspec; 3,015	% of the HP coho rt	73.7
						4,093	CHP; 2,308	Per 100, 000 perso ns	0.91 (95% CI, 0.8– 1.0) - 1.7 (95% CI, 1.6– 1.8)

							4,093	FHP; 941	Per 100, 000 perso ns	0.41 (95% CI, 0.3– 0.5) - 0.8 (95% CI, 0.7– 0.9)
LAM / PLCH	Duchemann 2017	France	CS, Rtr	Sources of case identificatio n from separate sources of patients were used: Public and private hospital s Comm unity Pulmon ologists and General Practitioners	MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.	1,194,601			PLCH, 4 LAM 4	0.8 per 100,00 0 person s 0.3 per 100,00 0 person s
	Duchemann 2017									
Silicosis	Knight 2015	South Africa	Prosp	Baseline survey using a random sample of approximatel y 1,000 miners in each cluster (a subset of the total workforce)		Radiographs were read for silicosis according to the ILO Classification. All radiographs were read by an experienced	11,557	1,374	NR	11.9%

				was conducted, staggered over calendar time (2006–2009).		nursing health professional for silicosis. For the purpose of the silicosis study an experienced occupational medicine physician and trained reader read All radiographs reported as silicosis ILO grade $\geq 0/1$ by reader I,				
	Siribadda 2016	Srilanka	CS	Workers at a silica factory in Matale District		A screening questionnaire was used to assess symptoms of silicosis as well as tuberculosis which is a recognized complication of silicosis. The chest radiographs were reported according to the guidelines of ILO and the diagnosis of silicosis was made according to the following criteria as per the guidelines of the ILO Chest imaging (usually a conventional chest radiograph) that shows opacities consistent with silicosis.	250	14	30;13	5.6%
	Silanum 2017	Thailand	CS	Stone carving workers	The questionnaire (that was developed from reviewing silicosis relevant documents) was performed by interviewing		315	80	NR	25.4%

				prior to performing the chest x-ray. Reading and interpreting chest films were done by NIOSH B reader according to ILO film.					
Souza 2017	Brazil	CS	Mineworkers in semiprecious stone mining	The diagnosis of silicosis was made by a physician based on exposure history and alterations typical of the disease shown on the chest X-ray. The chest X-rays were interpreted independently by a radiologist and a pulmonologist according to the ILO of Pneumoconioses. Silicosis was diagnosed in those patients whose chest X-ray showed rounded opacities with profusion equal to or greater than 1/0.		348	129	48; NR	37%
Casey 2019	USA	Rtr	Medicare beneficiaries	Three case definitions were used to identify beneficiaries with silicosis: - Broad: Any claim that included ICD-9-CM code 502, pneumoconiosis due to other silica or silicates, listed in any position during 1999-2014 - Intermediate : Among broad		4,99,23,987	Broad: 19,696	65-84 years : 84; 19	39.5 per 100,000

				definition cases those who used i) at least one inpatient, skilled nursing, or home health agency claim or ii) at least two outpatient provider claims within 365 days of each other. Narrow: Among intermediate cases, a narrow definition applied included, that cases have a chest X-ray or CT scan 30 days before or 30 days after a silicosis claim.					
Hoy 2021	Australia	Rtr	V- SHARP is a population-based health assessment programme for current and former stonemasons	HRCT, lung function tests			Interm: 11, 469	65-84 years : 85; 11	23 per 100,000
PoinenRughoputh 2021	China	CS	Silica dust-exposed workers employed for at least 1 year during the study period in any of eight Chinese pottery factories	Diagnosis of silicosis were based on 1986 Chinese pneumoconiosis Roetgen diagnostic criteria which classified silicosis as stages 1, II or III similar to ILO of 1/1 or greater			Narrow: 10,026	65-84 years : 84; 16	20.1 per 100,000
Requena-Mullor 2021	Spain	CS	Workers in the stone sector handling artificial stone or silica compacts that were exposed to crystalline silica dust daily	□ NR		239	86		36%
Souza 2021	Brazil	CS	The targeted population was composed of 776 workers from	□ NR, self-reported		8,887	1,219	40;25	13.7%

				277 mines registered in A cooperative that organizes the mining operation in the region						
							311	64	39;0	20.6%
							258	27	53;0	10.5%
	Lee 2015	South Korea	CS	Residents from 5 villages that are within 2km from mines	Chest radiographs were evaluated by radiologists Reports were classified by ILO classification of pneumoconioses	35	27	Age (yr) 40-49: 0 50-59: 11% 60-69: 37% 70-79: 48% > 80: 4%; 48		77.1%
	Wickram-atillake 2019	Sri Lanka	CS	Participants from 9 provinces where asbestos factories, tsunami cleanup, construction industry were located	Chest radiographs were independently Reports were classified by ILO classification of pneumoconioses	250	16	NR		7%
Pneumococ niosis	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: • Public and private hospitals • Community Pulmonologists and General Practitioners	A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be	1,194,601	42			3.5 per 100,000

					established based on available investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.					
	Simsek 2021	Turkey	CS	Workers from 19 workplaces that were randomly selected that use hard metal tools in metal machining processes in organized industrial zone	CXRs were evaluated by two readers separately and independently according to the ILO international classification of radiographs of pneumoconiosis,	139	54			38.8% of study population
Other	Choi 2015	South Korea	Longitudinal	Indium-handling workers from seven factories	PFTs and HRCTs to confirm ILD	50	12		36;16	24% of indium workers
Pulmonary Sarcoid	Duchemann 2017	France	CS	Sources of case identification from separate sources of patients were used: - Public and private hospitals - Community Pulmonologists and General Practitioners	A MDD of clinicians, radiologist and pathologists discussed each case, after a systematic review of medical charts, HRCT, and SLB, when applicable. ILD aetiology was determined when a diagnosis could be attributed in consensus according to the current guidelines, including 2011 ATS/ERS/JRS/ALAT diagnostic criteria for IPF. ILD aetiology remained undetermined when a diagnosis could not be established based on available	1,194,601	361			30.2 per 100,000

					investigations after MDD. For unreviewable cases, information on the aetiological diagnosis was taken from the completed questionnaires or the ICDs-10 codes.				
	Coquart 2015	Guadeloupe	Rtr	2 hospitals	Cases were identified from medical records using search term 'sarcoidosis'	72	49		68.1 %; 14.4 per 100,000 inhabitants (95% CI, 10.8 to 17.8)

Table S4 List of studies and detailed characteristics of studies reporting the **prevalence** of IPF, Sorted by Publication Year

Author, year	Study design	Country	Study population	ILD assessment	Sample size	No. of IPF cases by ILD definition	Mean age; Female (%)	Reported Prevalence (per 100,000 persons)
Kaunisto 2015	CS	Finland	University Hospital	- ICD codes J84.1 and J84.9 - The pulmonary physician re-evaluated the patients 'data by reading through the patient charts. 70–80 % of patients 'diagnoses did not meet the clinical criteria of idiopathic pulmonary fibrosis. - Baseline CT scans were centrally re-evaluated by chest radiologist and radiology resident. - HRCT scans were evaluated and SLB was performed in 27 (22 %) patients.	NR	111	74;40	8.6
Lee 2016	Rtr	Korea	HIRA database	Patients with IPF were defined as those with a disease code of K-J84.18.	51,026,868	2010, 2011, 2012, 2013;16,325	NR	30.4,30.1,32.4,35
Harari 2016	Rtr	Italy	Healthcare administrative database of Lombardy	<u>GCD</u> : Defined as IPF cases all individuals with at least one hospitalization with diagnosis of IPF or at least one outpatient visits with diagnosis of IPF (ICD-9-CM code 516.3) during the study period	Approx. 10 million	5,441	Majority >65 years; 46	35.51 (35.02–36.00)
				<u>BCD</u> : Defined IPF cases those patients that satisfied the GCD and had no claims (inpatient or outpatient) with a diagnosis code for any other type of ILDs on or after the date of the last claim with IPF diagnosis		3,573	Majority >65 years; 43	22.39 (21.9–22.8)
				<u>NCD</u> : Defined IPF cases those patients that satisfied the BCD and had one or more claim with a procedure code for surgical lung biopsy, transbronchial lung biopsy or computed tomography of the thorax,		2,097	Majority >65 years; 43	12.55 (12.3–12.8)

				on or before the date of the last claim with a diagnosis code for IPF.				
Hopkins 2016	Rtr	Canada	National databases: DAD and NACRS	Broad definition: by excluding cases with an ICD-10 CA code for another ILD after the J84.1 code, like other studies	NR	14,259	NR	41.8
				Narrow definition further excluded cases that did not have chest CT, bronchus or lung biopsy, or bronchoscopy prior to establishing a J84.1 code.		6,390	NR	20
Raghu 2016	Rtr	USA	Optum Clinformatics	Primary cohort: IPF patients were first identified if they had at least one claim with ICD-9-CM diagnosis code 516.3 (idiopathic fibrosing alveolitis, which is used by providers to identify IPF) during the study period. Two subgroups of the primary cohort were further identified by applying additional criteria to form more restrictive cohorts of prevalent IPF patients:	9,375,020	2005, 2006, 2007, 2008, 2009,2010; 7,671	NR	15.1, 16.9, 17.5, 18.4, 18.2, 13.4
				Broad case definition excluded patients if they had a claim with the diagnosis code 515 after the last diagnosis code 516.3.		2005, 2006, 2007, 2008, 2009,2010; 2,670	NR	8.4, 9.7, 10.6, 10.8, 11.6, 11.3
				Narrow case definition further restricted the broad case subgroup patients by requiring a claim for a surgical lung biopsy, transbronchial lung biopsy, or computed tomography scan of the thorax prior to the last diagnosis code 516.3.		2005, 2006, 2007, 2008, 2009,2010; 1,685	NR	4.6, 5.6, 6.1, 6.2, 6.9, 6.7
Raimundo 2016	CS	USA	Optum Humedica	Patients with at least one IPF inpatient claim, or two IPF outpatient claims with [ICD-9-CM] code 516.3 anytime in that calendar year (with no other ILD claim)	4,138,796	2009, 2010, 2011; 1,191, 1292, 1136	70-71;49-52	28.8, 28.1, 19.8
Strongman 2018	Rtr	UK	CPRD	Broad IPF -clinical syndrome definition included the following three additional Read codes: H563100 (Diffuse pulmonary fibrosis), H563200 (Pulmonary fibrosis), and H563.11 (Hamman-Rich syndrome).	9,748,108	2010 to 2013;4,527	NR	19.94 (18.48, 21.47) to 38.82 (37.04, 40.66)
Author, year	Study design	Country	Study population	ILD assessment	Sample size	No. of IPF cases by ILD definition	Mean age; Female (%)	Reported Prevalence (per 100,000 persons)
Kaul 2021	Rtr	USA	Veterans Health Administration (VHA) electronic health record	Broad: Patients who had a diagnosis code for IPF (ICD-9-CM code 515, 516.3, 516.31 or ICD10-CM code J84.111, J84.112, J84.89, J84.9, J84.10, J84.17) recorded during study period • Patients were considered to have IPF if they did not have any other diagnosis code for an alternative interstitial lung disease	10.7 million	NR	NR	1,160
				Narrow: Patients who had a diagnosis code for IPF (ICD-9-CM code 515, 516.3, 516.31 or ICD10-CM code J84.111, J84.112, J84.89, J84.9, J84.10, J84.17) recorded during study period Patients were considered to have IPF if they did not have any other diagnosis code for an alternative interstitial lung disease Patients who had procedure code for a lung biopsy or a CT scan of the thorax before the last IPF diagnosis		NR	NR	725

9.2. Appendix 2. Supplementary material to Chapter 4

Approved Independent Scientific Advisory Committee (ISAC) protocol for chapter 4 (ISAC protocol 20_000068)



Medicines & Healthcare products
Regulatory Agency



1 General information
Protocol reference Id 20_000068
Study title An investigation of the reliability and granularity of the recording of interstitial lung diseases in primary and secondary care datasets
Research Area Disease Epidemiology

Does this protocol describe an observational study using purely CPRD data?

Yes

Does this protocol involve requesting any additional information from GPs, or contact with patients?

No

Role	Chief Investigator
Title	Reader
Full name	Jennifer Quint
Affiliation/organisation	Imperial College London
Email	j.quint@imperial.ac.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Corresponding Applicant
Title	Post-doctoral Research Associate
Full name	Ann Morgan
Affiliation/organisation	Imperial College London
Email	a.morgan15@imperial.ac.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Consultant
Full name	Peter George
Affiliation/organisation	Royal Brompton Hospital
Email	p.george@rbht.nhs.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Doctoral student
Full name	Rikisha Shah Gupta
Affiliation/organisation	Imperial College London
Email	r.shah20@ic.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

Sponsor

Imperial College London

Funding source for the study

Is the funding source for the study the same as Chief Investigator's affiliation?

Yes

Funding source for the study

Imperial College London

Institution conducting the research

Is the institution conducting the research the same as Chief Investigator's affiliation?

Yes

Institution conducting the research

Imperial College London

Method to access the data

Indicate the method that will be used to access the data

Institutional multi-study licence

Is the institution the same as Chief Investigator's affiliation?

Yes

Institution name

Imperial College London

Extraction by CPRD

Will the dataset be extracted by CPRD

No

Multiple data delivery

This study requires multiple data extractions over its lifespan

No

Data processors

Data processor is	Same as the chief investigator's affiliation
Processing	Yes
Accessing	Yes
Storing	Yes
Processing area	UK

Data processor is	Same as the chief investigator's affiliation
Processing	Yes
Accessing	Yes
Storing	Yes
Processing area	UK

Primary care data

CPRD GOLD
CPRD Aurum

Do you require data linkages

Yes

Patient level data

HES Accident and Emergency
HES Admitted Patient Care
HES Diagnostic Imaging Dataset
HES Outpatient
ONS Death Registration Data

NCRAS data**Covid 19 linkages****Area level data**

Do you require area level data?
No

Practice level (UK)

Patient level (England only)

Withheld concepts

Are withheld concepts required?

No

Linkage to a dataset not listed

Are you requesting a linkage to a dataset not listed?

No

Patient data privacy

Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?

No

Lay Summary

Interstitial lung disease (ILD) is a collective term for a large group of conditions that are typically characterized by scarring (fibrosis) of the lungs. Idiopathic pulmonary fibrosis or IPF is the most common type of ILD, with approximately 5,000 new cases recorded in the UK each year. There is evidence that the number of people with ILD, and IPF in particular, is increasing, but whether this is a real increase or simply a reflection of better diagnosis or recording is not clear. The prognosis for people who develop ILD such as IPF is extremely poor; average survival from diagnosis is just three years and until recently there has been little in the way of effective treatment.

Some people with other ILDs, for example, hypersensitivity pneumonitis and auto-immune related ILDs also suffer from progressive lung fibrosis, similar to that seen in IPF, and have equally poor survival prospects. However in general we know far less about the incidence, prevalence and patient outcomes for these rarer types of ILD. Research in this area is hampered by clinical challenges in diagnosing the different types of ILD, name changes and reclassifications over time, and uncertainties surrounding the quality of the recording of specific types of ILD in patients' health records.

Through this study we hope to develop a better understanding of the way in which the different types of interstitial lung disease are recorded in primary healthcare databases and thereby support further research into the full spectrum of ILD and its consequences for patients.

Technical Summary

The validity of research based on CPRD data relies on the quality of the information recorded by GPs. Validation studies have suggested that the validity of clinical diagnoses of specific conditions is generally quite good, as evidenced by strong measures of positive predictive value (PPV), sensitivity and specificity. This is especially true of chronic diseases that are managed in a primary care setting, but less so for acute conditions that require hospitalisation.

Previous work has established the reliability of a diagnosis of IPF in GP records (PPV=95%). Given that IPF and other ILDs are almost always diagnosed in secondary care, this finding is not unexpected and it is generally accepted that a patient with an ILD diagnosis in their primary care record will indeed have the disease. However, questions remain about the completeness of the capture of cases of ILD in primary care data, the granularity of the recording in terms of ILD subtype, and whether GP recording practices have altered over time.

This validation study is designed to complement efforts to better characterise the incidence, prevalence and mortality associated with a number of important ILD subtypes, data for which are lacking at the present time but which are needed in order to support further research into this group of lung diseases. In the first instance, we will conduct a descriptive analysis in order to inform our understanding of how ILD is coded in primary and secondary care. We will then carry out investigations to see how well we can identify cases of specific ILDs using simple diagnostic code lists, and/or whether we need to use combinations of codes to improve case capture. We will also explore the possibility of developing an algorithm to identify patients with progressive fibrosing ILD, this being a patient group of particular clinical significance.

Outcomes to be measured

Number (and frequency) of ILD code occurrences (Read codes in CPRD GOLD and AURUM, and ICD-10 codes in HES/ONS);

- Crude estimates for prevalence and incidence of all ILD, and specific ILD subtypes (based on case identification in CPRD, using clinical diagnostic codes only);
- alternative strategies or algorithms for identifying cases of ILD by subtype in CPRD data, should reliance on simple lists of diagnostic clinical codes prove to be suboptimal.

Objectives, specific aims & rationale

This study is intended to support a wider programme of research into interstitial lung disease, which has an increased focus on its rarer manifestations, i.e. the non-IPF ILDs. Relative to IPF, much less is known about the other forms or subtypes of ILD, although there is evidence to suggest that for some ILDs at least, UK incidence and prevalence is increasing. This particular piece of validation work has been designed to support population-based estimation of UK trends the prevalence and incidence of, and mortality due to, ILD, disaggregated by subtype, over the past 10 years for which a separate ISAC form will be submitted subsequently.

The primary objective of this study is to assess the quality and reliability of the recording of interstitial lung diseases, by subtype, in UK primary care data. More specifically, we wish to address the following questions:

- i) How “granular” is the recording of ILD in CPRD data, that is to say, do GPs use specific medcodes to code diagnoses of various subtypes of ILD [i.e. IPF, and the non-IPF ILDs such as the auto-immune ILDs (RA-ILD; SSc-ILD), sarcoidosis, exposure-related ILDs (e.g. silicosis, asbestosis) and hypersensitivity pneumonitis (HP)] or do they tend to rely on more generic, less specific codes to code cases of non-IPF ILD?
- ii) Have there been any discernible changes in ILD coding practices over time?
- iii) How reliable (i.e. complete) is the capture of incident/prevalent cases of ILD by subtype in UK primary care databases?
- iv) What is the best strategy for identifying cases of specific subtypes of ILD, i.e. can we reliably identify patients with certain specific types of ILD using appropriate diagnostic codes alone or can capture in CPRD data be improved by using combinations of medcodes (i.e. “indicator markers”) and/or linked hospital data?

Given the therapeutic significance, and depending on our findings, we will also consider the possibility of developing a code set or algorithm to identify patients whose ILD may be described as being of the “progressive fibrosing phenotype” (i.e. PF-ILD).

Study background

Interstitial lung disease (ILD) is a large heterogeneous group of over 200 diffuse parenchymal lung diseases, characterised by non-infectious infiltration of the interstitium and alveoli, and in most cases lung fibrosis, which over time results in reduced lung function, impaired gas exchange and progressive breathlessness. On average ILD affects around 80 in every 100,000 men and around 70 in every 100,000 women worldwide, although there are marked variations between countries (1).

Included under the ILD umbrella are diseases of unknown cause (e.g. idiopathic pulmonary fibrosis (IPF), sarcoidosis and idiopathic nonspecific interstitial pneumonias (INSIPs)), as well as conditions with a known aetiology, such as the so-called auto-immune ILDs (e.g. RA-ILD and SSc-ILD which are associated with rheumatoid arthritis and systemic sclerosis, respectively) and those linked with environmental exposures (e.g. hypersensitivity pneumonitis, asbestosis and silicosis). The two most common ILDs are idiopathic pulmonary fibrosis (IPF) and sarcoidosis. Of these, IPF has received by far the greatest research attention to date due to its progressive and aggressive nature and evidence of rising prevalence and incidence, both here in the UK and elsewhere (2-4).

Interest in ILD has increased in recent years, in part due to evidence of increasing disease burden, and in part due to the approval of two new antifibrotic medications (nintedanib and pirfenidone) which has revolutionised the care of people diagnosed with IPF. Both of these drugs have been shown in clinical trials to modulate the natural history of IPF and improve survival, for which hitherto conventional drug therapies have proved to be largely ineffective (5) or indeed harmful (6). The arrival of new and effective therapies for IPF has prompted speculation that these

therapies might also be effective in patients diagnosed with other forms of ILD, in particular in those whose disease has similar biological and clinical characteristics to IPF, that is to say, self-sustaining fibrosis, progressive decline in lung function and early mortality (Figure 1). The INBUILD study of nintedanib in patients with non-IPF progressive fibrosing ILD has demonstrated that antifibrotic therapy slows the rate of disease progression with an effect similar to that observed in patients with IPF (7). A recent subanalysis has demonstrated that this treatment effect is consistent across a range of ILDs and is not specific to any one subtype (8).

While the natural history and epidemiology of IPF is relatively well described, the incidence, prevalence and outcomes for patients with other, non-IPF ILDs is less well documented in the literature. The data that do exist tend to come from small-scale registry-type studies conducted in secondary care populations; very few, if any, studies have been carried out in large, population-based cohorts. The estimation of the disease burden of non-IPF ILDs using large population-based datasets is challenging for a myriad of reasons, not least of which is the difficulties associated with establishing accurate diagnoses of ILD by subtype based on GP coding practices in existing UK primary care datasets.

insert figure: see attached file:

Figure 1: Schematic representation of types of interstitial lung disease (ILD) that may be associated with a progressive fibrosing phenotype. Connective tissue disease (CTD)-ILDs include rheumatoid arthritis-associated ILD, systemic sclerosis-associated ILD, mixed CTD-associated ILD and other autoimmune ILDs. Other ILDs include exposure-related ILDs (asbestosis and silicosis), non-idiopathic pulmonary fibrosis (IPF) idiopathic interstitial pneumonias (desquamative interstitial pneumonia, etc.), and others. g/f PF: genetic and/or familial pulmonary fibrosis; HP: hypersensitivity pneumonitis; iNSIP: idiopathic nonspecific interstitial pneumonia; IPAF: interstitial pneumonia with autoimmune features; uILD: unclassifiable ILD.

Source: Cottin et al., 2018 (9)

External validation of the recording of ILD in UK primary care databases is, as far as we are aware, limited to one study published in 2000 by Hubbard et al. (9). Analysis of detailed GP records for a random subset of 36 patients indicated high positive predictive value (PPV) for diagnoses of cryptogenic fibrosing alveolitis (a former name for IPF) in CPRD data (PPV=95%). Based on the findings of this study, and given the fairly prescriptive criteria for a diagnosis of ILD which is almost always made in secondary care, it is not unreasonable to assume that the PPV for other ILDs is similarly high, and certainly acceptable for studies which rely on sets of purely diagnostic codes to identify cases of ILD, and IPF in particular.

Aside from an acceptable PPV, there are however other aspects of the coding of ILD diagnoses by GPs that warrant attention, and which this study is designed to explore. The first concerns the granularity of the recording of ILD diagnoses by subtype. Previous work conducted by this group which examined the diagnostic accuracy of the recording of stroke in CPRD found that GPs tended to code strokes using more “generic” and less specific codes (e.g. cerebrovascular accident or CVA), even when more detailed diagnostic information was available for that same stroke event in other data sources, namely HES data. The concern therefore is that similar coding practices are being adopted for ILD, which will limit our ability to identify cases of certain subtypes of ILD, especially the rarer non-IPF ILDs.

A second concern is the completeness of capture of cases of ILD by primary care records. Studies which have examined the completeness of the primary care recording of acute conditions and those that are predominantly managed in secondary care have demonstrated that a reliance on CPRD data alone is likely to result in an underestimation of cases, especially incident cases (10). For example, a study by Pascoe et al. (2008) which looked at the recording of cancer cases found that 19% of cancer-registry patients had no code for cancer in their GP records (11). However, by combining codes for malignancy, abnormal test results and prescriptions, the same authors were able to identify 93% of the cancer patients in their sample. A more recent study confirmed that reliance on diagnostic codes alone to identify cancer cases in CPRD without linkage to alternative data sources (i.e. HES and registry data) risked missing a substantial

proportion of cases; interestingly in this particular study, completeness was high for breast and prostate cancers (90% and 78%, respectively) which are managed in primary care, but much lower for cancers such as lung, pancreatic and kidney which are not typically managed by GPs (12).

Other studies have cast doubt on the timeliness of case identification, and have highlighted the fact that the date of the first appearance of a diagnostic code (the standard strategy for defining an incident case) often bears little relationship to the actual date of diagnosis (13). Again this is likely to be a more significant concern for conditions like ILD that are diagnosed and managed in secondary care: often the earliest record of the confirmed diagnosis appears in free text or in the letter from the specialist attached to a consultation record but that consultation is coded using an administrative code (e.g. letter received). Another factor which is likely to have a bearing on the estimation and interpretation of trends in prevalence, incidence and mortality associated with

various types of ILD, is probable changes in coding practices over time. In the context of ILD coding this is likely to reflect name changes and disease reclassifications, as well as changes in GP coding preferences over the last decade or so.

Study type

This is a cross-referencing study to explore the quality, in terms of completeness and diagnostic accuracy, of the recording of ILD – by subtype – in primary care databases.

Study design

Cohort study.

Feasibility counts

A previous study designed to examine the causes of mortality in people with chronic respiratory diseases identified over 10,700 cases of ILD among the HES/ONS-linked CPRD GOLD population who were alive and contributing data between 1 January 2004 and 31 December 2015 (14).

It is anticipated that additional use of the CPRD AURUM dataset will more than double the number of patients that are eligible for inclusion in this study.

Sample size considerations

N/A

Planned use of linked data and benefit to patients in England and Wales

Access to HES and ONS data is requested in order to capture patients who have ILD but who may not have been identified as such in CPRD. HES APC data may provide more granular diagnostic information on ILD subtype that is not accurately reflected in primary care data.

HES OP data will allow us to confirm whether a patient attended as an out-patient for a respiratory reason. Frequency of A&E attendance is considered to be an important predictor of the progressive fibrotic ILD phenotype. HES DID data will allow us to confirm that a patient has attended for a chest/lung CT scan, which is a key step in the diagnosis of most forms of ILD. Frequency of imaging may be indicative of rapidly progressing disease.

ONS data may also offer an opportunity to identify patients with ILD and whose diagnosis is not reflected in their primary care record. It is considered unlikely that ONS data will offer much in the way of additional information pertaining to ILD subtype.

We anticipate our data requests to be as follows:

- i) extraction of CPRD data (GOLD and AURUM) for all patients with an ILD-related medcode (medcodeid), and HES/ONS data for this patient population (HES APC, A&E and OP) for the subset of patients with linked records;
- ii) HES (APC) and ONS data for people with a ILD ICD-10 code; and
- iii) CPRD (GOLD and AURUM) data for those people with linked records and with ILD ICD-10 codes in HES/ONS (some of whom may not have been captured in the initial CPRD cut, i.e. (i) above.

The primary aim of this study is to improve the identification of cases of ILD in primary care data to support further epidemiological research into this group of lung diseases, which in turn will feed through to earlier diagnosis and better treatment of patients with ILD. This will be especially true for patients with non-IFP ILD whose disease is progressive but unresponsive to standard pharmacological therapies but who may benefit from treatment with newly licensed antifibrotic medicines.

Definition of the study population

The intention will be to use both GOLD and AURUM data sets, as the starting point for the identification of a cohort of patients with an ILD. The rationale for including the latter is to capture more patients with more recent data.

Our study population will be drawn from subjects who have either an ILD medcode in their primary care record (see Appendix 1), or an ILD ICD-10 code in their hospitalization record (either as the primary reason for hospitalisation or as a comorbidity) and/or an ILD ICD-10 code in their mortality record.

The minimum eligibility criteria for inclusion in the base study population will be specified as follows:

- Acceptable patient status in CPRD
- Male or female gender
- Aged over 18 years at study entry
- At least 1 year of “useable” follow up during the study period (nominally 1 January 2008 –31 December 2019).
- At least one code denoting an ILD diagnosis in either CPRD, HES or ONS.

Subjects will be followed from 1 January 2008, date of current registration, date of their 18th birthday, or practice up to standard date, whichever is later. Follow up will end at death, transfer out date, last collection date from the practice, or 31 December 2019, whichever is earliest.

Selection of comparison groups/controls

NA

Exposures, outcomes and covariates

Exposure

Our main exposure of interest is ILD, categorised by subtype (where subtyping is based on the aetiology of disease), provisionally as follows:

- Idiopathic interstitial pneumonias (i.e. pulmonary fibrosis with no known cause and which includes IPF, iNSIP);
- Auto-immune ILDs (PF linked to other diseases such as RA, lupus and systemic sclerosis; also called connective tissue disease (CTD- ILDs) as well as vasculitis);
- Drug-induced ILD (PF brought on by drugs typically chemotherapy drugs such as amiodarone or nitrofurantoin);
- Hypersensitivity pneumonitis (allergic response to inhaled antigens such as birds, mould, metal worker's fluid etc.);
- Pneumoconiosis (lung damage due to exposure to mineral dusts including asbestosis, silicosis)
- Sarcoidosis (chronic inflammatory lung disorder of unknown cause) -both pulmonary and extra-pulmonary
- Smoking-related ILD (including RB-ILD, DIP and LCH); and
- Other ILD (including unclassified ILD, rare lung conditions such as Lymphangioleiomyomatosis, Birt-Hogg-Dube, lymphocytic interstitial pneumonia and amyloidosis).

Demographics

Age, sex, smoking status, geographical region.

Other variables/covariates of potential interest

From CPRD

- Signs and symptoms (e.g. cough, breathlessness, finger clubbing)
- Prescriptions (e.g. pirfenidone, nintedanib, prednisolone, mycophenolate mofetil, azathioprine, hydroxychloroquine)
- Evidence of a diagnosis of various auto-immune conditions (so RA, SSc, myositis (dermatomyositis, polymyositis) vasculitis, lupus, mixed connective tissue disease, and any other rheumatological comorbidity)
- Evidence of prescriptions for medications that are linked to drug-induced ILD (i.e. often chemotherapy drugs, such as amiodarone, nitrofurantoin)
- Evidence of radiation exposure (i.e. cancer radiotherapy)
- Evidence of referrals to secondary care (specifically for respiratory investigations)
- Evidence of further diagnostic tests performed (CT scan, lung biopsy, blood test – but not everything is necessarily ILD specific)
- Spirometry (FVC measurements, including % predicted)
- Evidence of oxygen prescription (may be poorly recorded in primary care)
- Evidence of smoking or smoking cessation
- Evidence of rheumatological condition or follow up

- Any evidence of occupation/occupational exposures and or civil compensation (if available)

From HES/ONS:

- HES OP evidence of a consultation with a respiratory consultant (inc frequency of consultations)

- from HES APC/A&E evidence of hospitalisation for/with ILD
- from ONS evidence of death due to an ILD (underlying cause) or with ILD involvement (contributory factor).

Data/statistical analysis

Descriptive analysis

Our initial analyses will be largely descriptive in nature. The purpose of the descriptive analyses will be to establish the frequency of the use of diagnostic ILD medical codes in primary care data (medcodes in CPRD GOLD data (and SNOMEDct codes in CPRD AURUM data). The frequency of use of diagnostic codes for specific subtypes of ILD will also be catalogued (IPF, other idiopathic interstitial pneumonias; hypersensitivity pneumonitis; auto-immune ILD; drug-induced ILD/radiation-induced ILD; smoking-induced ILD; pneumoconiosis; sarcoidosis; other/unclassified ILD) . A similar exercise will be undertaken for HES and ONS data.

We will also examine whether the use of certain diagnostic codes have increased or decreased over time, and consider what changes in GP coding practices may have contributed to any of the trends observed. It is noted that since 2000 there have been changes in ILD nomenclature and subtype classification that may complicate this aspect of the analysis (15) and for this reason we will rely on the assistance of an experienced ILD physician in order to interpret our findings.

Analysis 1: External validation: comparison of CPRD incidence/prevalence rates with published estimates

Based on the results of our descriptive analysis, we will draw up sets of medical codes which we consider represent an optimal set of diagnostic codes for identifying patients with ILD, both any ILD, and by individual subtype. Based on the patient cohorts identified by these derived code lists, we will compute crude estimates for prevalence and incidence of all ILD and each main ILD aetiological subtype, by calendar year for the 12-year period of our study (2008–2019), disaggregated by gender, age group, and geographical region if possible.

We will compare our computed prevalence and incidence rates with those reported in the literature (i.e. estimates based on non-CPRD data sources); this will in theory provide a means of assessing the completeness of the capture of cases of various ILD subtypes in CPRD based on a simple set of diagnostic codes. It is anticipated our CPRD-estimated rates for all ILD might not be too dissimilar to those reported in the literature, but those for some ILD subtypes may well be too low. If this is the case, we might infer that there is a tendency among GPs to use more generic, non-specific codes to code for ILD, and that reliance on simple diagnostic codes may be a suboptimal strategy for identifying patients with specific subtypes of ILD.

Analysis 2: Internal validation of the coding of ILD by subtype in CPRD primary care datasets

Our aim with this second piece of data analysis is to internally validate a CPRD diagnosis of a given ILD subtype using one or both of the following strategies:

- (i) looking for confirmatory evidence of disease status in CPRD–ONS/HES linked records; and/or
- (ii) using a broader set of indicative related codes or “indicator markers” (i.e. symptom, referral, test or treatment codes) to confirm the presence of a given ILD subtype.

To a certain extent, the course of this work will depend on the outcome of the above-mentioned analyses (the Descriptive analysis and Analysis 1), and therefore some of the elements of the data analysis we set below should be regarded as provisional.

For the purposes of this analysis, we will create a smaller “internal validation” study cohort, which will be drawn from our initial (i.e. base) CPRD ILD study population (see Analysis 1). This study cohort will be limited to individuals registered with linked GP practices who have at least one code indicative of an ILD diagnosis in either their primary care record or their HES APC record, and/or have died with or of an ILD. Additional inclusion/exclusion criteria will be applied to ensure overlapping periods of useable follow-up; this subset of individuals will form the basis of our internal validation study cohort (provisional cases).

Depending on their ILD coding, study subjects will be assigned to one of nine ILD subgroups as follows: IPF; other idiopathic interstitial pneumonias; hypersensitivity pneumonitis; auto-immune ILD; drug-induced ILD/ radiation-induced ILD; pneumoconiosis; sarcoidosis; other/unclassified ILD; ILD not otherwise specified (for those patients whose ILD diagnosis is not subtyped). For patients with multiple ILD codes in their CPRD and HES records, ILD subtype classification will be on the basis of the most specific code recorded. For instance, if a patient's record contains both non-specific ILD codes as well as codes indicating a diagnosis of say RA-ILD, it will be assumed that that patient has an auto-immune ILD. It should be noted that at this stage the proposed categorisation of ILD is provisional, and the remaining analyses largely exploratory. Depending on the granularity of the recording, we accept the possibility that categorising of ILD much beyond IPF-ILD and non-IPF-ILD may not be a viable proposition.

First-level validation: using HES/ONS as a "GOLD standard": For patients who have ILD codes in both CPRD and HES/ONS, we will compare the diagnostic accuracy of the coding in CPRD against that in HES/ONS. Based on previous work on the recording of stroke (ISAC protocol number: 16_095R), we anticipate that the coding of ILD in HES (but not necessarily in ONS) will be more specific, that is to say, we expect to see codes for various subtypes of ILD used more frequently than in CPRD.

As a further step, and depending on our findings, we may investigate the "mismatched" cases further in order to better understand the reasons for the discrepancies in recording between the three datasets, in terms of both completeness of CPRD case capture and diagnostic accuracy. It is hoped that this strategy might identify additional CPRD codes, which when combined with diagnostic codes, may improve the capture and identification of cases of certain types of ILD, and thus feed into the second-level internal validation (see below).

Second-level internal validation: using "indicator" or "marker" codes". For each patient in our validation cohort, or a subset depending on numbers, we will create a descriptive "medical profile", based on the appearance of additional coded information in patient records (i.e. symptoms, prescriptions, test results, referrals), focusing on those characteristics that might support/suggest a clinical diagnosis of a certain type of ILD (see Section N: Covariates). This exercise may extend to secondary care, should we find that we cannot find reliable evidence of referrals to secondary care for a respiratory reason from primary care data alone.

We will then compare the profiles of any generically-coded ILD cases against those with a specific subtype coding, and thereby attempt to derive a series of "indicative codes" (i.e. sets of combined codes) which could potentially represent an improvement on case capture relative to a single diagnostic codes. It is anticipated that these code sets will comprise a maximum of four or five additional or "marker" codes (identified on the basis of clinically plausibility and frequency of use).

Given the cost implications of alternative strategies (i.e. questionnaires to GPs) we will assess the validity our indicator code list sets for ILD subtypes, by expert adjudication. This is to say, we will ask two experienced ILD clinicians to independently assign an ILD subtype diagnosis to a patient based on their full clinical profile and compare the clinician assignment with that based on our indicator codes.

Finally, in light of the recent interest in patients whose ILD may be characterized as being of the progressive fibrosing phenotype, we will explore possible options for identifying patients in CPRD data sets who fall into this category. Patients with PF-ILD tend to have a more rapid loss of FVC, more frequent consultations and early mortality. It may for example be possible to identify patients with non-IPF PF-ILD and who may therefore benefit from the new antifibrotic medications by conducting some form of latent trajectory-type analysis (i.e. group-based trajectory modelling or longitudinal cluster analysis) utilising the medical profiles we have created and the above listed outcomes that characterise the PF-ILD phenotype (16). Again, any such analysis would likely be largely exploratory in nature, and contingent on our findings.

Based on the results of our combined analyses, we will derive what we consider to be the “best”

code list for ILD case ascertainment – by subtype – in primary care databases, with a focus on SNOMED codes. This code list will be utilized to establish the epidemiology of ILD in the UK, work which will be the subject of a separate ISAC.

Plan for addressing confounding

NA

Plans for addressing missing data

We anticipate that data for some variables of interest, for example, spirometry, will be incompletely recorded in CPRD. Where we find data are missing, we will investigate the possible reasons and mechanisms underpinning the missingness. Given the aims and objectives of the proposed study, we consider using multiple imputation methods to be inappropriate, and we will conduct our analyses on a complete-case basis. We also acknowledge that ideally any derived code sets and algorithms for identifying cases of ILD subtypes in primary care should not rely on variables for which there is a high degree of missingness (> 40%).

Patient or user group involvement

Findings from this study will be shared with relevant patient groups.

Plans for disseminating & communicating

The results of this study will be disseminated in peer-reviewed journals and at scientific conferences as soon as possible following the completion of the analyses.

Conflict of interest statement

No conflict of interest is indicated

Limitations of study design

Our main limitation is that, in the absence of a linked ILD registry-type dataset, we lack a good gold standard against which to compare the recording of ILDs by subtype in CPRD.

Secondly, at present we are uncertain as to quality of coding of specific ILDs in CPRD. We may well find that ILD diagnostic coding is not sufficiently granular to permit categorisation of ILD by subtype beyond a relatively simple IPF-ILD and non-IPF ILD. Furthermore, given that GP coding practices vary and CPRD data may not necessarily reflect the breadth of this variation, any inferences drawn about the nature of ILD coding by GPs on the basis of an analysis of linked practices alone may have limited generalisability.

References

1. Smith LJ QJ, Brown B. Interstitial lung disease. *Respiratory medicine*. London: JP Medical Publishers; 2015. p. 167-84.
2. Fernández Pérez ER, Daniels CE, Schroeder DR, St Sauver J, Hartman TE, Bartholmai BJ, et al. Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: a population-based study. *Chest*. 2010;137(1):129-37.
3. Navaratnam V, Fleming KM, West J, Smith CJ, Jenkins RG, Fogarty A, et al. The rising incidence of idiopathic pulmonary fibrosis in the U.K. *Thorax*. 2011;66(6):462-7.
4. Strongman H, Kausar I, Maher TM. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. *Advances in therapy*. 2018;35(5):724-36.
5. Lancaster L, Crestani B, Hernandez P, Inoue Y, Wachtlin D, Loaiza L, et al. Safety and survival data in patients with idiopathic pulmonary fibrosis treated with nintedanib: pooled data from six clinical trials. *BMJ open respiratory research*. 2019;6(1):e000397.
6. Idiopathic Pulmonary Fibrosis Clinical Research Network RG, Anstrom KJ, King TE Jr, Lasky JA, Martinez FJ. Prednisone, Azathioprine, and N-Acetylcysteine for Pulmonary Fibrosis. *N Engl J Med*. 2012;366(21):1968-77.
7. Flaherty KR, Wells AU, Cottin V, Devaraj A, Walsh SLF, Inoue Y, et al. Nintedanib in Progressive Fibrosing Interstitial Lung Diseases. *N Engl J Med*. 2019;381(18):1718-27.
8. Wells AU, Flaherty KR, Brown KK, Inoue Y, Devaraj A, Richeldi L, et al. Nintedanib in patients with progressive fibrosing interstitial lung diseases—subgroup analyses by interstitial lung disease diagnosis in the INBUILD trial: a randomised, double-blind, placebo-controlled, parallel-group trial. *The Lancet Respiratory Medicine*. 2020;8(5):453-60.
9. HUBBARD R, VENN A, LEWIS S, BRITTON J. Lung Cancer and Cryptogenic Fibrosing Alveolitis. *American Journal of Respiratory and Critical Care Medicine*. 2000;161(1):5-8.
10. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ : British Medical Journal*. 2013;346.
11. Pascoe SW, Neal RD, Heywood PL, Allgar VL, Miles JN, Stefoski-Mikeljevic J. Identifying patients with a cancer diagnosis using general practice medical records and Cancer Registry data. *Family Practice*. 2008;25(4):215-20.
12. Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. *Epidemiology (Cambridge, Mass)*. 2018;29(2):308-13.
13. Nicholson A, Ford E, Davies KA, Smith HE, Rait G, Tate AR, et al. Optimising Use of Electronic Health Records to Describe the Presentation of Rheumatoid Arthritis in Primary Care: A Strategy for Developing Code Lists. *PLoS ONE*. 2013;8(2):e54878.
14. Gayle AV, Axson EL, Bloom CI, Navaratnam V, Quint JK. Changing causes of death for patients with chronic respiratory disease in England, 2005-2015. *Thorax*. 2019;74(5):483-91.
15. Wells AU, Brown KK, Flaherty KR, Kolb M, Thannickal Victor J. What's in a name? That which we call IPF, by any other name would act the same. *Eur Respir J*. 2018;51(5):1800692.
16. Rao A, Jones A, Bottle A, Darzi A, Aylin P. A retrospective cohort study of high-impact users among patients with cerebrovascular conditions. *BMJ Open*. 2017;7(6):e014618.

Appendices

 appendix-1a.pdf

 appendix-1b.pdf

 figure-1ild-isac.pdf

Grant ID

Codelists for Comorbidities

1. Codelist for Chronic Pulmonary Obstructive Disease

aurum_medcode	readcode	Description
264537018	663K.00	Airways obstructn irreversible
1484924013	66YB.00	Chronic obstructive pulmonary disease monitoring
1484971019	66YD.00	Chronic obstructive pulmonary disease monitoring due
1488423019	66YL.00	Chronic obstructive pulmonary disease follow-up
8.45451E+14	66YL.11	COPD follow-up
1488424013	66YM.00	Chronic obstructive pulmonary disease annual review
1780380013	66YS.00	Chronic obstructive pulmonary disease monitoring by nurse
1780381012	66YT.00	Chronic obstructive pulmonary disease monitoring by doctor
2.99001E+14	66Yg.00	Chronic obstructive pulmonary disease disturbs sleep
2.99031E+14	66Yh.00	Chronic obstructive pulmonary disease does not disturb sleep
4.05121E+14	679V.00	Health education - chronic obstructive pulmonary disease
2.19926E+15	8CMW500	Chronic obstructive pulmonary disease care pathway
9.66841E+14	8CR1.00	Chronic obstructive pulmonary disease clini management plan
1.81387E+15	9NgP.00	On chronic obstructive pulmonary disease supprtv cre pathway
1.81388E+15	9NgP.11	On COPD (chr obstruc pulmonary disease) supportv cre pathway
2.21955E+15	9Nk7000	Seen in chronic obstructive pulmonary disease clinic
9.67531E+14	9Oi..00	Chronic obstructive pulmonary disease monitoring admin
1.89761E+14	9Oi0.00	Chronic obstructive pulmonary disease monitoring 1st letter
1.98471E+14	9Oi1.00	Chronic obstructive pulmonary disease monitoring 2nd letter
1.99481E+14	9Oi2.00	Chronic obstructive pulmonary disease monitoring 3rd letter
9.67571E+14	9Oi3.00	Chronic obstructive pulmonary disease monitoring verb invite
9.67581E+14	9Oi4.00	Chronic obstructive pulmonary disease monitor phone invite
475431013	H3...00	Chronic obstructive pulmonary disease
5.55471E+14	H3...11	Chronic obstructive airways disease
285100019	H312100	Emphysematous bronchitis
301455018	H312z00	Obstructive chronic bronchitis NOS
6.40491E+14	H32..00	Emphysema
301460019	H320.00	Chronic bullous emphysema
27096010	H320200	Giant bullous emphysema
301468014	H320z00	Chronic bullous emphysema NOS
9337016	H321.00	Panlobular emphysema
113497011	H322.00	Centrilobular emphysema
396110016	H32yz00	Other emphysema NOS
301477019	H32z.00	Emphysema NOS
457168017	H36..00	Mild chronic obstructive pulmonary disease
457169013	H37..00	Moderate chronic obstructive pulmonary disease
457171013	H38..00	Severe chronic obstructive pulmonary disease
5.16801E+14	H39..00	Very severe chronic obstructive pulmonary disease
216596014	H3A..00	End stage chronic obstructive airways disease
301539010	H3y..00	Other specified chronic obstructive airways disease
1222334016	H3y..11	Other specified chronic obstructive pulmonary disease
301545019	H3z..00	Chronic obstructive airways disease NOS
1222335015	H3z..11	Chronic obstructive pulmonary disease NOS

301836011	Hyu3100	[X]Other specified chronic obstructive pulmonary disease
-----------	---------	--

2. Codelist for Asthma

medcode	readcode	snomedctconceptid	Description
5867	173A.00	31387002	Exercise induced asthma
22752	173c.00	57607007	Occupational asthma
73522	173d.00	57607007	Work aggravated asthma
41017	1780	4.08E+08	Aspirin induced asthma
102341	1781	3.41E+14	Asthma trigger - pollen
103955	1782	3.41E+14	Asthma trigger - tobacco smoke
102952	1783	3.41E+14	Asthma trigger - warm air
103952	1784	2.01E+14	Asthma trigger - emotion
103945	1785	2.01E+14	Asthma trigger - damp
103321	1786	2.01E+14	Asthma trigger - animals
102301	1787	2.01E+14	Asthma trigger - seasonal
103813	1788	2.01E+14	Asthma trigger - cold air
102449	1789	2.01E+14	Asthma trigger - respiratory infection
103944	178A.00	3.41E+14	Asthma trigger - airborne dust
102871	178B.00	3.41E+14	Asthma trigger - exercise
11370	102..00	4.01E+08	Asthma confirmed
93736	388t.00	3.02E+14	Royal College of Physicians asthma assessment
103631	388t000	4.48E+08	Royal College Physician asthma assessment 3 question score
98185	38DL.00	4.43E+08	Asthma control test
100397	38DT.00	7.63E+08	Asthma control questionnaire
102209	38DV.00	7.64E+08	Mini asthma quality of life questionnaire
105674	661M100	8.12E+14	Asthma self-management plan agreed
105420	661N100	8.11E+14	Asthma self-management plan review
81	663..11	2.76E+08	Asthma monitoring
7416	663N.00	1.71E+08	Asthma disturbing sleep
30815	663N000	1.71E+08	Asthma causing night waking
38146	663N100	1.71E+08	Asthma disturbs sleep weekly
13175	663N200	1.71E+08	Asthma disturbs sleep frequently
38143	663O000	1.71E+08	Asthma never disturbs sleep
7191	663P.00	1.71E+08	Asthma limiting activities
102713	663P000	7.72E+14	Asthma limits activities 1 to 2 times per month
102888	663P100	7.72E+14	Asthma limits activities 1 to 2 times per week
103998	663P200	7.72E+14	Asthma limits activities most days
7378	663U.00	4.06E+08	Asthma management plan given
13064	663V.00	1.71E+08	Asthma severity
3458	663V000	3.7E+08	Occasional asthma
3018	663V100	3.7E+08	Mild asthma
13065	663V200	3.7E+08	Moderate asthma
3366	663V300	3.7E+08	Severe asthma
24479	663d.00	7.08E+08	Emergency asthma admission since last appointment
25181	663e.00	1.71E+08	Asthma restricts exercise
26861	6.63E+02	1.71E+08	Asthma sometimes restricts exercise
26506	#####	1.71E+08	Asthma severely restricts exercise
26504	663f.00	1.71E+08	Asthma never restricts exercise
10487	663j.00	3.12E+08	Asthma - currently active
47337	663m.00	7.08E+08	Asthma accident and emergency attendance since last visit
19520	663n.00	3.7E+08	Asthma treatment compliance satisfactory

19519	663p.00	3.7E+08	Asthma treatment compliance unsatisfactory
42824	663q.00	3.74E+08	Asthma daytime symptoms
39570	663r.00	3.7E+08	Asthma causes night symptoms 1 to 2 times per month
26501	663s.00	3.7E+08	Asthma never causes daytime symptoms
31225	663t.00	3.7E+08	Asthma causes daytime symptoms 1 to 2 times per month
24884	663u.00	3.7E+08	Asthma causes daytime symptoms 1 to 2 times per week
26503	663v.00	3.7E+08	Asthma causes daytime symptoms most days
38144	663w.00	3.7E+08	Asthma limits walking up hills or stairs
38145	663x.00	3.7E+08	Asthma limits walking on the flat
9018	663y.00	3.67E+08	Number of asthma exacerbations in past year
9552	66Y5.00	3.91E+08	Change in asthma management plan
9663	66Y9.00	3.91E+08	Step up change in asthma management plan
18223	66YA.00	3.91E+08	Step down change in asthma management plan
41020	66YC.00	3.91E+08	Absent from work or school due to asthma
10043	66YJ.00	3.95E+08	Asthma annual review
13176	66YK.00	3.95E+08	Asthma follow-up
31167	66YP.00	3.95E+08	Asthma night-time symptoms
19167	66YQ.00	4.01E+08	Asthma monitoring by nurse
30458	66YR.00	4.01E+08	Asthma monitoring by doctor
102170	66Yp.00	7.54E+14	Asthma review using Roy Colleg of Physicians three questions
102400	66Yq.00	7.72E+14	Asthma causes night time symptoms 1 to 2 times per week
102395	66Yr.00	7.72E+14	Asthma causes symptoms most nights
103612	66Ys.00	4.73E+08	Asthma never causes night symptoms
107167	66Yu.00	8.11E+14	Number days absent from school due to asthma in past 6 month
100107	679J000	5.27E+14	Health education - asthma self management
100740	679J100	5.27E+14	Health education - structured asthma discussion
103318	679J200	5.27E+14	Health education - structured patient focused asthma discuss
24506	8791	1.83E+08	Further asthma - drug prevent.
16785	8794	1.83E+08	Asthma control step 1
16667	8795	1.83E+08	Asthma control step 2
18224	8796	1.83E+08	Asthma control step 3
20886	8797	1.83E+08	Asthma control step 4
20860	8798	1.83E+08	Asthma control step 5
10274	8B3j.00	3.95E+08	Asthma medication review
99793	8CMA000	5.27E+14	Patient has a written asthma personal action plan
25791	8CR0.00	7.36E+08	Asthma clinical management plan
7058	8H2P.00	1.83E+08	Emergency admission, asthma
100509	9NNX.00	6.99E+08	Under care of asthma specialist nurse
46529	9OJ1.00	1.86E+08	Attends asthma monitoring
19539	9OJA.00	2.7E+08	Asthma monitoring check done
8355	9OJA.11	2.7E+08	Asthma monitored
5798	H312000	1.96E+08	Chronic asthmatic bronchitis
78	H33..00	1.96E+08	Asthma
1555	H33..11	1.96E+08	Bronchial asthma
7146	H330.00	3.89E+08	Extrinsic (atopic) asthma
2290	H330.11	3.89E+08	Allergic asthma
15248	H330.13	2.34E+08	Hay fever with asthma
7731	H330.14	2.34E+08	Pollen asthma
14777	H330000	63088003	Extrinsic asthma without status asthmaticus
5627	H330011	2.34E+08	Hay fever with asthma
27926	H330100	1.09E+15	Extrinsic asthma with status asthmaticus

6707	H330111	7.08E+08	Extrinsic asthma with asthma attack
45782	H330z00	4.25E+08	Extrinsic asthma NOS
5267	H331.00	2.66E+08	Intrinsic asthma
3665	H331.11	2.34E+08	Late onset asthma
29325	H331000	12428000	Intrinsic asthma without status asthmaticus
58196	H331100	1.09E+15	Intrinsic asthma with status asthmaticus
18323	H331111	7.08E+08	Intrinsic asthma with asthma attack
45073	H331z00	2.66E+08	Intrinsic asthma NOS
185	H333.00	7.08E+08	Acute exacerbation of asthma
40823	H334.00	2.25E+08	Brittle asthma
106805	H335.00	8.67E+14	Chronic asthma with fixed airflow obstruction
4442	H33z.00	1.96E+08	Asthma unspecified
4892	H33z000	7.35E+08	Status asthmaticus NOS
233	H33z011	7.08E+08	Severe asthma attack
232	H33z100	7.08E+08	Asthma attack
8335	H33z111	7.08E+08	Asthma attack NOS
12987	H33z200	2.34E+08	Late-onset asthma
16070	H33zz00	1.96E+08	Asthma NOS
4606	H33zz11	31387002	Exercise induced asthma
21232	H33zz12	3.89E+08	Allergic asthma NEC
93353	H35y600	23315001	Sequoiosis (red-cedar asthma)
39478	H35y700	56968009	Wood asthma
47684	H47y000	41553006	Detergent asthma

3. Codelist for Lung cancer

medcode	readcode	snomedctconceptid	Description
13243	B22..00	4.31E+08	Malignant neoplasm of trachea, bronchus and lung
12870	B221.00	3.72E+08	Malignant neoplasm of main bronchus
17391	B221000	1.88E+08	Malignant neoplasm of carina of bronchus
33444	B221100	93827000	Malignant neoplasm of hilus of lung
21698	B221z00	3.72E+08	Malignant neoplasm of main bronchus NOS
10358	B222.00	2.69E+08	Malignant neoplasm of upper lobe, bronchus or lung
31700	B222000	1.88E+08	Malignant neoplasm of upper lobe bronchus
25886	B222100	1.88E+08	Malignant neoplasm of upper lobe of lung
44169	B222z00	2.69E+08	Malignant neoplasm of upper lobe, bronchus or lung NOS
31268	B223.00	1.88E+08	Malignant neoplasm of middle lobe, bronchus or lung
41523	B223000	1.88E+08	Malignant neoplasm of middle lobe bronchus
39923	B223100	1.88E+08	Malignant neoplasm of middle lobe of lung
54134	B223z00	1.88E+08	Malignant neoplasm of middle lobe, bronchus or lung NOS
31188	B224.00	1.88E+08	Malignant neoplasm of lower lobe, bronchus or lung
18678	B224000	1.88E+08	Malignant neoplasm of lower lobe bronchus
12582	B224100	1.88E+08	Malignant neoplasm of lower lobe of lung
42566	B224z00	1.88E+08	Malignant neoplasm of lower lobe, bronchus or lung NOS
36371	B225.00	1.09E+08	Malignant neoplasm of overlapping lesion of bronchus & lung
38961	B22y.00	3.63E+08	Malignant neoplasm of other sites of bronchus or lung
3903	B22z.00	3.63E+08	Malignant neoplasm of bronchus or lung NOS
2587	B22z.11	93880001	Lung cancer
4137	B570.00	94391008	Secondary malignant neoplasm of lung
40595	Byu2000	1.09E+15	[X]Malignant neoplasm of bronchus or lung, unspecified
32246	ZV10111	4.15E+08	[V]Personal history of malignant neoplasm of bronchus

29284	ZV10112	4.15E+08	[V]Personal history of malignant neoplasm of lung
-------	---------	----------	---

4. Codelist for Ischemic stroke

medcodeid	cleansedreadcode	snomedctconceptid	Description
1.01E+14	F051200	21258007	thrombosis lateral sinus
1.05E+08	G6...00	62914000	cerebrovascular disease
1130181000000113	G679.00	4.44E+08	small vessel cerebrovascular disease
1160471000000111	662M100	5.2E+14	stroke 6 month review
1.19E+08	G640.00	71444005	cerebral thrombosis
11920121000006117		2.66E+08	transient cerebral ischemia
11930671000006110		8.71E+14	qstroke 10 year risk of stroke calculator
1.21E+09		3.74E+08	occlusive stroke
1.22E+09	7000000	14247003	hemispherectomy
1.22E+09	7012.11	3.02E+08	other operations on cerebroventricular shunt
1.22E+09	G64..00	20059004	cerebral arterial occlusion
1.22E+14	G64..13	2.31E+08	stroke due to cerebral arterial occlusion
1.22E+14	G66..12	2.31E+08	stroke unspecified
1.23E+09	ZV12511	2.76E+08	[v]personal history of stroke
1.23E+09	G673.00	42994005	cerebral aneurysm, nonruptured
1.23E+09	C1zy200	75968004	cerebral gigantism
1.23E+09	C1zy211	75968004	sotos syndrome
1.24E+09	G67A.00	95455008	cerebral vein thrombosis
1.25E+08	G641.00	75543006	cerebral embolism
12626401000006115		1110981000000104	qof (quality and outcomes framework) stroke and transient ischaemic attack quality indicator-related care invitation
1.26E+14	G63..12	1055001	stenosis of precerebral artery
12734411000006116		1.43E+14	qof (quality and outcomes framework) stroke and transient ischaemic attack quality indicator-related care invitation using preferred method of communication
13630291000006119		7.83E+08	aneurysm of anterior cerebral artery
13630331000006114		7.83E+08	aneurysm of middle cerebral artery
13717171000006112		16661971000119104	cerebrovascular accident due to stenosis of bilateral carotid arteries
1.39E+14	9h2..00	7.16E+14	exception reporting - stroke quality indicators
1.39E+14	9h21.00	7.17E+14	excepted from stroke quality indicators - patient unsuitable
1.39E+14	9h22.00	7.16E+14	excepted from stroke quality indicators - informed dissent
13909861000006114		7.89E+08	cerebral ischaemic stroke due to dissection of artery
13909881000006116		7.89E+08	cerebral ischaemic stroke due to global hypoperfusion with watershed infarct
13909891000006118		7.89E+08	cerebral ischaemic stroke due to hypercoagulable state
13937171000006112		8.38E+08	stenosis of cerebral artery
13960481000006114		8.66E+08	at risk of stroke
14014251000006116		16644681000119102	cerebrovascular accident due to occlusion of bilateral pontine arteries
14182621000006113		16891111000119104	cryptogenic stroke
14194161000006118		40076141000119107	acute stroke due to occlusion of left posterior cerebral artery
14194191000006114		46583221000119102	acute cerebrovascular accident due to stenosis of left carotid artery
14194211000006110		63230211000119103	acute cerebrovascular accident due to occlusion of left middle cerebral artery
14194281000006115		137592291000119101	acute cerebrovascular accident due to occlusion of right posterior cerebral artery
14194331000006112		272202251000119109	acute cerebrovascular accident due to occlusion of right middle cerebral artery

14194491000006118		407573811000119100	acute cerebrovascular accident due to occlusion of right carotid artery
14194571000006115		502836591000119106	acute cerebrovascular accident due to occlusion of left carotid artery
14194611000006113		595899961000119100	acute cerebrovascular accident of basal ganglia
14194641000006112		652287331000119104	acute cerebrovascular accident of brainstem
14194651000006114		652287331000119104	acute stroke of brainstem
14194661000006111		693158221000119105	acute cerebrovascular accident due to stenosis of right carotid artery
14194741000006118		738478141000119101	acute cerebrovascular accident due to occlusion of right cerebellar artery
14194801000006111		762005171000119101	acute cerebrovascular accident due to ischaemia
14194811000006114		762005171000119101	acute ischaemic stroke
14194941000006115		931164671000119105	acute cerebrovascular accident due to embolism of right middle cerebral artery
14194971000006111		936648941000119105	acute cerebrovascular accident due to embolism of left middle cerebral artery
1.48E+09	7A23800	3.91E+08	percutaneous coil embolisation of cerebral artery aneurysm
1.5E+14	G68X.00	1.95E+08	sequelae of stroke,not specfd as h'morrhage or infarction
1.5E+14	G68W.00	1.95E+08	sequelae/other + unspecified cerebrovascular diseases
1548241000006113	7A20500	2.25E+14	high-flow interposition extracranial to intracranial bypass from external carotid artery to middle cerebral artery
1548251000006110	7A20600	4.27E+08	bypass of carotid artery by anastomosis of superficial temporal artery to middle cerebral artery
1558671000006117	7P08300	4.26E+08	computed tomography of cerebral vessels
1562581000006114	9Om4.00	7.17E+14	stroke/transient ischaemic attack monitoring telephone invitation
1576261000006112		1576261000006108	cause of death- cerebral infarct
1.62E+14	9N4X.00	8.86E+13	did not attend stroke clinic
1682791000006112		1682791000006108	mitoch myopath/encephalopath/lactic acidosis/stroke-like episode
1716621000000117	1JA1000	7.69E+14	suspected cerebrovascular accident
1716631000000115	1JA1011	7.69E+14	suspected stroke
1726581000006116		1726581000006100	central post-stroke pain
1757281000006116		1757281000006100	suspected stroke
1768371000006119	7N4B.00	88556005	structure of cerebral artery
1773171000006119		1773171000006103	cha2ds2-vasc(chf,hyp,age2,dia,stroke2,vas dis,age,sex)risk score
1773371000006117		1773371000006101	cerebrovascular accident care plan
1.78E+09	P247.00	2.53E+08	dysplasia of cerebral cortex
1.78E+09	662e.00	6.99E+08	cva annual review
1809201000006113		1809201000006109	at risk of cerebrovascular disease
1809211000006111		1809211000006107	high risk of cerebrovascular disease
1.81E+14	7012100	2.31E+08	removal of cerebroventricular shunt
1823911000006110		1823911000006106	cerebrovascular disorder confirmed
1855541000006116		1855541000006100	cerebrovascular disease monitoring admin.
1855551000006119		1855551000006103	cerebrovascular disease monitoring in primary care
1855561000006117		1855561000006101	cerebrovascular disease monitoring in secondary care
1855571000006112		1855571000006108	cerebrovascular disease monitoring default
1857051000006118		1857051000006102	cerebrovascular disease management
1863321000006113		1863321000006109	referral by stroke nurse specialist
1870501000006114	C373M11	45502001	cerebral amyloidosis
1994761000006112		1994761000006108	referral to stroke rehabilitation service
2008521000006110		2008521000006106	referral to community stroke service
2115821000000118	661N700	8.11E+14	stroke self-management plan review
2118011000000114	661M700	8.12E+14	stroke self-management plan agreed
2126181000000114	662e.11	6.99E+08	stroke annual review

2.17E+08	6F...00	1.36E+08	stroke prevention
2.17E+08	6F...11	1.36E+08	cva prevention
2.17E+08	388I.00	1.36E+08	stroke risk
2.19E+14	G64..12	4.33E+08	infarction - cerebral
2185361000000119	8OAC.00	8.42E+14	provision of written information about stroke
2249731000000117	38Go.00	1085311000000106	qstroke risk calculator score
2.35E+13	7012y00	3.02E+08	other operation on connection from ventricle of brain os
2.36E+14	9Om..00	7.14E+14	stroke/transient ischaemic attack monitoring administration
2.41E+13	7A24.11	1.75E+08	other open operations on cerebral artery
2.42E+13	7A24.12	1.75E+08	other open operations on circle of willis
2436731000000112	5C13.00	9.55E+14	old cerebral infarction on imaging
2.45E+13	7A24z00	1.75E+08	other open operation on cerebral artery/circle of willis nos
2.45E+13	7A24y00	1.75E+08	other open operation on cerebral artery/circle of willis os
2.47E+09	8HBJ.00	4.13E+08	stroke / transient ischaemic attack referral
2.47E+09	8HTQ.00	4.13E+08	referral to stroke clinic
2.47E+09	G64z400	4.13E+08	infarction of basal ganglia
2.47E+09	13YA.00	4.13E+08	stroke group member
2.48E+09	12C4.11	2.75E+08	fh: cva
2.48E+09	14A7.12	2.76E+08	h/o: stroke
2.48E+09	44y3.00	1021781000000108	blood b-galactocerebrosidase level
2.5E+08	1225.11	1.6E+08	no fh: cva/stroke/tia
2.5E+08	1225	1.6E+08	no fh: stroke/tia
2.53E+09	9Om2.00	4.16E+08	stroke/transient ischaemic attack monitoring third letter
2.53E+09	9Om1.00	4.16E+08	stroke/transient ischaemic attack monitoring second letter
2.53E+09	9Om0.00	4.16E+08	stroke/transient ischaemic attack monitoring first letter
2.54E+09	F28y300	4.14E+08	ventriculomegaly
2539131000006116		2517002	stroke rehabilitation
2550771000006114		3258003	arteriography of cerebral arteries
2550801000006111		3258003	cerebral arteriography
2572431000006118		4605009	cerebral function
2583131000006114		5228007	right cerebral hemisphere structure
2583141000006116		5228007	right cerebral hemisphere
2.61E+08	481..00	1.68E+08	csf examination - general
2.61E+08	481Z.00	1.68E+08	csf exam. - general nos
2.61E+08	48Z..00	1.68E+08	csf examination nos
2.64E+08	662M.00	1.71E+08	stroke monitoring
2.66E+08	7004300	10458001	evacuation of intracerebral haematoma
2.66E+08	7011000	1.72E+08	maintenance of proximal catheter of cerebroventricular shunt
2.66E+08	7011100	1.72E+08	maintenance of distal catheter of cerebroventricular shunt
2.66E+08	7012411	2.82E+08	conversion of cerebroventricular shunt nec
2.66E+08	7011400	1.72E+08	renewal of valve of cerebroventricular shunt
2.66E+08	7012	3.02E+08	other operations on connection from ventricle of brain
2.66E+08	7012200	1.72E+08	aspiration of cerebroventricular shunt
2.66E+08	7012500	4.5E+08	attention to cerebroventricular shunt
2.66E+08	7012z00	3.02E+08	other operation on connection from ventricle of brain nos
2.67E+14	G677400	1.95E+08	occlusion and stenosis of multiple and bilateral cerebral arteries
2.67E+14	G677.00	1.95E+08	occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction

2.69E+09	B51..00	4.28E+08	malignant neoplasm of brain
2.72E+08	7A2..00	1.75E+08	carotid, cerebral and subclavian artery operations
2.72E+08	7A21.00	1.75E+08	other open operations on carotid artery
2.72E+08	7A21y00	1.75E+08	other specified other open operation on carotid artery
2.72E+08	7A21z00	1.75E+08	other open operation on carotid artery nos
2.72E+08	7A23000	1.75E+08	excision of aneurysm of cerebral artery
2.72E+08	7A23400	2.31E+08	ligation of aneurysm of cerebral artery nec
2.72E+08	7A23500	2.31E+08	ligation of aneurysm of circle of willis nec
2.72E+08	7A23600	4.51E+08	obliteration of aneurysm of cerebral artery
2.72E+08	7A24000	1.75E+08	reconstruction of cerebral artery
2.72E+08	7A24200	1.75E+08	anastomosis of cerebral artery
2.72E+08	7A24400	1.75E+08	open embolectomy of cerebral artery
2.72E+08	7A24600	1.75E+08	open embolisation of cerebral artery
2.72E+08	7A25000	1.75E+08	percutaneous transluminal embolisation of cerebral artery
2.72E+08	7A25200	2.33E+08	embolisation of cerebral artery nec
2.72E+08	7A27y00	1.75E+08	other specified other open operation on subclavian artery
2.72E+08	7A27z00	1.75E+08	other open operation on subclavian artery nos
2.72E+08	7A2z.00	1.75E+08	carotid and/or cerebral and/or subclavian artery operations
2725641000006119		14247003	cerebral hemispherectomy
2729551000000111		1089411000000104	cerebral infarction due to occlusion of cerebral artery
2729571000000119		1089421000000105	cerebral infarction due to stenosis of cerebral artery
2773351000006111		17232002	middle cerebral artery
2.86E+08	A270300	1.86E+08	listerial cerebral arteritis
2870551000006114		23238000	common variable agammaglobulinaemia
2870571000006116		23238000	cvag
2870621000006118		23238000	acquired hypogammaglobulinaemia
2.87E+09	G729.00	7.13E+08	dissection of precerebral artery
2.87E+08	A840000	1.87E+08	plasmodium falciparum malaria with cerebral complications
2.87E+08	A94y600	1.87E+08	rupture of syphilitic cerebral aneurysm
2.88E+08	AB65400	1.87E+08	cerebral cryptococcosis
2.88E+08	AB67100	1.87E+08	rhinocerebral mucormycosis
2.89E+08	B510100	1.88E+08	malignant neoplasm of cerebral cortex
2.89E+08	B515100	1.88E+08	malignant neoplasm of floor of cerebral ventricle
2.89E+08	B515z00	3.63E+08	malignant neoplasm of cerebral ventricle nos
2.89E+08	B517000	1.88E+08	malignant neoplasm of cerebral peduncle
2.89E+08	B51y.00	4.28E+08	malignant neoplasm of other parts of brain
2.89E+08	B51yz00	4.28E+08	malignant neoplasm of other part of brain nos
2.89E+08	B51z.00	4.28E+08	malignant neoplasm of brain nos
2.89E+08	B521000	1.88E+08	malignant neoplasm of cerebral dura mater
2.89E+08	B521100	1.88E+08	malignant neoplasm of cerebral arachnoid mater
2.89E+08	B521200	1.88E+08	malignant neoplasm of cerebral pia mater
2901571000006113		25133001	completed stroke
2.91E+08	B7F2000	1.89E+08	cerebral meningioma
2.93E+08	C327100	1.91E+08	gaucher disease
2.95E+08	E030400	1.92E+08	acute confusional state, of cerebrovascular origin
2.95E+08	E031400	1.92E+08	subacute confusional state, of cerebrovascular origin
2.97E+08	F101z00	1.93E+08	cerebral lipidoses nos
2.97E+08	F102z00	1.93E+08	cerebral degeneration in lipidosis nos
2.97E+08	F103.00	1.93E+08	cerebral degeneration in diseases ec
2.97E+08	F103z00	1.93E+08	cerebral degeneration in disease nos

2.97E+08	F11xz00	1.93E+08	cerebral degeneration other disease nos
2.97E+08	F28y100	41075004	cerebral fungus
2.99E+08	Fyu5600	2.31E+08	lacunar infarction
3E+08	G633.00	1.95E+08	multiple and bilateral precerebral arterial occlusion
3E+08	G63y.00	2.66E+08	other precerebral artery occlusion
3E+08	G63y000	1.95E+08	cerebral infarct due to thrombosis of precerebral arteries
3E+08	G63y100	1.95E+08	cerebral infarction due to embolism of precerebral arteries
3E+08	G63z.00	2.66E+08	precerebral artery occlusion nos
3E+08	G640000	1.95E+08	cerebral infarction due to thrombosis of cerebral arteries
3E+08	G641000	1.95E+08	cerebral infarction due to embolism of cerebral arteries
3E+08	G65y.00	2.66E+08	other transient cerebral ischaemia
3E+08	G65zz00	2.66E+08	transient cerebral ischaemia nos
3E+08	G65z000	1.95E+08	impending cerebral ischaemia
3E+08	G65z100	1.95E+08	intermittent cerebral ischaemia
3E+08	G667.00	1.95E+08	left sided cva
3E+08	G668.00	1.95E+08	right sided cva
3E+08	G67..00	62914000	other cerebrovascular disease
3E+08	G671z00	3.03E+08	generalised ischaemic cerebrovascular disease nos
3E+08	G671000	29322000	acute cerebrovascular insufficiency
3E+08	G673000	7.13E+08	nonruptured cerebral aneurysm due to dissection of cerebral artery
3E+08	G677000	1.95E+08	occlusion and stenosis of middle cerebral artery
3E+08	G677100	1.95E+08	occlusion and stenosis of anterior cerebral artery
3E+08	G677200	1.95E+08	occlusion and stenosis of posterior cerebral artery
3E+08	G67y.00	62914000	other cerebrovascular disease os
3E+08	G67z.00	62914000	other cerebrovascular disease nos
3E+08	G68..00	1.95E+08	late effects of cerebrovascular disease
3E+08	G683.00	1.95E+08	sequelae of cerebral infarction
3E+08	G6y..00	62914000	other specified cerebrovascular disease
3E+08	G6z..00	62914000	cerebrovascular disease nos
3.01E+08	Gyu6400	4.33E+08	[x]other cerebral infarction
3.01E+08	Gyu6500	2.66E+08	[x]occlusion and stenosis of other precerebral arteries
3.01E+08	Gyu6600	20059004	cerebral artery occlusion
3.01E+08	Gyu6700	62914000	[x]other specified cerebrovascular diseases
3.01E+08	Gyu6800	28366008	[x]cerebral arteritis in infectious and parasitic diseases
30303010	F28y000	17944005	cerebral calcification
3.07E+14	9N0p.00	2.02E+14	seen in stroke clinic
3.08E+08	L440000	6594005	puerperal cerebrovascular disorder unspecified
3.08E+08	L440100	2E+08	puerperal cerebrovascular disorder - delivered
3.08E+08	L440z00	6594005	puerperal cerebrovascular disorder nos
3.1E+13	S630000	4.5E+08	oth cerebral h'ge inj no open intracran wnd+unspec consc
31393014	SN20011	52072009	siriasis
3.14E+14	1JA1.00	2.05E+14	suspected cerebrovascular disease
3161611000006111		41075004	fungal infection of brain
3167661000006117		41440006	radioisotope brain imaging
3167671000006112		41440006	cerebral scan
3193191000006110		42994005	nonruptured cerebral aneurysm
3193201000006113		42994005	unruptured cerebral aneurysm
3.19E+14	7N43000	60176003	[so]anterior cerebral artery
3.21E+08	S610.00	2.1E+08	closed cerebral contusion

3.21E+08	S611.00	2.1E+08	open cerebral contusion
3.21E+08	S61z.00	2.69E+08	cerebral laceration and contusion nos
3.21E+14	7N43.00	88556005	[so]cerebral artery
3.21E+14	7N43z00	88556005	[so]cerebral artery nec
3.25E+08	SP10100	2.13E+08	cerebral anoxia complication
3.27E+14	7N43300	17232002	structure of middle cerebral artery
3.3E+14	7N43400	70382005	[so]posterior cerebral artery
3300291000006117		49453006	cerebral herniation
3.33E+14	7N43y00	88556005	[so]specified cerebral artery nec
3.46E+08	G63..11	2.31E+08	infarction - precerebral
3.46E+08	G64z.11	95457000	brainstem infarction nos
3.46E+08	G665.00	2.31E+08	pure motor lacunar syndrome
3.46E+08	G666.00	2.31E+08	pure sensory lacunar syndrome
3477271000006118		60176003	structure of anterior cerebral artery
3477281000006115		60176003	anterior cerebral artery
3.49E+08	7A27.13	1.75E+08	other open operations on vertebral artery
3.51E+09		3.29E+14	occlusion of left middle cerebral artery
3.51E+09		3.29E+14	occlusion of right middle cerebral artery
3.51E+09		3.3E+14	occlusion of left posterior cerebral artery
3.51E+09		7.34E+08	right middle cerebral artery thrombosis
3.51E+09		3.31E+14	cerebrovascular accident due to thrombus of left carotid artery
3.51E+09		3.3E+14	cerebrovascular accident due to thrombus of basilar artery
3.51E+09		3.3E+14	stroke due to basilar artery thrombus
3.51E+09		3.3E+14	cerebrovascular accident due to thrombus of right carotid artery
3.51E+09		16002031000119102	cerebrovascular accident due to thrombus of right middle cerebral artery
3.51E+09		16002031000119102	stroke due to thrombus of right middle cerebral artery
3.51E+09		16002111000119106	cerebrovascular accident due to thrombus of left middle cerebral artery
3.51E+09		16002111000119106	stroke due to thrombus of left middle cerebral artery
3.51E+09		16023991000119104	cerebrovascular accident due to occlusion of left pontine artery
3.51E+09		16000391000119104	cerebrovascular accident due to occlusion of right posterior cerebral artery
3.51E+09		16000511000119103	cerebrovascular accident due to occlusion of left middle cerebral artery
3.51E+09		16000351000119109	cerebrovascular accident due to occlusion of left posterior cerebral artery
3.51E+09		16000431000119109	cerebrovascular accident due to occlusion of right middle cerebral artery
3.51E+09		3.29E+14	left hemispheric cerebellar artery embolism with stroke
3.51E+09		3.3E+14	non-ruptured acquired cerebral aneurysm
3.51E+09		7.34E+08	thrombosis of right middle cerebral artery
3.51E+09		7.34E+08	thrombosis of left middle cerebral artery
3.51E+09		7.35E+08	embolus of right posterior cerebral artery
3.51E+09		7.35E+08	embolus of left middle cerebral artery
3.51E+09		7.35E+08	embolus of right middle cerebral artery
3.52E+09		16218291000119100	acute cerebral ischaemia
3.52E+09		7.36E+08	stroke care plan
3522691000006112		62914000	cerebrovascular lesion
3522701000006112		62914000	cvd - cerebrovascular disease
3.57E+08	F427K11	14210003	lipofuscinosis nec
3.58E+13	S631.00	4.5E+08	other cerebral h'ge after injury + open intracranial wound
3.58E+13	S630.00	4.5E+08	other cerebral h'ge after injury no open intracranial wound

3.61E+08	G671100	1.11E+08	chronic cerebral ischaemia
3615001000006115		68523003	cerebral white matter
3.64E+09		3.29E+14	cerebrovascular accident due to occlusion of right middle cerebral artery by embolus
3.64E+09		7.63E+08	occlusion of right middle cerebral artery by embolus
3.64E+14	7005500	4.3E+08	multiple subpial transections of cerebral cortex
3644591000006111		70382005	structure of posterior cerebral artery
3644601000006115		70382005	posterior cerebral artery
3683671000006119		72792008	left cerebral hemisphere structure
3683681000006116		72792008	left cerebral hemisphere
3.71E+14	Gyu6G00	1.25E+14	[x]cereb infarct due unsp occlus/stenos precerebr arteries
3.71E+14	Gyu6900	28366008	[x]cerebral arteritis in other diseases ce
3.71E+14	Gyu6300	20059004	[x]cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrts
3.71E+14	Gyu6.00	62914000	[x]cerebrovascular diseases
3.72E+14	7P24200	2.31E+14	delivery of rehabilitation for stroke
3.79E+14	7008500	4.29E+08	insertion of carmustine wafers in cerebral neoplasm
3818181000006114		81037000	lacunar infarct
3882941000006118		85028008	application of clamp to cerebral aneurysm, crutchfield or silverstone
3882971000006114		85028008	intracerebral aneurysm clipped
3939231000006110		88556005	cerebral artery
3.94E+08	7A24.00	1.75E+08	other open operations on cerebral artery or circle of willis
3.94E+08	7A27.00	1.75E+08	other open operations on subclavian artery
3.96E+08	G63..00	2.66E+08	precerebral arterial occlusion
3.96E+08	G64z.00	4.33E+08	cerebral infarction
3.96E+08	G65..12	2.66E+08	transient ischaemic attack
3.96E+08	G65..00	2.66E+08	transient cerebral ischaemia
3.97E+13	S631z00	4.5E+08	oth cereb h'ge inj + open intracran wnd+concussion unspc
3.97E+13	S630400	4.5E+08	oth cereb h'ge inj no open intracran wnd+>24hr loc +recovery
3.98E+13	S630z00	4.5E+08	oth cereb h'ge inj no open intracran wnd+concussion unspc
3.98E+13	S630600	4.5E+08	oth cereb h'ge inj no open intracran wnd+loc unspc duration
3.98E+13	S631300	4.5E+08	oth cerebral h'ge inj + open intracran wnd+1-24hr loss consc
3.98E+13	S630100	4.5E+08	oth cerebral h'ge inj no open intracranial wnd+no loss consc
3.98E+13	S61y.00	2.69E+08	oth cerebral laceration/contusion + open intracranial wound
3.99E+13	S61x.00	2.69E+08	oth cerebral laceration/contusion no open intracranial wound
3.99E+08	F427K00	14210003	lipofuscinosis
4.01E+08	P203.00	2.68E+08	meningocele - cerebral
4025321000006111		93748005	primary malignant neoplasm of cerebral ventricle
4.03E+08	S61..00	2.69E+08	cerebral laceration and contusion
4031711000006119		94248000	secondary malignant neoplasm of cerebrum
4.05E+08	G66..00	2.31E+08	cerebrovascular accident
4056971000006118		95455008	thrombosis of cerebral veins
4057031000006117		95457000	brain stem infarct
4057041000006110		95457000	brain stem stroke
4057051000006112		95457000	infarction of brain stem
4.07E+14	Gyu6A00	62914000	[x]other cerebrovascular disorders in diseases ce
4.11E+08	12C4.12	2.75E+08	fh: stroke
4.11E+08	L440.12	2.75E+08	stroke in the puerperium
4.12E+08	14A7.11	2.76E+08	h/o: cva

4.14E+08	1M4..00	2.77E+08	central post-stroke pain
4.16E+08	F11z.11	2.79E+08	cerebral atrophy
4.17E+14	Fyu5500	2.66E+08	transient cerebral ischemia
4.17E+14	Fyu5700	62914000	[x]other vascular syndroms/brain in cerebrovascul diseases
4193641000006111		1.11E+08	nonparalytic stroke
4.26E+14	Gyu6C00	1.95E+08	[x]sequelae of stroke,not specfd as h'morrhage or infarction
4.26E+14	Gyu6D00	1.95E+08	[x]sequelae/other + unspecified cerebrovascular diseases
4411861000006113		1.29E+08	cerebral arterial aneurysm
4.5E+08	8Hd6.00	3.07E+08	admission to stroke unit
4.51E+08	G64z200	3.08E+08	left sided cerebral infarction
4.51E+08	G64z300	3.08E+08	right sided cerebral infarction
4.51E+08	1477	3.08E+08	h/o: cerebrovascular disease
4.51E+08	14AK.00	3.08E+08	h/o: stroke in last year
4.56E+08	F483600	3.12E+08	cerebral akinetopsia
4.59E+13	7A23z00	2.66E+08	operation on cerebral artery/ circle of willis aneurysm nos
4.59E+13	7A23y00	2.66E+08	operation on cerebral artery/ circle of willis aneurysm os
4644401000006115		1.72E+08	insertion of antisyphon device into ventricular shunt
4700881000006118		1.81E+08	entire mca - middle cerebral artery
47484015	G674.00	28366008	cerebral arteritis
4759821000006115		1.91E+08	glucocerebrosidosis
4778181000006119		1.95E+08	left sided cerebral hemisphere cerebrovascular accident
4778201000006118		1.95E+08	right sided cerebral hemisphere cerebrovascular accident
4778311000006112		1.95E+08	sequelae of cerebrovascular disease
4.78E+08	7012300	16900000	renewal of cerebroventricular shunt
47937012	F11y100	28634005	cerebral ataxia
4.83E+08	B515.00	3.63E+08	malignant neoplasm of cerebral ventricles
4.91E+14	7A25400	3258003	arteriography of cerebral artery
4.99E+14	7011.11	3.02E+08	attention to component of cerebroventricular shunt
5010981000006119		2.31E+08	stroke
5011041000006118		2.31E+08	anterior cerebral circulation infarction
5011051000006116		2.31E+08	total anterior cerebral circulation infarction
5011061000006119		2.31E+08	taci - total anterior cerebral circulation infarction
5011071000006114		2.31E+08	total anterior cerebral circulation stroke
5011081000006112		2.31E+08	tacs - total anterior cerebral circulation stroke
5011091000006110		2.31E+08	partial anterior cerebral circulation infarction
5011101000006116		2.31E+08	paci - partial anterior cerebral circulation infarction
5011111000006118		2.31E+08	partial anterior cerebral circulation stroke
5011121000006114		2.31E+08	pacs - partial anterior cerebral circulation stroke
5011131000006112		2.31E+08	posterior cerebral circulation infarction
5011141000006119		2.31E+08	posterior cerebral circulation stroke
5011161000006115		2.31E+08	lacunar stroke
5011171000006110		2.31E+08	laci - lacunar infarction
5011191000006111		2.31E+08	pure motor lacunar infarction
5011211000006112		2.31E+08	pure sensory lacunar infarction
5011231000006118		2.31E+08	pure sensorimotor lacunar infarction
5011451000006114		2.31E+08	stroke of uncertain pathology
5011461000006111		2.31E+08	anterior circulation stroke of uncertain pathology
5011471000006116		2.31E+08	posterior circulation stroke of uncertain pathology
5011861000006114		2.31E+08	asymptomatic cerebrovascular disease

5013671000006119		2.31E+08	removal of shunt tubing from ventricle of brain
5049061000006114		2.33E+08	embolisation of intracranial artery
5049071000006119		2.33E+08	embolization of intracranial artery
5058411000006114		2.34E+08	ruptured cerebral aneurysm
5.06E+08	7A23200	85028008	clipping of aneurysm of cerebral artery
5165331000006118		2.42E+08	mri of cerebral vessels
5.24E+14	S61yz00	2.69E+08	brain cont + open intracranial wound + concussion unspec
5.24E+14	S61y600	2.69E+08	brain cont + open intracranial wound + loc unspec duration
5.24E+14	S61y000	2.69E+08	brain cont + open intracranial wound + unspec state consc
5.24E+14	S61xz00	2.69E+08	brain cont no open intracranial wound + concussion unspec
5.24E+14	S61x600	2.69E+08	brain cont no open intracranial wound + loc unspec duration
5.25E+14	G64z000	95457000	brain stem infarction
5291741000006114		2.51E+08	cerebral perfusion pressure
5291751000006111		2.51E+08	cpp - cerebral perfusion pressure
5326301000006115		2.53E+08	cortical dysplasia
5.39E+14	7A2y.00	1.75E+08	carotid, cerebral and subclavian artery operations os
5.42E+14	G6W..00	1.25E+14	cerebral infarction due to occlusion of precerebral artery
5.43E+14	7A23.00	2.66E+08	operations on aneurysm of cerebral artery or circle of willis
5.43E+14	7A23.11	2.66E+08	cerebral artery aneurysm operations
5.43E+14	F23..14	1.28E+08	cerebral atonia
5.43E+14	F11x.00	1.93E+08	cerebral degeneration associated with another disorder
5.43E+14	G641.11	75543006	cerebral embolus
5.43E+14	F101.00	14210003	cerebral lipidoses
5.43E+14	B583200	94248000	cerebral metastases
5.43E+14	G6X..00	20059004	cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrts
5.43E+14	L440.00	6594005	cerebrovascular disorder in the puerperium
5455801000006110		2.63E+08	contusion of cerebrum
5455811000006113		2.63E+08	cerebral contusion
5475241000006111		2.64E+08	cerebral function analysing monitor
5484131000006117		2.65E+08	removal of cerebral ventricular shunt
5492171000006112		2.66E+08	tia
5492201000006111		2.66E+08	tia - transient ischaemic attack
5492221000006118		2.66E+08	transient ischemic attack
5.58E+14	7A23.12	2.66E+08	circle of willis aneurysm operations
5582311000006110		2.76E+08	occipital cerebral infarction
5582901000006111		2.76E+08	unruptured aneurysm of middle cerebral artery
5587541000006115		2.77E+08	cerebral irritation
5591251000006111		2.77E+08	primary cerebral lymphoma
5597481000006114		2.77E+08	ruptured aneurysm of anterior cerebral artery
5597491000006112		2.77E+08	ruptured aneurysm of middle cerebral artery
5599881000006118		2.77E+08	cerebral ventricular distension
5611061000006113		2.78E+08	entire cerebral hemisphere
5649201000006113		2.81E+08	extension of cerebrovascular accident
5653361000006116		2.82E+08	conversion of ventricular shunt
5726831000006118		2.88E+08	cerebral ischaemia
5866391000006119		2.99E+08	stroke test positive
5866401000006117		2.99E+08	stroke test negative
5870291000006116		3E+08	finding of cerebral dominance

5905541000006118		3.02E+08	procedure on a brain ventricular shunt
5.94E+14	7011300	2.82E+08	conversion of cerebroventricular shunt nec
5960821000006117		3.07E+08	referral to stroke service
5960971000006110		3.07E+08	seen by stroke service
5967391000006111		3.07E+08	multiple lacunar infarcts
5.98E+14	S610200	2.1E+08	cortex cont no open intracranial wnd + <1hr loss of consc
5.98E+14	S610400	2.1E+08	cortex cont no open intracranial wnd + >24 hr loc + recovery
5.98E+14	S610300	2.1E+08	cortex cont no open intracranial wnd + 1-24hrs loss of consc
5.98E+14	S610z00	2.1E+08	cortex cont no open intracranial wnd + concussion unspec
5.98E+14	S610600	2.1E+08	cortex cont no open intracranial wnd + loc unspec duration
5.98E+14	S610100	2.1E+08	cortex cont no open intracranial wnd + no loss of consc
5.98E+14	S610000	2.1E+08	cortex cont no open intracranial wnd + unspec state consc
5.98E+14	S610500	2.1E+08	cortex cont no open intracranial wnd +>24hr loc not restored
5994741000006112		3.1E+08	stroke unit
5996861000006112		3.1E+08	stroke service
6044961000006114		3.14E+08	branch of middle cerebral artery
6.05E+14	G64..11	2.31E+08	cva - cerebrovascular accident due to cerebral artery occlusion
6.05E+14	L440.11	2.75E+08	cva - cerebrovascular accident in the puerperium
6.05E+14	G66..13	2.31E+08	cva - cerebrovascular accident
6.06E+14	G66..11	2.31E+08	cva unspecified
6321781000006118		3.69E+08	left posterior cerebral artery
6322771000006112		3.69E+08	right middle cerebral artery
6322791000006113		3.69E+08	left middle cerebral artery
6348291000006115		3.71E+08	thrombotic stroke
6348301000006119		3.71E+08	embolic stroke
6.41E+14	F032000	53622003	encephalitis due to malaria
6837051000006119		4.13E+08	basal ganglion stroke
6837061000006117		4.13E+08	basal ganglion infarct
6837371000006110		4.13E+08	stroke / transient ischemic attack referral
6837381000006113		4.13E+08	stroke/transient ischemic attack referral
6846141000006118		4.14E+08	cardioembolic stroke
6847241000006113		4.14E+08	cerebral ventriculomegaly
6908391000006111		4.18E+08	stroke/transient ischaemic attack monitoring invitation
6990981000006114		4.23E+08	ischaemic stroke
6990991000006112		4.23E+08	ischemic stroke
7054171000006115		4.26E+08	acute lacunar infarction
7067811000006116		4.27E+08	cerebral vasculitis
7071901000006111		4.27E+08	thalamic infarction
7071911000006114		4.27E+08	thalamic stroke
7217961000006118		4.39E+08	percutaneous balloon assisted coil embolisation of cerebral aneurysm using fluoroscopic guidance
7285511000006116		4.44E+08	cerebral small vessel disease
7289121000006119		4.44E+08	recurrent transient cerebral ischaemic attack
7289131000006116		4.44E+08	recurrent transient cerebral ischemic attack
7.31E+14	F032011	53622003	malarial encephalitis
7339271000006112		4.47E+08	concussion injury of cerebrum
7339281000006110		4.47E+08	cerebral concussion
7382631000006110		4.51E+08	national institutes of health stroke scale
7.4E+14	F23..13	1.28E+08	littles disease

7515281000006115		6.99E+08	multiple lacunar infarcts
7519441000006112		6.99E+08	cerebrovascular accident annual review
7525971000006116		7E+08	embolism of middle cerebral artery
7568281000006111		7.03E+08	stroke clinic
75861010	C373M00	45502001	cerebrovascular amyloidosis
7597121000006117		7.05E+08	stroke impact scale version 3.0
7598981000006119		7.05E+08	cerebral infarction due to embolism of middle cerebral artery
7598991000006116		7.05E+08	thrombosis of middle cerebral artery
7599001000006116		7.05E+08	cerebral infarction due to thrombosis of middle cerebral artery
7618821000006118		7.08E+08	stroke impact scale version 3.0 score
7699511000006117		7.13E+08	dissection of cerebral artery
7703601000006113		7.13E+08	post stroke pain
7.73E+14	7011200	1.72E+08	insertion of antisyphon device into cerebroventricular shunt
78268012	S630.11	46966000	cerebral compression due to injury
7839991000006115		7.23E+08	silent cerebral infarct
7855631000006111		7.24E+08	cerebral ischaemic stroke due to small artery occlusion
7861341000006115		7.25E+08	epilepsy due to stroke
7.87E+14	SN20111	52072009	ictus solaris
8.01E+14	G671.00	3.03E+08	diffuse cerebrovascular disease
8011371000006117		2.37E+13	sequela of ischaemic cerebral infarction
8026451000006117		8.86E+13	dna (did not attend) stroke clinic
8044481000006118		1.41E+14	ischaemic stroke without coma
8.09E+14	14A7.00	2.76E+08	history of cerebrovascular accident
8195431000006118		5.2E+14	stroke/cerebrovascular accident 6 month review
8196061000006110		5.21E+14	did not attend stroke review
8221321000006118		7.2E+14	stroke and tia
8264201000006110		8.16E+14	management of stroke rehabilitation
8265471000006116		8.18E+14	acute stroke care
8274811000006116		8.28E+14	stroke medicine
8.41E+14	G678.00	3.91E+08	cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy
8453821000006113		1078671000000108	emergency hospital admission to stroke service
8.57E+14		8.57E+14	no h/o: cva/stroke
8.58E+14		8.58E+14	no fh: cva/stroke
8.65E+14	7925.99	1.75E+08	heart: int. mammary a implant
8.66E+14	7A23299	85028008	clipping - aneurysm
8.84E+14	G63z.99	5.84E+14	precerebral a. occlusion nos
8.85E+14	G64z.99	6.83E+14	cerebral a. occlusion nos
8.85E+14	G66..98	6.86E+14	stroke/cva - undefined
8.85E+14	G66..99	6.86E+14	stroke
8.85E+14	G673.99	42994005	cerebral aneurysm-non ruptured
8.85E+14	G68..99	1.95E+08	cerebrovasc. dis. late effects
89147017	A849100	53622003	cerebral malaria
8.97E+14	S61..99	2.69E+08	cerebral laceration/contusion
8.97E+14	S62..99	82894007	cerebral h'ge after injury
8.97E+14	S64z.99	6.84E+14	intracerebral injury nos
8.99E+13	7A25z00	2.66E+08	transluminal operation on cerebral art/circle of willis nos
8.99E+13	7A25y00	2.66E+08	transluminal operation on cerebral art/circle of willis os
9E+14	SP00399	2.13E+08	bypass graft mech. complic.
9E+13	7A25.00	2.66E+08	transluminal operations on cerebral artery or circle of willis

9.03E+14	7012299	1.72E+08	therap.brain ventric.shunt asp
9.03E+14	7012399	16900000	ex-cran.ventric.shunt replaced
9.06E+14		9.06E+14	[rfc] long-term effect of cerebral heamorrhage
9.08E+14		9.08E+14	[rfc] stroke/cva
9.08E+14		9.08E+14	[rfc] stroke
9.09E+14		9.09E+14	[rfc] stroke
9.09E+14		9.09E+14	[rfc] cva
9.29E+14	C315100	39925003	melas - mitochondrial encephalopathy, lactic acidosis and stroke-like episodes
9.4E+14		9.4E+14	stroke monitoring administration
9.4E+14		9.4E+14	stroke monitoring invite 1
9.4E+14		9.4E+14	stroke monitoring invite 2
9.4E+14		9.4E+14	stroke monitoring invite 3
9.59E+13	G65z.00	2.66E+08	transient cerebral ischaemia nos
9.68E+14		9.68E+14	annual stroke/cva blood test
9.87E+14	9Om3.00	4.17E+08	stroke/transient ischaemic attack monitoring verbal invitation
9.89E+14	G65..99	2.66E+08	transient ischaemic attacks
9.9E+14	S61z.99	5.67E+14	intracerebral injury nos
9912271000006112		16371781000119100	cerebellar stroke
13909901000006119		7.89E+08	Cerebral ischaemic stroke co-occurrent with subarachnoid haemorrhage
1573101000006112		1573101000006108	Cerebral infarction with haemorrhagic transformation
2127781000000110	662M200	8.17E+14	Stroke initial post discharge review
2.55E+09	662o.00	4.18E+08	Haemorrhagic stroke monitoring
3E+08	G617.00	1.95E+08	Intracerebral haemorrhage, intraventricular
3E+08	G660.00	1.95E+08	Middle cerebral artery syndrome
3E+08	G661.00	1.95E+08	Anterior cerebral artery syndrome
3E+08	G662.00	1.95E+08	Posterior cerebral artery syndrome
3E+08	G664.00	1.95E+08	Cerebellar stroke syndrome
3.46E+08	G619.00	2.31E+08	Lobar cerebral haemorrhage
3.55E+09		16549441000119102	Radionuclide scan of cerebrospinal fluid flow of brain using indium 111
4777861000006119		1.95E+08	Intracerebral haemorrhage with intraventricular haemorrhage
4777881000006112		1.95E+08	Intracerebral haemorrhage, multiple localised
5011291000006119		2.31E+08	Haemorrhagic cerebral infarction
5011331000006114		2.31E+08	Anterior cerebral circulation haemorrhagic infarction
5011351000006119		2.31E+08	Posterior cerebral circulation haemorrhagic infarction
5.25E+14	G663.00	1.95E+08	Brainstem stroke syndrome
9.67E+14	8HHM.00	1.13E+14	Referral to multidisciplinary stroke function improvement service

5. Codelist for Heart Failure

aurum_medcode	readcode	description
1488804017	1O1..00	Heart failure confirmed
2352391000000117	2JZ..00	On optimal heart failure therapy
7.84191E+14	33BA.00	Impaired left ventricular function
8.33381E+14	388D.00	New York Heart Assoc classification heart failure symptoms
2159197017	585f.00	Echocardiogram shows left ventricular systolic dysfunction
2117931000000116	661M500	Heart failure self-management plan agreed
216184014	662T.00	Congestive heart failure monitoring
1484918019	662W.00	Heart failure annual review
2616471011	662g.00	New York Heart Association classification - class II

2616472016	662h.00	New York Heart Association classification - class III
2616473014	662i.00	New York Heart Association classification - class IV
4.04741E+14	662p.00	Heart failure 6 month review
2645623019	679X.00	Heart failure education
3.08011E+14	67D4.00	Heart failure information given to patient
451426015	8B29.00	Cardiac failure therapy
3.08301E+14	8CL3.00	Heart failure care plan discussed with patient
1746171000000119	8CMK.00	Has heart failure management plan
2205951000000117	8CMW800	Heart failure clinical pathway
1784061000006118	8CeC.00	Preferred place of care for next exacerbation heart failure
2549208013	8H2S.00	Admit heart failure emergency
1484917012	8HBE.00	Heart failure follow-up
2549697018	8HHb.00	Referral to heart failure nurse
3.08041E+14	8HHz.00	Referral to heart failure exercise programme
216246012	8HTL.00	Referral to heart failure clinic
2256811000000114	8HTL000	Referral to rapid access heart failure clinic
3.11561E+14	8Hk0.00	Referred to heart failure education group
2549243014	9N0k.00	Seen in heart failure clinic
2548316014	9N2p.00	Seen by community heart failure nurse
3.03361E+14	9N6T.00	Referred by heart failure nurse specialist
2.26181E+14	9On..00	Left ventricular dysfunction monitoring administration
2533628012	9On0.00	Left ventricular dysfunction monitoring first letter
3.08231E+14	9Or..00	Heart failure monitoring administration
4.06801E+14	9Or1.00	Heart failure monitoring telephone invite
4.07041E+14	9Or2.00	Heart failure monitoring verbal invite
4.07061E+14	9Or3.00	Heart failure monitoring first letter
4.07081E+14	9Or4.00	Heart failure monitoring second letter
4.07101E+14	9Or5.00	Heart failure monitoring third letter
7.28671E+14	G210100	Malignant hypertensive heart disease with CCF
5.04901E+14	G211100	Benign hypertensive heart disease with CCF
7.41681E+14	G21z100	Hypertensive heart disease NOS with CCF
7.41701E+14	G232.00	Hypertensive heart&renal dis wth (congestive) heart failure
7.89941E+14	G234.00	Hyperten heart&renal dis+both(congestv)heart and renal fail
299796018	G343.00	Ischaemic cardiomyopathy
142397010	G55..00	Cardiomyopathy
59588016	G554200	Familial cardiomyopathy
300040013	G554400	Primary dilated cardiomyopathy
300041012	G554z00	Other primary cardiomyopathy NOS
138527014	G555.00	Alcoholic cardiomyopathy
300052018	G558.00	Cardiomyopathy in disease EC
300057012	G558z00	Cardiomyopathy in diseases EC NOS
1705371000000116	G55A.00	Tachycardiomyopathy
2674238014	G55A.11	Tachycardia-induced cardiomyopathy
395768019	G55y.00	Secondary cardiomyopathy NOS
34556015	G55y.11	Secondary dilated cardiomyopathy
300062013	G55z.00	Cardiomyopathy NOS
139475013	G58..00	Heart failure
139482012	G58..11	Cardiac failure
70653017	G580.00	Congestive heart failure

493287011	G580.11	Congestive cardiac failure
510016018	G580.14	Biventricular failure
18472010	G580000	Acute congestive heart failure
147247018	G580100	Chronic congestive heart failure
300179017	G580200	Decompensated cardiac failure
300180019	G580300	Compensated cardiac failure
141306010	G581.00	Left ventricular failure
1495417010	G581.11	Asthma - cardiac
300190010	G581000	Acute left ventricular failure
94251011	G582.00	Acute heart failure
1647701000000118	G583.00	Heart failure with normal ejection fraction
1661371000000112	G583.11	HFNEF - heart failure with normal ejection fraction
2227501000000110	G583.12	Heart failure with preserved ejection fraction
395772015	G58z.00	Heart failure NOS
2.23981E+14	G58z.12	Cardiac failure NOS
216207010	G5yy900	Left ventricular systolic dysfunction
1489358014	G5yyA00	Left ventricular diastolic dysfunction
6978012	G5yyC00	Diastolic dysfunction
2694523019	G5yyD00	Left ventricular cardiac dysfunction
350484012	SP11111	Heart failure as a complication of care
8.33381E+14	ZRad.00	New York Heart Assoc classification heart failure symptoms

6. Codelist for Ischemic heart disease

aurum medcode	readcode	description
1430	G33..00	angina pectoris
240	G3â€¦00	Ischaemic heart disease
241	G30..00	acute myocardial infarction
1792	G3â€¦13	IHD - Ischaemic heart disease
1677	G30..15	MI - acute myocardial infarction
14658	G30z.00	Acute myocardial infarction NOS
68357	G31y100	Microinfarction of heart
66388	G33z000	Status anginosus
19542	662K000	angina control - good
10562	G307100	acute non-st segment elevation myocardial infarction
737	792..11	Coronary artery bypass graft operations
1344	G340.12	Coronary artery disease
1431	G311.13	unstable angina
1676	G3z..00	Ischaemic heart disease NOS
2901	7928	Transluminal balloon angioplasty of coronary artery
11983	G311500	acute coronary syndrome
7347	G311100	Unstable angina
8942	7929400	Insertion of coronary artery stent
12229	G30X000	Acute ST segment elevation myocardial infarction
5904	792..00	Coronary artery operations
28554	G33zz00	angina pectoris nos
1678	G308.00	Inferior myocardial infarction NOS
7137	7920y00	Saphenous vein graft replacement of coronary artery OS
43939	793G.00	Perc translumin balloon angioplasty stenting coronary artery
4017	G32..00	Old myocardial infarction

1655	G340.11	Triple vessel disease of the heart
5413	G340.00	Coronary atherosclerosis
105184	792E.00	Percutaneous coronary intervention
1414	G33z300	angina on effort
25842	G33z.00	angina pectoris nos
12804	G33z700	stable angina
3999	G340000	Single coronary vessel disease
4656	G311.11	crescendo angina
1204	G30..14	Heart attack
5674	ZV45K11	[V]Presence of coronary artery bypass graft - CABG
5744	7927500	Open angioplasty of coronary artery
732	7928z00	Transluminal balloon angioplasty of coronary artery NOS
5254	G340100	Double coronary vessel disease
2491	G30..12	Coronary thrombosis
3704	G307.00	acute subendocardial infarction
12734	SP07600	Coronary artery bypass graft occlusion
36523	G311.00	preinfarction syndrome
14897	G301z00	Anterior myocardial infarction NOS
7442	7920200	Saphenous vein graft replacement of three coronary arteries
18118	G311400	worsening angina
15373	662K100	angina control - poor
9276	G31y000	Acute coronary insufficiency
35674	14A3.00	H/O: myocardial infarct <60
18249	7920	Saphenous vein graft replacement of coronary artery
5030	ZV45K00	[V]Presence of coronary artery bypass graft
20095	G330.00	angina decubitus
20416	G340.12	Atherosclerotic heart disease
7320	G343.00	Ischaemic cardiomyopathy
5387	G301.00	other specified anterior myocardial infarction
42304	7929500	Insertion of drug-eluting coronary artery stent
40399	14A4.00	H/O: myocardial infarct > 60
12139	G300.00	acute anterolateral infarction
15349	662Kz00	Angina control NOS
17307	G311200	angina at rest
11610	7920300	Saphenous vein graft replacement of four+ coronary arteries
8935	G302.00	acute inferolateral infarction
18670	7928000	Percut transluminal balloon angioplasty one coronary artery
7634	7920100	Saphenous vein graft replacement of two coronary arteries
8312	7920.11	Saphenous vein graft bypass of coronary artery
36854	G332.00	coronary artery spasm
23078	G34y100	Chronic myocardial ischaemia
10603	792z.00	Coronary artery operations NOS
32450	G33z400	ischaemic chest pain
17872	G301100	acute anteroseptal infarction
18889	G34z000	Asymptomatic coronary heart disease
22828	7929000	Percutaneous transluminal laser coronary angioplasty
9507	G307000	acute non-q wave infarction
60067	793G000	Perc translum ball angio insert 1-2 drug elut stents cor art
12986	G331.00	prinzmetal's angina

33471	792Dz00	Other bypass of coronary artery NOS
24783	G3â€¸11	Arteriosclerotic heart disease
34963	792D.00	Other bypass of coronary artery
22383	G3y..00	Other specified ischaemic heart disease
23892	G304.00	posterior myocardial infarction nos
28138	G34..00	Other chronic ischaemic heart disease
19655	G311.14	angina at rest
17133	G30A.00	mural thrombosis
29421	G344.00	Silent myocardial ischaemia
15661	G310.11	Dressler's syndrome
14898	G305.00	Lateral myocardial infarction NOS
26863	G33z600	new onset angina
8679	7920000	Saphenous vein graft replacement of one coronary artery
29643	G303.00	Acute inferoposterior infarction
17689	G30..17	silent myocardial infarction
13571	G30..16	Thrombosis - coronary
13566	G30..11	Attack - heart
33461	7924	Revision of bypass for coronary artery
33735	7928100	Percut translum balloon angioplasty mult coronary arteries
15754	G34z.00	Other chronic ischaemic heart disease NOS
16408	G32..11	healed myocardial infarction
18842	G35..00	subsequent myocardial infarction
21844	G31y300	Transient myocardial ischaemia
46017	G30yz00	other acute myocardial infarction nos
9414	7921	Other autograft replacement of coronary artery
51515	7920z00	Saphenous vein graft replacement coronary artery NOS
36609	G342.00	Atherosclerotic cardiovascular disease
34803	G30y.00	other acute myocardial infarction
18125	G330000	nocturnal angina
33650	7929100	percut transluminal coronary thrombolysis with streptokinase
9555	G33z500	Post infarct angina
85947	793G200	Perc translum balloon angioplasty insert 1-2 stents cor art
32272	G38..00	postoperative myocardial infarction
3159	792Dy00	Other specified other bypass of coronary artery
31571	792y.00	Other specified operations on coronary artery
61208	793Gz00	Perc translum balloon angioplasty stenting coronary art NOS
30421	G30..13	Cardiac rupture following myocardial infarction (MI)
27951	G31..00	Other acute and subacute ischaemic heart disease
29758	G30X.00	acute transmural myocardial infarction of unspecif site
41221	G30y200	Acute septal infarction
10209	7921200	Autograft replacement of three coronary arteries NEC
32651	7922.11	Allograft bypass of coronary artery
34328	G311300	refractory angina
11048	G331.11	variant angina pectoris
57062	14AJ.00	H/O: Angina in last year
87849	793G100	Perc tran ball angio ins 3 or more drug elut stents cor art
24540	G34y000	Chronic coronary insufficiency
50372	14AH.00	H/O: Myocardial infarction in last year
40429	G301000	acute anteroapical infarction

9413	G31y.00	Other acute and subacute ischaemic heart disease
30330	G309.00	acute q-wave infarct
23579	G310.00	Postmyocardial infarction syndrome
39693	G31y200	Subendocardial ischaemia
28736	G30y000	acute atrial infarction
27977	G31yz00	Other acute and subacute ischaemic heart disease NOS
34633	G34y.00	Other specified chronic ischaemic heart disease
32854	G30B.00	acute posterolateral myocardial infarction
42462	7928200	Percut translum balloon angioplasty bypass graft coronary a
45809	G350.00	Subsequent myocardial infarction of anterior wall
37657	G362.00	Ventric septal defect/curr comp fol acut myocardal infarctn
38609	G351.00	subsequent myocardial infarction of inferior wall
44585	792Bz00	Repair of coronary artery NOS
63467	G306.00	true posterior myocardial infarction
7696	G33z200	Syncope anginosa
35713	G34yz00	Other specified chronic ischaemic heart disease NOS
29553	G366.00	Thrombosis atrium,auric append&vent/curr comp foll acute MI
18913	ZV45700	[V]Presence of aortocoronary bypass graft
23708	G361.00	Atrial septal defect/curr comp folow acut myocardal infarct
36423	G36..00	Certain current complication follow acute myocardial infarct
41835	G384.00	postoperative subendocardial myocardial infarction
72562	G353.00	subsequent myocardial infarction of other sites
59189	G363.00	Ruptur cardiac wall w/out haemopericard/cur comp fol ac MI

7. Myocardial infarction

medcodeid	aurum	Description	snomedtconceptid
2729671000000118		Acute transmural myocardial infarction	1089471000000109
14218681000006110		PACS - primary angle closure suspect	1128111000000103
12122331000006116		Acute ST segment elevation myocardial infarction of inferolateral wall	12238111000119106
12122341000006114		Acute ST segment elevation myocardial infarction of inferoposterior wall	12238151000119107
4354131000006110		Left coronary artery occlusion	123641001
4354151000006115		Right coronary artery occlusion	123642008
4.86121E+14	SK0y.11	Anterior tibial compartment syndrome	12694001
2.12081E+14	G382.00	Postoperative transmural myocardial infarction other sites	129574000
9843961000006118		Acute ST segment elevation myocardial infarction of anterolateral wall	15712881000119105
9843981000006111		Acute ST segment elevation myocardial infarction of lateral wall	15712921000119103
9843991000006114		Acute ST segment elevation myocardial infarction of anteroseptal wall	15712961000119108
9844011000006119		Acute ST segment elevation myocardial infarction of posterior wall	15713041000119103
9844051000006118		Acute ST segment elevation myocardial infarction due to right coronary artery occlusion	15713121000119105
1576271000006117		Cause of Death- Myocardial Infarction	1576271000006101
1576301000006115		Cause of Death- Acute Myocardial Infarction	1576301000006104
9868571000006116		Mural thrombus of left ventricle following acute myocardial infarction	15960981000119105
9.67931E+14	G30B.00	Acute myocardial infarction of posterolateral wall	15990001
256460011	323Z.00	ECG: myocardial infarct NOS	164865005
256453017	3231	ECG: no myocardial infarction	164866006
256454011	3232	ECG: old myocardial infarction	164867002
14015031000006119		Myocardial infarction due to demand ischaemia	16837681000119104
4032016	G32..11	Healed myocardial infarction	1755008
1790421000006113		ACSC	1790421000006109

1793491000006116		FNACS PO	1793491000006100
1849601000006118		No FH: Myocardial infarction	1849601000006102
299712015	G306.00	True posterior myocardial infarction	194802003
299742017	G312.00	Coronary thrombosis not resulting in myocardial infarction	194821006
299745015	G31y000	Acute coronary insufficiency	194823009
1.18831E+14	G35X.00	Subsequent myocardial infarction of unspecified site	194856005
299811016	G350.00	Subsequent myocardial infarction of anterior wall	194857001
299812011	G351.00	Subsequent myocardial infarction of inferior wall	194858006
8.13961E+14	G360.00	Haemopericardium as current complication following acute myocardial infarction	194862000
4.98031E+14	G361.00	Atrial septal defect as current complication following acute myocardial infarction	194863005
1.58601E+14	G363.00	Rupture of cardiac wall without haemopericardium as current complication following acute myocardial infarction	194865003
1.58611E+14	G364.00	Rupture of chordae tendinae as current complication following acute myocardial infarction	194866002
1.59001E+14	G365.00	Rupture of papillary muscle as current complication following acute myocardial infarction	194867006
1.00681E+14	G366.00	Thrombosis of atrium, auricular appendage, and ventricle as current complications following acute myocardial infarction	194868001
37443015	G30..14	Heart attack	22298006
8011041000006111		Acute myocardial infarction due to right coronary artery occlusion	2.3311E+13
299710011	G304.00	Acute posterior myocardial infarction	233838001
5056411000006110		Old anterior myocardial infarction	233839009
5056421000006119		Old inferior myocardial infarction	233840006
5056431000006116		Old lateral myocardial infarction	233841005
5056441000006114		Old posterior myocardial infarction	233842003
5056461000006113		MI - Silent myocardial infarction	233843008
6.7081E+13	G362.00	Ventricular septal defect as current complication following acute myocardial infarction	233846000
5.37751E+14	G30..13	Cardiac rupture after acute myocardial infarction	233847009
2878041000006119		Coronary spasm	23687008
397701010	12C5.00	FH: Myocardial infarction	266897007
2.17381E+14	12C2.12	FH: Myocardial infarction at less than 60	275122004
2.17421E+14	12C3.12	FH: Myocardial infarction at greater than 60	275123009
411056017	12C5.11	FH: Coronary thrombosis	275124003
5605381000006117		Anxiety about having a heart attack	277834005
8087551000006113		History of acute ST segment elevation myocardial infarction	2.85721E+14
8087751000006116		Acute ST segment elevation myocardial infarction involving left anterior descending coronary artery	2.85981E+14
447324018	G309.00	Acute Q-wave infarct	304914007
451369010	14AH.00	H/O: Myocardial infarction in last year	308065005
2.12061E+14	G380.00	Postoperative transmural myocardial infarction of anterior wall	311792005
2.12071E+14	G381.00	Postoperative transmural myocardial infarction of inferior wall	311793000
455422014	G384.00	Postoperative subendocardial myocardial infarction	311796008
6043771000006118		Non-Q wave myocardial infarction	314207007
6.16081E+14	889A.00	Diabetes mellitus insulin-glucose infusion in acute myocardial infarction	315287002
460138014	12CA.00	FH myocardial infarction male first degree age known	315619001
460139018	12CB.00	FH myocardial infarction male first degree age unknown	315620007
460140016	12CC.00	FH myocardial infarction female first degree age known	315621006
460141017	12CD.00	FH myocardial infarction female first degree age unknown	315622004
3024781000006112		Past myocardial infarction diagnosed on ECG AND/OR other special investigation, but currently presenting no symptoms	32574007
5.43291E+14	G36..00	Myocardial infarction with complication	371068009
6546111000006118		ACS - Acute coronary syndrome	394659003
6546751000006118		First myocardial infarction	394710008
6601121000006118		Coronary artery thrombosis	398274000

1738171000006114	14AT.00	History of myocardial infarction	399211009
1780491019	G30X000	Acute ST segment elevation myocardial infarction	401303003
1780501013	G307100	Acute non-ST segment elevation myocardial infarction	401314000
2160096015	SP07600	Coronary artery bypass graft occlusion	408546009
7084801000006112		Mixed myocardial ischaemia and infarction	428196007
7093091000006111		Recent myocardial infarction	428752002
2692615014		Patient facsimile number	429524004
8193681000006118		Family history of myocardial infarction in second degree relative less than 50 years	5.15711E+14
8193701000006115		Family history of myocardial infarction in first degree relative less than 60 years	5.15741E+14
299709018	G301z00	Anterior myocardial infarction NOS	54329005
94884017	G30..00	Acute myocardial infarction	57054005
299711010	G305.00	Acute lateral myocardial infarction	58612006
3459561000006119		Acute myocardial infarction of apical-lateral wall	59063002
4.55651E+14	G301100	Acute anteroseptal myocardial infarction	62695002
3536581000006113		Coronary occlusion	63739005
3551171000006114		Acute myocardial infarction of high lateral wall	64627002
4.57531E+14	G302.00	Acute inferolateral myocardial infarction	65547006
109915012	G310.00	Post-myocardial infarction syndrome	66189004
7510331000006117		History of non-ST segment elevation myocardial infarction	698593009
4.55641E+14	G300.00	Acute anterolateral myocardial infarction	70211005
7571581000006113		Acute anterior ST segment elevation myocardial infarction	703164000
7571611000006117		Acute ST segment elevation myocardial infarction of anterior wall involving right ventricle	703165004
7572221000006117		Subsequent ST segment elevation myocardial infarction of inferior wall	703209002
7572301000006119		Acute myocardial infarction during procedure	703212004
7572331000006110		Acute STEMI (ST elevation myocardial infarction) of inferior wall	703213009
7574041000006111		Mitral valve regurgitation due to acute myocardial infarction	703326006
7574481000006111		Subsequent non-ST segment elevation myocardial infarction	703360004
7597161000006111		Inferior ST segment elevation	705007006
7662871000006115		Inducible ischaemia manifest on stress test post myocardial infarction	710031008
1741581000006112	9hM0.00	Excepted from myocardial infarction quality indicators - informed dissent	7.15971E+14
1741591000006110	9hM1.00	Excepted from myocardial infarction quality indicators - patient unsuitable	7.15981E+14
1741571000006114	9hM..00	Exception reporting - myocardial infarction quality indicators	7.17541E+14
299714019	G308.00	Acute myocardial infarction of inferior wall	73795002
1234005010	G303.00	Acute inferoposterior infarction	76593002
13930031000006112		Delayed postmyocardial infarction pericarditis	827164008
13935161000006114		Acute myocardial infarction of right ventricle	836293000
13935211000006117		Acute myocardial infarction of inferolateral wall with posterior extension	836295007
2173281000000112	14AW.00	H/O acute coronary syndrome	8.37091E+14
13941091000006115		Acute ST segment elevation myocardial infarction due to proximal left anterior descending coronary artery occlusion	840309000
13941141000006111		Acute ST segment elevation myocardial infarction due to mid left anterior descending coronary artery occlusion	840312002
13941231000006112		Acute STEMI (ST elevation myocardial infarction) due to distal LAD (left anterior descending) coronary artery occlusion	840316004
13945211000006116		Acute ST segment elevation myocardial infarction due to occlusion of anterior descending branch of left coronary artery	840609007
13947721000006111		Acute ST segment elevation myocardial infarction due to occlusion of intermediate artery	846683001
13961371000006114		Acute ST segment elevation myocardial infarction due to occlusion of mid portion of right coronary artery	868220007
13961451000006111		Acute STEMI (ST elevation myocardial infarction) due to PDA (posterior descending artery) branch of RCA (right coronary artery) occlusion	868225002
13961461000006113		Acute ST segment elevation myocardial infarction due to occlusion of posterior lateral branch of right coronary artery	868226001
14132941000006118		Myocardial infarction with non-obstructive coronary artery	879955009
9.05351E+14		[RFC] Myocardial infarction (MI)	9.05351E+14

12737271000006118		Primary angle closure suspect	1128111000000103
2701851000006110		ACS - Anterior tibial compartment syndrome	12694001
455423016	G38z.00	Postoperative myocardial infarction, unspecified	129574000
208365015	G38..00	Postoperative myocardial infarction	129574000
2.12091E+14	G383.00	Postoperative transmural myocardial infarction unspec site	129574000
256452010	323..00	ECG: myocardial infarction	164865005
4031011	G32..00	Old myocardial infarction	1755008
6.82481E+14	G311000	Myocardial infarction aborted	194821006
4775931000006111		Aborted myocardial infarction	194821006
2619484018	G311011	MI - Myocardial infarction aborted	194821006
9.31961E+14	G31y099	Acute coronary syndrome	194823009
300882013	Gyu3600	[X]Subsequent myocardial infarction of unspecified site	194856005
299813018	G353.00	Subsequent myocardial infarction of other sites	194856005
300881018	Gyu3500	[X]Subsequent myocardial infarction of other sites	194856005
299808017	G35..00	Subsequent myocardial infarction	194856005
2855351000006111		Myocardial infarct	22298006
2855341000006114		MI - Myocardial infarction	22298006
2.19521E+14	G30..11	Attack - heart	22298006
2855301000006112		Myocardial infarction	22298006
350376014	G30..17	Silent myocardial infarction	233843008
5056501000006113		Post-infarction ventricular septal defect	233846000
39773018	G332.00	Coronary artery spasm	23687008
3.39631E+14	ZV17312	Family history of myocardial infarction	266897007
5495521000006113		Family history: Myocardial infarction	266897007
2.17371E+14	12C2.11	FH: Myocardial infarction < 60	275122004
5569701000006117		Family history: Myocardial infarction at greater than 60	275123009
2.17411E+14	12C3.11	FH: Myocardial infarction > 60	275123009
5935321000006111		Acute Q wave myocardial infarction	304914007
1488382011	G311500	Acute coronary syndrome	394659003
6601131000006115		CT - Coronary thrombosis	398274000
1786198013	G30..16	Thrombosis - coronary	398274000
1786197015	G30..12	Coronary thrombosis	398274000
2.30021E+14	G32..12	Personal history of myocardial infarction	399211009
6651221000006117		STEMI - ST elevation myocardial infarction	401303003
6651391000006114		NSTEMI - Non-ST segment elevation MI	401314000
299707016	G301.00	Acute anterior myocardial infarction	54329005
3381601000006117		Acute myocardial infarction of anterior wall	54329005
3.62461E+14	Gyu3400	[X]Acute transmural myocardial infarction of unspecif site	57054005
8.84141E+14	G30..98	Coronary thrombosis	57054005
2.19531E+14	G30..15	MI - acute myocardial infarction	57054005
8.84151E+14	G30..99	Myocardial Infarction	57054005
299721019	G30z.00	Acute myocardial infarction NOS	57054005
4.60681E+14	G30X.00	Acute transmural myocardial infarction of unspecif site	57054005
299720018	G30yz00	Other acute myocardial infarction NOS	57054005
3427201000006111		AMI - Acute myocardial infarction	57054005
299718016	G30y.00	Other acute myocardial infarction	57054005
3452181000006112		Acute myocardial infarction of lateral wall	58612006
3536601000006115		Coronary artery occluded	63739005
3565871000006113		Acute myocardial infarction of inferolateral wall	65547006

109916013	G310.11	Dressler's syndrome	66189004
3576371000006117		Postmyocardial infarction syndrome	66189004
3641641000006116		Acute myocardial infarction of anterolateral wall	70211005
7571601000006115		Acute STEMI (ST elevation myocardial infarction) of anterior wall	703164000
7572341000006117		Acute inferior ST segment elevation myocardial infarction	703213009
7572321000006112		Acute ST segment elevation myocardial infarction of inferior wall	703213009
7574491000006114		Subsequent NSTEMI (non-ST segment elevation myocardial infarction)	703360004
3699921000006110		Acute inferior myocardial infarction	73795002
3745741000006117		Acute myocardial infarction of inferoposterior wall	76593002
13930051000006117		Dressler's syndrome	827164008
13941161000006110		Acute STEMI (ST elevation myocardial infarction) due to mid LAD (left anterior descending) coronary artery occlusion	840312002
13941211000006118		Acute ST segment elevation myocardial infarction due to distal left anterior descending coronary artery occlusion	840316004

8. Codelsit for Gastroesophageal reflux

medcodeid	cleansedreadcode	snomedetconceptid	term
353140018	J10y413	698065002	acid reflux
4.52891E+14	J101111	698065002	acid reflux
5485511000006116		265357009	anti-reflux surgery
394128017	760L200	265359007	antireflux fundoplication using abdominal approach
394127010	760L000	265358004	antireflux fundoplication using thoracic approach
394129013	760L300	265360002	antireflux gastropexy
4.87911E+14	760L311	9429009	antireflux gastropasty
269068011	760Lz00	265357009	antireflux operation nos
269047010	760L100	265357009	antireflux operation using thoracic approach nec
4.87941E+14	760L.00	265357009	antireflux operations
4.87951E+14	760L400	265357009	antireflux procedure and gastropasty hfq
1221523017	760L111	265357009	antireflux procedure using thoracic approach nec
5485521000006112		265357009	antireflux surgery
3686781000006118		72950008	bile reflux gastritis
2758911000006117		16331000	burning reflux
3082581000006116		36151004	corrosive oesophagitis
5272481000006112		249496004	esophageal reflux
5272491000006110		249496004	esophageal reflux finding
5272521000006112		249496004	esophageal reflux observation
5485491000006110		265357009	esophageal reflux operation
5492771000006113		266433003	esophageal reflux with esophagitis
302369017	J101400	196601009	gangrenous oesophagitis
339001014	1957	225587003	gastric reflux
5082461000006117		235595009	gastro-esophageal reflux
5082481000006110		235595009	gastro-esophageal reflux disease
5492781000006111		266433003	gastro-esophageal reflux disease with esophagitis
5082571000006111		235596005	gastro-esophageal reflux disease with stricture
5492861000006111		266435005	gastro-esophageal reflux disease without esophagitis
5492871000006116		266435005	gastro-esophageal reflux disease without oesophagitis
353143016	J10y412	235595009	gastro-oesophageal reflux
5082501000006117		235595009	gastro-oesophageal reflux disease
5082561000006116		235596005	gastro-oesophageal reflux disease with stricture
7.99341E+14	J101112	266433003	gastro-oesophageal reflux with oesophagitis

5082451000006119		235595009	gastroesophageal reflux disease
5492811000006113		266433003	gastroesophageal reflux disease with esophagitis
5492821000006117		266433003	gastroesophageal reflux disease with oesophagitis
5492881000006118		266435005	gastroesophageal reflux disease without esophagitis
5082441000006116		235595009	gastroesophageal reflux disease
5082541000006115		235595009	gerd - gastro-esophageal reflux disease
5082491000006113		235595009	gor - gastro-esophageal reflux
5082531000006113		235595009	gor - gastro-oesophageal reflux
5082471000006112		235595009	gord - gastro-esophageal reflux disease
5082521000006110		235595009	gord - gastro-oesophageal reflux disease
2532987010	J10y500	414581006	laryngopharyngeal reflux
372246012	J10y411	249496004	oesophageal reflux
5272471000006114		249496004	oesophageal reflux finding
5272511000006116		249496004	oesophageal reflux observation
5485501000006119		265357009	oesophageal reflux operation
2.67111E+14	760L.11	265357009	oesophageal reflux operations
11904971000006117		265357009	oesophageal reflux operations
396339011	J101113	266433003	oesophageal reflux with oesophagitis
28383015	J101.00	16761005	oesophagitis
302381018	J101z00	16761005	oesophagitis nos
2.66601E+14	J10y400	266435005	oesophageal reflux without mention of oesophagitis
269067018	760Ly00	265357009	other specified antireflux operation
302378011	J101y00	16761005	other specified oesophagitis
269073017	760My00	173704005	other specified revision of antireflux operation
497946018	J101114	266433003	peptic oesophagitis
302370016	J101500	196602002	phlegmonous oesophagitis
194802019	J101300	3482005	postoperative oesophagitis
9.60261E+14		9.60261E+14	reflux
3.03321E+14	171J.00	2.00151E+14	reflux cough
5492761000006118		266433003	reflux esophagitis
396340013	J101100	266433003	reflux oesophagitis
198466017	J101115	266433003	regurgitant oesophagitis
269074011	760Mz00	173704005	revision of antireflux operation nos
269069015	760M.00	173704005	revision of antireflux operations
498970016	J101600	439955006	ulcerative oesophagitis



Medicines & Healthcare products
Regulatory Agency



1

General
information

Protocol reference Id

20_000108

Study title

To investigate the prevalence, incidence and mortality for Interstitial Lung Disease (ILD) in the UK for the study period 2009 to 2019

Research Area

Disease Epidemiology

Does this protocol describe an observational study using purely CPRD data?

Yes

Does this protocol involve requesting any additional information from GPs, or contact with patients?

No

Role	Chief Investigator
Title	Reader
Full name	Jennifer Quint
Affiliation/organisation	Imperial College London
Email	j.quint@imperial.ac.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Corresponding Applicant
Title	Doctoral student
Full name	Rikisha Shah Gupta
Affiliation/organisation	Imperial College London
Email	r.shah20@ic.ac.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Consultant
Full name	Peter George
Affiliation/organisation	Royal Brompton Hospital
Email	p.george@rbht.nhs.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Post-doctoral Research Associate
Full name	Ann Morgan
Affiliation/organisation	Imperial College London
Email	a.morgan15@imperial.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

Sponsor

Imperial College London

Funding source for the study

Is the funding source for the study the same as Chief Investigator's affiliation?

Yes

Funding source for the study

Imperial College London

Institution conducting the research

Is the institution conducting the research the same as Chief Investigator's affiliation?

Yes

Institution conducting the research

Imperial College London

Method to access the data

Indicate the method that will be used to access the data

Institutional multi-study licence

Is the institution the same as Chief Investigator's affiliation?

Yes

Institution name

Imperial College London

Extraction by CPRD

Will the dataset be extracted by CPRD

No

Multiple data delivery

This study requires multiple data extractions over its lifespan

No

Data processors

Data processor is	Same as the institution conducting the research
Processing	Yes
Accessing	Yes
Storing	Yes
Processing area	UK

Primary care data

CPRD GOLD

CPRD Aurum

Do you require data linkages

Yes

Patient level data

HES Admitted Patient Care

ONS Death Registration Data

NCRAS data**Covid 19 linkages****Area level data****Do you require area level data?**

No

Practice level (UK)**Patient level (England only)**

Withheld concepts

Are withheld concepts required?

No

Linkage to a dataset not listed

Are you requesting a linkage to a dataset not listed?

No

Patient data privacy

Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?

No

Lay Summary

Interstitial Lung Diseases (ILD) are a large group of conditions characterised by scarring of the lungs. In some cases, the cause for the lung scarring is clear. For example, problems with the immune system such as rheumatoid arthritis or working with asbestos. However, in many cases, the cause is unknown (Idiopathic pulmonary fibrosis, IPF). The burden of all-encompassing ILD and its various types is not well established in the UK.

For this research, we propose to use information from primary care (Clinical Practice Research Datalink) and secondary care databases (Hospital Episode Statistics and Office of National Statistics) to investigate how many people in the UK have interstitial lung disease, and what type of ILD diseases they have. We will also explore the number of people diagnosed with specific types of ILD over the past 10 years, ascertain any trends during this time and whether these disease patterns differ by different types of ILD and how many have died as a result of their ILD. We hope that through this study we will develop a better understanding of burden of the various conditions that come under the heading of interstitial lung disease and in so doing support the further research that is needed in this area.

Technical Summary

ILD is an umbrella term which encompasses several conditions with different clinical pathways leading to varied health outcomes. Pharmacological intervention varies depending on the underlying diagnosis, disease severity and trajectory and patient choice. In this study, we will identify an ILD cohort which represents the range of different conditions and use it to estimate the prevalence and incidence of all ILD. Given the heterogenous nature of the conditions and potential inaccuracies in diagnostic coding it is considered likely that reliance on simple lists of clinical diagnostic codes may lead to potential under- and over- ascertainment of cases. For this reason, this research will be conducted in parallel with an ILD code validation study (ID number 20_000068) to assess the reliability of identifying ILD in primary and secondary care data. Further, we plan to stratify the overall ILD cohort by ILD subtypes.

Preliminary literature review demonstrates incidence and prevalence estimates vary by ILD disease type and geography. The current challenges in ILD diagnoses is due to its various entities and added complexity due to lack of understanding. As well, recent papers have reported an increase in the observed overall ILD cases. Whether this is an actual increase in ILD cases or due to improved coding practices through the years is unknown.

We plan to estimate the incidence and prevalence by overall ILD and by its predefined subtypes. We will also estimate and all-cause mortality in overall ILD cohort and its sub groups.

To comprehend the seriousness of ILD burden in the UK, we aim to enhance ILD evidence using CPRD-HES APC-ONS linked data. We hope that this research will generate epidemiological evidence to fill in the current evidence gaps in the UK.

Outcomes to be measured

The primary outcome of this study is to establish the epidemiology of ILD by predefined aetiological subgroups in the UK for the period from 2009 to 2019 defined by incidence, prevalence and all-cause mortality in ILD cohort (overall ILD and its subgroups).

Objectives, specific aims & rationale

The primary objective is to generate epidemiological evidence for overall and predefined ILD subgroups in the UK. The objectives and aims for this study are listed below:

- 1) Calculate the incidence of overall ILDs and by aetiological subgroups for the 10-year study period.
- 2) Estimate mortality rates in the overall ILD cohort and by subgroups over the study period.
- 3) Calculate the prevalence of all ILDs and for the subgroups for the 10-year study period.

Rationale: This study will help us to better understand the disease burden of all-encompassing ILD and its various subtypes in the UK.

Study background

Interstitial lung disease (ILD) encompasses a heterogeneous group of disorders, ranging from conditions which completely resolve without requirement for pharmacological intervention through to fibrotic lung diseases which inexorably progress to respiratory failure and death despite treatment. Hence ILD is often used as an umbrella term to represent a diverse group of lung conditions with different aetiologies, unpredictable progression and varying survival times. Estimation of the disease burden of ILD, in particular of specific disease subtypes, is challenging for several reasons such as inaccuracy in coding practices for less prevalent ILDs. Table 1 (below) demonstrates the aetiological ILD classification [1].

- i) Idiopathic Interstitial Pneumonia (IIP) : IPF, Idiopathic pulmonary non specific interstitial Pneumonia(iNSIP)
- ii) Exposure related ILDs : Hypersensitivity pneumonitis (HP), Asbestosis, Silicosis, Pneumoconiosis
- iii) Autoimmune ILDs: Rheumatoid arthritis ILD (RA-ILD), Systemic sclerosis ILD (SSc-ILD)
- iv) Sarcoidosis
- v) Other ILDs : Drug related ILDs, Cystic lung diseases (Lymphangioleiomyomatosis, Birt-Hogg-Dube, lymphocytic interstitial pneumonia)

The most common ILD with progressive fibrosis is IPF, IPF has attracted the most research interest in recent years. This is likely because not only is it the most prevalent of the ILDs but also has a universally progressive nature and a poor prognosis. IPF is regarded as a useful model for other progressive fibrosing ILDs [2], the lack of specific epidemiological evidence relating to other non-IPF ILDs reflects there is much unknown.

Estimation of the disease burden of ILD, in particular of specific disease subtypes, is challenging for several reasons, such as, absence of standardization, variable methods for data collection and diagnostic changes over time in the study of ILD epidemiology. In the context of electronic health records these difficulties are compounded by inaccuracies and inconsistencies in coding practices, especially for the less prevalent ILDs. This has resulted in lack of adequate published literature to quantify ILD burden. According to Guo B, Wang L, Xia S, et al. the incidence of ILD ranges from 4.6 to 31.5 per 100,000 persons in Europe and North America [3]. A systematic literature review by Olson et al, reported overall a prevalence range of IPF in Europe as 0.63 – 7.6 per 100,000 (extracted from 4 studies) [4]. Disparity in the IPF and ILD estimates across Europe affirm the above-mentioned challenges.

According to GBD, incidence estimates for ILD increased by approximately 18 % (5.6 per 100,000 persons in 2007 to 6.6 per 100,000 in 2017) from 2007 – 2017[5]. Another study conducted in Denmark, reported a rise in incidence from 38 to 66 per 10,000 persons during the six-year observation period (2003 - 2009). This new rise in cases it is expected to increase the burden on health care resources. However, it is difficult to assess whether the rise in ILD cases is an actual surge or attributed to better diagnostic methods, enhanced awareness and improved classification.

The primary aim of this study is to establish the epidemiological outcomes, such as, incidence, prevalence and mortality in ILD patients in the UK, using CPRD-HES APC-ONS linked data. We hope to fill in the gap in understanding of ILD burden and variation across different ILD conditions, beyond IPF.

Study type

Exploratory and descriptive

Study design

This will be a retrospective observational cohort study

Feasibility counts

In previous study using CPRD GOLD data from 2006 – 2016 investigating a cohort of patients with ILD, there were 10,700 patients who were followed up for a mean of 2.2 years[6].

Sample size considerations

Table below estimates the sufficient sample size for calculating the prevalence of ILD in the UK, with marginal error not exceeding 1% (95% confidence interval). Due to lack of UK data, the calculations are based on estimates for the European Union (EU) from the GBD study. ILD prevalence rate for EU is reported as 13.98 per 10,000 persons (2017). To ensure we have enough sample for most frequent as well as rare ILD conditions, we have calculated sample size for prevalence ranging from 5 per 10,000 persons to 20 per 10,000 persons. Power Literature Prevalence rate, 0.6 0.7 0.8 (estimated sample size) 5 288 336 384 10 576 672 768 15 864 1,008 1,152 20 1,552 1,345 1,537 *formula used for estimating sample size[7]: $n = (1.962 \times \text{published prevalence rate} \times \text{power}) / (\text{marginal error})^2$ [7] Based on the sample size calculation we have an adequate sample to get an 80% power.

Planned use of linked data and benefit to patients in England and Wales

Access to HES APC and ONS data is requested to identify ILD patients not captured in the primary care data. Primary care data may be lacking the detail for identifying ILD subtype, therefore CPRD-HES APC linked data will be used to identify additional cases and classify ILD patients, by subtype. ONS will be accessed to identify ILD-related deaths not captured in the CPRD during the study period.

This work aims to provide the policy workers with a better understanding of ILD burden in England and to support with evidence for improved patient care.

Definition of the study population

The study population will be patients registered with a GP practice participating in the CPRD or patients identified from HES APC or ONS data. Inclusion and exclusion criteria are summarised below –

Inclusion criteria for ILD cohort –

Patients with diagnosis of ILD (provisional code list in Appendix 1)

Patients aged at least 18 years old on diagnosis of ILD

Patients will be eligible to enter the cohort after the latest of: 1 Jan 2009, continuous CPRD registration, index date of their ILD diagnosis and their up to standard registration date (this is applicable for CPRD GOLD only. For CPRD AURUM, additional data checks will be conducted).

Exclusion criteria for ILD cohort –

Patients that do not meet the acceptable quality standard

Patients observed outside the study period, 2009 - 2019

Patients observed after the last data collection date of the contributing practice

Incident cases will be reported after one year of data collection, this will ensure the recorded diagnosis is new.

An accurate ILD cohort will be derived using the diagnosis codes that will be validated in a separate ILD code validation study.

Selection of comparison groups/controls

This is a descriptive study to establish epidemiological evidence within ILD subgroups. There is no plan for comparison.

Exposures, outcomes and covariates

Demographic information [Patient Data] : Gender, year of birth, patient registration date, date of death and practice location
Date of death [ONS Death Registration Data] : Death date and cause of death
Practice [Practice] : Information about the general practice (primary care setting), quality of data entry from the contributing practice.
Clinical information [Clinical] : Diagnosis of ILD, ILD subtypes
Hospitalisation [HES Admitted Patient Care] : Reason for hospitalisation
Referral information [Referral] : Diagnosis and reason for referral

Data/statistical analysis

All the data programming will be carried out in R and/or Stata. The statistical analyses as they refer to each objective of the study are presented below –

Objective 1 Incidence:

The number of new cases for the 10-year study period and for each calendar year will be estimated. New cases will be included after allowing one year of data collection from the contributing practice. Incidence will be reported as number of cases per 100,000 person-years for ILD subgroups. Depending on data availability incidence trend for the 10-year period will be presented.

Objective 2 Mortality:

The all-cause mortality rate for the 10-year period and by calendar year will be calculated in the predefined ILD cohort. If data permits, the mortality trend for the study period will be reported. The rate will be further stratified into aetiological subgroups depending on the data availability.

Objective 3 Period Prevalence:

Prevalent cases will be observed for the study period. Prevalence will be reported as number of prevalent cases at the end of each year per 100,000 persons.

Plan for addressing confounding

This is an exploratory, no regression or other multifaceted analysis is planned for this research.

Plans for addressing missing data

This study will be carried out in parallel with ILD code validation study (submitted, ID number 20_000068) to allow a definitive overall ILD and its disease types cohort identification. We intent to use complete cases only.

Patient or user group involvement

Findings from this study will be shared with relevant patient groups. The key findings from this study will be discussed in ILD clinics for better understanding of ILD.

Plans for disseminating & communicating

The results of this study will be disseminated in peer-reviewed journals and at scientific conferences as soon as possible following the completion of the analyses.

Conflict of interest statement

No conflict of interest

Limitations of study design

This study is conducted using clinical definitions based on exploratory findings, literature review and diagnosis code validation study to identify patients with ILD in CPRD. The possibility of a lack of sufficient granularity in diagnostic coding to permit accurate ILD subgroup classification cannot be excluded. These limitations will be noted in any publications/reports arising from the study.

References

1. Flaherty, K. R., Wells, A. U., Cottin, V., Devaraj, A., Walsh, S. L. F., Inoue, Y., . . . Investigators, I. T. (2019). Nintedanib in Progressive Fibrosing Interstitial Lung Diseases. *N Engl J Med*, 381(18), 1718-1727. doi:10.1056/NEJMoa1908681
2. Raghu, G., Nyberg, F., & Morgan, G. (2004). The epidemiology of interstitial lung disease and its association with lung cancer. *Br J Cancer*, 91 Suppl 2, S3-10. doi:10.1038/sj.bjc.6602061
3. Guo, B., Wang, L., Xia, S., Mao, M., Qian, W., Peng, X., Zheng, Z., Chen, R., Han, Q., & Luo, Q. (2020). The interstitial lung disease spectrum under a uniform diagnostic algorithm: a retrospective study of 1,945 individuals. *Journal of thoracic disease*, 12(7), 3688–3696. <https://doi.org/10.21037/jtd-19-4021>
4. Olson, A., Hartmann, N., Schlenker-Herceg, R., & Wallace, L. (2018). Prevalence of progressive fibrosing interstitial lung disease. *European Respiratory Journal*, 52(suppl 62), PA3030. doi:10.1183/13993003.congress-2018.PA3030
5. <http://ghdx.healthdata.org/gbd-results-tool>
6. Strongman H, Kausar I & Maher TM. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. *Adv Ther*. 2018 May;35(5):724-736
7. Hajian-Tilaki K. (2011). Sample size estimation in epidemiologic studies. *Caspian journal of internal medicine*, 2(4), 289–298

Appendices

 appendix-1-ild-epidemiology-isac.pdf

Grant ID

9.4. Appendix 4. Supplementary material to Chapter 6

Approved Independent Scientific Advisory Committee (ISAC) protocol for chapter 5 (ISAC protocol 23_003093)



Medicines & Healthcare products
Regulatory Agency



1 General information

Protocol reference Id

23_003093

Study title

Changes in survival, and causes of death in people with idiopathic pulmonary fibrosis in England, 2010 to 2022

Research Area

Disease Epidemiology

Methodological

Does this protocol describe an observational study using purely CPRD data?

Yes

Does this protocol involve requesting any additional information from GPs, or contact with patients?

No

Role	Chief Investigator
Title	Reader
Full name	Jennifer Quint
Affiliation/organisation	Imperial College London
Email	j.quint@imperial.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

Role	Corresponding Applicant
Title	Doctoral student
Full name	Rikisha Shah Gupta
Affiliation/organisation	Imperial College London
Email	r.shah20@ic.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

Role	Collaborator
Title	Consultant
Full name	Peter George
Affiliation/organisation	Royal Brompton Hospital
Email	p.george@rbht.nhs.uk
Will this person be analysing the data?	No
Status	Confirmed

Role	Collaborator
Title	Post-doctoral Research Associate
Full name	Ann Morgan
Affiliation/organisation	Imperial College London
Email	a.morgan15@imperial.ac.uk
Will this person be analysing the data?	Yes
Status	Confirmed

Sponsor

Imperial College London

Funding source for the study

Is the funding source for the study the same as Chief Investigator's affiliation?

No

Funding source for the study

Boehringer Ingelheim GmbH

Institution conducting the research

Is the institution conducting the research the same as Chief Investigator's affiliation?

Yes

Institution conducting the research

Imperial College London

Method to access the data

Indicate the method that will be used to access the data

Institutional multi-study licence

Is the institution the same as Chief Investigator's affiliation?

Yes

Institution name

Imperial College London

Extraction by CPRD

Will the dataset be extracted by CPRD

No

Multiple data delivery

This study requires multiple data extractions over its lifespan

No

Data processors

Data processor is	Same as the chief investigator's affiliation
Processing	Yes
Accessing	Yes
Storing	Yes
Processing area	UK

Primary care data

CPRD Aurum

Do you require data linkages

Yes

Patient level data

HES Admitted Patient Care

ONS Death Registration Data

NCRAS data**Covid 19 linkages****Area level data****Do you require area level data?**

Yes

Practice level (UK)**Patient level (England only)**

Patient Level Index of Multiple Deprivation

Withheld concepts

Are withheld concepts required?

No

Linkage to a dataset not listed

Are you requesting a linkage to a dataset not listed?

No

Patient data privacy

Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?

No

Lay Summary

Idiopathic pulmonary fibrosis (IPF) is a lung disease where scar tissue forms in the lungs, making it hard to breathe. As the disease progresses, the lung tissue becomes thickened and stiff, making it difficult for oxygen to pass into the bloodstream. The exact cause is still not fully understood, although it is believed to involve a combination of genetic and environmental factors. It primarily affects older adults. Most survive for about three to five years after diagnosis. Treatments, called antifibrotics aim to slow progression and improve quality of life. Despite the seriousness of the condition and poor prognosis until recently there hasn't been much research. Until a decade ago there were no guidelines on how to diagnose Idiopathic Pulmonary Fibrosis. We conducted a prior study to understand how general practitioners (GPs) record Idiopathic Pulmonary Fibrosis and after observing changes in coding practices calculated updated incidence and prevalence. We hope through this study to develop a better understanding of disease survival and assess the impact of treatments on survival. We will use linked primary and secondary care data and mortality data to calculate the mortality rates of people with idiopathic pulmonary fibrosis and see how this has changed over time as well as explore what people with idiopathic pulmonary fibrosis die of and how this has changed over time. This will help us better understand how we can improve survival in this population.

Technical Summary

Idiopathic pulmonary fibrosis (IPF) is a lung disease where scar tissue forms in the lungs, making it hard to breathe. People diagnosed with IPF have a historic median survival of less than 3-5 years from diagnosis. We conducted a prior study to understand how general practitioners (GPs) record IPF (ID number 20_000068) and from this derived four algorithms with varying accuracy from which to determine someone has a diagnosis. These 4 algorithms are: i) individuals with a diagnosis of pulmonary fibrosis of any aetiology (not necessarily IPF diagnosis) in Aurum and a primary or secondary discharge diagnosis of IPF in HES APC, known as 'Aurum and HES' ii) individuals with an evidence of at least one clinical code strongly indicative of an IPF diagnosis in Aurum, known as 'Aurum narrow' iii) individuals with a diagnosis of pulmonary fibrosis of any aetiology (not necessarily IPF diagnosis) in Aurum (evidence of at least one clinical code with a possible diagnosis of IPF) known as 'Aurum broad' and, iv) individuals discharged with a primary or secondary diagnosis of IPF in HES APC, known as 'HES APC'.

In this study, we will use linked Aurum, hospital episode statistics inpatient data (HES APC) Office for national statistics (ONS) mortality data and deprivation data (IMD). We will investigate changes in survival and causes of death over the past 10 years using the definitions from the validation study. Each of the following aims will be assessed:

- i) Determine median survival in patients with IPF
- ii) Determine year-on-year all-cause mortality
- iii) Investigate changes in causes of death

Understanding the causes of death and survival over time will better help us understand how these people are currently managed and will inform us whether treatment developments may have had an impact on the IPF field.

Outcomes to be measured

Our overall study period is from 1 January 2010 to the end of the available linked data. Our study period will include people who have a new diagnosis of IPF (incident cases). Findings from another study (ISAC protocol 20_000068) will be used to define 4 IPF study cohorts. Outcomes that will be assessed in these 4 IPF study cohorts are

1. Year-on-year survival over the study period
2. Year-on-year all-cause mortality
3. Change in causes of death over the study period

Objectives, specific aims & rationale

• Objectives

To investigate changes in survival and causes of death in people diagnosed with IPF over the last 10 years in England.

• Specific aims

These aims will be assessed for the 4 pre-specified IPF study cohorts:

1. To determine year-on-year median survival
2. To determine year-on-year all-cause mortality rates
3. To investigate changes in causes of death

• Rationale

As per our knowledge, there is no recent data to verify whether there has been an improvement in survival and mortality of patients with IPF after the availability of the two antifibrotic treatments in the UK. Whether the developments in the IPF field over the past decade have translated to better patient management, improved disease outcomes and better quality of life is still unknown.

Understanding changes in the causes of death and survival over time in people with IPF will help us to better understand whether we see any influence of the antifibrotic treatments in patients living with this chronic condition.

Study background

IPF is the most common, severe, progressive, and fibrotic interstitial lung disease (ILD) with significant morbidity and poor survival. This disease is characterized by recurrent exacerbations and progressive symptoms resulting in a poor quality of life. Patients diagnosed with IPF have a historic median survival of 3-5 years¹. Pharmacological intervention varies depending on the underlying diagnosis, disease severity and trajectory, and patient choice. A recent analysis [ISAC protocol 20_000108] showed IPF specific mortality rate (per 100,000 person-years) standardized to the 2018 England population increased from 5.2 per 100,000 person-years in 2008 to 7.93 in 2018 per 100,000 person-years. We observed a 1.5-fold increase in IPF-specific mortality in 2018 as compared to 2008. Strongman et al^[1] reported for the incident cases diagnosed from 2000 to 2012, the median survival was about 3.0 years. These authors reported the all-cause mortality rate in incident IPF patients was higher in patients who were diagnosed in 2000 than in the later years. However, the authors did not observe any improvements in survival over the same period. Whereas Navaratnam et al^[2] reported IPF-specific mortality rates for IPF from 1979 to 2016 was 4.68 per 100,000 person-years and the IPF mortality rose by 5% per year after adjusting for age. Comorbidities have been reported in up to 89% of patients with IPF^[3]. These have a negative impact on patients with IPF due to increased risk of mortality^[4-6], significant effect on health-related quality of life^[7], and growing healthcare costs^[8]. The most commonly observed conditions are COPD, GERD and cancer, some of these conditions can develop as a consequence of IPF (e.g., pulmonary hypertension), and others may be explained as risk factors for IPF (e.g. lung cancer, GERD)^[3].

Historically, treatment options, awareness of the disease, guidelines for diagnosis, and its management have been limited; however, with reimbursement approval of antifibrotic treatments firstly pirfenidone in 2013 and nintedanib in 2015; release of ATS-ERS IPF guidelines followed by the publication of National Institute for Health and Care Excellence (NICE) guidelines on the diagnosis and management of suspected IPF^[9-13], UK has seen a number of important developments in the IPF field. Despite these developments in the IPF field for the past 10 years, it is still unknown whether there have been changes in overall patient survival over time. As per our knowledge, there is no recent data to study the impact of this improvement in IPF after 2012. There is a paucity of data on mortality in patients with IPF after the availability of anti-fibrotic treatments and whether they had an impact on survival.

Study type

This study is primarily a descriptive study

Study design

This is a retrospective cohort study using CPRD Aurum, HES APC, and ONS-linked data (Appendix1).

Feasibility counts

Based on previous work, we anticipate that our base study population will comprise a cohort of around 25,000 people with a diagnosis of IPF.

Sample size considerations

This is a descriptive study; we are not testing a hypothesis. We plan to include all individuals who meet this study criteria. We will ensure that we preserve confidentiality at the reporting stage. • In literature, the median survival of IPF patients is reported between 2.7 (95% CI 2.5-3.0) to 3 (95% CI 2.8-3.1) years. • Assuming 95% confidence interval, CI width (confidence interval) ranges from 0.3 (i.e. 2.8 to 3.1) to 0.5 (i.e. 2.5 to 3.0) please see below the expected sample size to give us reasonable precision. CI width N 0.2 1,629 0.3 754 0.4 431 0.5 284 Based on our previous work we anticipate that our base study population will comprise a cohort of around 25,000 people with a diagnosis of IPF. We, therefore, have a sufficient sample size for our analyses.

Planned use of linked data and benefit to patients in England and Wales

HES APC provides data on all admissions to NHS hospital hospitals in England (acute, mental and primary care trusts, including private and charitable hospitals). Linked HES APC data are required to identify patients who were admitted to the hospital for IPF. The ONS dataset contains information on the date, place, and cause of death extracted from death certificates for all deaths registered in England and Wales. ONS data includes information on both the underlying cause of death as well as other contributory causes of death. Linked ONS data is required to evaluate the outcomes of interest. Reliance on the primary database CPRD Aurum alone, may not be sufficient to find the IPF cases and assess the mortality outcome, hence their linkage to secondary databases will provide a comprehensive view. IMD data provide information on socioeconomic status.

Phase 2 and 3 trials assessed whether the treatment with antifibrotic treatments improved lung function or reduced disease progression over time[14-17]. Improvement in patient survival was not a primary endpoint. But a post-marketing study using IPF – PRO registry in the United States reported a reduction in the risk of death [HR, 0.53] in patients who received antifibrotic treatment[18]. This work will be of benefit to patients in England and Wales as it will provide the policymakers with a better understanding of IPF mortality and if there are potentially modifiable areas for clinical improvement.

Definition of the study population

The study population will include all men and women 40 years of age and over who are registered with a GP in England and have data available in the Aurum database. Individuals were included if they were i) at least 40 years at the date of IPF diagnosis, ii) met either of the 4 IPF case definitions (as described in the section exposure, outcomes and covariates), and iii) alive at any time between 1 January 2010 and 31 December 2022. Patients will be considered eligible if they were registered to a practice, the practice they were registered to had passed CPRD quality check, had no missing sex information, and if their first IPF diagnosis was recorded during the study period (incident cases only). Patients will be followed, or the patient start date will be the latest of the following dates: study start date, the patient's eighteenth birthday, the date a patient joined the practice or the date of first diagnosis with IPF [Appendix 2]

Selection of comparison groups/controls

We have no comparison or controls for this study. This descriptive study will assess the aims in four pre-defined IPF cohorts.

Exposures, outcomes and covariates

1. Exposure

Our main exposure of interest is IPF including possible and strongly indicative diagnoses. We will use validated diagnostic-code-based algorithms from a previous analysis to detect 'true' IPF cases (ISAC 20_000068) [Appendix 3]. The validation study recommended two algorithms that had a high positive predictive value. The first algorithm is based on a combination of a broad list of clinical codes in the Aurum database [Appendix 4] along with primary or secondary discharge diagnosis in any "episodes" within a hospital spell in HES database (ICD 10 codes J84.1, J84.8 or J84.9). Second algorithm is based on a focused list of clinical codes to detect IPF in the Aurum database [Appendix 3]. Additionally, we will also include two broader case definitions with higher sensitivity, i) a set of more generic, less focused clinical codes denoting a diagnosis of pulmonary fibrosis of any aetiology, not necessarily IPF in the primary care database, defined as 'broad Aurum cohort' [Appendix 4] and, ii) ICD-10 codes, J84.1, J84.8 and J84.9 to find hospitalizations for IPF in a secondary hospital database, defined as the 'HES APC' cohort.

2. Outcomes: Death date and cause of death (underlying and other causes of death).

We will determine the commonest causes of death within each IPF cohort (reporting n and %) using information derived by ONS. We will use information from all causes mentioned on the death certificate (proposed causes). We will categorize the causes of death by ICD-10 chapters of interest– i) disease of the respiratory system, ii) neoplasms, iii) disease of the circulatory system, iv) disease of the musculoskeletal system and connective tissue.

3. Covariates:

For the eligible study population, we will assess the prevalence of various risk factors and comorbidities as recorded in Aurum. A total of 12 risk factors and comorbidities will be assessed.

All codelists are available via GitHub

(<https://github.com/NHLI-Respiratory-Epi/Validation-of-the-recording-of-Idiopathic-Pulmonary-Fibros>)

Sex, age, region, ethnicity, socioeconomic status

- Smoking status (current/ex-/never)

Risk factors and comorbidities:

- Respiratory conditions: Chronic obstructive pulmonary disease, asthma, lung cancer
- Cardiovascular conditions: heart failure, ischemic heart disease, myocardial infarction, pulmonary arterial hypertension
- Gastrointestinal conditions: gastroesophageal reflux
- Cerebrovascular conditions: ischemic stroke
- Others: Smoking, BMI

Covariates will be assessed at 3 study time points to see at the population level whether the distribution of comorbidities and risk factors may or may not have changed over time.

Data/statistical analysis

Some will be descriptive analyses. All statistical analyses will be performed in STATA (StataCorp, College Station, Texas, USA).

We will have 4 groups: i) individuals with a diagnosis of pulmonary fibrosis of any aetiology (not necessarily IPF diagnosis) in Aurum and a primary or secondary discharge diagnosis of IPF in HES APC, known as 'Aurum and HES' ii) individuals with an evidence of at least one clinical code strongly indicative of an IPF diagnosis in Aurum, known as 'Aurum narrow' iii) individuals with a diagnosis of pulmonary fibrosis of any aetiology (not necessarily IPF diagnosis) in Aurum (evidence of at least one clinical code with a possible diagnosis of IPF) known as 'Aurum broad' and, iv) individuals discharged with a primary or secondary diagnosis of IPF in HES APC, known as 'HES APC'. Each of the aim will be assessed in these 4 IPF groups.

Aim 1: To determine median survival in patients with IPF

Using linked data, we will calculate the median survival for each of the 4 IPF cohorts. We will calculate median survival and compare by incident year. For example, we will compute the survival time for patients with diagnosed in 2012 to 2022 and then compare whether the patients diagnosed in later years had higher median survival than patients diagnosed in 2012. Patients will be followed from the latest of the following dates: the study start date, the patient's eighteenth birthday, the date a patient joined the practice or the date of first diagnosis with IPF. Patients will be censored at the study end date; the study end date is the earliest of the date they left the practice or the last date that the practice contributed data to Aurum prior to analysis.

Aim 2: To determine year-on-year all-cause mortality rates

Using linked data, we will analyze the all-cause mortality trends from 2010 through to 2022 among people with IPF for each of the 4 IPF cohorts. We will assess whether there is an increasing trend in mortality overall compared to the general population. We will calculate comparative mortality rates in women and men in England and Wales to mortality rates in the general population from ONS data. We will calculate crude all-cause mortality rates by gender and age categories (< 75 years and 75 years) for each IPF cohort by calculating the number of deaths in each IPF cohort in the linked dataset and dividing this by the midpoint population for that year. We will compare results to ONS published data on all-cause mortality figures for the same gender and age bands in England and Wales in the given years. Additionally, using direct standardization (age-specific rates), we will calculate comparative mortality figures separately for men and women and by socioeconomic status.

Aim 3: To investigate changes in causes of death

We will determine the proportions of deaths due to specific causes in people with IPF for each of the 4 cohorts. We will specifically investigate the following ICD 10 chapters: i) disease of the respiratory system, ii) neoplasms, iii) disease of circulatory system, iv) disease of the musculoskeletal system and connective tissue. We will determine the commonest causes of death within each IPF cohort (reporting n and %) using information derived by ONS. We will use information from all causes mentioned on the death certificate (proposed causes).

Crude and age-adjusted mortality rates in men and women will be calculated. Cause-specific mortality rates for above mentioned categories will be determined by dividing the number of deaths due to each cause by the total person-years of follow-up. We will determine the cumulative incidence function to allow adjustment for competing risks and calculate the cumulative probability of dying from a specific cause at each time point having survived to that time without a death from any other cause. This will be derived from baseline functions and estimates of hazards from cause-specific Cox regression models. Trends in the cause of death will be assessed using 2 tests.

Plan for addressing confounding

As this is a descriptive study, we will not be addressing confounders but as numbers allow, we will stratify for age, sex, or socioeconomic status.

Plans for addressing missing data

The number and proportions of missing data will be described for demographics and other covariates variables at baseline. No plan to apply any imputation method.

Patient or user group involvement

Findings from this study will be shared with relevant patient groups. The key findings from this study will be discussed in ILD clinics for a better understanding and awareness of IPF.

Plans for disseminating & communicating

The results of this study will be disseminated in peer-reviewed journals and at scientific conferences as soon as possible following the completion of the analyses. We will target external scientific meetings, for example for the American and European Respiratory Society Meetings, and within academic institutions in general. There are no restrictions on the extent or timing of publication. We will not include any cells with counts less than five due to anonymity concerns.

Conflict of interest statement

Rikisha Shah is an employee of Gilead Sciences however this work will be conducted as part of her Ph.D. studies at Imperial College London.

Limitations of study design

- We relied on GP recording or ICD-10 codes recorded at hospital admission to identify IPF cases. We applied the finding from a separate ILD validation study to define patients with IPF in the linked Aurum, HES, and ONS databases. But we cannot confirm the validity of the diagnosis. We are limited by the acumen of the reviewing clinician recording the diagnosis. However, previous validation studies have reported chronic conditions are well recorded in Aurum [19,20].
- Due to changes to IPF diagnostic criteria there is a possibility that some incident cases were misclassified and were prevalent cases.
- Misclassification of covariates may be possible. Depending on how often this occurs, in such cases, we may use an unknown category.
- All these limitations will be mentioned in all reports of study findings but in view of the limited data on this population, we feel this study will still offer value to the public, even after taking these limitations into account.

References

1. Strongman H, Kausar I, Maher TM. Incidence, prevalence, and survival of patients with idiopathic pulmonary fibrosis in the UK. *Adv Ther.* 2018;35(5):724–736.
2. Navaratnam V, Hubbard RB. The Mortality Burden of Idiopathic Pulmonary Fibrosis in the United Kingdom. *Am J Respir Crit Care Med.* 2019 Jul 15;200(2):256-258. doi: 10.1164/rccm.201902-0467LE. PMID: 30973756.
3. Prior TS, Hoyer N, Hilberg O, Shaker SB, Davidsen JR, Rasmussen F, Bendstrup E. Clusters of comorbidities in idiopathic pulmonary fibrosis. *Respir Med.* 2021 Aug-Sep;185:106490. doi: 10.1016/j.rmed.2021.106490. Epub 2021 Jun 1. PMID: 34130097.
4. M. Kreuter, S. Ehlers-Tenenbaum, K. Palmowski, J. Bruhwylter, U. Oltmanns, T. Muley, C.P. Heussel, A. Warth, M. Kolb, F.J.F. Herth, Impact of comorbidities on mortality in patients with idiopathic pulmonary fibrosis, *PloS One* 11 (2016), e0151425, <https://doi.org/10.1371/journal.pone.0151425>.

5. O. Hilberg, E. Bendstrup, A. Løkke, R. Ibsen, A. Fløe, C. Hyldgaard, Co-morbidity and mortality among patients with interstitial lung diseases: a population-based study, *Respirology* 23 (2018) 606–612, <https://doi.org/10.1111/resp.13234>.
6. C. Hyldgaard, O. Hilberg, E. Bendstrup, How does comorbidity influence survival in idiopathic pulmonary fibrosis? *Respir. Med.* 108 (2014) 647–653, <https://doi.org/10.1016/j.rmed.2014.01.008>.
7. M. Kreuter, J. Swigris, D. Pittrow, S. Geier, J. Klotsche, A. Prasse, H. Wirtz, D. Koschel, S. Andreas, M. Claussen, C. Groh'e, H. Wilkens, L. Hagmeyer, D. Skowasch, J.F. Meyer, J. Kirschner, S. Gl'aser, F.J.F. Herth, T. Welte, C. Neurohr, M. Schwaiblmair, M. Held, T. Bahmer, M. Frankenberger, J. Behr, Health related quality of life in patients with idiopathic pulmonary fibrosis in clinical practice: insights-IPF registry, *Respir. Res.* 18 (2017) 139, <https://doi.org/10.1186/s12931-017-0621-y>.
8. H.R. Collard, A.J. Ward, S. Lanes, D. Courtney Hayflinger, D.M. Rosenberg, E. Hunsche, Burden of illness in idiopathic pulmonary fibrosis, *J. Med. Econ.* 15 (2012) 829–835, <https://doi.org/10.3111/13696998.2012.680553>.
9. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, et al. ATS/ERS/JRS/ALAT Committee on Idiopathic Pulmonary Fibrosis. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med.* 2011;183(6):788–824.
10. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, et al; American Thoracic Society/European Respiratory Society/Japanese Respiratory Society/Latin American Thoracic Society. Diagnosis of idiopathic pulmonary fibrosis. an official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med.* 2018;198(5):e44–e68.
11. National Institute for Health and Care Excellence. Idiopathic pulmonary fibrosis in adults: diagnosis and management [Internet] [London]: NICE; 2013 [updated 2017 May; cited 2021 Sept 27]. (Clinical guideline [CG163]). [Available from: <https://www.nice.org.uk/guidance/cg163>.
12. National Institute for Health and Care Excellence. Pirfenidone for treating idiopathic pulmonary fibrosis [London]: NICE; 2018 [reviewed 2021 May; cited 2021 Sept 27]. (Technology appraisal guidance [TA504]). [Available from: <https://www.nice.org.uk/guidance/ta504>.
13. National Institute for Health and Care Excellence. Nintedanib for treating idiopathic pulmonary fibrosis [London]: NICE; 2016 [cited 2021 Sept 27]. Technology appraisal guidance [TA379]) [Available from: <https://www.nice.org.uk/guidance/ta379>.
14. Flaherty KR, Wells AU, Cottin V, Devaraj A, Walsh SLF, Inoue Y, Richeldi L, Kolb M, Tetzlaff K, Stowasser S, Coeck C, Clerisme-Beaty E, Rosenstock B, Quaresma M, Haeufel T, Goeldner RG, Schlenker-Herceg R, Brown KK; INBUILD Trial Investigators. Nintedanib in Progressive Fibrosing Interstitial Lung Diseases. *N Engl J Med.* 2019 Oct 31;381(18):1718-1727. doi: 10.1056/NEJMoa1908681. Epub 2019 Sep 29. PMID: 31566307.
15. Ikeda S, Sekine A, Baba T, Kato T, Katano T, Tabata E, Shintani R, Yamakawa H, Oda T, Okuda R, Kitamura H, Iwasawa T, Takemura T, Ogura T. Randomized phase II study of nintedanib with or without pirfenidone in patients with idiopathic pulmonary fibrosis who experienced disease progression during prior pirfenidone administration. *Medicine (Baltimore).* 2022 Jun 3;101(22):e29232. doi: 10.1097/MD.00000000000029232. PMID: 35665728; PMCID: PMC9276254.
16. King TE Jr, Bradford WZ, Castro-Bernardini S, Fagan EA, Glaspole I, Glassberg MK, Gorina E, Hopkins PM, Kardatzke D, Lancaster L, Lederer DJ, Nathan SD, Pereira CA, Sahn SA, Sussman R, Swigris JJ, Noble PW; ASCEND Study Group. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med.* 2014 May 29;370(22):2083-92. doi: 10.1056/NEJMoa1402582. Epub 2014 May 18. Erratum in: *N Engl J Med.* 2014 Sep 18;371(12):1172. PMID: 24836312.
17. Noble PW, Albera C, Bradford WZ, Costabel U, Glassberg MK, Kardatzke D, King TE Jr, Lancaster L, Sahn SA, Szwarcberg J, Valeyre D, du Bois RM; CAPACITY Study Group. Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials. *Lancet.* 2011 May 21;377(9779):1760-9. doi: 10.1016/S0140-6736(11)60405-4. Epub 2011 May 13. PMID: 21571362.
18. de Andrade JA, Neely ML, Hellkamp AS, Culver DA, Kim HJ, Liesching T, Lobo LJ, Ramaswamy M, Safdar Z, Bender S, Conoscenti CS, Leonard TB, Palmer SM, Snyder LD; IPF-PRO Registry investigators. Effect of Antifibrotic Therapy on Survival in Patients With Idiopathic Pulmonary Fibrosis. *Clin Ther.* 2023 Apr;45(4):306-315. doi:

10.1016/j.clinthera.2023.03.003. Epub 2023 Mar 28. PMID: 36997445.

19. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol*. 2010;69(1):4–14.

20. Quint JK, Müllerova H, DiSantostefano RL, Forbes H, Eaton S, Hurst JR, Davis K, Smeeth L. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ Open. 2014 Jul 23;4(7):e005540.

Appendices

 appendix-1.pdf

 appendix-2.pdf

 appendix-3.pdf

 appendix-4.pdf

Grant ID

9.5. Appendix 5. Publications Associated with the Thesis

CHAPTERS	ASSOCIATED LINKS TO THE PUBLICATIONS AND ABSTRACT
CHAPTER 2	Incidence and prevalence of interstitial lung diseases worldwide: a systematic literature review - https://doi.org/10.1136/bmjresp-2022-001291
CHAPTER 4	Validation of the recording of idiopathic pulmonary fibrosis in routinely collected electronic healthcare records in England - https://doi.org/10.1186/s12890-023-02550-0
CHAPTER 5	Incidence, prevalence and mortality of idiopathic pulmonary fibrosis in England from 2008 to 2018: a cohort study - https://doi.org/10.1136/thorax-2023-220887
CHAPTER 6	Changes in survival in people with IPF in England 2010 – 2020 - https://doi.org/10.1183/13993003.congress-2024.PA5146 (Conference abstract)