# Harmonising respiratory EHR data across England, Wales & Scotland: curating research-ready datasets for Asthma, COPD & ILD

Sara Hatam[1], Sean Scully[2], Sarah Cook[3], Hywel T Evans[2], Alastair Hume[1,4], Constantinos Kallis[3,5], Ian Farr[2], Chris Orton[2], Aziz Sheikh[1], Jennifer K Quint[3,5]

(1) DataLoch, Usher Institute, The University of Edinburgh, UK; (2) Population Data Science, Swansea University Medical School, Swansea, UK;
(3) School of Public Health, Imperial College London, UK; (4) EPCC, The University of Edinburgh; (5) National Heart and Lung Institute, Imperial College London, UK

## BACKGROUND

- Asthma, chronic obstructive pulmonary disease (COPD) and interstitial lung disease (ILD) are chronic respiratory conditions associated with substantial disability and mortality globally
- Electronic healthcare records (EHRs) are powerful resources for health researchers to improve patient outcomes across these diseases
- However, consistent approaches in data curation are needed to enable valid comparative studies

### Which databases were used?

- Clinical Practice Research Datalink (CPRD) Aurum – England
- Secure Anonymised Information Linkage (SAIL) Databank – Wales
- DataLoch – Scotland (Lothian)

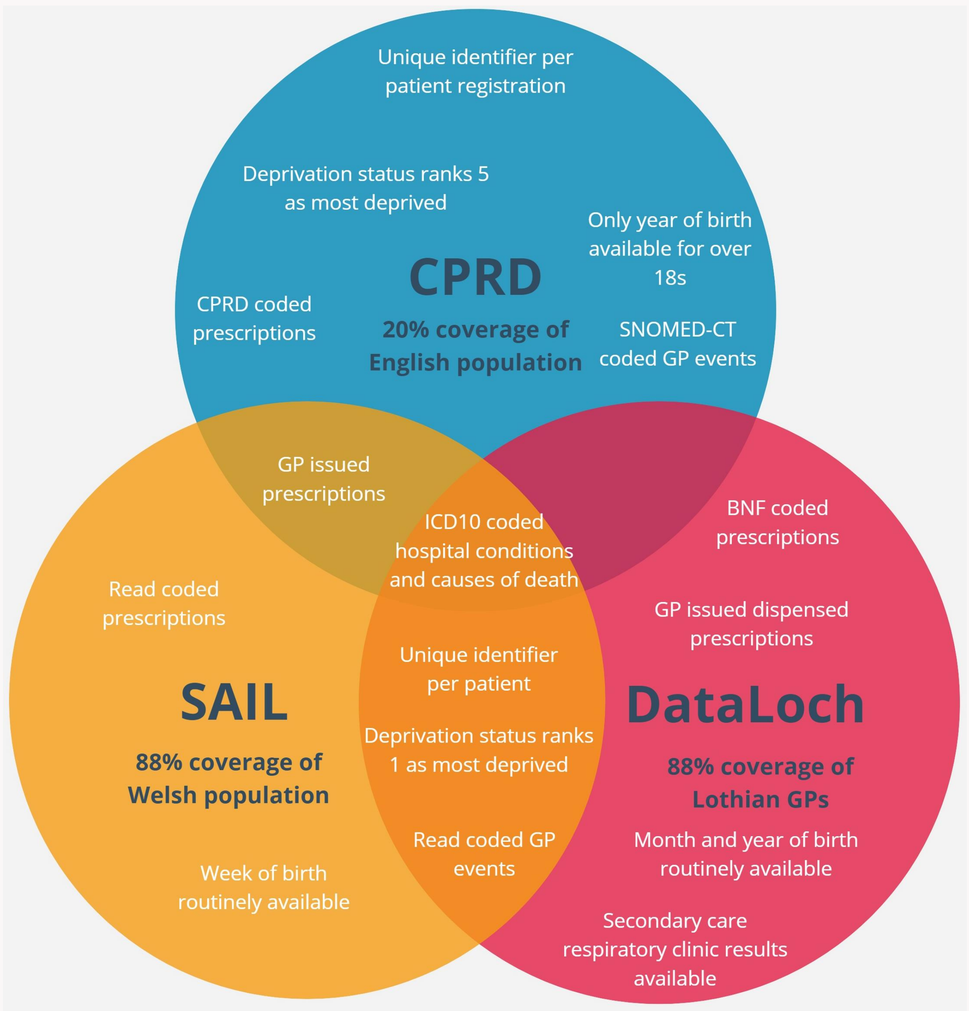


**Fig 1.** Venn diagram showing key overlaps and differences in data between CPRD Aurum, SAIL and DataLoch.

## METHODS

Using the same codelists and algorithms across all three databases, we:

- Created asthma, COPD and ILD patient cohorts from primary care records
- Derived common variables used for respiratory research

### Patients were eligible for inclusion if:

- they had any follow-up time from 2004
- at least one valid code for the condition prior to end of 2019

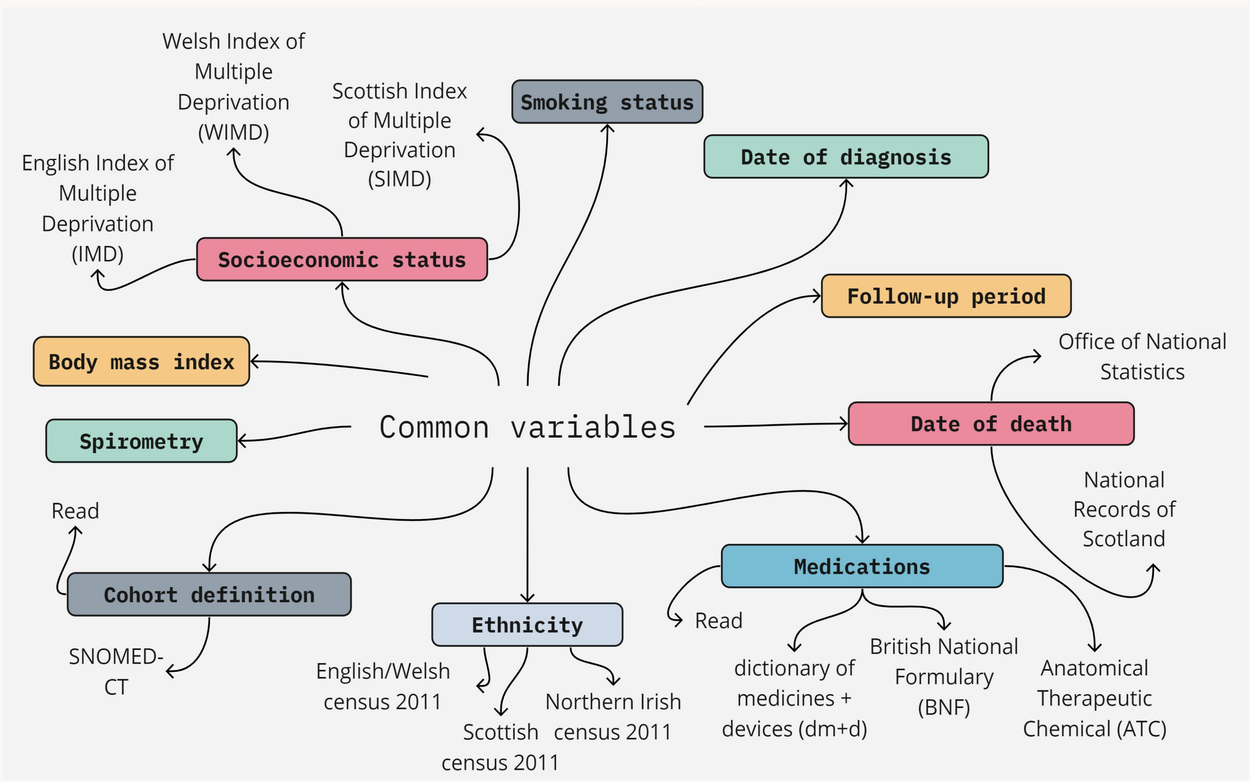


**Fig 2.** Diagram showing common variables we derived across CPRD Aurum, SAIL and DataLoch including key differences in sources or coding systems for consideration.

## RESULTS

Asthma, COPD and ILD cohorts were generated using CPRD Aurum, SAIL and DataLoch using parallelised methodology for cohort selection.

| | Asthma | COPD | ILD |
|---|---|---|---|
| England (CPRD) | n = 2,173,370 | n = 602,295 | n = 58,118 |
| Wales (SAIL) | n = 572,271 | n = 163,792 | n = 20,869 |
| Scotland (DataLoch) | n = 163,570 | n = 41,414 | n = 5,163 |

**Fig 3.** The number of patients in each cohort (asthma, COPD and ILD) across each EHR database (CPRD Aurum, SAIL and DataLoch).

Figure 4 below shows population distribution by age at earliest mention of condition and sex:

- All three EHRs had similar distributions for COPD and ILD
- But for asthma, CPRD had a strong bimodal distribution that was different to SAIL and DataLoch

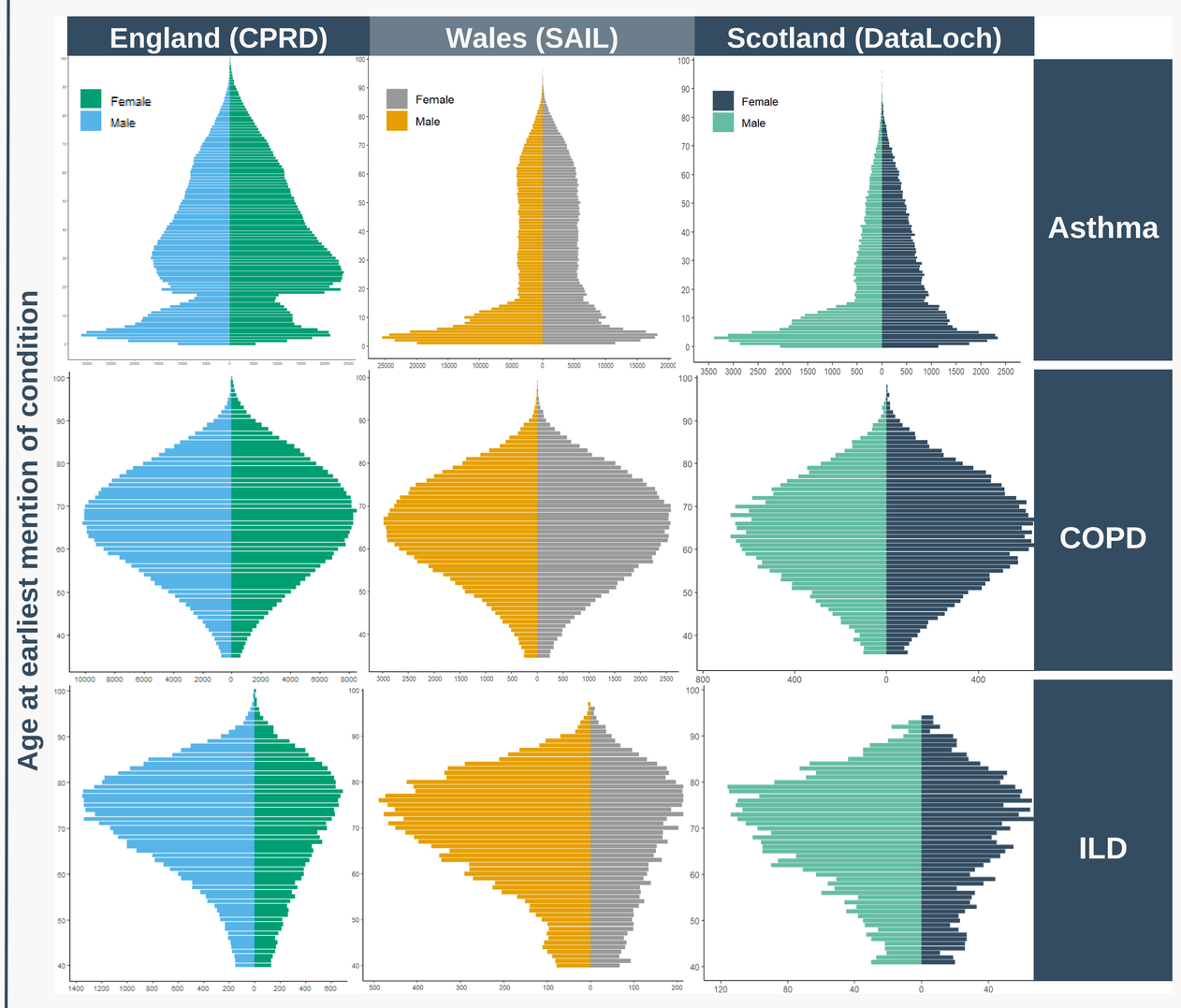


**Fig 4.** The number of patients in each cohort (asthma, COPD and ILD) across each EHR database (CPRD Aurum, SAIL and DataLoch) stratified by sex and age at earliest mention of condition.

### Why was asthma different to COPD/ILD and how can we mitigate?

- For COPD and ILD, patients had to be aged ≥35 and ≥40 at condition event, respectively, whereas asthma had no age cut-offs
- Likely due to patients moving practice aged 18-30 and informing their new GP of asthma history or resolved childhood asthma returning in adulthood
- Patients given unique ID per GP registration in CPRD, whereas patient identifiers same across registrations in SAIL and DataLoch
- Highlighted need for diagnosis date variable that discards first year of registration for chronic conditions and enable better comparisons across EHR populations

## DISCUSSION

- We identified and overcame obstacles in coding and recording between the databases to enable valid comparisons
- Codelists and metadata generated can be re-utilised to develop cohorts for different time periods (see GitHub)
- These resources provide foundation for curation of respiratory datasets in other EHR databases, and potentially other disease domains