BREATHE Health Data Research Hub

HDRUK Health Data Research UK

UK Research and Innovation

# BREATHE dataset curation

Dr Sarah Cook, Sara Hatam, Sean Scully

Dr Hywel Evans, Dr Chris Orton, Prof Jennifer Quint

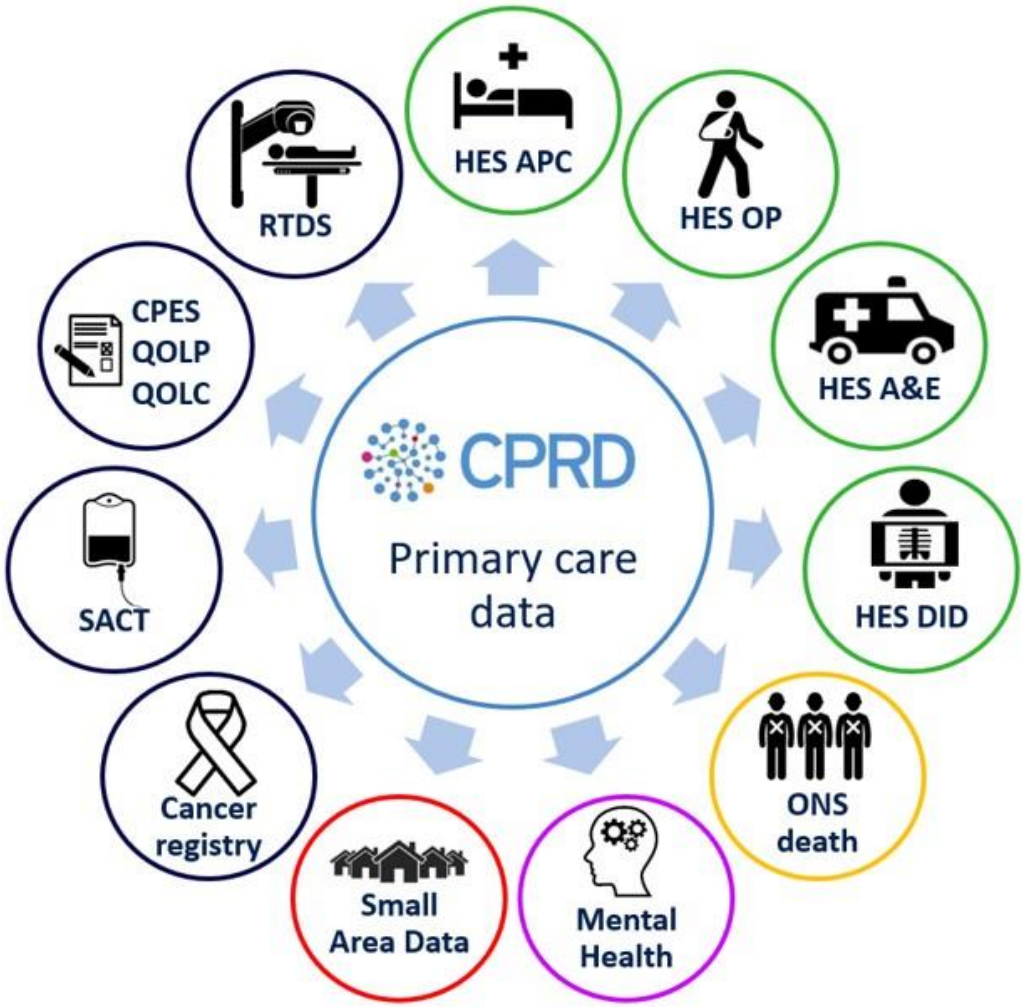
Imperial College London

THE UNIVERSITY of EDINBURGH

Swansea University Prifysgol Abertawe

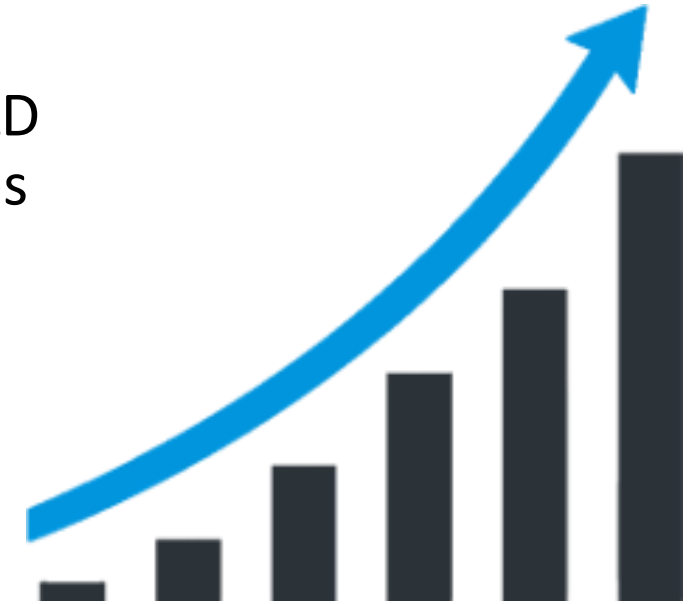
# Background

- Success of Welsh Asthma Observatory (WAO), expand to ILD and COPD in Wales using SAIL Databank
- Curation of datasets for asthma, ILD and COPD in England using Clinical Practice Research Datalink (CPRD)



# Aims

- To speed up research
- Create reproducible datasets
- Generate consistent phenotype definitions
- Document data cleaning methods
- Understand incidence and prevalence for asthma, ILD and COPD from 2005-2020 across England and Wales



# Methods

- All code is being made legible and easy to use to allow for future adaptation by less experienced researchers and easy analysis
- For SAIL, data firstly collected from WLGP and PEDW datasets for each condition to form the Patient Cohorts
- All members of the cohort had a COPD or ILD event between 2005 and 2020, an identifiable gender code (not unknown), and linkage with death records

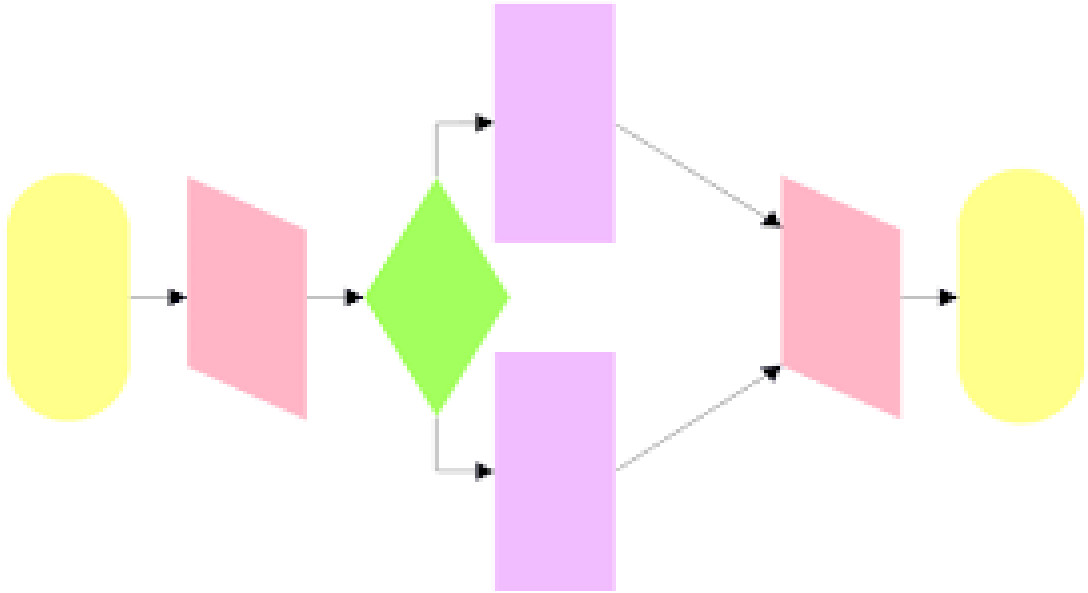


SQL

# Methods – Part 2

- A table of all COPD/ILD events has then been produced for the cohort, followed by a table of preselected comorbidity event flags
- Close co-ordination between the two teams has been essential to ensure records and code lists remain consistent
- Patient, Observability, Measurement and Vaccination tables for the cohorts to be produced down the line, with the aim of expanding the functionality of the datasets and the range of research that can be conducted

| medcodeid | cleansedre~e | snomedctconceptid | snomedctdescrip~d | emiscodeca~d | term |
|---|---|---|---|---|---|
| 55910017 | J045200 | 33505004 | 55910017 | 32 | malocclusion due to mouth breathing |
| 112574019 | H585300 | 67782005 | 112574019 | 32 | adult respiratory distress syndrome |
| 198599018 | 1736.00 | 55442000 | 198599018 | 27 | paroxysmal nocturnal dyspnoea |
| 252384017 | 1731.00 | 161938003 | 252384017 | 27 | no breathlessness |
| 252385016 | 1732.00 | 161939006 | 252385016 | 27 | breathless - moderate exertion |
| 252386015 | 1733.00 | 161940008 | 252386015 | 27 | breathless - mild exertion |
| 252419019 | 175..00 | 161956003 | 252419019 | 27 | breath symptom |
| 253357012 | 1D88.00 | 162481009 | 253357012 | 27 | breathing aggravates symptom |



# Future Work

- Describe epidemiology of asthma, COPD and ILD in England and Wales from 2005 to 2020
- Complete curation of datasets for research use (to prevent replication of effort in data cleaning and management)
- Publish curation methodology/codelists in HDRUK Phenotype Library/GitHub

HDRUK Phenotype Library

GitHub

# Use Cases

- None yet as the project still in the early stages
- In future, datasets can be re-used for wide range of epidemiological research into asthma, COPD and ILD e.g.
  - Pharmacoepidemiology
  - Health service utilization
  - Predictive models
  - Paediatric research (asthma)
- Data can be maintained and updated long-term with multiple data refreshes

# Any Questions?