Making Data Research Ready: implementing HDR UK CALIBER phenotypes within the DataLoch repository

Zengyi Huang[1], Alastair Hume[1,2], Atul Anand[1,3], Franz Gruber[1], Chloe Brook[1,2], Elvina Gountouna[1], Jennifer Daub[1], Amy Tilbrook[1], Pamela Linksted[1,4], Kathy Harrison[1], Nicholas Mills[1,3]

[1] DataLoch, Usher Institute, University of Edinburgh [2] EPCC, University of Edinburgh [3] Centre for Cardiovascular Science, University of Edinburgh [4] NHS Lothian

Introduction

The Health Data Research (HDR) UK CALIBER phenotype set defines algorithms for over 300 physical and mental health conditions (Kuan *et al.* 2019). We aimed to:

- implement HDR UK CALIBER phenotyping within the DataLoch repository
- streamline the identification of health conditions

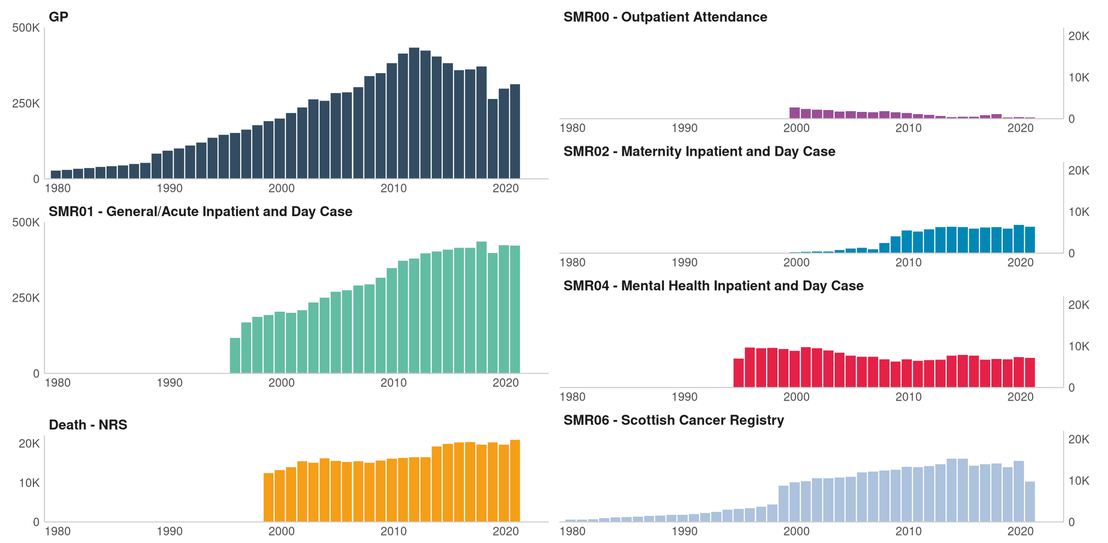
DataLoch phenotype datasets

We applied the HDR UK CALIBER phenotyping algorithms to all individuals in the DataLoch repository.

- Source datasets include GP records, Scottish Morbidity Records (SMR00, SMR01, SMR02, SMR04 and SMR06) and the national death registry.
- Three phenotype datasets (GP read code events with phenotype, SMR code with phenotype and NRS death code with phenotype) have been created to link GP Read codes, ICD-10 and OPCS-4 codes from SMR and ICD-10 codes from deaths.
- A phenotype event dataset was generated for the DataLoch cohort, combining the three aforementioned datasets (using CHI as the linkage identifier).
- This implementation harmonises Read, ICD-10, and OPCS-4 codes across primary and secondary care systems.

Number of phenotype events per year by data source

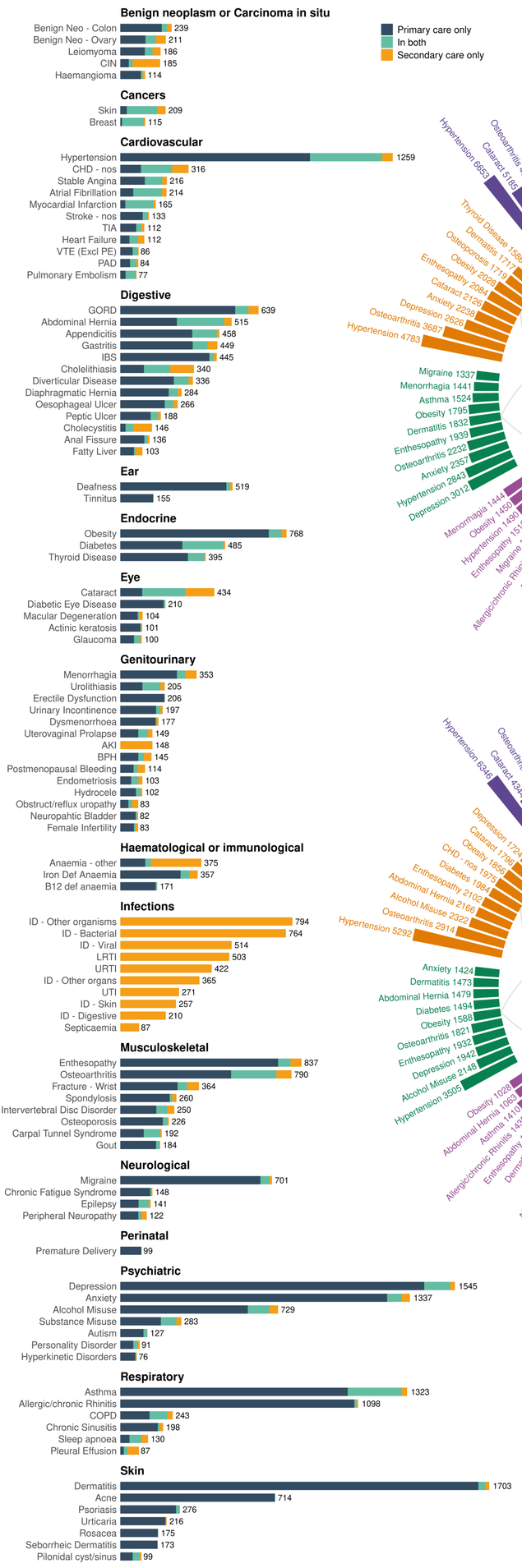
(1.6 million current and past Lothian residents in DataLoch)



Note: A phenotype event is a diagnosis or procedural (Read, ICD-10 or OPCS-4) code mapped to a CALIBER phenotype on a specific date.

Prevalence of the most common conditions in Lothian

(Sample population: 870,000 individuals in DataLoch. Point prevalence per 10,000 on 01/01/2023)



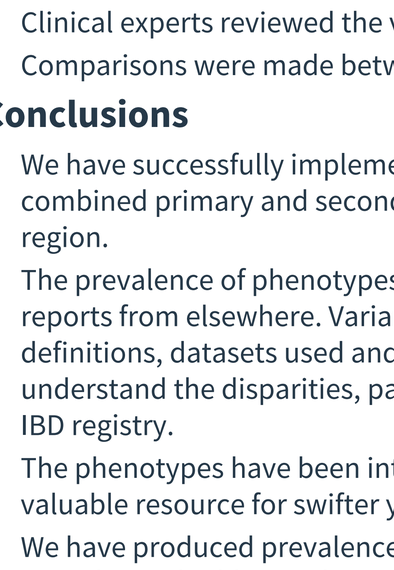
ID=infectious disease. AKI=acute kidney injury. BPH=benign prostatic hyperplasia. CHD=Coronary heart disease. CIN=cervical intraepithelial neoplasia. COPD=chronic obstructive pulmonary disease. GORD=gastro-oesophageal reflux disease. HDLC=high-density lipoprotein cholesterol. IBS=irritable bowel syndrome. LRTI=lower respiratory tract infection. PAD=Peripheral arterial disease. PCOS=polycystic ovary syndrome. PID=pelvic inflammatory disease. TIA=transient ischaemic attack. URTI=upper respiratory tract infection. UTI=urinary tract infection. VTE=Venous thromboembolic disease

Note: CALIBER phenotypes based on blood test results (e.g., chronic kidney disease) were not included in this analysis.

Comparison of DataLoch phenotype prevalence with:

(1) the IBD study by Jones et al. (2019)

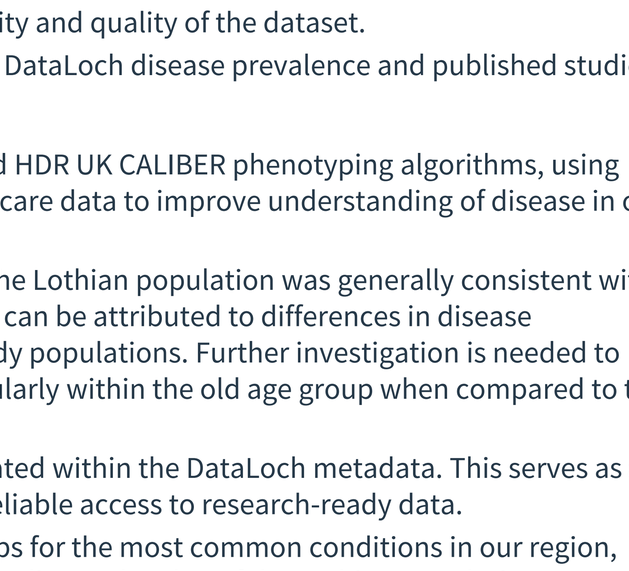
IBD prevalence per 10,000 on 31/08/2018



Notes:
1. DataLoch study population: ~570,000
2. The Jones study used secondary care data across Lothian residents, including pathology reports, hospital coding and prescribing, but excluding GP data.
3. The 'missed' cases in the Jones study were estimated by capture-recapture methods.

(2) the multimorbidity study by Barnett et al. (2012)

Prevalence of 30 conditions on 31/03/2007



Notes:
1. GP records for 1.75M people registered with 314 practices in Scotland was used in the Bartnett study.
2. The differences in the bottom 5 conditions are due to lifetime (DataLoch) vs active prescribing coding (Barnett).
3. The codeset for COPD in the Barnett study is broader than CALIBER codelist for the condition.

Quality assessment

- Clinical experts reviewed the validity and quality of the dataset.
- Comparisons were made between DataLoch disease prevalence and published studies.

Conclusions

- We have successfully implemented HDR UK CALIBER phenotyping algorithms, using combined primary and secondary care data to improve understanding of disease in our region.
- The prevalence of phenotypes in the Lothian population was generally consistent with reports from elsewhere. Variances can be attributed to differences in disease definitions, datasets used and study populations. Further investigation is needed to understand the disparities, particularly within the old age group when compared to the IBD registry.
- The phenotypes have been integrated within the DataLoch metadata. This serves as a valuable resource for swifter yet reliable access to research-ready data.
- We have produced prevalence maps for the most common conditions in our region, providing valuable insights into the disease burden of the Lothian population.

References

1. Kuan V, Denaxas S, Gonzalez-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digital Health 2019; 1: e63–77
2. Jones GR et al. IBD prevalence in Lothian, Scotland, derived by capture–recapture methodology. Gut 2019; 68:1953-1960
3. Barnett K, Mercer S, Norbury M, Watt G, Wyke S, Guthrie B. The epidemiology of multimorbidity in a large cross-sectional dataset: implications for health care, research and medical education. Lancet. 2012; 380: 37-43