Phenotyping UK electronic health records from 15 million individuals for precision medicine: the CALIBER resource

Spiros DENAXAS[a,b,c1] and Arturo GONZALEZ-IZQUIERDO[a,b,c] and Natalie FITZPATRICK[a,b,c] and Kenan DIREK[a,b,c] and Harry HEMINGWAY[a,b,c]
[a] *Institute of Health Informatics, University College London, United Kingdom*
[b] *Health Data Research UK, University College London, United Kingdom*
[c] *The National Institute for Health Research University College London Hospitals Biomedical Research Centre, United Kingdom*

Abstract. Electronic health records (EHR) are increasingly being used for observational research at scale. In the UK, we have established the CALIBER research resource which utilizes national primary and hospital EHR data sources and enables researchers to create and validate longitudinal disease phenotypes at scale. In this work, we will describe the core components of the resource and provide results from three exemplar research studies on high-resolution epidemiology, disease risk prediction and subtype discovery which demonstrate both the opportunities and challenges of using EHR for research.

Keywords. electronic health records, phenotyping, data linkage, prognosis, biomedical informatics

1. Introduction

Electronic Health Records (EHR) are a rich source of information on human diseases [1]. EHR are generated during routine patient interactions in primary or secondary healthcare. EHR can contain information on diagnoses, symptoms. surgical procedures and interventions, prescriptions, laboratory biomarkers (e.g. high-density lipoprotein cholesterol) and physiological measurements (e.g. blood pressure (BP), body mass index). Linking EHR which span primary care and hospital healthcare settings in the United Kingdom (UK) can enable researchers to create longitudinal phenotypes that accurately capture disease onset, severity, and progression [2]. The process of defining disease phenotypes in EHR data however is challenging and time-consuming since EHR are variably structured, fragmented, curated using different clinical terminologies and collected for purposes other than medical research [3].

[1] Institute of Health Informatics, 222 Euston Road, University College London, NW1 2DA, London, United Kingdom; E-mail: s.denaxas@ucl.ac.uk

2. Objective

Here we present and describe a state-of-the-art phenomics resource, CALIBER, for developing, validating and sharing reproducible phenotypes in national structured EHR in the UK. We additionally briefly describe contemporary research exemplars using CALIBER data for translational research: a) disaggregating disease endpoints through high resolution clinical epidemiology, b) disease risk prediction using supervised machine learning approaches, and c) subtype discovery using unsupervised learning.

3. Methods

3.1 CALIBER phenomics resource

We implemented and applied a rule-based phenotyping framework [4] for extracting information on diseases (status, severity, onset), lifestyle risk factors and biomarkers and applied it to a sample of 15 million individuals. CALIBER utilizes data from three national EHR sources: a) primary care EHR from the Clinical Practice Research Datalink (CPRD), b) administrative data on diagnoses and procedures during admission to hospitals from Hospital Episode Statistics (HES), and c) cause-specific mortality information from the Office for National Statistics (ONS) death register. Data were recorded using five controlled clinical terminologies: a) Read (primary care), b) ICD-10 (hospital diagnoses, causes of death), c) ICD-9 (causes of death <1999), and d) OPCS-4 (surgical procedures), and d) DM+D (prescriptions in primary care).

3.2 Contemporary research exemplars

We present three contemporary research exemplars utilizing the CALIBER resource and phenotyping framework: a) high resolution epidemiology: we calculated Hazard ratios (HRs) based on disease-specific Cox models with time since study entry as the timescale, adjusted for baseline age and stratified by sex and primary care practice and report the associations of systolic and diastolic BP with 12 different cardiovascular diseases (CVD), b) disease risk prediction: using a global vectors model[5], we trained clinical concept embeddings from hospitalization diagnosis and procedure information recorded in HES and evaluated them for predicting for the risk of admission to hospital in heart failure (HF) patients, and c) subtype discovery: we applied dimensionality reduction using multiple correspondence analysis and data clustering using k-means to a cohort of Chronic Obstructive Pulmonary Disease (COPD) patients in order to identify and characterize novel and clinically-meaningful disease subtypes.

4. Results

4.1 CALIBER phenomics resource

We created an iterative, rule-based EHR phenotyping approach which combined domain expert input with data exploration. We curated >90,000 ontology terms from five clinical terminologies and created 51 phenotyping algorithms (35

diseases/syndromes, ten biomarkers, six lifestyle risk factors). Phenotype validation is a critical step in the process, and we provided up to six approaches for validating phenotypes: a) the ability to replicate aetiological and prognostic associations reported from non-EHR studies, b) case note review for Positive Predictive Value (PPV) reporting, c) the ability to replicate associations with genetic variants from non-EHR Genome-Wide Association Studies, d) algorithm performance in external populations, and e) cross-EHR-source concordance and stratification of populations.

For each phenotype, we created a textual description with details on the implementation logic, the pre-processing steps and implementation steps. For some algorithms, we generated flowchart descriptions to describe how different components are combined to form the finalized phenotype and for facilitating the translation to machine-code (e.g. SQL) for execution and data extraction. Algorithms are curated on an open-access resource, the CALIBER Portal (https://www.caliberresearch.org/portal), [6,7] and have been used in >60 publications from national and international research groups. Each phenotype page on the Portal[2] contains sufficient implementation and validation information for external researchers to re-use the algorithm.

4.2 Contemporary research exemplars

High resolution epidemiology: In a cohort of 1.25 million patients, we reported [8] highly-heterogeneous associations between BP and CVD disease endpoints: high systolic BP was more strongly associated with stable angina, Hazard Ratio (HR) 0.41 [95% CI 1.36-1.46] than diastolic whereas diastolic BP had the strongest association with abdominal aortic aneurysm (HR 1.45 [95% CI 1.34–1.56]). We have undertaken similar analyses in other conventional CVD risk factors e.g. smoking [9], type-II diabetes [10], alcohol [11], social deprivation [12], heart rate [13], sex [14].

Disease risk prediction: We trained clinical concept embeddings [15] from 2,447 ICD-9, 10,527 ICD-10 and 6,887 OPCS-4 terms across 2,779,598 hospitalizations in the UK Biobank. In the UK Biobank, we identified 4,581 HF cases (using the CALIBER HF phenotype [16,17])(30.52% female) and matched them to 13,740 controls. Clinical concept embeddings performed marginally better (AUROC 0.6965) than one-hot encoding of hospitalization data for predicting admission to hospital due to HF.

Disease subtype discover: In the CPRD [18], we identified 30,961 current and former smokers diagnosed with COPD and extracted 15 clinical features including risk factors and comorbidities. Using clustering, we identified five clinically-meaningful COPD clusters with distinct dominant clinical profiles (e.g. anxiety/depression, frailty, CVD, obesity and atopy) and different healthcare utilization and exacerbation profiles.

5. Conclusions

In this manuscript, we described the CALIBER resource as a framework for using national EHR from primary and secondary health care, disease and national mortality

[2] For example, www.caliberresearch.org/portal/phenotypes/heartfailure

registries. Challenges remain with regards to scaling the phenotyping efforts to thousands of diseases and for recreating the life course of disease [19] .

References

[1] Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. Eur Heart J. 2018;39: 1481–1495.

[2] Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: opportunities and challenges. Eur Heart J Qual Care Clin Outcomes. 2015;1: 9–16.

[3] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20: 117–121.

[4] Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012;41: 1625–1638.

[5] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2014. pp. 1532–1543.

[6] Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One. 2014;9: e110900.

[7] Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick N, Banerjee A, Dobson R, et al. UK phenomics platform for developing and validating EHR phenotypes: CALIBER. bioRxiv. 2019. doi:10.1101/539403

[8] Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. Lancet. 2014;383: 1899–1911.

[9] Pujades-Rodriguez M, George J, Shah AD, Rapsomaniki E, Denaxas S, West R, et al. Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease in 1937360 people in England: lifetime risks and implications for risk prediction. Int J Epidemiol. 2015;44: 129–141.

[10] Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, et al. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. The Lancet Diabetes & Endocrinology. 2015;3: 105–113.

[11] Bell S, Daskalopoulou M, Rapsomaniki E, George J, Britton A, Bobak M, et al. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population-based cohort study using linked health records. BMJ. 2017;356: j909.

[12] Pujades-Rodriguez M, Timmis A, Stogiannis D, Rapsomaniki E, Denaxas S, Shah A, et al. Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: implications for risk prediction and prevention. PLoS One. journals.plos.org; 2014;9: e104671.

[13] Archangelidi O, Pujades-Rodriguez M, Timmis A, Jouven X, Denaxas S, Hemingway H. Clinically recorded heart rate and incidence of 12 coronary, cardiac, cerebrovascular and peripheral arterial diseases in 233,970 men and women: A linked electronic health record study. Eur J Prev Cardiol. 2018;25: 1485–1495.

[14] George J, Rapsomaniki E, Pujades-Rodriguez M, Shah AD, Denaxas S, Herrett E, et al. How Does Cardiovascular Disease First Present in Women and Men? Incidence of 12 Cardiovascular Diseases in a Contemporary Cohort of 1,937,360 People. Circulation. ncbi.nlm.nih.gov; 2015;132: 1320–1328.

[15] Denaxas S, Stenetorp P, Riedel S, Pikoula M, Dobson R, Hemingway H. Application of Clinical Concept Embeddings for Heart Failure Prediction in UK EHR data [Internet]. arXiv [cs.CL]. 2018.

[16] Uijl A, Koudstaal S, Direk K, Denaxas S, Groenwold RHH, Banerjee A, et al. Risk factors for incident heart failure in age- and sex-specific strata: a population-based cohort using linked electronic health records. Eur J Heart Fail. 2019; doi:10.1002/ejhf.1350

[17] Koudstaal S, Pujades-Rodriguez M, Denaxas S, Gho JM, Shah AD, Yu N, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. Eur J Heart Fail. Wiley; 2017;19: 1119–1127.

[18] Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC Med Inform Decis Mak. 2019;19: 86.
[19] Kuan V, Denaxas S. et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the National Health Service. Lancet Digital Health. 2019 (in press).