HDR UK National Phenomics Resource Workshop
11 March 2021

The HDR UK CALIBER Phenotype Library: What is it? How can I use it? and opportunities for research

Workshop Summary

Emerging opportunities in electronic health record (EHR) data allow us for the first time to develop consistent ways to define and understand risk across a wide range of diseases. However, EHR data are not primarily generated for research purposes, are stored in disparate sources often using different formats and require a significant amount of pre-processing. Funded by HDR UK, the HDR UK CALIBER Phenotype Library is the largest national open-access library of reproducible phenotyping algorithms for defining human disease, lifestyle risk factors and biomarkers in EHR.

On 11 March 2021, researchers, health data scientists and clinicians came together to learn how the Phenotype Library can benefit their research, improve research reproducibility and hear how research that has benefited from the Library has been carried out to improve patient health and care.

HDR UK CALIBER Phenotype Library

Led by Prof Spiros Denaxas at UCL and developed by researchers across the UK, the Phenotype Library facilitates the dissemination of methods and tools for defining disease and health-related conditions in a consistent, reproducible fashion across data modalities (structured, text, imaging, sensors and wearables). The Library provides the research and clinical community with an open platform for storage, dissemination, re-use, evaluation and citation of their own curated phenotyping algorithms and metadata in order to reduce duplication of effort and improve research reproducibility.

The Library contains definitions for hundreds of diseases: currently 353 phenotype algorithms and 951 codelists are available for re-use and another 569 phenotypes are due to go live

imminently. A growing number of phenotype collections are also being stored in the Library including BREATHE (respiratory phenotypes collated by the HDR UK Hub for Respiratory Health for use in research) and the British Heart Foundation Data Science Centre COVID-19 phenotypes, and we are working on expanding the Phenotype Library to include a number of new collections (HDR UK Hubs for critical care (PIONEER) and cancer (DATA-CAN) and ClinicalCodes.org and UK Biobank).

The research community is invited to contribute their own phenotypes (details are available on the Phenotype Library website). To promote discoverability, open-source tools and resources from the Library have been integrated into the HDR UK Innovation Gateway

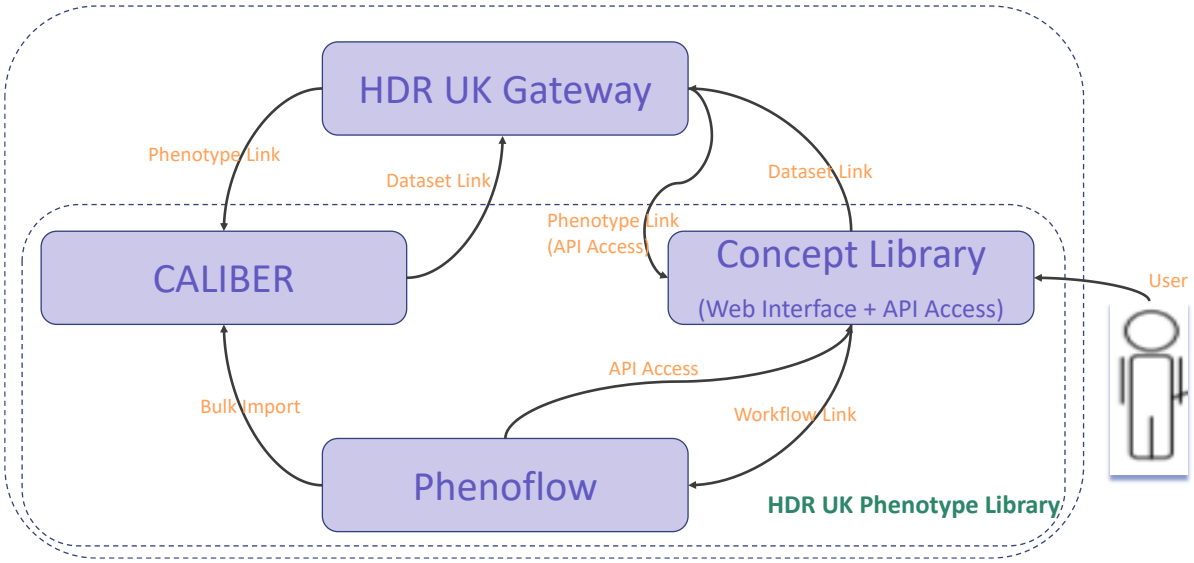

Figure: Active development to enhance and integrate existing tools

For more information on the Phenotype Library, contact Spiros Denaxas (s.denaxas@ucl.ac.uk)

The Concept Library: Creation, Use, and Sharing of Research Knowledge

The Secure Anonymised Information Linkage (SAIL) Databank of anonymised data is a Wales-wide research resource focused on improving health, well-being and services. All SAIL Databank users must agree to share their computer code and scripts through the channels SAIL make available for this purpose.

A "concept" is the definition of a single entity that will be used in a research project and can include a disease, treatment, test result or anything else that may be defined within the data. Often definitions that are created are of interest to researchers for many studies, but there are barriers to easily sharing them.

Similar to CALIBER, the Concept Library is a resource for phenotype definitions. The Concept Library is a system for storing, managing, sharing and documenting clinical code lists in electronic health research. The aim of the Concept Library is to create a system that describes research study designs in a machine-readable format to facilitate rapid study development, higher quality research, easier replication and sharing of methods between researchers, institutions, and countries. It is currently available to users of the SAIL Databank and the team is working to develop links to the HDR UK CALIBER Phenotype Library, Phenoflow and the HDR UK Innovation Gateway.

For more information on the Concept Library, contact Dan Thayer (d.s.thayer@swansea.ac.uk)

Phenotype modelling and validation

A phenotype can take two different forms: (1) a phenotype definition is the logic of the phenotype expressed as a non-computational abstract layer (i.e. it is a mechanism for identifying a given cohort). (2) a phenotype computable form is realised from that definition and will execute and run on a computer.

A high-quality phenotype definition will be ***reproducible*** (a definition contains sufficient information that it can be **accurately** implemented across multiple sites and datasets), ***portable*** (a definition contains sufficient information that it can be **easily** implemented across multiple sites and datasets), and ***valid*** (a definition accurately captures the disease or condition modelled).

A phenotype model relates to the choices that are made about which information is required for that definition and how it is structured. Structuring a definition according to a standard model improves reproducibility and portability. The design a phenotype model to enable high quality definitions that are reproducible consist of 3 steps. First, you should apply the model consistently to improve reproducibility and portability. The model should be designed based on a high-level modelling language that is domain specific (e.g. CQL) rather than a directly executable language. Second, it should include implementation information (implementation language and structure) to provide a clear pathway for implementation which will improve portability. Third, it should build in redundancy by expressing the same technical and non-technical descriptions in multiple ways to make a definition clearer and easier to implement.

Validity is also important when considering definition quality but approaches must scale. Phenotype validation, such as in the CALIBER Phenotype Library, involves use of GOLD standards but where such validation is not possible we should be thinking about hybrid and new approaches including SILVER standards which are being developed to facilitate this process.

For more information on Phenotype modelling, contact Martin Chapman (martin.chapman@kcl.ac.uk)

Developing computational phenotypes: Phenoflow

Despite the excellent resources for phenotype definitions that are available (CALIBER and the Concept Library) and the HDR UK Innovation platform as a discoverability tool, there is an unmet need around providing executable forms for phenotype definitions and implementation of phenotypes, for example if a user wants to use a phenotype definition to identify a cohort in a local dataset that is not available on the Gateway.

Researchers at King's College London have developed a new multi-layer model for a phenotype definition, which is realised as a workflow, and can be combined with different implementation units in order to produce a computable form and support the generation of structured phenotyping algorithms and their interoperability.

The Library is supported by a workflow based, multi-layer phenotype model which consists of 3 layers: (1) Abstract layer which expresses the logic of a phenotype through a set of simple, sequential nested steps (2) Functional layer which specifies the metadata of entities passed between the operations within the abstract layer (3) Computational layer which defines an environment for

the execution of one or more implementation units for each step in the abstract layer, providing a template for development.

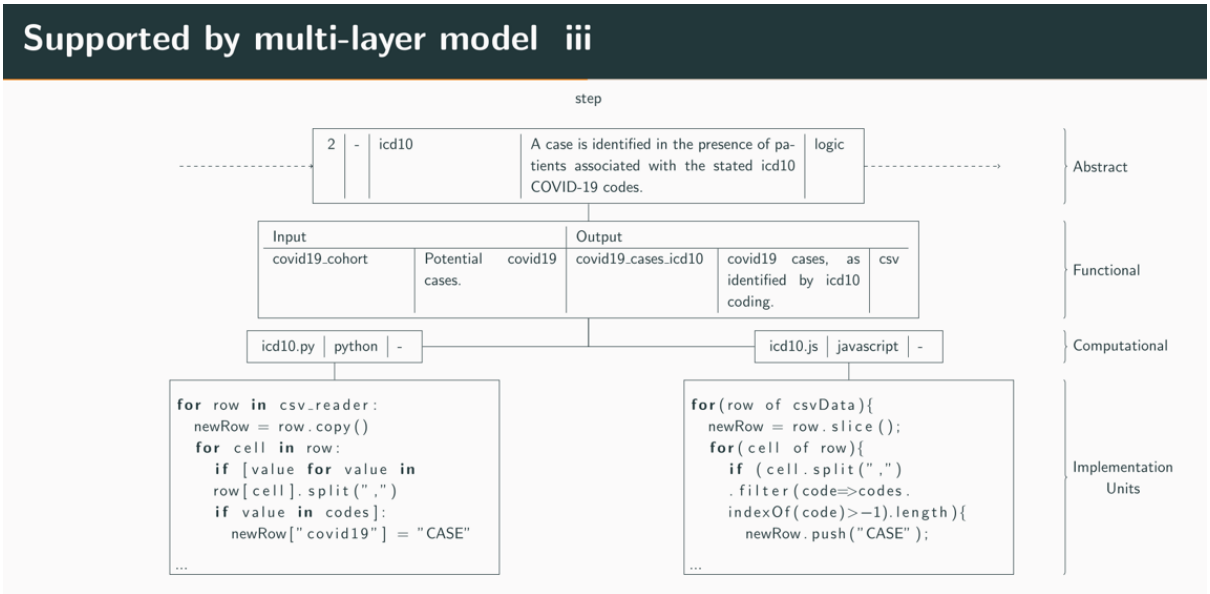

Figure. Individual step of COVID-19 structured phenotype definition and new implementation units

Future work will focus on how we can leverage the multi-level model to express relationships between phenotypes e.g. sub-phenotypes at each level of the model; publish more implementations for complex disease-specific phenotypes e.g. COVID-19, stroke (currently there are 333 phenotype implementations in Phenoflow); and explore the use of Phenoflow as a validation tool.

For more information on Phenoflow, contact Vasa Curcin (vasa.curcin@kcl.ac.uk)

Using the Library to demonstrate clinical impact and benefit to patients: use cases from across the UK

The Phenotype Library has facilitated research across the UK. During this session, we heard from researchers and clinicians from across the HDR national sites showing how they have used diverse data to answer clinically meaningful research questions.

***OurRisk.CoV** is an online risk calculator (http://covid19-phenomics.org/PrototypeOurRiskCoV.html) for patients and the public which provides 1-year COVID-19 mortality risks for common conditions by age and sex in the context of current guidelines. The risk calculator was developed with considerable input from members of the public. The risk calculator has had 1.3 million pageviews from 636K users across the world (64% of users were from the UK, 21% were from the United States and 5% from Australia).

For more information, contact Ami Banerjee (a.banerjee@ucl.ac.uk)

***Using epidemiology to quantify the applicability of trial evidence to inform guideline development** There can be good reasons to exclude people from trials for example patient safety and ethical concerns, however, exclusions are not always justified. This project is

carrying out a large-scale comparison of trial eligible versus trial ineligible populations across a broad range of conditions to inform guideline development. The tool is built on top of data from CPRD linked to hospital and mortality data. The Febuxostat versus Allopurinol Controlled Trial (FACT) is a randomised controlled trial randomising 762 people with gout to receive febuxostat 80mg, febuxostat 120mg or allopurinol 300mg. The trial compared characteristics of FACT eligible vs ineligible gout patients based on 308 phenotype definitions in the HDR UK CALIBER Phenotype Library. It is hoped that guideline developers can make use of this information to understand the impact recommendations and think more broadly about comorbidities in general, rather than being very disease focused. Future work will look to build phenotypes specific to certain trial criteria that could be added to the portal and feed into guideline development.

For more information, contact Daniel Morales (d.r.z.morales@dundee.ac.uk)

*** Mental Health Phenotyping** In mental health, people with more severe disease are less likely to participate in research and are much more likely to be lost to follow up. Primary care data in Wales seemed to suggest that diagnosis of depression was reducing. This presentation took us through the team's journey to define depression and anxiety phenotypes through a series of mental health projects, and highlighted the importance of researchers using the same case definitions.

For more information, contact Ann John (A.John@swansea.ac.uk)

***Using the Mauro Data Mapper to represent phenotypes** Mauro is an open-source "metadata catalogue" for modeling data assets, and data specifications focused on data semantics, and interoperability. The team have been working with NHS Digital NHS digital supporting the NHS Data Dictionary, and integration with fhir.nhs.uk and new NHS terminology server.

For more information, contact James Welch (james.welch@cs.ox.ac.uk)

***CovidSurg Risk calculator** (https://covidsurgrisk.app/) is a data-driven model based on real-world prospective patient data. The tool uses patient and clinical factors that are highly associated with mortality to predict the risk of death for patients undergoing surgery with COVID-19. A machine learning technique was used to generate the CovidSurg Risk calculator which has the potential to better inform surgeons, patients and healthcare decision makers to reach a better understanding of risk when treating surgical patients during the COVID-19 pandemic.

For more information, contact Laura Bravo (LXB732@student.bham.ac.uk)

For more information on the HDR UK National Phenomics Resource please contact Natalie Fitzpatrick, HDR UK Phenomics Programme Manager (n.fitzpatrick@ucl.ac.uk).

Speakers:

- **Prof Spiros Denaxas**, Professor in Biomedical Informatics, UCL
- **Prof Emily Jefferson**, Director of the Health Informatics Centre, University of Dundee
- **Dr Shahzad Mumtaz**, Health Data Scientist, University of Dundee
- **Dr Susheel Varma**, Director of Engineering, Health Data Research UK
- **Dan Thayer**, Senior Data Scientist, Swansea University
- **Dr Martin Chapman**, Research Associate, King's College London
- **Dr Ami Banerjee**, Associate Professor in Clinical Data Science, UCL

- **Dr Daniel Morales**, GP and Wellcome Trust Clinical Research Development Fellow, University of Dundee
- **Prof Ann John**, Professor in Public Health and Psychiatry, Swansea University Medical School
- **Dr James Welch**, Researcher/Software Architect, University of Oxford
- **Laura Bravo**, Wellcome Trust PhD student in Clinical Bioinformatics, University of Birmingham
- **Prof George Gkoutos**, Professor of Clinical Bioinformatics, University of Birmingham
- **Dr Vasa Curcin**, Reader in Health Informatics, King's College London
- **Dr Martin Chapman**, Research Associate, King's College London

Note: Slides from the workshop are available on the UCL Institute of Health Informatics webpage in the 'links' section