HDRUK
Health Data Research UK

# Health Data Scientist

## Applicant pack

### Job description and person specification

**Post:** BHF Data Science Centre – Health Data Scientist (based within HDR UK) We have multiple positions available for this role. Therefore, it is not essential that each candidate meets all requirements as there is some flexibility in the focus of each position

**Location:** Flexible. Primary base could be anywhere in the UK, but must be willing and able to travel to London and elsewhere (mainly UK) in connection with the role

**Salary:** **£50,000**

**Duration:** To 31st Dec 2024 with the possibility of extension

**Reporting to:** Senior Health Data Scientist, BHF Data Science Centre

## About Health Data Research UK

Health Data Research UK (HDR UK) is the national Institute for data science in health. Our UK team of experts develop and apply cutting-edge data science approaches to clinical, biological, genomic and other multi-dimensional health data to address the most pressing health research challenges facing the public.

Our mission is to make game-changing improvements in the health of patients and populations through data science research and innovation.

For the first time we are bringing together our unique nationwide data assets and specialists across academia, research and healthcare to unlock knowledge and deliver new insights from molecule to man. By undertaking research at scale, across a population of up to 65 million people, we have an unrivalled opportunity to use data to the highest ethical standards to drive breakthroughs in medical research. This unleashes the potential to improve the way we are able to prevent, detect and diagnose diseases such as cancer, heart disease and asthma.

At Health Data Research UK, we employ talented individuals who bring their own unique skills and experience to support the vision and benefit the whole team.

We have been central to the UK's response to COVID-19, enabling a trustworthy, national approach to using health data, drawing on the full capabilities of UK research, enabling health data for research into understanding the virus, clinical trials for treatments (including Dexamethasone), symptom trackers, risk calculators and impacts on vulnerable groups, including cancer patients.

### HDR UK's strategy

Our strategy focuses on three core areas:

1. **Research Data Infrastructure and Services** - providing the UK-wide and global co-ordination and leadership of health data infrastructure and services required to make health-relevant data FAIR (Findable, Accessible, Interoperable and Reusable). This builds on the convening, collaborative and co-ordinating role of the UK Health Data Research Alliance and includes the Health Data Research Innovation Gateway and the Health Data Research Hubs
2. **Research Driver Programmes** - advancing research discoveries through high impact UK-wide programmes that address major health and societal challenges, guide the development of the infrastructure and services for the benefit of other researchers and are outward-looking with global reach.
3. **One Institute Partnerships** - through national leadership with a clear vision and ambition to assemble a health data research ecosystem with enduring benefits for all researchers. As an innovative distributed UK-wide and increasingly global Institute, we act as a flagship for team science, drawing on skills, resources, and expertise from academic, NHS, industry and government partners.

## About the BHF Data Science Centre

The **British Heart Foundation (BHF) Data Science Centre**, is building on a £10m initial investment from the BHF to deliver the data and data science needed to address some of the most pressing challenges in heart and circulatory health research.

The centre works in partnership with patients, the public, NHS, researchers and clinicians to promote the safe, ethical and scientifically robust use of data for research into the causes, prevention and treatment of all diseases of the heart and circulation (including, for example, heart attacks, heart failure, heart rhythm disorders, stroke, peripheral vascular disease and vascular dementia).

The centre also hosts the diabetes data science catalyst, which enables access to and use of data from people with diabetes. Research enabled by the catalyst will enhance our knowledge of the links between diabetes and cardiovascular disease; facilitate a deeper understanding of the causes and progression of diabetes as a major cardiovascular risk factor; and drive improvements in treatment and prevention of diabetes, with associated reductions in cardiovascular disease.

Extensive and ongoing engagement with key stakeholders has shaped the development of the centre's six thematic areas across which cardiovascular and diabetes research will be supported:

- Better access to and use of **structured health data** (e.g., recorded using controlled clinical terminologies and medical ontologies) UK population-wide for cardiovascular research

- Better access to and use of **unstructured health data** (including imaging data) at scale for cardiovascular research
- Enabling large-scale use of **personal monitoring data** in a wide range of cardiovascular research
- Developing and refining **computable cardiovascular phenotypes** for different applications
- Supporting discoveries of cardiovascular disease causes, prediction, early detection, prognostic tools and treatments using **disease-based cohorts**
- Developing methods and infrastructure for large, efficient, **data-enabled cardiovascular trials**

The BHF Data Science Centre does not hold data itself. Instead, it works with relevant data custodians, including through the UK Health Data Research Alliance and Health Data Research Innovation Gateway, to provide knowledge and expertise to help researchers from the NHS, academia and industry find, access, understand, connect and analyse the UK's unique cardiovascular 'big data' that is distributed across national registries, NHS electronic medical records and other relevant datasets.

As part of the "structured health data" thematic area, the BHF Data Science Centre is coordinating the **CVD-COVID-UK/COVID-IMPACT** programme of work which, in collaboration with NHS Digital, has for the first time linked de-identified data across an individual's healthcare journey for 96% of the English population, to enable vital research into the relationship between cardiovascular disease and COVID-19. This equates to >10 billion of rows of data spanning birth to death and covering across primary care, hospitalisation, medication, COVID-19 test and vaccination data and specialist cardiovascular audit and registry data. This data is brought together in a secure environment that protects patient privacy but allows safe researcher collaboration to answer urgent health questions. The knowledge and expertise from this work will underpin and support work across the other thematic areas.

## Purpose of the post

The post holder will be a key member of the BHF Data Science Centre team, working on one or more of the six thematic areas and diabetes data science catalyst. This will involve working closely with the Senior Health Data Scientists, Director, relevant Associate Directors, wider BHF Data Science Centre team, HDR UK staff, data custodians and researchers / data scientists from academia, NHS organisations and industry.

The main responsibility of this role will be to provide the data management and data curation methods for processing and preparing research datasets from linked national hospital, primary care, mortality, COVID-19 test data, vaccination data and specialist audit and registry datasets from 65 million patients and billions of data points.

The post-holder will contribute to the development and application of re-usable data curation pipelines, algorithms to assess the data quality within and across data sources, and developing approaches to define novel phenotyping algorithms to derive clinically important markers from complex health records (e.g., disease outcomes, phenotypes, medications).

The post holder will also apply knowledge gained about these data to support the aims of other thematic areas, such as defining outcome measures from routinely collected health data for cardiovascular clinical trials; linking personal monitoring data with other types of health data; use of imaging data in research; access to diabetes specific datasets. An understanding of the challenges and limitations, as well as the opportunities, in accessing and using different types of health data will be important. Where necessary the post holder will also undertake analyses to answer research questions in collaboration with other research groups.

This post would suit a health data/computer scientist with significant experience in the curation, management and analysis of health data from different types of research studies (e.g. clinical trials, epidemiological studies).

## Main responsibilities

- Providing data management and curation methods in the trusted research environments (TREs) for the CVD-COVID-UK/COVID-IMPACT programme of work.
- Lead data curation pipeline development under the supervision of the Senior Health Data Scientists.
- Provide data science support and expertise / research software engineering across the Centre's thematic areas. Input into and deliver across multiple parallel projects, communicating progress, challenges, and escalating issues where necessary.
- Develop approaches for assessing data quality and data utility of the various routinely collected health datasets across the four devolved nations. Carry out technical validation checks on linked data sources (e.g. duplicates, linkage errors) and develop functions to check these data rigorously for errors and inconsistencies.
- Summarise and disseminate findings and lessons from within and across data quality and data utility comparisons to inform research and contribute to discussions of where routinely collected data can be used in research studies, across the Centre's thematic areas, or the need for further guidance (e.g. comparing trial-specific data collection with routinely collected health data)
- Write, organise and curate support documentation for linked data resources (e.g. data dictionaries, variable mapping tables, data access process documentation, Git repositories).
- Work with relevant researchers to identify and apply appropriate existing and new phenotype definitions and algorithms from linked national health data.
- Prepare numerical and graphical summaries to communicate findings to researchers when curating data.
- Where necessary work with relevant researchers to undertake collaborative data analysis to answer agreed research questions.
- Prepare and present results in oral and written reports and publications
- Be an active participant and attend the regular Centre and project meetings, reporting on progress and presenting analytical results
- We are committed to open source, transparent and reproducible research and the post will be releasing tools, algorithms, and approaches under an open-source licence.

### Planning and organising

The postholder will join a small team that is embedded in HDR UK and the post holder will be responsible for planning and day-to-day management of their own workload across diverse and complex work programmes and projects.  At the same time, the post holder will require a flexible approach to work to changing demands, particularly external changes.

## Problem solving

The postholder will have some expertise but also the need to develop expertise in the data quality and utility of national routinely collected coded health data as well as other complex (and often less structured) health data e.g. imaging, personal monitoring data, and how it can be used in different types of research.
The postholder will apply analytic approaches to complex health-related data requiring prior technical data science knowledge. They will also need to be able to resolve complex data curation and analysis challenges, discussing as required with other members of the team and external colleagues.
The post holder will make an effective judgement on when to escalate issues to senior colleagues' attention and with what urgency.

## Decision making

In collaboration with the Senior Health Data Scientists, the postholder will make decisions about the most appropriate tools and approaches for querying, analysing, maintaining and documenting complex health-related data.

## Continuous improvement

HDR UK is dedicated to continuous improvement through our quality management system and has achieved ISO 9001 accreditation. The post-holder will review, analyse, identify and implement opportunities for quality improvement within their role and as part of the wider team through our strategy development and internal audit processes.

## Key contacts/relationships

The post holder will work in close conjunction with the core BHF Data Science Centre team but primarily with the BHF Data Science Centre Senior Health Data Scientists, BHF Data Science Centre Director and Associate Directors. They will also work with external research analysts, data managers, data custodians, clinicians, health data scientists and epidemiologists across a number of programmes and projects.

They will build and maintain effective working relationships across multiple HDR UK teams (Uniting the Data, Improving the Data and Using the Data), partners in the British Heart Foundation, the wider cardiovascular and health data science communities, partners in substantive HDR UK Hubs and sites, and other key stakeholders.

## Knowledge, skills and experience

**Experience**

- Good first degree and/or higher degree or equivalent experience in one of the following subjects: bioinformatics, biostatistics, computer science, mathematics, statistics, data science, informatics, epidemiology.
- Strong data manipulation and analysis skills, including:
  - Scripting skills and experience in writing re-usable code in at least one programming language, in particular SQL, Python/PySpark

    - Advanced skills in at least one statistical software package (e.g. R, Stata), particularly in manipulating large datasets.
- Strong relevant experience working with large-scale health-related longitudinal data, deriving variables from electronic health records and preparing analysis-ready datasets.
- Understanding of information governance, privacy, and security issues with using NHS health records.
- Understanding of different study designs and analysis methods (e.g. in epidemiological studies, clinical trials).
- Understanding of sources of routinely collected health data and their application to different types of research studies
- Knowledge of commonly used terminologies in health data (e.g. ICD10, SNOMED) and existing phenotyping algorithms, such as those developed by **CALIBER.**
- Experience of using Git and fundamental concepts in source code revision.
- Writing, presenting and explaining technical and/ or scientific reports to a wide range of scientific and lay audiences.
- Ability and track record of working independently and co-operatively as part of a team

### Skills

- Committed to open source, reproducible, research
- Ability to work accurately, with attention-to-detail
- Excellent networking skills and experience of working in multidisciplinary teams
- Ability to work collaboratively in a small team
- Excellent written and verbal communication skills with the ability to communicate effectively and confidently with people at all levels
- Ability to clearly communicate technical concepts to a non-technical audience
- Excellent report writing and presentation skills
- Excellent organisational and time management skills, with the ability to work independently as well as manage competing priorities and issues under time pressures
- Experience of working in a fast-paced and evolving environment.

### Dimensions

- This is a full-time role. Travel in the UK may be required to partner organisations
- HDR UK is a national institute, and our activities take place across the UK.

## Application Process

Interested applicants are encouraged to contact the BHF Data Science Centre Senior Health Data Scientists to discuss further: bhfdsc@hdruk.ac.uk

**How to apply:** Unless specified, please apply using our online portal. We use a recruitment process that is based on finding out more about the relevant skills and knowledge an applicant has and to help us with this we ask you 3 or 4 skills-based questions as part of the application process.

We will ask you to upload your CV and covering letter, **please do this in one document**, this will be anonymised and will be used if you are shortlisted.

Please contact **recruitment@hdruk.ac.uk** if you have any queries regarding your application

HDRUK
Health Data Research UK

## Equal Opportunities Policy Statement

Health Data Research UK is an equal opportunities employer, and as such aims to treat all employees, consultants and applicants fairly. It is our policy to provide employment equality to all, irrespective of:

- Gender, including gender reassignment
- Marital or civil partnership status
- Having or not having dependants
- Religion or belief
- Race (including colour, nationality, ethnic or national origins)
- Disability
- Sexual orientation
- Age

We are opposed to all forms of unlawful and unfair discrimination.  All job applicants and employees who work for us will be treated fairly and will not be unfairly discriminated against on any of the above grounds. Decisions about recruitment and selection, promotion, training or any other benefit will be made objectively and without unlawful discrimination.