# Big Data for Complex Disease Programme
## A National Population-level Resource for Exploring Data-Driven Cancer Research Questions

A. Webb[1], S. Boylan[1], A. Hodgkinson[1], A. Timmons[1], F. Taggart[1,2], E. Morris[1,3], G. Abel[1,4], A. Wood[1,5], M. Lawler[1,2], & the BDCD consortium

[1]Health Data Research UK, UK [2]Queen's University Belfast, UK [3]University of Oxford, UK, [4]University of Exeter, UK [5]University of Cambridge, UK

HDRUK Health Data Research UK

British Heart Foundation Data Science Centre
Led by Health Data Research UK

## Background

Big Data for Complex Disease (BDCD) is one of 5 research driver programmes funded by Health Data Research UK (HDRUK).

BDCD aims to use linked, de-identified population-level electronic health and health-related data to better understand:

- Causes of complex diseases, including cancers
- Who is most at risk and why
- Which current treatments are best and what differences there are in how the treatments work
- How to improve prevention and care for patients so that there are fewer deaths and less serious health problems

## Working with Patients and the Public

Our PPIE strategy has been developed in collaboration with our public contributors, researchers and stakeholders

1. Embedding patient & public involvement & engagement (PPIE) across all areas to strengthen our research.
2. Engaging and involving diverse audiences through varied and innovative methods.
3. Taking a collaborative approach to PPIE, sharing knowledge, insights and learnings within and outside the programme.
4. Creating long-term impact through dissemination methods that consider the public, healthcare professionals, and policy makers.

Amy Hodgkinson (PPIE Manager), Fez Awan, Mandy Rudczenko, and Jude Beng (Public Contributors) at Inequalities Workshop

- Co-created a lay summary
- Reviewed fellowship applications and sat on interview panel
- Reviewed abstracts
- Spoke at our Longitudinal Data Modelling Symposium
- Co-developed a PPIE session for inequalities workshop
- Helped recruit new PPIE manager
- Attended BDCD strategic meeting and HDR UK Conference

## Programme Structure

**Dataset curation:** UK-wide de-identified, linked, longitudinal population-level health data

SAIL Databank | NHSE SDE | Scottish Safe Haven | DataLoch | NI HBS

**Programme workstreams:**

- WS1 Disease Prediction
- WS2 Relationship between diseases
- WS3 Impact of inequalities

**Individual projects:** PhD Students and Fellows | Wider Consortium

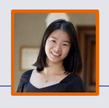
BDCD works with data custodians to provide access to data across each of the 4 Nations of the UK. BDCD researchers also gain access to data curation tools and support through the BHF Data Science Centre.

BDCD can help facilitate new projects from cancer researchers using health data to answer questions about:

- Cancer risk
- Interactions between cancer and other diseases or comorbidities, like cardiovascular disease or diabetes
- Underlying causes of inequalities in cancer prevention, diagnosis, treatment, and survival, and mechanisms to mitigate them

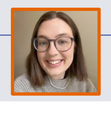
Predictive value of non-specific symptoms in the diagnosis of cancer and other diseases

Repurpose and enrich cardiovascular risk scores for cancer



Prediction models using blood tests to identify patients in primary care for cancer investigation

Impact of in-hospital diagnosis of cancer on patient relevant and NHS outcomes



Multi-state models predicting transition from severe mental illness to cancer

Health economic modelling of disparities in cancer screening

## Data Access and Available Tools

**Research question requiring whole population data**

Disease Risk | Relationships | Inequalities

**Identify datasets of interest available in BDCD**

Wales – SAIL | Scotland – National Safe Haven and DataLoch | England – NHSE SDE | Northern Ireland – Honest Broker Service

**Data access application support through BDCD**

Wales - IGRP | Scotland – HSC-PBPP | England – TBC | Northern Ireland – HBS-DAC

**Research support through BDCD and Health Data Science Team**

Feasibility | Reproducible pipelines and code | Data curation | Team science

**Tutorials**
- Step-by-step coding guidance for TREs

**Data summary and data insight dashboards**
- Analyse structure, quality, and utility of datasets

**Common functions**
- Standardized scripts to streamline data management

**Data curation pipelines**
- Reusable pipelines for transforming raw data to increase reproducibility

**Curated assets**
- User-friendly data tables generated by cleaning and integrating datasets

**Phenotypes**
- Standardised phenotype algorithms and comorbidity indices

## Acknowledgements

This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities, including Cancer Research UK. This work appreciatively uses data provided by patients and collected by the NHS as part of their care and support.

## Get Involved

Scan the QR code to find out more about the wide range of primary care, secondary care, and specialist disease datasets, including cancer registries, across national secure data environments (SDEs) that BDCD supports.