An introduction to the Health Data Research UK Gateway: facilitating discovery and access to health data

06/07/22

Susheel Varma, Ex- Chief Technology Officer, HDR UK
Current: Head of AI and Data Science, ICO
Our mission

HDR UK’s mission is to unite the UK’s health data to enable discoveries that improve people’s lives

Our 20-year vision is for large scale data and advanced analytics to benefit every patient interaction, clinical trial, biomedical discovery and enhance public health.

Core funders

Programme funders

BILL & MELINDA GATES FOUNDATION

THE HEALTH FOUNDATION

UK RESEARCH AND INNOVATION
One Institute Partnerships

Three programmes working with local, regional, national and international partners across NHS, academic and industrial ecosystems:

UK Regional Networks

HDR Global

Institute Office

Drawing on health data science expertise and infrastructure across the UK and globally. Delivering the transformative potential of health data research to impact everyone, everywhere
Our future strategy focuses on three integrated areas of activity

Uniting health data
Advancing research discoveries
Promoting partnerships
A trustworthy health data research ecosystem – differentiated talent, emerging technology, robust datasets, all connected through a platform

To transform science practice and health globally
1. Research Data Infrastructure and Services

Uniting health data through delivery of FAIR Research Data Infrastructure and Services
Leading healthcare and research organisations united to establish best practice, standards and tools.

Our aim is for the Alliance to be internationally recognised as a powerful collaboration that is enabling trustworthy use of health-relevant data at scale and networked to deliver public benefit.
The UK Health Data Research Alliance (70 members and growing...)
people don't understand the shocking state our data ecosystems and how hard that is to fix ... AI looks like a magic wand ... but it’s like putting in a flush toilet to a house that has no plumbing/sewage ... just doesn’t work.
We have built a UK-wide FAIR data infrastructure to accelerate and streamline health data research at scale.
Our role in the landscape of UK-wide health data and research

HDR UK is one part of the ecosystem & to succeed in our mission our role is to enable and convene other organisations.
The road to hell is paved with good intentions...

JA Taylor et al. BMJ Open 2021;11:e047575
Streamlining access to health data → Accelerate research cycle times

HRA and NHS Digital Checklist for Researchers (health data only – new applications)

1. HRA web tool
   - In your study research?
     - Yes: Contact sponsor
     - No: MRC “Data tool” (in progress)

2. MRC “Data tool” (in progress)
   - What data do you want?
     - Only NHS data
     - Will include OHIS data at some stage
   - NHS data needed?
     - Yes: Identify DATA CONTROLLER
     - No: Map the data flows

3. NHS DIGITAL pre-application requirements
   - Does the DATA CONTROLLER have an NHS Digital Sharing Permissions Consent (DSPC)?
     - Yes: MARS confirmation of approval letter
     - No: Contact NHS Digital to apply for DSPC
   - NHS Data requirements fulfilled?
     - Yes: Request to access data via NHS DARS Online
     - No: Contact NHS Digital for recommendation for approval from HRA

4. DARS ONLINE
   - Apply
     - Select required datasets
     - Articulate purpose for request
   - Work with NHS D via detail (eg Purpose for review)
   - Recommendation for approval from HRA

5. Approve DSA
   - Data issued
   - IDENTIFIABLE DATA
     - NHS ETHICS
     - $251 (CAG)
     - HRA
   - NON-NHS ETHICS
     - Contact materials
     - MRC Approval
   - ONS DATA
     - Enquiries@NHSdigital.net

Outstanding questions:
- A data only study – pseudo only data? Any role for HRA?
Building the Gateway for researchers and innovators to access health data

www.healthdatagateway.org

• Access to request over **774 datasets**, 188 tools, 199 educational courses and 2000+ publications
• A dedicated collection of datasets for each Hub – **140 in total**
• Facilitated **584 requests to access health datasets**, in particular actively supporting the government’s National Core Studies into COVID-19 ("Data and Connectivity")

“Really impressed with this resource. I think as a gateway to search by data type and indication, it’s a really powerful tool.”

David Leather, GSK
A platform already used by custodians to make their data available

- An independent Alliance of data custodians dedicated to improving human health
- 50 member organisations (incl. national bodies, hospital trusts, charities, cohorts)
- Guided by Principles for Participation based on transparency and public benefit
- Collaborating to design best practice and standards for the use of health data
Accelerate FAIR-ness in a federated world – Metrics, Cookbooks, Data Utility Framework, Maturity levels

<table>
<thead>
<tr>
<th>Findable</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1. (meta)data are assigned a globally unique and eternally persistent identifier.</td>
<td>No URI or PID and no documentation</td>
<td>PID without metadata or documentation</td>
<td>PID with limited metadata, just enough to understand the data</td>
<td>PID with standardised metadata, registered or indexed in a trusted data repository</td>
<td>Extensive metadata and rich additional documentation available and searchable in a trusted data repository</td>
</tr>
<tr>
<td>F2. data are described with rich metadata.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3. (meta)data are registered or indexed in a searchable resource.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F4. metadata specify the data identifier.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accessible</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 (meta)data are retrievable by their identifier using a standardized communications protocol.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1.1 the protocol is open, free, and universally implementable.</td>
<td>No user license / unclear conditions of reuse / metadata nor data are accessible</td>
<td>No metadata and user Access restrictions apply with only bespoke access</td>
<td>Appropriately licensed and limited (meta)data retrievable using standardised protocols</td>
<td>Public access (after registration) with (meta)data accessible (even when data is no longer available)</td>
<td>Open Access (unrestricted)</td>
</tr>
<tr>
<td>A1.2 the protocol allows for an authentication and authorization procedure, where necessary.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2 metadata are accessible, even when the data are no longer available.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interoperable</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.</td>
<td>Proprietary, non-open format data</td>
<td>Proprietary format accepted by certified and trusted data repository</td>
<td>Non-proprietary, open format (archival format)</td>
<td>Data additionally harmonised/data standardised using a standard vocabulary</td>
<td>Data is additionally linked to other data to provide context</td>
</tr>
<tr>
<td>I2. (meta)data use vocabularies that follow FAIR principles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I3. (meta)data include qualified references to other (meta)data.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Re-usable</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R1. meta(data) have a plurality of accurate and relevant attributes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1.1. (meta)data are released with a clear and accessible data usage license.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1.2. (meta)data are associated with their provenance.</td>
<td>No clear provenance of data (to facilitate replication and reuse)</td>
<td>Explanation of how data was or can be used is available with user access restrictions</td>
<td>Data automatically usable by machines and (meta)data meet domain-relevant community standards</td>
<td>Data stored in a trusted data repository</td>
<td>Data is reliable and tested against gold standard (reference data)</td>
</tr>
<tr>
<td>R1.3. (meta)data meet domain-relevant community standards.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Solution:** Cohort Discovery enables researchers to **discover, assess and request access to potential datasets** that exactly match the research project cohort definition using standardised **inclusion & exclusion criteria** and **co-variates**

Datasets with female patients between 18-35 who have asthma and diabetes and who are not smokers and not pregnant

Researchers are able to **reuse the cohort query** to define their research protocol when submitting their data access request

Researchers will be able to **reuse and compare cohort definitions** between similar protocols

Cohort definitions will be able to **reuse phenotype definitions** (asthma, diabetes) without resorting to using ICD-10, Read, SNOMED-CT codes

**Statistical Disclosure Control Policies:**
- User validation (e.g. Bona-fide Researcher)
- Low number suppression (e.g. >50)
- Query Count binning (e.g. 50, 60, 70, -)
- Query Rate limiting

**Total Patients:** 23K
- Dataset 1: 20K
- Dataset 2: 3K
- Dataset 3: 1K

Custodian controlled Cohort Beacons

Layout for illustration purpose only
Improving data access – our work so far

✓ Worked with Alliance members to develop a common data access form based on the Five Safes* for datasets listed in the Gateway
✓ We went through an harmonisation process to identify a set of essential questions common to all custodians
✓ Has launched on the Gateway for 283 datasets and 12 custodians (2 more soon). Look to continue to encourage datasets listed in the Gateway to use the new form
✓ Had 584 data access requests using the new form (requesting 90 datasets)

# Data Governance – Risk Profiles based on the Five SAFE principles

<table>
<thead>
<tr>
<th></th>
<th>SAFE People</th>
<th>SAFE Project</th>
<th>SAFE Data</th>
<th>SAFE Setting</th>
<th>SAFE Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Data</td>
<td>None</td>
<td>None</td>
<td>Low</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Aggregated Data</td>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>De-Identified Data</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Sensitive Data</td>
<td>High</td>
<td>Very High</td>
<td>Very High</td>
<td>Very High</td>
<td>High</td>
</tr>
<tr>
<td>Identifiable Data</td>
<td>Very High</td>
<td>Very High</td>
<td>Very High</td>
<td>Very High</td>
<td>Very High</td>
</tr>
</tbody>
</table>

## Data Sensitivity

- **Public Data**: None
- **Aggregated Data**: Low
- **De-Identified Data**: Medium
- **Sensitive Data**: High
- **Identifiable Data**: Very High

## Example Users

- **General Public**: None
- **Data Journalists**: Low
- **Researchers**: Medium
- **Hospital Analysts**: High
- **Clinical Staff**: Very High

---

**Data Governance – Risk Profiles based on the Five SAFE principles**

- **SAFE**
- **Unsafe**

---

**Example Users**

- **USE**
- **People**
- **Project**
- **Data**
- **Settings**
- **Outputs**
Data Use Registers

Getting the greatest use and benefit from the UK’s rich health data is an important goal— but it is essential that those whose data is being used can see what is being done with it.

The Gateway data use register functionality clearly shows how datasets published on Gateway are being used, by whom and most importantly for what purpose.

What benefits does the data use register offer?

**Automated process:** Data custodians who use the Gateway data access management system are automatically told about use of their datasets. This cuts down on administration time and provides the public with real-time data.

**Alignment to Five Safes Framework:** To show that only qualified people in secure settings can access the data, and for purposes that benefit the public, the content and structure of individual data uses follows the Fives Safes framework.

**Closing the loop on impact:** By providing a clear link between data use and research outputs we can better demonstrate the impact and value of using health data for research.

**Best practice:** The register implements a [national standard for data use registers](https://www.hdruk.ac.uk/access-to-health-data/data-use-registers/), developed after a wide consultation with data custodians and the public.

View the Data Use Register
For more information, visit the webpage [https://www.hdruk.ac.uk/access-to-health-data/data-use-registers/](https://www.hdruk.ac.uk/access-to-health-data/data-use-registers/)
What is a TRE?

A TRE is a Trusted Research Environment. Also known as ‘Data Safe Havens’, TRES are highly secure computing environments that provide remote access to health data for approved researchers to use in research that can save and improve lives.

Why are they important?

TREs make research safer. Making data available through a TRE means that people can be confident that their personal health data is accessed securely and their privacy protected. TREs help make research efficient, collaborative and cost effective, providing rich data that enables deep insights which will go on to improve healthcare and save lives. TREs provide approved researchers with a single location to access valuable datasets. The data and analytical tools are all in one place, a bit like a secure reference library.

How is my data safeguarded?

Health data should always be kept safe and secure, and used responsibly to ensure privacy. Health Data Research UK ensures these high standards are met by promoting the use of the ‘Five Safes’ model across all TRES.

- **Safe People**
  - Only trained and specifically accredited researchers can access the data

- **Safe Projects**
  - Data is only used for ethical, approved research with the potential for clear public benefit

- **Safe Settings**
  - Access to data is only possible using secure technology systems – the data never leaves the TRE

- **Safe Data**
  - Researchers only use data that have been de-identified to protect privacy

- **Safe Outputs**
  - All research outputs are checked to ensure they cannot be used to identify subjects

Learn more about TREs and discover examples of how TREs are being used to enable life-saving health research.
Federated TRE User Journey – COVID-19 National Core Studies – Data & Connectivity

Innovation Gateway
- Search Metadata
- Sign up/Login to Gateway
- Cohort Discovery
- Request access to data
- Transfer code & configuration
- Submit Task/Workflow
- Receive aggregated results

Federated Services
- Federated ID + Accreditation Status
- Cohort query engine
- Data Access Workflow
- Coordinate data/code transfer
- Workflow Execution Service
- Coordinate results aggregation

Trusted Research Environment
- Federated Catalogues
  - TRE 1
  - TRE 2
  - TRE n
- Auth + Accreditation Status
- Cohort Query Execution
- DAB* Processing
- Provision Workspace
- Task execution Service
- Airlock/Escrow Service

User action

Automated action

Process

Authenticated user action
Trusted Research Environments Landscape View (research only – non-exhaustive)

National TREs
- HSCNI - Honest Broker Service
- PHS - Scottish National Safe Haven
- NHS Digital
- UKSeRP
- ONS - SRS
- OpenSAFELY

Commercial TREs
- AIMES
- AridhiaDRE
- AzureTRE
- AWS ServiceBench
- Lifebit
- ...

International TREs
- Terra.bio
- TEHDAS
- MedCo
- ICODA
- DNASTack
- ...

Domain/Location-Specific TREs (some being created)
- Genomics England
- QResearch
- EMBL-EBI*
- CPRD
- Turing Institute
- MRC-CLIMB (COG-UK)
- PIONEER
- eMedLab
- DISCOVER-NOW
- UKHSA
- CIPHA
- Our Future Health*
- ...

The future of TREs is already here, just unevenly distributed & fragmented
Open, Federated and Interoperable Technology Stack for Trusted Research Environments

**Identity Federation**
Provides authenticated, authorized and auditable access to federated resources using standardized single sign-on and identity federation.

**Analytics Federation**
Reuse and combine portable tools and workflows to enhance healthcare delivery with advanced data-driven translational insights.

**{Meta}Data Federation**
Discover, explore, organize and securely access federated data for accelerating the translation into healthcare.

**Infrastructure Federation**
Immediate access to advanced & flexible hybrid cloud-based computational resource including access to specialized accelerators and container orchestration services.
FOREST – Federation Capability Maturity Model (Unbundling TREs)
## Data Governance Structures / Design Patterns

<table>
<thead>
<tr>
<th>Governance Structures</th>
<th>Relationship</th>
<th>Data Availability</th>
<th>Permissions</th>
<th>Governance Challenges</th>
<th>IP, License &amp; Contracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed</td>
<td>One → None</td>
<td>Very Low</td>
<td>Very Low</td>
<td>Limits to collaboration</td>
<td>Enshrined in Law</td>
</tr>
<tr>
<td>Restricted</td>
<td>One → Some</td>
<td>Very Low</td>
<td>Low</td>
<td>Limits to collaboration</td>
<td>Enshrined in Law</td>
</tr>
<tr>
<td>One-to-one</td>
<td>One → One</td>
<td>Low</td>
<td>Low</td>
<td>Power asymmetry</td>
<td>DSA, Contract</td>
</tr>
<tr>
<td>Clique</td>
<td>Some → Some</td>
<td>Low/Medium</td>
<td>Medium</td>
<td>Trusteeship</td>
<td>Consortium Agreement</td>
</tr>
<tr>
<td>Trusted Research Environments</td>
<td>Some → Many</td>
<td>High</td>
<td>Low/Medium</td>
<td>Trusteeship / user/project validation</td>
<td>DSA, Contract</td>
</tr>
<tr>
<td>Federated Analytics</td>
<td>Some → Many</td>
<td>High</td>
<td>Low/Medium</td>
<td>Trusteeship / user/project validation</td>
<td>DSA, Contract</td>
</tr>
<tr>
<td>Federated Query</td>
<td>Some → Many</td>
<td>High/Medium</td>
<td>Low/Medium</td>
<td>Trusteeship / user validation</td>
<td>DSA, Contract</td>
</tr>
<tr>
<td>Open Access</td>
<td>Some → Many</td>
<td>High</td>
<td>High</td>
<td>Revocation of rights</td>
<td>License</td>
</tr>
<tr>
<td>Citizen Science</td>
<td>Many → Many</td>
<td>High</td>
<td>High</td>
<td>Uneven capacity for analysis</td>
<td>Contract or License</td>
</tr>
</tbody>
</table>

Adapted from SBN Green Paper
Operationalizing Data Governance at scale – TREs are not a silver bullet

- **Principles** – FAIR, CARE?, SAFE?
- **Requirements, Roles & Responsibilities, Rules of participation**
- **Interoperable Standards, Processes, Policies, Frameworks**
- **Modular Design Patterns, Software, Libraries**
- **Exemplars & Extensible Use Cases, Training, Awareness, Community Engagement**

We need a set of **PRIME Directives** to help operationalize, interoperate and federate across data governance patterns.
The future - building on these excellent foundations and partnerships to deliver a step change in benefits for UK science and population

- **Data as Infrastructure** – building blocks to create interoperable global networks focused on collaboration and sharing
- **Clinical/Scientific Driver Programmes** that help guide our development efforts and pilot tools
- **UK wide** and at the centre of an international collaborative network of science
- **Diverse data types beyond just NHS data** (omics, biomedical, wearables, social sciences)
- **Thought leadership and expertise for innovative approaches to trust, governance and standards** for data collection and access
- **Exemplar of team science and partnership** working across industry, academia and healthcare
- **Patient Involvement and Engagement** at all stages of the research/innovation lifecycle
- **Training** the current and next generation of scientists to responsible, ethical and equitable use of AI
DARE UK - Key needs and opportunities
dareuk.org.uk

1. **Data and discovery**, including technical standards such as data standards, metadata catalogues and common data models

2. **Access and accreditation**, including governance standards, rules and frameworks for enabling data access e.g., accreditation of TREs and researchers

3. **Digital research infrastructure**, including the physical and software infrastructure

4. **Capability and capacity**, including shortages of data scientists and data engineers, and career pathways for those maintaining infrastructure

5. **Maintaining trust**, namely demonstrating trustworthiness, gaining public and organisational trust, and addressing risk aversion of those holding data

6. **Funding and incentives**, including the current research culture, need for sustained funding of infrastructure and research, and the responsibilities of different groups involved

COVID-19 National Core Studies – Data & Connectivity

Data available for COVID-19 research across the UK

August 2021

This document sets out the data sets available for COVID-19 research across a network of UK Trusted Research Environments (TREs), made available by the

Funded by: Health Data Research UK, in partnership with Office for National Statistics.

Key outputs across the NCS programme to date

- 489 Publications in academic journals, with 124 per print
- 93 High-quality datasets made available via the Health Data Innovation Gateway, described in this brochure
- 190 NCS-linked data use in the Data Use Register, a new standard for transparent public reporting shared in this HDRUK white paper
- 847 Researchers supported through 86 projects

The COVID-19 National Core Studies (NCS) are a crucial part of the UK’s ongoing pandemic response. They enable the UK to use health data and research to inform both the near and long-term responses to COVID-19, as well as accelerating progress to establish a world-leading health data and research infrastructure for the future.

The 6 National Core Studies are:
- Epidemiology and Surveillance led by Imran Diamond (Office for National Statistics) collects and analyses data to inform restrictions and protection against infection outbreaks.
- Clinical Trials Infrastructure led by Patrick Chilvers (Medical Research Council) accelerates delivery of large scale COVID-19 trials for drugs and vaccines.
- Transmission and Environment (also known as PROJECT) led by Andrew Curran (Health and Safety Executive) improves understanding of COVID-19 virus transmission in different settings and environments.
- Immunity led by Paul Moss (University of Birmingham) supports research to improve understanding of immunity against COVID-19, to inform back-to-work policies.
- Longitudinal Health and Wellbeing led by Ruth Charlwood (University College London) and Jonathan Sterne (University of Bristol) uses data from longitudinal studies to address the impact of COVID-19 and inform mitigating strategies.
- Data and Connectivity led by Freda Morris (Health Data Research UK) in partnership with Office for National Statistics makes UK-wide health and administrative data available to catalyze COVID-19 research.

This new quarterly report aims to:
- Communicate the impact the National Core Studies are having on COVID-19 responses
- Promote NCS Open Science data, tools and resources to ensure they are taken up
- Highlight where advances and learning gained during NCS translates into a legacy of stronger health threat preparedness

Managed by:

bit.ly/HDRUK-NCS
Are there examples of research being conducted in TREs?

1,168 COVID-19 pre-prints and 98 published papers. Better Care, Understanding Causes of Disease, Clinical Trials & Public Health

293 COVID-19 research projects using national data custodian data

300 Health Data Research Hub contracts with industry & academia

>22,000 patient and public participants in health data research

HDR Standards: Trusted Research Environments, Data Utility, Federation

Streamlined data access request process, harmonised across TREs. Time from application to active research: 3-14 days

701 discoverable datasets

>16,000 monthly searches

Example: Zoe COVID-19 symptom tracker dataset

Example: Hubs + CVD - COVID-19 For the first time, linked health data resource covering 54.4 million people

112 datasets set up in 5 national trusted research environments – by National Core Studies Data & Connectivity

1 Wales and Scotland data only
Thank you for listening
Here’s some obligatory holiday snaps from the peak district

bit.ly/HDRUK-TRE
hdruk.ac.uk
healthdatagateway.org

Email: enquires@hdruk.ac.uk
Twitter: @HDR_UK