AI for Multiple Long-term Conditions
Research Support Facility

Research and Data Infrastructure Horizon Scan

The Alan Turing Institute

THE UNIVERSITY of EDINBURGH

Swansea University
Prifysgol Abertawe

FUNDED BY
NIHR | National Institute for Health and Care Research

This document is up to date from 1st December 2024.

This document will be updated on a regular basis.

If the information in this document has changed, or there are suggestions for additional data sets, or you have a general query about the content – please use the contact details below.

Corresponding Author:
Chris Orton
Swansea University Medical School
C.Orton@swansea.ac.uk

Authors:

Chris Orton
Swansea University

Reece Labrom
Swansea University

Rachael Stickland
The Alan Turing Institute

Beth Collop
University of Edinburgh

Monica Fletcher OBE
University of Edinburgh

TABLE OF CONTENTS

INTRODUCTION

This Research and Data Infrastructure Horizon Scan provides a comprehensive overview of data providers, infrastructures, and platforms that can support research into multiple long-term conditions (MLTC). With a focus on enabling advanced analytics, including machine learning and artificial intelligence, the document is designed to facilitate groundbreaking scientific discoveries by consolidating high-level information on key data sources in one accessible resource. The document includes details on 18 nationally and regionally representative data providers and infrastructures. These entities host extensive health and administrative datasets that are crucial for research into MLTC. By offering an easily accessible summary, the document serves as a preliminary resource for researchers to identify potential data options aligned with their objectives. The information presented is derived from publicly available sources and is deliberately kept at a high level, with links provided for users to delve deeper into specifics. The focus is on data providers already known to researchers in the MLTC domain or those with the potential to complement existing datasets, thus enhancing the scope of current or future research. As the MLTC research landscape evolves, this document is intended to be regularly updated, incorporating additional data providers to reflect emerging opportunities and maintain its relevance. It acts as a strategic guide for researchers navigating the complexities of data availability and integration in the pursuit of transformative MLTC research.

The AI for Multiple Long-term Conditions Research Support Facility (AIM RSF) located at the Turing Institute in collaboration with Swansea University and the University of Edinburgh, is funded by the NIHR AIM Research Support Facility (NIHR202647). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

https://cprd.com/
CPRD Datasets

Geographical Area Coverage
UK – Primarily England, some Welsh providers

Description
Real-world research service supporting public health and clinical studies, incorporating a collection of nationally representative primary care records.

Population Coverage
18m (currently registered), up to 60 million using longitudinal records

Data Types
- Anonymised primary care data from EMIS (Aurum) and VISION (Gold) software systems. Linkable to Hospital Episode Statistics, mortality records, COVID-19 specific datasets, and other NHS/ONS registers.
- GP data is coded in SNOMED-CT, with associated CPRD medcode and Read code lookup.

Access method(s)/permutations
- Physical transfer of data via secure download.
- Reliant on research teams having available infrastructure to host the data and share across the research team.
- Data must be selected in detail as part of the application process (variable selection)
- Currently developing a TRE for remote access to data on a secure system moving forwards.

For more: https://www.cprd.com/data-access

Cost Implications
Cost recovery model based on an annual institutional licence needed on a multi or single study basis, also dependent on nominated users accessing the data. Extra costs for linked secondary care, mortality and other information such as ethnicity or COVID-19 specific data.

Discussion
- Data is provided in prepared statistical formats, and has unique linkages across events where patients are assigned multiple IDs.
- Suitable for MLTC research due to the data linkage, history, and all-cause nature of the data.

DATA LOCH

https://dataloch.org/
DataLoch Datasets

Geographical Area Coverage
South-east Scotland (predominantly Lothian).

Description
A data service and TRE developed between University of Edinburgh and NHS Lothian to provide access to deidentified data collected as part of routine interaction with health and social care.

Population Coverage
900k

Data Types
GP, hospital, lab, mortality and other outcomes, registry, demographics.

Access method(s)/permutations
TRE – developed with Public Health Scotland and the EPCC at the University of Edinburgh.
For more: https://dataloch.org/data/how-to-access

Cost Implications
Cost recovery for support (e.g. data preparation) and infrastructure (provision of computational resources within TRE).

Discussion
- DataLoch provides one of the only large-scale primary care datasets in Scotland which is widely available for research into any condition.
- Specific disease registries have been set up, and DataLoch prepare data into research-ready datasets before provisioning them in their TRE for use (all curation of raw data undertaken by their analyst team).

DEMENTIAS PLATFORM UK (DPUK)

https://www.dementiasplatform.uk/
https://portal.dementiasplatform.uk/

Geographical Area Coverage
📍 UK-wide, some international data

Description
TRE for longitudinal cohort studies and other data collections wishing to share data for dementia, neuroscience, and ageing research, also able to support MLTC research where relevant.

Population Coverage
👥 ~3m participant records across all contributing cohorts.

Data Types
Datasets are primarily consented cohort or clinical studies with a dementia, neuroscience or ageing focus, which typically consist of survey, clinical, omics, and imaging data.

Access method(s)/permutations

- TRE – 2FA virtual desktop infrastructure hosted by SeRP UK at Swansea University.
- Data are typically provided in files, mostly statistical formats such as delimited files (csv, tsv), SPSS, and Stata. Many software are made available to curate and analyse data in the environment.
- Datasets must be selected and justified as part of the application.

For more: https://portal.dementiasplatform.uk/Apply/ApplicationProcess

Cost Implications
💰 Some cohorts have modest cost recovery charges to access data which are paid to the study directly upon approval of an application, however the TRE itself is mostly free to access. There are limited cases where costs would be incurred, for example where a user needs an uplifted technology specification.

Discussion
Whilst having a neuroscience and dementias theme, this topic is particularly relevant to MLTCs in ageing, and the data providers to DPUK have highly detailed records on ageing populations.
Cohort datasets typically represent far fewer people than routine outcomes data, however collect granular information on lifestyle, health, and administrative data such as economic status. They can therefore sometimes run to thousands of variables which presents data management complexity over and above that of routine health records, but provide extra specificity into the insights of longitudinal population study.

DISCOVERNOW

https://discover-now.co.uk/
DiscoverNOW Datasets

Discover-NOW
Health Data Research Hub for Real World Evidence

Geographical Area Coverage
📍 North-west London (soon to be all London), England.

Description
Health data research hub for real-world evidence, supporting clinicians, researchers and scientists in recruiting patients to new studies, develop treatments, manage and prevent health conditions, and test and evaluate innovations.

Population Coverage
👥 ~2.7m (growing to 10.8m)

Data Types
North-west London deidentified dataset – linked coded primary care, secondary, acute, mental health, community health, and social care. Same data types will be present in the planned scale-up to all-London.

Access method(s)/permutations
- Remote access to an NHS-hosted analytical environment.
- Data Access Committee reviews are based on an application detailing clinical sponsorship, data needed, legal basis, and other FAIR processing guidelines.

For more: https://discover-now.co.uk/how-to-access-the-data/

Cost Implications
💰 Tiered pricing model for access based upon services provided to any project.

Discussion
- DiscoverNOW is a deidentified mirror of the North West London Integrated Care System (WSIC). Presence of patients in the data is based on their registration with a GP, much as with other national and regional health record providers for research. Links to the WSIC dashboarding system and can provide pre-built algorithms such as in frailty, self-care, and long-term conditions as defined by QOF.
- Full information provided on datasets here: https://discover-now.co.uk/the-data/

Electronic Data Research and Innovation Service (EDRIS) (Public Health Scotland (PHS))

Public Health Scotland

Public Health Scotland

Public Health Scotland Datasets

Geographical Area Coverage

Scotland

Description

Data service for access to national records representing Scotland. Administers the Scottish National Safe Haven for TRE access to data, and can also provide data physically subject to infrastructure satisfying PHS guidelines.

Population Coverage

~5.5m

Data Types

National data for the Scottish population including audits/registers, cancer data, child health, healthcare utilisation, hospital activity statistics, screening and mortality.

Access method(s)/permutations

- Physical data share subject to review of proposed infrastructure to house Scottish data.
- TRE – Scottish National Safe Haven (VPN and 2FA).

For more: https://publichealthscotland.scot/services/data-research-and-innovation-services/electronic-data-research-and-innovation-service-edris/what-data-is-available-and-how-it-can-be-accessed/

Cost Implications

Cost recovery based on support and effort in facilitating approvals, extracting and linking data, and facilitating use of the National Safe Haven where relevant.

Discussion

- Currently does not provide access to GP records, although PHS did facilitate GP record use throughout the pandemic for observation and surveillance purposes.
- Mirrors data held by the English, Welsh, and Northern Irish data providers for national level hospital records, disease registries, and mortality data.

HONEST BROKER SERVICE (HBS)
HEALTH AND SOCIAL CARE NORTHERN IRELAND (HSCNI)

https://hscbusiness.hscni.net/services/2454.htm
HSCNI Datasets

Geographical Area Coverage
Northern Ireland

Description
Trusted Research Environment for HSCNI providing anonymised patient level data for research.

Population Coverage
~2m

Data Types
Predominantly secondary care data encompassing inpatient and outpatient data as well as laboratory results and cancer pathways. Ongoing discussions with GPs across Northern Ireland about mobilising primary care data for research.

Access method(s)/permutations
TRE – either via the physical safe setting in Belfast, or remotely using the HDRNI UK SeRP remote platform, provided by Swansea University (2FA virtual desktop infrastructure).
For more: https://bso.hscni.net/directorates/digital/honest-broker-service/honest-broker-service-researcher-access/

Cost Implications
Cost recovery based on support services for providing data (data itself carries no charge) and costs for providing the analytical platforms for use on various research projects.

Discussion
HBS are fairly early in their service provision of data for research, however are currently able to extract data for predominantly secondary care research and some linked data across care pathways for wider health analyses. The largest difference to the rest of the UK are the secondary data use laws and the relative separation of each arm of the healthcare system which means integrating data from primary, secondary, and tertiary care is currently a work in progress.

NHS England (NHSE)

https://digital.nhs.uk/
NHS England Datasets

Geographical Area Coverage
England

Description
National data provider for NHS England patients, including all routinely collected hospital data and associated administrative records, whole-nation mortality, mental health data, and some specific COVID-19 datasets including whole-nation primary care.

Population Coverage
59.6m (based on 2021 census).

Data Types
- Routinely available: secondary care, associated admin records for care pathways, lab results, mental health data, maternity services, COVID-19 records.
- Subject to COVID-19 use cases and legal bases: Primary care.

Access method(s)/permutations
- Physical share – data transfer is subject to the process set out by Data Access Request Service (DARS) and having sufficient infrastructure to pass NHS scrutiny.
- NHS England Secure Data Environment (cloud-based remote access) is currently available via the British Heart Foundation Data Science Centre, and in some limited other cases, moving to a more standard model of access for users in the near future.
- Data is typically shared in delimited format (csv, tsv).

For more: https://digital.nhs.uk/services/data-access-request-service-dars

Cost Implications
Costs are incurred on an annual basis and are based on costs per dataset, per linkage, per extraction. Further sharing of data to third parties as part of an approved application incurs a further sub-licence fee.

Discussion
- Whilst hospital records and associated data are routinely shared and maintained, primary care data from NHS England is not yet available for anything other than COVID-19 research, although some elements of MLTC research can also be performed in conjunction with the COVID-19 broad topic area.
- The application process is in one stage but reviewed by a Professional Advisory Group and by what is currently an Interim Data Advisory Group.

NHS England sub-national Secure Data Environments (SDEs) *In development*

England SECURE DATA ENVIRONMENT | Part of the NHS Research Secure Data Environment Network

https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/

https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/accessing-data-for-research-and-analysis/

Geographical Area Coverage

📍 Regions of England, each made up typically of several NHS trusts.

Description

As part of the transformation of NHS England, and the creation of the NHS England national SDE, regional networks in England are being established to enable data sharing and data integration across NHS trusts in specific areas. From a research perspective, it is expected that GP, hospital, lab, registry, and administrative health data will be available through the integrated regional SDEs, however they are also being established to improve direct healthcare.

Population Coverage

👥 Several million per region.

Data Types

Fully integrated care records – GP, hospital, lab, and community health data.

Access method(s)/permutations

SDE (Secure Data Environment) – via remote access to a (likely) cloud-based virtual system.

Cost Implications

💰 Likely to follow a similar pattern to the main NHS England national service.

Discussion

These SDEs will provide access to regional data in England, and the integration of the trusts with the administration of the systems is intended to improve the data quality and coverage available – including potential uplift in GP data availability.

Office for National Statistics (ONS) Secure Research Service (SRS)

https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice
Access the data securely - Office for National Statistics (ons.gov.uk)

Geographical Area Coverage
📍 UK

Description
A TRE providing access to deidentified, unpublished data, which are predominantly administrative records but some health-related data.

Population Coverage
👥 Varies across datasets, but data is typically UK-level.

Data Types
Nationally collected administrative records on the population of the UK, predominantly collected by government departments.

Access method(s)/permutations
- TRE – SafeRoom, SafePod, or Assured Organisational Connectivity (AOC) – the latter is the most likely for UK-wide researchers. AOC is established through an agreement between the researcher's institution and ONS.
- Users must become an accredited researcher under the Digital Economy Act to access data in the SRS.
- For more: https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/onsresearchanddataaccesspolicy

Cost Implications
💰 Currently free at the point of use.

Discussion
The ONS SRS was particularly active during the midst of the COVID-19 pandemic due to the linkages between the administrative records held and some NHS England data which was onboarded. It has now largely returned to an administrative data provider, including records from ONS and other government departments, but linkage of this data to health records is still possible subject to processing standards and legal bases.

OPENSAFELY

https://www.opensafely.org/

Geographical Area Coverage

England

Description

Software platform for analysis of electronic health record data, currently deployed within secure data centres of primary care providers TPP and EMIS. Created during the COVID-19 pandemic and currently data is only available for access for COVID-19 related research.

Population Coverage

58m

Data Types

Primary care data held on primary care systems providers TPP and EMIS, with linkage to HES/SUS, Emergency Care, ONS data such as the COVID-19 infection survey, household identifiers and mortality.

Access method(s)/permutations

- Via the OpenSAFELY platform, a federated software solution where researchers prepare code which is then deployed through a connection to a 'Jobs site' to OpenSAFELY, dependent on having software such as Python and Docker installed.
- Applications for data access are approved by the data controller (i.e. NHS England) which then permits queries to run against the real data via the Jobs site.
- For more: https://docs.opensafely.org/data-access-policy/

Cost Implications

Currently free to access but only for onboarding of initial users.

Discussion

- OpenSAFELY provides a federated solution to deploy analytical queries to primary care data held primarily within the TPP and EMIS systems providers, with linkages to secondary care and COVID-19 data also possible.
- OpenSAFELY works differently to most other TREs, whereby there is no direct access to individual level data, and all code has to be developed using metadata or dummy data and then converted to a script to deploy through their jobs site to return results and outputs. Whilst this may suit some analytical approaches, it means that iteration and visibility of data manipulation will be more complex, and will need specific skillsets from users, particularly R, Python, Git, and potentially in wrapping code into Docker containers.
- One of the rules that users must follow is in transparency of code and executed analytics, where all code used and deployed through OpenSAFELY must be published and open-sourced for the community to review and access.
- Currently, the data is only available for COVID-19 research and external pilot users, rather than being generally available for any research application.

PIONEER – THE HEALTH DATA RESEARCH HUB FOR ACUTE CARE

https://pioneerdatahub.co.uk
Pioneer Hub HDR Gateway Collection
Pioneer Hub Datasets

Geographical Area Coverage

West Midlands, England

Description

PIONEER is the Health Data Research Hub for Acute Care, led by the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, in partnership with West Midlands Ambulance Service. Acute care is the provision of unplanned medical care; from out of hours primary care, ambulance assessment, emergency medicine, surgery and intensive care.

Population Coverage

6m

Data Types

Electronic health record data from multiple data providers for children and adults encompassing physiology, acuity scores, laboratory results, imaging, pathology samples, diagnostics, drug administrations, interventions, and previous and subsequent healthcare contacts which led to (or resulted from) an acute healthcare event. Coverage is from 1st January 2000 and incorporates ongoing data collection.

Access method(s)/permutations

- Via the PIONEER TRE, or dissemination to an approved TRE of equivalent security for use of PIONEER data
- Application process via an online form, subject to approval by the Data Trust Committee and University Hospitals Birmingham NHS Foundation Trust (as PIONEER data controller)
- Data licensed only for the approved use case, not permitted to be used for wider purposed than the agreed project
- Applications open (but not limited) to NHS providers, academics, commercial organisations including SMEs

For more: https://www.pioneerdatahub.co.uk/data/data-request-process/

Cost Implications

Services are costed per request. Typically, this includes use of the TRE for accessing data on a licensed basis, as well as any other support requested such as data and infrastructure consultancy including linkage, synthetic data generation, specialist data such as imaging, AI algorithm expertise, software development, and data insight and problem solving.

Discussion

- Given the disease-inclusive nature of the PIONEER Hub and presence of symptoms and demographics unrestricted by age, ethnicity, and multimorbidity status, the data would be suitable for MLTC research with multiple deployment methodologies.
- PIONEER can provide dedicated support for use of their data and systems as part of their service catalogue.

QRESEARCH

https://www.qresearch.org/
QResearch Datasets
QResearch HDR Gateway Collection

Geographical Area Coverage
UK – participating GP practices using EMIS.

Description
Consolidated database derived from the anonymised health records of over 35 million patients (longitudinal data collection going back to 1989).

Population Coverage
35m (including current and historical patients since 1989 – cumulative figure).

Data Types
GP records linked to mortality, cancer, and hospital data.

Access method(s)/permutations

- Open to only academic researchers from UK universities – access via remote desktop over VPN with 2FA authentication. User machines must meet established criteria as per the QResearch Workstation Setup Requirements, as VPN connection is locked to specific user machines.
- Infrastructure is provided by an on-premise cluster based at University of Oxford's Data Centre. As standard, the virtual desktops include Stata, R, and Python.
- Datasets are provisioned based on variable selection and sample size needed for the research question.

For more: https://www.qresearch.org/information/information-for-researchers/

Cost Implications

- Cost recovery based on all aspects of the administration and operation of the QResearch project lifecycle, more information at https://www.qresearch.org/information/information-for-researchers/.
- The approximate cost of 1 user accessing data for one project for one year, with GP data only and up to 1m records, is £30,000 + VAT.

Discussion
QResearch contains the records of certain EMIS providing practices across the UK, and whilst the total population of data going back to 1989 is estimated to be 35m, the live and registered population will be smaller for projects needing more recent records. The pathway to approvals depends on a set of steps needed to initiate the formal application, which includes a pre-submission protocol to enquire as to the feasibility and cost of a project, securing funding (unless already secured), and receiving scientific committee approval prior to data extract readiness in the TRE. QResearch describes the most suitable studies for their data as: Case control, cross-sectional, cohort, feasibility, and sample size calculations, which could prove useful in identifying cohorts within MLTC patients and analysing these using ML/AI methods in the provided software.

ROYAL COLLEGE OF GENERAL PRACTITIONERS RESEARCH AND SURVEILLANCE CENTRE (RCGP RSC)

https://www.rcgp.org.uk/representing-you/research-at-rcgp/research-surveillance-centre
RCGP HDR Gateway Entry

Geographical Area Coverage
England

Description
Linkable research resource containing pseudonymised primary care records, linkable to hospital, death, and other nationally collected data.

Population Coverage
4m (currently registered patients across ~1700 GP practices), over 13m patient records longitudinally.

Data Types
Primary care data including demographics, diagnoses and symptoms, drug exposures, vaccination history, laboratory tests, and referrals to hospital and specialist care.

Access method(s)/permutations
- Access via the ORCHID database system, hosted by the Nuffield Department of Primary Care Health Sciences at the University of Oxford.
- Data access process entails justification of ethical approval, SQUIRE-based research protocol, and variable selection prior to extract release into ORCHID.
- Access is via VPN and 2FA (DUO app) into the ORCHID remote desktop system, analytics being supported by various desktop specification choices and as standard containing RStudio (Linux-based) and SQL server.

For more: https://www.rcgp.org.uk/representing-you/research-at-rcgp/research-surveillance-centre/supporting-research-teams

Cost Implications
Cost recovery is implemented based on the needs of each project – based on the support of the team at the RCGP RSC, ORCHID, and in the volume and complexity of data being extracted and accessed.

Discussion
- The periodicity of the data refreshes of the RCGP RSC data (twice-weekly) means that near real-time surveillance and research can be achieved on granular primary care records across providing practices, which is near unparalleled in UK data providers. However, linkage to wider datasets is not provided as standard alongside the primary care datasets and further applications for hospital, disease-specific, and mortality data are likely needed.
  The technical systems provided to access the data support the use of R, Stata, and direct curation in the SQL server where data is presented.

SAIL DATABANK

https://saildatabank.com
SAIL Databank Datasets

Geographical Area Coverage
Wales

Description
Trusted Research Environment (TRE) for Wales, also incorporating some UK-wide and international datasets

Population Coverage
3.3m (Welsh population), up to 5.5m using longitudinal records.

Data Types
GP data for ~90% of the Welsh population, 100% coverage of hospital data, lab data, cancer records, and linked administrative records such as Census.
GP data is coded in Read v2/3 and some local codes.

Access method(s)/permutations
- TRE (SAIL Gateway – 2FA virtual desktop infrastructure)
- Data provided via a DB2 database. SQL, R, and Python are most commonly used in connecting to the database or using files extracted from the database. After database interaction, other software (SPSS, Stata) also available. A universal linkage key (ALF) is used across datasets, but for deep dive linkage into data such as hospital records, linking on different fields is needed.
- Datasets must be selected and rationale given for their inclusion in any application, however variable selection within each dataset is not needed other than in specific cases such as ethnicity, geography, deprivation etc. as data related to these is provisioned separately.

For more: https://saildatabank.com/data/apply-to-work-with-the-data/

Cost Implications
Currently free to download data or use the SecureLab subject to eligibility checks.

Discussion
- Data of sufficient granularity to run studies on the majority of the population of Wales, and use within UK-wide studies, either as feasibility for scale-up or as part of a full study.
- Data curation is a large part of using SAIL – taking raw health records and shaping them for statistical analysis or model training.
- Suitable for MLTC research due to the data linkage, history, and all-condition nature of the data.

UK BIOBANK

https://www.ukbiobank.ac.uk/
UK Biobank Data

Geographical Area Coverage
England (89%), Scotland (7%), Wales (4%)

Description
Large-scale biomedical database and research resource containing in-depth genetic and health information. Participants were aged 40-69 when they joined between 2006-2010 and provided consent for long-term follow-up.

Population Coverage
500k

Data Types
Questionnaire data on health and lifestyle, sample data derived from blood, urine, and saliva, physical measurements (height, weight, spirometry, blood pressure, heel bone density), MR brain and heart imaging, activity monitoring, and genetics. UK Biobank participants also consent to linkage with health-related outcomes data: primary care data is available for 230k participants, hospital inpatient data available for the full cohort, with derived 'first occurrence' data fields for codes related to medical conditions and outcomes, cancer data, and mortality.

Access method(s)/permutations
- Via the UK Biobank Research Analysis Platform (RAP), a cloud-based TRE enabled by DNAnexus technology and powered by Amazon Web Services (AWS). Remote analysis platform pre-loaded with tools for genomic, statistical, and imaging analysis.

For more: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access

Cost Implications
UK Biobank has a tiered access fee system dependent on data being requested and the length of retention/access. For example, Tier 1 contains core data only (questionnaires, linked health records, phenotypes), whereas Tier 3 contains all Biobank data including genomics and imaging. The RAP also has a separate costing mechanism based on a user's usage of the cloud infrastructure over and above the standard platform – i.e. needing extra compute, storage, or software in addition to that of the basic platform provided under the Tier costs.

Discussion
UK Biobank offers one of the world's largest and most in-depth cohort studies, and also has the world's largest imaging study as part of its offering. Given the age of the participants, MLTC research is possible due to the average age of those experiencing MLTCs.

UK Birth Cohort Data

🌐 Multiple sites for different datasets
https://nshd.mrc.ac.uk/
https://ncds.info/
https://bcs70.info/
https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

Geographical Area Coverage
📍 UK – Primarily England

Description
Birth cohorts are longitudinal studies of a set of participants who are recruited at birth and followed-up at regular intervals for either set periods of time, or the rest of their lives.

Population Coverage
👥 Typically, between 15-20,000 consented participants per study.

Data Types
The principal British cohort studies, which collect medical and social data on their participants via survey, clinical testing, and linkage, are the National Survey of Health and Development (NSHD), National Child Development Study (NCDS), 1970 British Cohort Study (BCS70), and the Millennium Cohort Study (MCS).

Access method(s)/permutations
Primarily through data download from services such as the UK Data Service or dedicated secure transfer from the study themselves. UKLLC (see next row) also provide access to these data through their TRE (remote access), currently only for COVID-19 related research.

Cost Implications
💰 Most cohort data is currently free to access or has modest access costs (low £000's at most). UKLLC also operate as free at the point of access currently.

Discussion
These data are typically supplied in pre-formatted statistical files, apart from via UKLLC where they will be via SQL database.
Whilst these datasets are small in participant numbers, some of their granularity and linkage between Patient/Participant Reported Outcome Measures (PROMs), wider survey, health data, and administrative records is unparalleled given the connection between the participant and their data. They will, due to the age of the participants in most of the studies, often contain multimorbidity and can also inform wider patterns that can be later tested in larger real-world data holdings.

UK Data Service (UKDS)

https://ukdataservice.ac.uk/
https://ukdataservice.ac.uk/find-data/access-conditions/
https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/costing/

Geographical Area Coverage
UK Wide

Description
Data provider, archive, and secure lab setting for economic, population, and social data.

Population Coverage
Dependent on dataset of interest (wide array of records).

Data Types
Datasets deposited by ESRC funded studies, cohort studies, surveys, research outputs, business data, and administrative records.

Access method(s)/permutations

- Data is categorised into Open (downloadable without registration), Safeguarded (downloadable with agreement to an end user licence), and Controlled (only accessible via the Secure Lab remote environment, subject to an application process).
- Data is primarily in statistical formats such as SPSS and Stata, or delimited files (csv, tsv).

For more: https://ukdataservice.ac.uk/find-data/access-conditions/

Cost Implications
Currently free to download data or use the SecureLab subject to eligibility checks.

Discussion
UK Data Service acts as the gateway to many useful datasets which have been derived from studies set up to monitor certain populations, or administrative records with the same aim. The data held by UKDS tends to be less health orientated, other than the clinical collections of the cohort studies themselves, however can introduce useful administrative and non-health information such as social status and other cultural factors into research associated with MLTCs.

UK Longitudinal Linkage Collaboration (UKLLC)

https://ukllc.ac.uk/
UKLLC Data

Geographical Area Coverage
UK Wide

Description
TRE housing and combining 21 longitudinal population studies who have currently contributed COVID-19 survey and clinical data, linked to routine health records.

Population Coverage
Combined ~250,000 participants, plans to increase to up to 2m over the next 5 years.

Data Types
- The studies which contribute to UKLLC can be found here: https://ukllc.ac.uk/partner-studies/
- All studies who have provided demographics will have linked health records from NHS England (COVID-19 sets such as GP, infection and vaccination data, as well as wider HES and mental health data). Applications to Wales (approved) and Scotland for linked records are ongoing.

Access method(s)/permutations
- TRE – 2FA virtual desktop infrastructure hosted by SeRP UK at Swansea University.
- Data is provided via a SQL database.

For more: https://ukllc.ac.uk/access

Cost Implications
Currently free at the point of use for researchers, extra compute power and systems in addition to standard setups may be chargeable.

Discussion
- As with the birth cohorts, longitudinal population studies have high granularity of data collection which can aid and enhance the use of larger scale health records.
- Whilst UKLLC is currently COVID-19 only, this is pivoting to all-cause and increasing linkage to health and administrative records beyond England to all-UK.

CODELISTS

Some examples of codelists which could be used in conjunction with the various data and infrastructure listed above, in order to aid MLTC research and wider use, are as follows:

- https://opencodelists.org
- https://phenotypes.healthdatagateway.org/
- https://clinicalcodes.rss.mhs.man.ac.uk/
- https://www.phpc.cam.ac.uk/pcu/research/research-groups/crmh/cprd_cam/codelists/v11/
- https://zenodo.org/records/12705968

GLOSSARY

#123

2FA – Two Factor Authorisation

A

AOC - Assured organisational connectivity

AWS – Amazon Web Services

B

BCS70 - 1970 British Cohort Study

BHF – British Heart Foundation

C

CPRD – Clinical Practice Research Data

D

DARS - Data Access Request Service

DPUK – Dementias Platform UK

E

EDRIS – Electronic Data Research and Innovation Service

EMIS – Egton Medical Information Systems (former name)

EPCC – Edinburgh Parallel Computing Centre

ESRC – Economic and Social Research Council

H

HBS – Honest Broker Service

HES – Hospital Episode Statistics

HSCNI – Health and Social Care Northern Ireland

M

MCS - Millennium Cohort Study

MLTC – Multiple Long Term Conditions

N

NCDS - National Child Development Study

NHS – National Health Service

NSHD - National Survey of Health and Development

O

ONS - Office for National Statistics

P

PHS – Public Health Scotland

PROMs - Patient/Participant Reported Outcome Measures

Python – Common programming language

Q

QOF – Quality of Outcomes Framework

R

R - Common programming language

RAP – (UK Biobank) Research Analysis Platform

RCGP – Royal College of General Practitioners

RSC – Research and Surveillance Centre

S

SAIL – Secure Anonymised Information Linkage

SDE - Secure Data Environment

SNOMED-CT – Structured clinical vocabulary for use in an electronic health record

SQL – Common programming language

SQUIRE – Standards for QUality Improvement Reporting Excellence

SRS - Secure Research Service

T

TPP – The Phoenix Partnership

TRE - Trusted Research Environment

U

UKLLC - UK Longitudinal Linkage Collaboration

V

VPN – Virtual Private Network

W

WSIC - Northwest London Integrated Care System