Institute of Education

UCL

Novel Health Record Linkages

By Danielle Gomes

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Summary

- This review scopes health-related data sets at the national/country-level, in order to inform future recommendations for additional health record linkages in the CLS cohorts. It includes health data that are not already subject to linkage applications within CLS. It covers England, Scotland and Wales. Given the difficulties in orchestrating linkages from data held by multiple local authorities into national studies, it considers only linkages to national registers.
- It documents a wide range of data sets covering specific conditions, including cancers (screening, diagnosis, treatment), diabetes, mental health, as well as other types of health-related data such as Child Measurement and Community Services. For each data source it provides information on population coverage, data available, years covered and how to access the data.
- The data in England can be accessed via NHS-Digital or Public Health England; in Wales it can be accessed via the Secure Anonymised Information Linkage (SAIL); and in Scotland, it can be accessed via The Electronic Data Research and Innovation Service (eDRIS).

Contents

Novel Health Record Linkages

England

In England, we have scoped the possibility of linking and accessing a wide range of novel health data provided by Public Health England and NHS Digital. PHE is an operationally autonomous executive agency of the Department of Health. Data owned by PHE is made available for research via the Office for Data Release (ODR) which is the body responsible for providing a common governance framework for responding to requests to access PHE data for secondary purposes. NHS-Digital is responsible for collecting and publishing data and information across the health and social care system in England and makes data available via the Data Access Request Service (DARS) online. Some PHE data may be accessible via the NHS-Digital –DARS platform.

We have listed here datasets available from NHS-Digital or PHE that are not already subject to linkage applications within CLS. The latter include Hospital Episode Statistics and mortality data. We note that a summary of all data sets is provided in Table 1.

Data available via NHS-Digital

- **Mental Health Services Data Set (MHSDS)**

The MHSDS is a patient level, output based, secondary uses data set, which delivers robust, comprehensive, nationally consistent and comparable person-based information for people in contact with Mental Health Services. As a secondary uses data set it intends to re-use clinical and operational data for purposes other than direct patient care. The MHSDS is unique in its coverage, because it covers not only services provided in hospitals, but also in outpatient clinics and in the community, where the majority of people in contact with these services are treated.

Population coverage – children, young people and adults who are in contact with Mental Health Services.

Data included - MHSDS brings together key information from Adult and Children's mental health, learning disabilities or autism spectrum disorder, CYP-IAPT and early intervention care pathway that has been captured on clinical systems as part of patient care.

Years covered – 2016/17 onwards

Note that for the years 2006/07- 2015/16, data are available in the Mental Health Minimum Data Set- (MHMDS), subsequently renamed the Mental Health and Learning Disabilities Data Set (MHLDS) following an expansion in scope on the 1st September 2014 to include people in contact with learning disabilities and autism spectrum disorder services. The **MHSDS** dataset replaced the MHLDS and provides the most up to date information available about care given to users of NHS funded secondary mental health, learning disability and autism services for all ages in England.

- **Maternity Services Data Set**

The Maternity Services Data Set (MSDS) is a patient-level 'secondary uses' data set that re-uses clinical and operational data for purposes other than direct patient care, such as commissioning and clinical audit. It captures key information at each stage of the maternity

service care pathway in NHS-funded maternity services, including those provided by GP practices and hospitals.

Population coverage – mothers at each stage of the maternity service care pathway and babies' demographics and birth details.

Data included - The data include mother's demographics, booking appointment details (includes things like expected delivery date, previous pregnancies and outcomes, substance use status, cigarettes per day, alcohol units, mental health prediction and detection indicator, weight, height etc), complicating medical diagnosis at booking, previous complicating obstetric diagnoses at booking, family history diagnosis at booking, ultrasound scan (offered at booking and performed between 10 weeks and 13 weeks but can be undertaken at any time during pregnancy), mother's screening tests (e.g. ABO blood group and rhesus, rubella susceptibility, hepatitis B screening, Downs Syndrome, fetal anomaly, STIs, HIV etc), maternity care plan, medical and obstetric diagnoses admissions and re-admissions, , medical induction method, labour and delivery, pain relief and anaesthesia type in labour and delivery, maternal critical incident, genital tract trauma, fetus outcome, baby's demographics and birth details, baby complications at birth, neonatal resuscitation, neonatal critical care, diagnoses and screening tests (e.g. physical screening examination, hearing), mother's postpartum discharge from maternity services.

Years covered- 2015/16 onwards

- **National Diabetes Audit dataset**

The main National Diabetes Audit, known as the Core Audit, collects information from general practices and specialist diabetes out-patient services to look at whether people with diabetes are receiving their annual care checks, are achieving their treatment targets and looks at their health outcomes along with whether they have been offered and attended structured education.

Population coverage – people diagnosed with diabetes in GP practices and out-patient services

Data included- The data includes information on diabetes diagnosis and treatment, structured education offered, blood pressure, BMI and cholesterol readings.

Years covered- 2003/04 onwards

- **National Child Measurement Program**

The National Child Measurement Programme (NCMP) measures the height and weight of children to assess overweight and obesity levels in children within primary schools.

Population coverage – children in Reception class (aged 4 to 5) and year 6 (aged 10 to 11).

Data included- The data includes information on pupil's height, weight and body mass index scores.

Years covered – 2006/07-2017/18 (latest available)

- **Community Services Dataset**

The CSDS is a patient-level data set providing information relating to publicly funded community services for children, young people and adults. These services can include health centres, schools, mental health trusts, and health visiting services. The CSDS is an update to the Children and Young People's Health Services (CYPHS) data set standard. The CYPHS data set collected data for all patients aged 0 up until their nineteenth birthday.

Population coverage – people of all ages in receipt of publicly funded Community Services.

Data included- Personal and demographic information, social and personal circumstances, breastfeeding and nutrition, diagnoses including long-term conditions and disabilities, care events plus screening activities and scored assessments.

Years covered- CYPHS from 2015-2017; from 2017 onwards CSDS replaced CYPHS.

Data available via Public Health England

- **Cancer registration**

National Cancer Registration and Analysis Service (NCRAS) is run by Public Health England. It is responsible for cancer registration in England to support cancer epidemiology, public health, service monitoring, and research. Cancer registration is the systematic collection of data about cancer and tumour diseases. It brings together data from more than 500 local and regional data sets to build a picture of an individual's treatment from diagnosis.

Population coverage – cancer registrations in England.

Data included -Data on the patient, their diagnosis, tumour characteristics and details of the care and treatment received.

Years covered- 1985 onwards

- **Systemic Anti-Cancer Therapy Data set (SACT)**

The Systemic Anti-Cancer Therapy data set collects clinical information on patients receiving cancer chemotherapy in or funded by the NHS in England. SACT covers chemotherapy treatment for all solid tumour and haematological malignancies and those in clinical trials.

Population coverage – all cancer patients, both adult and paediatric, in acute inpatient, day case, outpatient settings and delivery in the community.

Data included- Information on all drug treatments with an anti-cancer effect, in all treatment settings, including traditional cytotoxic chemotherapy and all newer agents.

Years covered- 2012 onwards

- **Radiotherapy Data Set (RTDS)**

The Radiotherapy Data Set (RTDS) collates data from all NHS Acute Trust providers of radiotherapy services in England to provide consistent and comparable data on radiotherapy services in England.

Population coverage – every patient receiving teletherapy or brachytherapy.

Data included- Information on all radiotherapy of the following types: • Teletherapy • Brachytherapy given using automated remote after loading machines. All other brachytherapy given for the treatment of malignant disease delivered in England to patients in NHS facilities, or in private facilities where delivery is funded by the NHS.

Years covered- 2009 onwards

- **National Prostate Cancer Audit**

The aim of the National Prostate Cancer Audit (NPCA) is to assess the process and outcomes of prostate cancer care provided by the NHS in England and Wales.

Population coverage – all men with newly diagnosed prostate cancer.

Data included- Information on the characteristics of all men with newly diagnosed prostate cancer, how their cancer was detected, and the referral pathway • The crucial steps in the diagnostic and staging process • The planning of initial treatment • Initial treatments • Initial health outcomes. In addition, the NPCA collects patient-reported experience and outcome measures in the men included in the audit who underwent a radical treatment one year after diagnosis.

Years covered- 2014 onwards

- **GUMCAD** (Genitourinary Medicine Clinic Activity Dataset)

The GUMCAD is the mandatory surveillance system for sexually transmitted infections (STIs) in England the new name for the Genitourinary Medicine Clinic Activity Dataset (GUMCAD).It is a mandatory reporting system providing data on sexual health services and STI diagnoses from all commissioned Level 3 and Level 2 sexual health services in England. Note, it is not currently available for research.

Wales

In Wales, health data can be accessed via the Secure Anonymised Information Linkage (SAIL). SAIL databank is a Wales-wide research resource that receives core funding from the Welsh Government and Care Research Wales, and provides data linkage services and secure remote access to linkable anonymised datasets.

Data available via SAIL

We have listed here datasets available from SAIL that are not already subject to linkage applications within CLS. The latter include child health (including immunisations), Emergency Department, Patient Episode Database, Primary Care General Practice, Welsh Data Service (WDS),

- **Welsh Cancer Intelligence and Surveillance Unit (WCISU) –**

The Welsh Cancer Intelligence & Surveillance Unit (WCISU) is the National Cancer Registry for Wales and its primary role is to record, store and report on all incidence of cancer in Wales.

Population coverage – all incidences of cancer for the resident population of Wales wherever they are treated.

Data included- Occurrences of cancer in Welsh residents via direct or indirect submissions from Welsh Hospitals.

Years covered- 1972 onwards

- **National Child Community Child Health Database (NCCHDS) –**

Each health board in Wales has a Child Health System database which is managed locally and held by NHS trusts and used by them to administer child immunisation and health surveillance programmes. The records are collated from each of the local databases each quarter to create the NCCHD.

Population coverage – all children born, resident or treated in Wales.

Data included- Details relating to maternal and child health related indicators such as births, immunisation screening and safeguarding of children.

Years covered- 1987 onwards

- **Congenital Anomaly Register and information Service (CARIS)** –

CARIS is a programme which sits within the Health Intelligence Division of Public Health Wales.

Population coverage – any foetus or baby who has or is suspected of having a congenital anomaly and whose mother is normally resident in Wales at time of birth.

Data included- Data on congenital anomalies, including miscarriages. It includes babies in whom anomalies are diagnosed at any time from conception to the end of the first year of life. Categories of anomalies include blood disorders, central nervous system, chromosomal, congenital cardiac, disorders of sex development, ear anomalies & hearing loss, cleft lip & palate, infections, metabolic & endocrine, muscoskeletal system, and respiratory system.

Years covered- 1998- 2017

- **Bowel Screening Wales**

Administrative and clinical information for bowel screening.

Population coverage – offered to men and women resident in Wales aged between 60 and 74 years old

Data included- Invitation and screening test records for eligible individuals.

Years covered- 2008 onwards

- **Breast Test Wales**

Administrative and clinical information for breast screening. Older women can self-refer.

Population coverage – offered to women resident in Wales aged 50 to 70 years.

Data included- Assessment and screening test records for individuals who are eligible for breast screening: routine invitations, self-referrals and family history screening women.

Years covered- 1989- onwards

- **Cervical Screening Wales**

Administrative and clinical information for cervical screening. This dataset contains all individuals who are eligible and invited for cervical screening.

Population coverage – offered to women resident in Wales aged between 20 and 64 years old.

Data included- Invitation, screening tests and assessment records.

Years covered- 1990 onwards for invitation and screening test records. Assessment data available from 2011.

Scotland

In Scotland, health data can be accessed via The Electronic Data Research and Innovation Service (eDRIS). eDRIS is part of Information Services Division (ISD) and provides a single point of contact to assist in the completion of applications to the Public Benefit and Privacy Panel and assist researchers in study design, approvals and data access in a secure environment.

We have listed here datasets available from eDRIS that are not already subject to linkage applications within CLS. The latter include outpatient attendance; general / acute inpatient and day case; maternity inpatients/day cases; Scottish birth record/neonatal records; Scottish Immunisation and Recall System; Prescribing Information System; and Child Health Review

Data available via edris

- **Scottish Bowel Screening Programme (SBoSP) dataset-**

The SBoSP commenced in June 2007. The dataset is derived from two sources, BoSS information and Health Board information.

Population coverage – all eligible participants in the SBoSP i.e. 50-74 year olds invited to take part.

Data included- BoSS is the Bowel Screening IT System. This system is hosted by Atos and contains non-clinical information relating to the invites and test results of kits sent to the Scottish Bowel Screening Centre in Dundee. The Health Board information files contain clinical data.

Years covered- The SBoSP commenced a phased roll out in 2007 with all health boards participating by December 2009.

- **Scottish Breast Screening System (SBSS) –**

Data is collected by the 6 screening centres in Scotland using the breast screening system. The system is maintained by an external supplier on behalf of National Services Division (NSD) who manage the programme. The UK National Health Service Breast Screening Programme (NHSBSP) was introduced in 1988.

Population coverage – women in Scotland aged 50-64 years until 2003-04 since when the age range was extended to include women up to the age of 70 years.

Data included- The Scottish Breast Screening Programme (SBSP) Information System maintains an extensive data set holding information relating to each step as a woman moves through her screening episode. This includes mammographic images performed & exposure details, comprehensive radiological, diagnostic and treatment information including areas such as: pathology characteristics of screen detected lesions such as morphological type, grade, size, nodal involvement & disease extent. Cytology specimen type and localisation technique, type of surgical procedure, localisation techniques, lymph node procedures & details of other forms of treatment.

Years covered- 1990 onwards

- **Smoking Cessation dataset (SCD)**

SCD was set up to capture agreed national minimum dataset (mds) for NHS cessation services in Scotland.

Population coverage - all individuals who attempt to quit via NHS smoking cessation services.

Data included- The annual national cessation monitoring analyses provide evidence of the reach and quit success of NHS funded smoking cessation services in Scotland. This includes evidence on the reach and targeting of specific groups, such as: those living in the most deprived communities, pregnant women, younger age groups and those living in urban/rural areas. They also allow monitoring of different interventions and pharmacotherapies used and their effectiveness in helping clients to quit. The analyses support service planning in NHS boards, providing evidence on service 'reach' across different client groups and geographical variations within boards, via Local Authority level analyses, as well as evidence on what works best in helping clients to quit smoking.

Years covered- 2007 onwards

Charges

NHS-Digital - New application £1,030.

Permission to sublicense £10,000, valid for 12 months.

Annual review fee £500 + fee for linkage depending on number of cases to be linked.

Public Health England – All requests to access PHE data charged on a project by project basis to reflect the amount of work required. For the 2018/19 financial year, the unit charge was set at £378* per hour (excl. VAT).

SAIL- Charged on a project by project basis to reflect the amount of work required. Example provided by SAIL: an academic project, looking for less than 10 days of total support may be broken down as follows - 4 days of data linkage and 5 days of data preparation within SAIL approx. £3,000-3,500. An academic project, looking for 30 days or less of total support may be broken down as follows: 8 days of data linkage and 22 days of data preparation within SAIL approx. £9,500-10,000.

Edris – Small study £5,920; medium study £14,800; large study £22,200

Table 1: Summary of Health Data

Records	Data Controller	Access mechanism	Population coverage	Dates the records cover
Mental Health Services Data Set (MHSDS)	NHS-Digital	DARS- Online	England	2016-
Mental Health Minimum Data Set	NHS-Digital	DARS- Online	England	2006/07-2014/15
GUMCAD (Genitourinary Medicine Clinic Activity Dataset)	NHS-Digital	Not available through DARS request information	England	Not available
National Diabetes Audit	NHS-Digital	DARS- Online	England and Wales	2003/04-
Community Services Dataset	NHS Digital	DARS-online	England	2015-
National Child Measurement Program	NHS-Digital	DARS- Online	England	2006/07-2018/19
Maternity Services Dataset	NHS-Digital	DARS- Online	England	2015/16-2018/19
Diagnostic Imaging Dataset	NHS-Digital	DARS Online	England	Historic and latest
Hearing and vision – HES	NHS Digital	DARS Online	England	Available in the HES dataset
Cancer registration (patient,tumour treatment table)	PHE	Application to Office for data release ODR@phe.gov.uk	England	1985-2017

Route to diagnosis	PHE	Application to Office for data release ODR@phe.gov.uk	England	2006-2016
Systemic Anti-Cancer Therapy Data set (SACT)	PHE	Application to Office for Data Release ODR@phe.gov.uk	England	April 2012 onwards (six months behind current date)
Radiotherapy Data Set (RTDS)	PHE	Application to Office for Data Release ODR@phe.gov.uk	England	2009-2017
Welsh cancer Intelligence and Surveillance Unit (WCISU)	SAIL	As a restricted dataset, it requires the agreement of the data provider in addition to SAIL standard Information Governance procedure.	Wales	1972-
National Child Community Child Health Database (NCCHDS)	SAIL	Application to SAIL https://saildatabank.com/application-process/	Wales	1987-
Congenital Anomaly register and information Service (CARIS)	SAIL	As a restricted dataset, it requires the agreement of the data provider in addition to SAIL standard Information Governance procedure.	Wales	1998-2011
Bowel Screening Wales	SAIL	As a restricted dataset, it requires the agreement of the data provider in addition to SAIL standard Information Governance procedure	Wales	2008 -
Cervical screening	SAIL	As a restricted dataset, it requires the agreement of the data provider in addition to SAIL standard Information Governance procedure	Wales	January 1990- onwards for invitation and screening tests Assessment data available from April 2011
Breast test	SAIL	As a restricted dataset, it requires the agreement of the data provider in addition to SAIL	Wales	February 1989 onwards for invitation and screening tests

		standard Information Governance procedure		
Scottish Bowel Screening Programme (SBoSP) dataset	ISD Scotland - EDRIS https://www.isdscotland.org/Products-and-Services/EDRIS/	Enquiry contact form to nss.edris@nhs.net For main dataset information contact-NSS.isdcancerstats@nhs.net	Scotland	2007 -
Scottish Breast screening system (SBSS)	ISD Scotland - EDRIS https://www.isdscotland.org/Products-and-Services/EDRIS/	Enquiry contact form to nss.edris@nhs.net For main dataset information contact-NSS.isdcancerstats@nhs.net	Scotland	1990 -
Smoking Cessation dataset (SCD)	ISD Scotland - EDRIS https://www.isdscotland.org/Products-and-Services/EDRIS/	Enquiry contact form to nss.edris@nhs.net For main dataset information contact-NSS.ISDSmokingCessation@nhs.net	Scotland	2007 -

NHS-Digital accessed 30.05.2019

https://digital.nhs.uk/data-and-information/data-insights-and-statistics

https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737140/PHE_data_catalogue.pdf

https://saildatabank.com/

http://www.publichealthwalesobservatory.wales.nhs.uk/ncchd

http://www.publichealthwalesobservatory.wales.nhs.uk/ncchd

http://www.caris.wales.nhs.uk/home

https://www.isdscotland.org/Products-and-Services/EDRIS/

https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data#odr-service-charges

https://digital.nhs.uk/data-and-information

Annex: Consultation with Understanding Society

As part of this review, CLS consulted with Understanding Society (Jon Burton and team) regarding their plans for data linkages. A summary is provided below.

DVLA

USoc is linking to information held about the vehicle, rather than the driver. So they are getting information on the vehicle (e.g., engine size) and they are doing this without explicit consent. They have collected information from participants about the registration number of the vehicle. However, rather than linking this using the DVLA, they are getting the information that is publicly-held online. So they are not linking individual-level data. For this reason, they are not planning to link to traffic offences, because this would require individual-level linkage (and consent). They have asked for the car registration details on the main Understanding Society survey and they will be making the administrative data available via the UK Data Service (not the registration itself – and the data might be Special License or above).

FCA

Usoc has asked for consent to link to the data from the Financial Conduct Authority in the Innovation Panel, and these data have been matched using name, address, date of birth. They are also asking the consent question on the main sample, and so the plans are to match these in the future. The data held by the FCA are provided by the different Credit Rating Agencies.

Employer data

As part of their Innovation Panel, Usoc asked for information (name, address) of the employer and have tried to use this to link to publicly-held information (held by councils) about the company. They did have a challenge in that the format, structure, and content of the data reported by participants may be different to how the administrative data are held. So for example, the address formatting, use of postcodes etc may be different in the two datasets and this makes matching more difficult. There is also the issue about the status of the address collected – e.g., is it a local branch of a chain or the registered head office? For example, would someone say they work at Starbucks in Colchester, but the administrative data covers the Starbucks head office? They have redesigned the way they ask it at the mainstage in an attempt to structure the responses better (a lot of the interviews are done on-line, and so they can't rely on an interviewer being able to format the address).

Twitter

Usoc have done some work using the Innovation Panel to collect consent to link to Twitter data, and for those who consent, they collected their Twitter handle. There is some basic information in this working paper that covers IP10: https://www.understandingsociety.ac.uk/research/publications/525086 The challenge for this data is how to link it in a way which maintains confidentiality, and how to then release the data for third-party use. For the IP10 experiment, the Twitter handles and a unique identifier

were passed to a member of the research team who worked in a different institution. The plan is that they harvest the Twitter information, and then code the Tweets on a number of conditions. Then the coded data and the identifier will be passed back to ISER and matched onto the survey data. This is something that is feasible on a small-scale, but a challenge would be in scaling it up to the full sample, if this were to be done. There is also the issue that the coding of the tweets are very project-specific, so this limits secondary analysis because every new project would want to go back to the original text of the tweets and code them differently.

Instagram

This is something that has been suggested, but USoc has no current plans to do this.

Home energy rates

Usoc asked for consent to link administrative data (The National Energy Efficiency Data) held by the Department of Energy and Climate Change (DECC). This would be linked by the address of the property. This linkage is currently going ahead and they are negotiating the necessary agreements with BEIS.

Voter histories

In the Innovation Panel there was an experiment about asking for consent to the electoral register. The working paper referenced above also includes a description of this. They are not intending to actually carry out linkage to this data, but the experiment was to try and improve the consent rate to try and improve the way the British Election Study asked.

Employer pensions

Usoc have asked for consent to link to the National Employment Savings Trust (NEST), this is the government's workplace pension scheme. This question is in the current wave of fieldwork and so they have not started the process of linkage. It is expected that they will use name, date of birth, and address to link.

Competitions and Market Authority (CMA)

This linkage will be done using name and address. CMA and regulators need to understand consumer outcomes in markets to fulfil their statutory functions and inform interventions. This is even more pertinent in the context of vulnerability. To effectively protect consumers who are most in need, they need a robust understanding of vulnerable consumer outcomes, how they change over time, and the extent to which different vulnerabilities interact/overlap and effect outcomes. This requires comprehensive, high quality, large scale, granular datasets but currently most evidence is small scale, one-off and high level; it is not possible to assess consumer outcomes (such as the 'loyalty penalty' paid by different consumer groups). The CMA is working closely with regulators to explore the feasibility of regulators collecting transaction and other data from suppliers, and linking this to robust ongoing survey data on consumer characteristics (USOC data), including those associated with vulnerability.

There are other data sources which they are considering. These are mainly publicly available data which do not refer to an individual, but are either household or address based, such as Council Tax bands, Zoopla, Transport for London.