SAIL Databank

First published: 01/02/2024

Last updated: 17/10/2024

Data source | Human | Primary care medical records | Birth registry

Death registry | Other

Administrative details

Administrative details

Data source ID

1111162

Data source acronym

SAIL

Data holder

Swansea University Medical School

Data source type

Primary care medical records

Birth registry

Death registry

Other

Data source type, other

NHS Wales hospital admissions (Inpatients and daycases)

Child Health System

Main financial support

National, regional, or municipal public funding

Funding from industry or contract research

Care setting

Primary care – GP, community pharmacist level

Hospital inpatient care

Primary care – specialist level (e.g. paediatricians)

Secondary care – specialist level (ambulatory)

Data source qualification

If the data source has successfully undergone a formal qualification process (e.g., from the EMA, ISO or other certifications), this should be described.

Yes

Description of the qualification

ISO27001

Data source website

https://saildatabank.com/

Contact details

Dan Thayer d.s.thayer@swansea.ac.uk

(Main)

d.s.thayer@swansea.ac.uk

Data source regions and languages

Data source countries

United Kingdom

Data source languages

English

Data source regions

Wales [Cymru GB-CYM]

Data source establishment

Data source established

15/06/2010

Data source time span

First collection: 15/06/2010

The date when data started to be collected or extracted.

Publications

Data source publications

[SAIL website](#)

Studies

List of studies that have been conducted using the data source

[Similarity-based approaches to identifying risk of future asthma attack using UK primary care data](#)

[Exploring new treatments and outcomes in type 2 diabetes](#)

[Methods for controlling by indication for prescriptions: application to medications for neuropathic pain](#)

[Studying drug exposure when disease is measured through accurate identification of an incident case: application to breast cancer in pregnancy (ConcePTION breast cancer demo)](#)

[Exposure to SSRI/SNRI and depression in pregnancy and long-term childhood outcomes: the effect of modifying factors](#)

[Data characterization of population-based data sources: ConcePTION pipeline](#)

Data elements collected

The data source contains the following information

Disease information

Does the data source collect information with a focus on a specific disease? This might be a patient registry or other similar initiatives.

Yes

Disease details (other)

Information on disease diagnosis, symptoms, presentation, treatment, etc. including disease specific registries for cancer, congenital anomalies, type 1 diabetes, and others; Covid-19 Test results are captured in one of the datasets

Rare diseases

Are rare diseases captured? In the European Union a rare disease is one that affects no more than 5 people in 10,000.

Yes

Pregnancy and/or neonates

Does the data source collect information on pregnant women and/or neonatal subpopulation (under 28 days of age)?

Yes

Hospital admission and/or discharge

Yes

ICU admission

Is information on intensive care unit admission available?

Yes

Cause of death

Captured

Cause of death vocabulary

ICD-10

Read

Prescriptions of medicines

Captured

Prescriptions vocabulary

ATC

other

Prescriptions vocabulary, other

Read

Dispensing of medicines

Not Captured

Advanced therapy medicinal products (ATMP)

Is information on advanced therapy medicinal products included? A medicinal product for human use that is either a gene therapy medicinal product, a somatic cell therapy product or a tissue engineered products as defined in Regulation (EC) No 1394/2007 [Reg (EC) No 1394/2007 Art 1(1)].

No

Contraception

Is information on the use of any type of contraception (oral, injectable, devices etc.) available?

Yes

Indication for use

Does the data source capture information on the therapeutic indication for the use of medicinal products?

Not Captured

Medical devices

Is information on medicinal devices (e.g., pens, syringes, inhalers) available?

Yes

Administration of vaccines

Yes

Procedures

Does the data source capture information on procedures (e.g., diagnostic tests, therapeutic, surgical interventions)?

Captured

Procedures vocabulary

OPCS

Healthcare provider

Is information on the person providing healthcare (e.g., physician, pharmacist, specialist) available? The healthcare provider refers to individual health professionals or a health facility organisation licensed to provide health care diagnosis and treatment services including medication, surgery and medical devices.

Yes

Clinical measurements

Is information on clinical measurements (e.g., BMI, blood pressure, height) available?

Yes

Genetic data

Are data related to genotyping, genome sequencing available?

Not Captured

Biomarker data

Does the data source capture biomarker information? The term "biomarker" refers to a broad subcategory of medical signs (objective indications of medical state observed from outside the patient), which can be measured accurately and reproducibly. For example, haematological assays, infectious disease markers or metabolomic biomarkers.

Not Captured

Patient-reported outcomes

Is information on patient-reported outcomes (e.g., quality of life) available?

Yes

Patient-generated data

Is patient-generated information (e.g., from wearable devices) available?

No

Units of healthcare utilisation

Are units of healthcare utilisation (e.g., number of visits to GP per year, number of hospital days) available or can they be derived? Units of healthcare utilisation refer to the quantification of the use of services for the purpose of preventing or curing health problems.

Yes

Unique identifier for persons

Are patients uniquely identified in the data source?

Yes

Diagnostic codes

Captured

Diagnosis / medical event vocabulary

ICD-10

Read

Medicinal product information

Captured

Medicinal product information collected

Brand name

Formulation

Strength

Route of administration

Active ingredient(s)

Medicinal product vocabulary

Other

If 'other,' what vocabulary is used?

Read

Quality of life measurements

Not Captured

Lifestyle factors

Captured

Lifestyle factors

Tobacco use

Alcohol use

Other

Sociodemographic information

Captured

Sociodemographic information collected

Age

Ethnicity

Gender

Socioeconomic status

Deprivation index

Health area

Quantitative descriptors

Population Qualitative Data

Population age groups

Paediatric Population (< 18 years)

Preterm newborn infants (0 – 27 days)

Term newborn infants (0 – 27 days)

Infants and toddlers (28 days – 23 months)

Children (2 to < 12 years)

Adolescents (12 to < 18 years)

Adults (18 to < 46 years)

Adults (46 to < 65 years)

Elderly (≥ 65 years)

Adults (65 to < 75 years)

Adults (75 to < 85 years)

Adults (85 years and over)

Estimated percentage of the population covered by the data source in the catchment area

100%

Description of the population covered by the data source in the catchment area whose data are not collected (e.g., people who are registered only for private care)

Anyone wishing to opt out of anonymised data related to them being sent to SAIL or used for other secondary purposes, should make an enquiry to the relevant data provider(s) about what options they may provide for allowing individuals to opt out. For primary care records, individuals can opt out by making a request to their GP. Opt outs are rare. Private healthcare is not included, but is a very small percentage of care in Wales, and it would be rare to only receive private care and not be registered with the NHS.

Family linkage

Family linkage available in the data source permanently or can be created on an ad hoc basis

Permanently

Family linkage available between the following persons

Household

Mother-child

Population

Population size

5564900

Active population size

3059900

Population by age group

Age group	Population size	Active population size
Paediatric Population (< 18 years)	1065100	585700
Preterm newborn infants (0 – 27 days)	300	200
Term newborn infants (0 – 27 days)	3600	2000
Infants and toddlers (28 days – 23 months)	95100	52300

Age group	Population size	Active population size
Children (2 to < 12 years)	516800	284100
Adolescents (12 to < 18 years)	449300	247100
Adults (18 to < 46 years)	1894200	1041500
Adults (46 to < 65 years)	1418200	779800
Elderly (≥ 65 years)	1187400	652900
Adults (65 to < 75 years)	615500	338400
Adults (75 to < 85 years)	423000	232600
Adults (85 years and over)	148900	81900

Median observation time

Median time (years) between first and last available records for unique individuals captured in the data source

21.00

Median time (years) between first and last available records for unique active individuals (alive and currently registered) capt

25.00

Data flows and management

Access and validation

Governance details

Documents or webpages that describe the overall governance of the data source and processes and procedures for data capture and management, data quality check and validation results (governing data access or utilisation for research purposes).

https://saildatabank.com/governance/

Biospecimen access

Are biospecimens available in the data source (e.g., tissue samples)?

No

Access to subject details

Can individual patients/practitioners/practices included in the data source be contacted?

No

Description of data collection

Datasets are provided by various organisations and are anonomysed before loading into SAIL Databank.

The data provider splits each dataset into two components:

1) Demographic Component – this holds the identifying information to be anonymised.

2) Content Component – this holds other details, such as diagnosis, medication, etc.

The demographic component (1) is sent to Digital Health and Care Wales, where it is validated and each record is anonymised and assigned a unique, non-identifiable code. This code, and minimal information on gender, area of residence and week of birth is then sent to SAIL Databank.

The content component (2) is sent directly to SAIL Databank where the two components of the dataset (1 and 2) are linked together. The complete de-identified dataset can now be accessed for research, subject to approvals.

Event triggering registration

Event triggering registration of a person in the data source

Birth

Practice registration

Residency obtained

Event triggering de-registration of a person in the data source

Death

Practice deregistration

Emigration

Event triggering creation of a record in the data source

Receiving healthcare; other events like prescriptions, letters sent to patient, etc; creation of disease registries.

Data source linkage

Linkage

Is the data source described created by the linkage of other data sources (prelinked data source) and/or can the data source be linked to other data source on an ad-hoc basis?

Yes

Linkage description, pre-linked

Linkage based on demographic details and NHS number.

Linkage description, possible linkage

Combination of exact (NHS number) and probabilistic linkage

Linked data sources

Pre linked

Is the data source described created by the linkage of other data sources?

Yes

Data source, other

Numerous datasets from multiple sources linked.

Linkage strategy

Probabilistic

Linkage variable

NHS number, name, date of birth, gender, postcode

Linkage completeness

Varies by dataset; >99% for key datasets

Pre linked

Is the data source described created by the linkage of other data sources?

No

Data source, other

SAIL's system enables linkage of additional datasets as acquired by SAIL or brought in for specific projects. Any Welsh data with required demographic fields can be processed and linked to existing datasets in the system.

Linkage variable

NHS number, name, date of birth, gender, postcode

Linkage completeness

Varies by dataset

Data management specifications that apply for the data source

Data source refresh

Monthly

Informed consent for use of data for research

Not Required

Possibility of data validation

Can validity of the data in the data source be verified (e.g., access to original medical charts)?

Yes

Data source preservation

Are records preserved in the data source indefinitely?

Yes

Approval for publication

Is an approval needed for publishing the results of a study using the data source?

Yes

Data source last refresh

01/04/2023

Common Data Model (CDM) mapping

CDM mapping

Has the data source been converted (ETL-ed) to a common data model?

Yes

CDM Mappings

CDM name

ConcepTION CDM

CDM website

https://www.imi-conception.eu/

CDM release frequency

6 months

Data source ETL status

Completed

Data source ETL specifications (link)

https://www.imi-conception.eu/

CDM name

OMOP

CDM website

https://www.ohdsi.org/Data-standardization/

Data source ETL CDM version

5.3.1

Data source ETL frequency

12,00 months

Data source ETL status

Completed