Millennium Cohort Study

Linked health administrative datasets – Welsh Health Data (SAIL)

User Guide (Version 1)

September 2024

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Sarah Kerry-Barnard, Karen Dennison, Emla Fitzsimons

How to cite this guide

Kerry-Barnard, S., Dennison K., Fitzsimons, E (2024) *Millennium Cohort Study: A guide to the linked health administrative datasets – Welsh Medical Records. User Guide (Version 1).* London: UCL Centre for Longitudinal Studies.

This guide was published in October 2024 by the UCL Centre for Longitudinal Studies.

Acknowledgements

You should also cite the data and acknowledge CLS following the guidance from https://cls.ucl.ac.uk/data-access-training/citing-our-data/

CLS would like to thank SAIL Databank for the provision of the linked data.

Centre for Longitudinal Studies

UCL Centre for Longitudinal Studies (CLS)
UCL Social Research Institute
University College London
20 Bedford Way, London WC1H 0AL
www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It is home to a unique series of UK national cohort studies.
For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies:
email: clsdata@ucl.ac.uk

Contents

About the Millennium Cohort Study

The Millennium Cohort Study (MCS) is a longitudinal birth cohort study, following a nationally representative sample of approximately 19,000 people born in the UK at the turn of the century.

The study has captured rich information about the different aspects of cohort members' lives, from birth to childhood and adolescence, and is continuing to keep up with them now they are adults.

As a multidisciplinary study, MCS is used by researchers working in a wide range of fields. Findings from MCS have influenced policy at the highest level, and today the study remains a vital source of evidence on the major issues affecting young people's lives.

Further details of the data available from the main surveys can be found on the CLS website www.cls.ucl.ac.uk/cls-studies/millennium-cohort-study/ and, in particular, the CLS discoverability webpage at www.cls.ucl.ac.uk/data-access-training/exploring-our-data/

1. Introduction

This guide describes the data linkage of health administrative records from the Welsh Health Data to survey data for cohort members in the Millennium Cohort Study (MCS). The Welsh Health Data were provided by SAIL Databank, which contains information about all hospital admissions in Wales.

The main aim of this data linkage exercise is to enhance the research potential of the study, by combining administrative health records with the rich information collected in the surveys. The MCS linked Welsh Health data include health records for those cohort members and parents/other adult household respondents who provided consent to health data linkage in the Age 7 sweep.

2. Consent to health data linkage

At the fourth survey, when the cohort were around 7 years old, adults with parental responsibility were asked to give consent (Appendix 1) to link information collected within MCS to their child's routine health records up until their fourteenth birthday. Parents and other adult household respondents were also asked to consent to their own health records being linked. Further information on the consent procedure and validation of the consents received has been previously reported (Shepherd, 2013) [1].

Why consent age 7: principle of Gillick competence

The linkage has been conducted on the basis of consent from the parent, when the cohort member was 7 years of age. In keeping with the principle of Gillick competence, no linked health records for cohort members after their 14th birthday have been included in the data. In one case, a cohort member was admitted to hospital before their 14th birthday but discharged afterwards; this case has been removed from the records, but there is a note with the data with the cohort member's MCSID and the admission date.

[1] https://cls.ucl.ac.uk/wp-content/uploads/2017/07/Consent-to-linkage-to-child-health-data-in-MCS-CLS-Feb-2013.pdf Accessed 10/09/2024

Any administrative records that contain details relating to GP registrations starting after the cohort member's 14th birthday have been removed. In addition, the end date for those starting before the cohort member's 14th birthday has been adjusted to the 1st day of the month they turned 14. This is flagged in the datasets with the variable 'd_flag'.

Cohort members were themselves asked for consent to health linkages during the 2018 sweep (age 17) MCS7, however MCS7 was collected after the SAIL linkage took place. For more information on the fieldwork and consent collection, the Millennium Cohort Study: Age 17 Sweep (MCS7) User Guide and Millennium Cohort Study Seventh Sweep (MCS7) Technical report can be found under 'Documentation' at https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/mcs-age-17-sweep/

3. Health data linkage

3.1 Welsh Health Data datasets from SAIL

There are 6 groups of linked heath datasets provided to CLS by SAIL Databank:

- Annual District Death Extract (ADDE, 2008-2020)
- Emergency Health Service (EDDS, 2009-2020)
- Outpatient Database for Wales (OPDW, 2004-2020)
- Patient Episode Dataset for Wales (PEDW, 1993-2020)
- Welsh Demographic Service Dataset (WDSD, 1954-2020)
- Welsh Longitudinal General Practice Dataset (WLGP, 1966-2021)

The years of data available in the dataset are the earliest years for which data was available from SAIL Databank.

3.2 Matching strategy

Welsh health records are held by the SAIL Databank. In 2019 CLS provided NHS Wales Informatics Service (NWIS - now called Digital Health Care Wales – SAIL Databank's Trusted Third Party for linkages) with the direct identifiers of those MCS cohort members whose parents had consented to the Cohort Members' health records up to age 14 and for whom consent had not been subsequently withdrawn.

CLS also provided direct identifiers of parents who had consented to their own health records being linked. NWIS matched those records to those held in the Welsh Demographic Service, anonymised them and assigned each one a unique, non-identifiable code. NWIS then sent this code, and minimal information on gender, area of residence and week of birth to SAIL Databank so that the data could be matched to the health records held in the SAIL Databank. Researchers can apply to access the health data linked to MCS survey responses through the SAIL Databank. In 2020 CLS also received approval from the SAIL Databank to extract the linked health data for data sharing via the UK Data Service. CLS received the extracted files in 2021.

3.3 Matching rates

During the fourth MCS sweep (age 7, 2008), the parents of 13,048 out of 14,043 cohort members consented to health linkages, a 92.9% consent rate. Accounting for withdrawals, 12,562 had consented at the time of deposit. 1,904 cohort members who had consent, also had an interview address in Wales in MCS4, MCS5, MCS6 or for current/last known address (as in 2019).

In 2019, CLS sent NWIS a matching file containing the information from all age 7 (MCS4) consenting cohort members and parents/other adult household members regardless of whether their MCS interview address had ever been in Wales to match as many records as possible. This included 13,043 cohort members and 19,866 parents/guardians who still had valid consents. The file included surname, forenames, addresses/postcodes (for MCS4, MCS5, MCS6 and current / last known address in 2019), date of birth and gender. A total of 2108 MCS cohort members were successfully matched to their Welsh health records. This figure is higher than the number of cohort members with interview addresses in Wales – these additional

cases are likely to be people who weren't interviewed in Wales but who moved to/from Wales between interviews. Table 1 below shows the number of successful matches to education records following data linkage.

Consent and overall linkage

The consent rate for cohort members was 92.9%, and for participants with interview address in Wales MC4-7 it was 94%. The total number with matched health data (prior to withdrawals) 2108 (minus withdrawals this is 2034), of these 1903 were interviewed in Wales MCS4-7 (minus withdrawals 1845), with a linkage rate of 99.9%. See table 1 for further details.

Table 1: Consent and linkage table – cohort members

	Whole cohort	Resident in Wales Sweeps 4-7
Took part in MCS4 (2008)	14,043	2,021
Sent for matching	13,043	1,904
Matched data returned	2,108	1,903
Matched data included in deposit (minus withdrawals)	2,034	1,845

The consent rate for parents and other adult respondents was 85.5%, this was 81.4% for participants with interview address in Wales MCS4-7. The total number with matched health data was 3070, minus withdrawals: 3065. For those with an interview in Wales MCS4-7, 2779 had matched data returned, corresponding to a linkage rate of 97.8 %. See table 2 for further information.

Table 2: Consent and linkage table – parents & other adult household respondents

	Whole cohort	Resident in Wales Sweeps 4-7

	Parents	Other adult respondents	All adults	Parents	Other adult respondents	All adults
Took part in MCS4 (2008)	13,797	9,429	23, 226	1,998	1,494	3,492
Sent for matching	11,698	8168	19,866	1,697	1,147	2,844
Matched data returned	1,807	1,263	3,070	1,647	1,132	2,759
Matched data included in deposit (minus withdrawals)	1,806	1,259	3,065	1646	1,128	2,774

Data are available and matched for a total 5099 cohort members across a number of different tables.

Note: these figures are true at the time of deposit, actual numbers of records available may decrease if data subjects withdrawal their consent for data sharing.

4. The research datasets

4.1 Licensing and data access

The SAIL data have been processed by CLS and supplied to the UK Data Service (UKDS) under Secure Access Licence. Applicants wishing to access this data need to:

- register with the UKDS at https://www.ukdataservice.ac.uk/get-data/how-to-access/registration.

- establish the necessary agreement with the UKDS and abide by the terms and conditions of the UKDS Secure Access licence,

- access the SAIL linked data via the UKDS Secure Lab, via the researcher's own institutional desktop PC or at the Safe Room at the UK Data Archive.

4.2 Datasets provided

The datasets are in long format, each data subject can have more than one record in each dataset, with the exception of the Consents table and the Annual District Death Extract. The Annual District Death Extract is in wide format, only one record is expected per subject, only subjects with a death recorded in this dataset should have a record in this dataset. The consent table includes all subjects for whom we have a consent status.

Table 3: SAIL health datasets

SAIL health data	**Dataset name and contents** Files have prefix: mcs_nhs_wales_*	**Years**
Annual District Death Extract (ADDE)	• ***ADDE_deaths:*** Death Record information	2008 - 2020
Emergency Health Service (EDDS)	• ***EDDS_emer:*** Emergency Care information	2009 - 2020
Outpatient Database for Wales (OPDW)	• ***OPDW_outpatients:*** Outpatients records • ***OPDW_diag:*** Diagnosis records • ***OPDW_oper:*** Operations records	May 2004 - July 2020
Patient Episode Dataset for Wales (PEDW, Hospital Inpatient Records)	• ***PEDW_diag:*** Diagnosis records • ***PEDW_episode:*** Hospital episodes • ***PEDW_oper:*** Operation records • ***PEDW_spell:*** Hospital spells (a group of episodes)	Jan 1993 - July 2020

SAIL health data	Dataset name and contents Files have prefix: mcs_nhs_wales_*	Years
	• ***PEDW_superspell:*** Super spell (groups of spells)	
Welsh Demographic Service Dataset (WDSD)	• ***WDSD_pers***: Demographic information • ***WDSD_pers_gp***: GP history • ***WDSD_pers_add:*** Registered address history • ***WDSD_clean_add_wales:*** Cleaned address history • ***WDSD_clean_lsoa_2001:*** Welsh Indices of Multiple Deprivation information, quintiles and deciles according to LSOA01 • ***WDSD_clean_lsoa_2011:*** Welsh Indices of Multiple Deprivation information, quintiles and deciles according to LSOA011 • ***WDSD_clean_add_wales:*** Cleaned address history	Feb 1954 - Aug 2020
Welsh Longitudinal General Practice Dataset (WLGP)	• ***WLGP_gp_dates_inclnonsail_median:*** GP registration history. GP practices are kept separate in this dataset. Known as WLGP_CLEAN_GP_REG_BY_PRAC_INCLNONSAIL_MEDIAN in the SAIL databank • ***WLGP_gp_dates:*** GP registration history. GP practices are aggregated. Known as WLGP_CLEAN_GP_REG_MEDIAN in the SAIL databank • ***WLGP_gp_event_alf:*** GP record event • ***WLGP_gp_event_cleansed:*** GP record event	Feb 1966 - May 2021

SAIL health data	Dataset name and contents Files have prefix: mcs_nhs_wales_*	Years
	• ***WLGP_patient_alf_cleansed:*** Information about the source of GP data for each individual	
CLS consent	• ***health_consent _deposit:*** consent information for health linkages at the time of this deposit	2024

Annual District Death Extract (ADDE)

The ADDE dataset covers years 2008-2020 and contains data from the ONS register of all deaths relating to Welsh residents, including those that have died outside of Wales. There is one cohort member and 35 parents and partners. The data includes cause of death, year of death and place of death.

More information on the dataset can be found here: https://web.www.healthdatagateway.org/dataset/15cf4241-abad-4dcc-95b0-8cd7c02be999

Emergency Health Service (EDDS)

The EDDS dataset runs from 2009 to 2020 and contains data from all Accident and Emergency department and Minor Injury Unit attendances to NHS Hospitals in Wales. The data contains 1262 cohort members and 1997 parents and partners.

More information on the dataset can be found here: Emergency Department Dataset (EDDS) (healthdatagateway.org)

Outpatient Database for Wales (Hospital Outpatient Records) (OPDW)

The OPDW dataset runs from May 2004 – July 2020 and contains data for outpatient attendances. Details of diagnoses and operations and procedures can be found in the respective datasets, OPDW_diag and OPDW_oper.

Outpatient tables must be joined using att_id_e, case_rec_num, attend_dt and prov_unit_cd.

More information on the dataset can be found here: https://web.www.healthdatagateway.org/dataset/d331159b-b286-4ab9-8b36-db39123ec229

- OPDW_outpatients: Outpatients records. The data contains 1494 cohort members and 2591 parents and partners.
- OPDW_diag: Diagnosis records. The data contains 56 cohort members and 521 parents and partners.
- OPDW_oper: Operations records. The data contains 193 cohort members and 1050 parents and partners.

Patient Episode Dataset for Wales (Hospital Inpatient Records) (PEDW)

The PEDW covers the hospital admissions (both inpatient and daycases) in NHS hospitals in Wales, starting in Jan 1993 until July 2020. The dataset is separated into diagnoses, episodes, operations, spells and superspells.

The diagnoses and operations tables can be linked to spell using the variable spell_num_e, in conjunction with the mcsid and pnum, because the spell_num_e is only unique to each provider.

More information on the dataset can be found here: https://web.www.healthdatagateway.org/dataset/4c33a5d2-164c-41d7-9797-dc2b008cc852

- PEDW_diag: Diagnosis records. The data contains 1456 cohort members and 2559 parents and partners.
- PEDW_episode: Hospital episodes. The data contains 1463 cohort members and 2563 parents and partners.
- PEDW_oper: Operation records. The data contains 628 cohort members and 2453 parents and partners.
- PEDW_spell: Hospital spells (a group of episodes). The data contains 1463 cohort members and 2563 parents and partners.

- PEDW_superspell: Super spell (groups of spells). The data contains 1463 cohort members and 2563 parents and partners.

Welsh Demographic Service Dataset (WDSD)

The Welsh Demographic Service Dataset contains records related to the GP registration of the data subjects. The data spans from Feb 1954 to Aug 2020.

More information on the dataset can be found here: https://web.www.healthdatagateway.org/dataset/8a8a5e90-b0c6-4839-bcd2-c69e6e8dca6d

- WDSD_pers: Demographic information. The data contains 2071 cohort members and 3054 parents and partners.
- WDSD_pers_gp: GP history. The data contains 2071 cohort members and 3054 parents and partners.
- WDSD_address: Registered address history. The data contains 2071 cohort members and 3054 parents and partners. The variable 'to_dt' is missing where the subject was still registered at the end of the research period, for cohort members, this is the month they turned 14, for parents, this is the end of the period of data capture, which is August 2020.
- WDSD_clean_add_wales: Cleaned address history. The data contains 2071 cohort members and 3054 parents and partners. Variable 'd_end_dt' is missing where the subject was still registered at the end of the research period, for cohort members, this is the month they turned 14, for parents, this is the end of the period of data capture, which is August 2020.
- WDSD_clean_lsoa_2001: Welsh Indices of Multiple Deprivation information, quintiles and deciles according to LSOA01. The data contains 2001 cohort members and 3028 parents and partners.
- WDSD_clean_lsoa_2011: Welsh Indices of Multiple Deprivation information, quintiles and deciles according to LSOA011. The data contains 2001 cohort members and 3028 parents and partners. https://web.www.healthdatagateway.org/dataset/8a8a5e90-b0c6-4839-bcd2-c69e6e8dca6d

Welsh Longitudinal General Practice Dataset (WLGP)

The Welsh Longitudinal General practice dataset gives information about GP registration and attendances, and can be used to determine whether the individual is registered with a practice that supplies SAIL with GP attendance records or not.

The data spans from Feb 1966 to May 2021. More information can be found here: https://web.www.healthdatagateway.org/dataset/73093d81-0a1a-4e67-9b9a-4c1d20eaacf2

The variables prac_cd_pe, loc_num_pe and source_extract should be used when joining tables together in this dataset.

- WLGP_gp_dates_inclnonsail_median: GP registration history. GP practices are kept separate in this dataset. The data contains 2047 cohort members and 3039 parents and partners
- WLGP_gp_dates: GP registration history. GP practices are aggregated. Known as WLGP_CLEAN_GP_REG_MEDIAN. The data contains 1791 cohort members and 2668 parents and partners
- WLGP_gp_event_alf: GP record event.
- WLGP_gp_event_cleansed: GP record event. The data contains 1785 cohort members and 2734 parents and partners. The variables prac_cd_pe, loc_num_pe and source_extract should be used when joining this table with WLGP_patient_alf_cleansed
- WLGP_patient_alf_cleansed: Information about the source of GP data for each individual. The data contains 1816 cohort members and 2746 parents and partners. The variables prac_cd_pe, loc_num_pe and source_extract should be used when joining this table with WLGP_gp_event_cleansed.

A list of the available variables can be found on the data dictionary, available via the UKDS website. This data dictionary provides further information such as variable names and variable descriptions, as well as a field to request the variables for data application.

Note that this is routinely collected data, so will come with some errors and outlying values.

Consent dataset

The SAIL_health_consent dataset details who in the cohort had valid consent to linkage at the time of data sharing via the UKDS.

4.3 Data documentation provided

Users need to use the HES datasets in conjunction with the data dictionaries and documents provided by CLS available via UKDS, as follows:

Table 4: Data documentation

Documentation file	File name
User guide	MCS_SAIL_Health_UserGuide_v1.pdf
ICD-10 codes	ICD-10: International statistical classification of diseases and related health problems-V1-eng.pdf ICD-10: International statistical classification of diseases and related health problems-V2-eng.pdf ICD-10: International statistical classification of diseases and related health problems-V3-eng.pdf
OCPCS-4 codes	OPCS48 Metadata File Description V1.0.pdf OPCS48 ToCE Analysis Nov 2016 V1.0.xlsx OPCS48 ToCE Specification V0.1.pdf
Read Codes	Read_code_WLGP.xlsx Read Code Abbreviations.doc

Acronyms

Users may find useful to become familiar with the following list of acronyms used in the data dictionary and data labels:

CLS: Centre for Longitudinal Studies

HCP: Health Care Provider

HDU: High Dependency Unit

ICU: Intensive Care Unit

SpeII: A collection of medical episodes, from admission to discharge.

UKDS: UK Data Service

International Classification of Disease v10 (ICD-10)

These supplementary files originate from the WHO website[2] and will only be made available for approved projects.

Researchers should refer to "ICD-10: International statistical classification of diseases and related health problems V1" to interpret the diagnostic codes in the APC and OP datasets, V2 and V3 may be of help in building lists of codes to search for by diagnosis.

OPCS4 Interventions and Procedures Classification System

To interpret the OPCS data, researchers need to use the following supplementary files[3]:

- OPCS48 ToCE Analysis Nov 2016 V1.0
- OPCS48 ToCE Specification V0.1
- OPCS48 Metadata File Description V1.0

The version of OPCS-4 used over time does change, so codes for a procedure performed in 2007 are not necessarily the same as the same procedure performed in 2012, for example. The file "OPCS ToCE Analysis Nov 2016 V1.0" provides codes for each of the versions below.

---

[2] International statistical classification of diseases and related health problems, 10th revision, Fifth edition, 2016 https://apps.who.int/iris/handle/10665/246208, Accessed 24th August 2020

[3] The OPCS Classification of Interventions and Procedures, codes, terms and text is Crown copyright (2019) published by Health and Social Care Information Centre, also known as NHS Digital and licenced under the Open Government Licence available at www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm.

Table 5: OPCS versions

Version	Time period
OPCS4.9	2020 until further notice
OPCS4.8	2017-20
OPCS4.7	2014-17
OPCS4.6	2011-14
OPCS4.5	2009-11
OPCS4.4	2007-09
OPCS4.3	2006-07
OPCS4.2	Up to 31 March 2006

Read Codes and Read Code Abbreviations

The WLGP episode level datasets use Read codes to record medical information, including diagnoses, symptoms, procedures and tests. The files 'Read Code Abbreviations.doc' and 'EXTR_WLGP_READ_CODES.csv' can be used to understand the data. The records where version code has value "L" use a coding system other than Read Code, for example, "EMI.." refers to an attachment in the electronic health record. The read codes have been truncated to in line with the truncation of the ICD-10 codes. Please refer to the read_codes.csv supplementary file to interpret these files.

4.4 Identifiers

Use of individual identifiers to merge with cohort study data

MCS data are identified with the same research IDs used for the rest of cohort data available at the UKDS. This enables the data to be easily merged with one another.

While merging Covid-19 survey data with cohort study data should be similarly straightforward using their respective identifiers, users should consult individual user guides for specific information beforehand.

For MCS, researchers need to use both the MCS family identifier (MCSID) and the two individual person identifiers (CNUM00/PNUM00) to merge on with other cohort

data. As CNUM00 and PNUM00 include the wave number they may need consistent naming across datasets beforehand depending on the method of merging used.

There are different ways the data of MCS can be merged depending on the focus of the research project (Parent/Carers, Cohort Members or family). Details, syntax and examples on merging is provided by the MCS Data Handling Guide.

4.5 Data processing

Variable names

Whilst every attempt has been made to apply the variable and value labels in full, sometimes this is not compatible with the SPSS format, so where a variable uses multiple dictionaries, the labels haven't been applied.

Variables that have been included in the dataset unchanged also have the same variable name as in the

Variables that have been altered, either by truncation, top coding, recoding or creation of a pseudonymised key are named with the prefix D_. For example, the diagnosis variable wob becomes D_wob as it has been reduced from precision to the nearest week to the nearest year.

Variable labels and value labels

The majority of the variable and value labels have come directly from the NHS Data Dictionaries. We have also made use of external look-ups such as to the international coding such as ICD-10, OPCS4 and other diagnostic and treatment look ups. The PEDW and OPDW datasets use ICD-10 codes for recording diagnoses (diag_cd_nn) and OPCS-4 to record operations and procedures (oper_cd_123 + oper_cd_4), please see section 4.2 in this document for advice on interpretation.

Note that not all codes could be matched to the lookup files, so some values remain unlabelled.

Variables referring to health administrative groups have been described using the lookups available from the Office of National Statistics[4].

The variables that describe the health organisations have been labelled using these external sources. These have been applied to the data, however not every value in the dataset could be matched to a label; where this occurs the label will say “not in dictionary (*given value*)”.

Missing data

Some of the variables may only contain data for a few cases and mostly missing cases. Further information about the variables can be found in the online data dictionary[5] – some variables are retired and will not be entered after their retirement date. This happens when there have been changes to the data set and certain variables are no longer populated after certain years.
The missing cases have been recorded with the coded ‘-1’ for most variables. These variables did not contain any useful information and were removed.

4.6 Data de-identification

CLS is committed to protect research participants’ rights and avoid data disclosure and re-identification of individuals using one or more variables in the dataset or in combination with other existing data. A number of measures, such as removal of

[4] *Health Authorities (HA):* Office for National Statistics: Health Authorities and Health Boards (December 2001) Names and Codes in Great Britain. https://geoportal.statistics.gov.uk/datasets/health-authorities-and-health-boards-december-2001-names-and-codes-in-great-britain

Primary Care Trusts (PCT): Office for National Statistics: Primary Care Organisations (October 2005) Names and Codes in England https://geoportal.statistics.gov.uk/datasets/primary-care-organisations-october-2005-names-and-codes-in-england

Strategic Health Authorities (SHA) 2004: Office for National Statistics: Strategic Health Authorities (February 2004) Names and Codes in England. https://geoportal.statistics.gov.uk/datasets/strategic-health-authorities-february-2004-names-and-codes-in-england

Strategic Health Authorities (SHA) 2010: Office for National Statistics: Strategic Health Authorities (December 2010) Names and Codes in England https://geoportal.statistics.gov.uk/datasets/strategic-health-authorities-december-2010-names-and-codes-in-england

[5] https://www.datadictionary.nhs.uk/index.html NHS Data Model and Dictionary

variables, truncation and recoding, were put in place to de-identify the data as much as possible.

Dates of birth, small geographical details and rare cases that could easily lead to data disclosure have been removed. Variables including specific GP and health providers were pseudonymised, see Appendix.

Variables that could be used in combination to derive a date of birth for a person have been removed from the database or truncated.

A detailed description of the de-identification to the variables can be found in the **Appendix** of this document.

4.7 Output Disclosure control: requirements

For data available via the UKDS Secure Lab, the UK Data Service will always perform a certain level of disclosure control on the outputs generated by researchers, as outlined in their SDC Handbook, which can be downloaded from https://securedatagroup.org/sdc-handbook/.

The two UK Data Service Secure Lab rules of thumb that users should apply to all outputs are:

- Threshold rule: No cells should contain less than 10 observations;
- Dominance rule: No observation should dominate the data to a huge extent.

Appendix 1. Modifications to the datasets

Table 6: Modifications to the Annual District Deaths dataset

Variable name	SAIL original variable name	Variable description	Modification
Dv_death_reg_dt	Death_reg_dt	Date of Death Registration	Year of Death Registration
Dv_death_dt	Death_dt	Date of Death	Year of Death
Dv_death_health_org_cd	Death_health_org_cd	Organisation Code for the Local Health Board (LHB) for the place of the deceased, based upon the previous census	Pseudonymised code used

Table 7: Modifications to the Emergency Care EDDS dataset

Variable name	SAIL original variable name	Variable description	Modification
D_prov_unit_cd	prov_unit_cd	Provider unit code	Pseudonymised code used
D_prov_site_cd	prov_site_cd	Provider site code	Pseudonymised code used
D_site_cd_of_treat	site_cd_of_treat	Treatment site code	Pseudonymised code used

Table 8: Modifications to the Outpatients OPDW dataset

Variable name	SAIL original variable name	Variable description	Modification
D_prov_unit_cd	prov_unit_cd	Provider unit code	Pseudonymised code used

D_prov_site_cd	prov_site_cd	Provider site code	Pseudonymised code used
D_site_cd_of_treat	site_cd_of_treat	Treatment site code	Pseudonymised code used
D_source_of_ref_cd	source_of_ref_cd	Source of referral code	Pseudonymised code used

Table 9: Modifications to the Diagnosis and Operations OPDW datasets

Variable name	SAIL original variable name	Variable description	Modification
D_prov_unit_cd	prov_unit_cd	Provider unit code	Pseudonymised code used

Table 10: Modifications to the Diagnosis, Operations and Superspell PEDW datasets

Variable name	SAIL original variable name	Variable description	Modification
D_prov_unit_cd	prov_unit_cd	Provider unit code	Pseudonymised code used

Table 11: Modifications to the Episodes PEDW dataset

Variable name	SAIL original variable name	Variable description	Modification
D_prov_unit_cd	prov_unit_cd	Provider unit code	Pseudonymised code used
D_prov_site_cd	prov_site_cd	Provider site code	Pseudonymised code used
D_site_cd_of_treat	site_cd_of_treat	Treatment site code	Pseudonymised code used
D_curr_prov_unit_cd	Curr_prov_unit_cd	Current provider unit code	Pseudonymised code used

D_curr_prov_site_cd	Curr_prov_site_cd	Current provider site code	Pseudonymised code used
D_ua_cd	Ua_cd	Unitary Authority	Pseudonymised code used
D_gmc_con_cd_pe	Gmc_con_cd_pe	Consultant code	Pseudonymised code used

Table 12: Modifications to the Spells PEDW dataset

Variable name	SAIL original variable name	Variable description	Modification
D_prov_unit_cd	prov_unit_cd	Provider unit code	Pseudonymised code used
D_curr_prov_unit_cd	Curr_prov_unit_cd	Current provider unit code	Pseudonymised code used
D_curr_ua_cd	Curr_ua_cd	Unitary Authority	Pseudonymised code used
D_res_dha_cd	Res_dha_cd	District Health Authority of residence	Pseudonymised code used
D_res_ward_cd	Res_ward_cd	Electoral Ward of residence	Pseudonymised code used
D_ref_org_cd	Ref_org_cd	Referring organisation code	Pseudonymised code used
D_curr_res_dha_cd	Curr_res_dha_cd	Current district health authority of residence	Pseudonymised code used
D_curr_res_ward	Curr_res_ward	Current electoral ward of residence	Pseudonymised code used

Table 13: Modifications to the Person, WDSD dataset

Variable name	SAIL original variable name	Variable description	Modification

D_wob	wob	Week of birth	Truncated to year
D_dod	dod	Date of death	Truncated to year

Table 14: Modifications to the LSOA01, LSOA11, Person WDSD dataset

Variable name	SAIL original variable name	Variable description	Modification
D_end_date	end_date	End date	Truncated to month CM turned 14, if later

Table 15: Modifications to the Address, WDSD dataset

Variable name	SAIL original variable name	Variable description	Modification
D_to_dt	To_dt	End date	Truncated to month CM turned 14, if later

Table 16: Modifications to the GP registration history (dates including non-SAIL), GP WLGP dataset

Variable name	SAIL original variable name	Variable description	Modification
D_end_date	end_date	End date	Truncated to month CM turned 14, if later

Table 17: Modifications to the Patient ALF cleansed, GP ALF event cleansed, dataset

Variable name	SAIL original variable name	Variable description	Modification

Dv_wob	Wob	Week of birth	Truncated to year