Millennium Cohort Study

Linked health administrative datasets – Hospital Episode Statistics (HES)

User Guide (Version 2)

April 2025

CENTRE FOR LONGITUDINAL STUDIES

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, Danielle Gomes, Aida Sanchez-Galvez, Emla Fitzsimons

How to cite this guide

Kerry- Barnard, S., Fitzsimons, E., Gomes, D., Sanchez-Galvez, A (2025) Millennium Cohort Studies: A guide to the linked health administrative datasets – Hospital Episode Statistics (HES). User Guide (Version 2). London: UCL Centre for Longitudinal Studies.

This guide was published in April 2025 by the UCL Centre for Longitudinal Studies.

Data citation and CLS acknowledgement

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

Where practicable, any outputs will acknowledge the source of the data as: “This work uses data provided by patients and collected by the NHS as part of their care and support.”

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 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 that 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 https://cls.ucl.ac.uk/data-access-training/exploring-ourdata/

1. Introduction

This guide describes the data linkage of health administrative records from the Hospital Episode Statistics (HES) to survey data for cohort members in the Millennium Cohort Studies (MCS). The main aim of this data linkage exercise is to enhance the research potential of the study, by combining administrative record with the rich information collected in the surveys.

In 2020 CLS made a request to NHS England to link all consenting MCS participants to their HES records, which covered the earliest years which data was available from NHS England up to March 2020.

In 2024, CLS updated the linkage to include data from April 2018 to March 2020 for the Accident and Emergency datasets and from April 2020 to March 2023 for the other four datasets (APC, OP, CC ECDS). The data linkage was carried out by the NHS England team for both linkages.

2. Consent to health data linkage

In the 2018 sweep (age 17) MCS7, which was a face to face interview, cohort members were asked for consent to link their health records to their MCS survey data. In order to obtain consent, a data linkage booklet was sent to the cohort member in advance of their interview once an appointment had been secured. It provided information on the purpose, types, value and process of data linkage, and encouraged cohort members to contact the CLS survey team with any questions they might have. During the interview, interviewers first showed the cohort member a video on their tablet explaining data linkage. After confirming that cohort members had read the information booklet (and in the event they had not, that the interviewer had read some key information out to them), consent to data linkage was collected on paper. Cohort members were left with a carbon copy of the consents they had given.

Detailed information on the fieldwork and consent collection can be found in the Millennium Cohort Study: Age 17 Sweep (MCS7) User Guide and Millennium Cohort

Study Seventh Sweep (MCS7) Technical report. All documents can be found under 'documentation' at https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/mcs-age-17-sweep/

At Age 23, CLS sought consent from participants who had not previously consented to data linkage at Age 17. While this consent is not part of the current data linkage, it will be available through the UK Longitudinal Linkage Collaboration (UKLLC). NHS data linked to CLS survey data can also be accessed remotely via the UKLLC at https://ukllc.ac.uk/apply

3. Health data linkage

3.1 HES datasets

The Hospital Episode Statistics (HES) is a database that contains information about all hospital admissions in England. The data holder is NHS England.

HES data is comprised of four datasets: Accident and Emergency episodes dataset (AE), Admitted Patient Care episodes dataset (APC), Adult Critical Care episodes dataset (CC) and Outpatients episodes dataset (OP). The Emergency Care Data Set (ECDS) was introduced in 2017 to gradually replace the A&E dataset.

The data cover diverse topics including diagnosis, maternity, mortality, mental health, types of therapies, treatment's length, Indices of Multiple Deprivation (IMD), service providers, organisations, and regional geographical location.

The NHS England website contains detailed information about each dataset, including quality reports on expected episodes that are missing, potential coding issues (i.e. where variables are not correctly coded), duplicate episodes have been observed and systemic problems that led to an absence in data. This information can be found here: Hospital Episode Statistics (HES) - NHS England

3.2 Matching strategy

In 2020 CLS made a request to NHS England to link all consenting MCS participants to their HES records. The previous version of this User guide included the first data linkage which covered the earliest years which data was available from NHS England up to March 2020.

In 2024, CLS requested NHS England to update the linkage to include data from April 2017 to March 2023.

Table 1: List of HES datasets linked by CLS

HES dataset	Contents
A&E	Attendance to Accident and Emergency Care [April 2007 - March 2020]
ECDS	Attendance to Emergency Care [April 2020 - March 2023]
APC	Attendance to Admitted Patient Care [April 1997 - March 2023]
CC	Attendance to Critical Care [April 2009 - March 2023]
OP	Attendance to Outpatient [April 2003 - March 2023]

A MCS cohort member was only matched when there was a record for them as a patient within the various databases, hence the difference in the numbers of matched cases for each type of dataset. The matching is subject to a quality indicator recorded in the variable 'match_rank' that allows the user to assess the quality of match, but this is not available in the refreshed data (2024 linkage).

a) Matching using the participant's personal information

CLS sent a matching file to NHS England containing the following information for cohort members recorded as agreeing to health linkages: Name (forename, middle name and surname, other surname), sex, date of birth, full current address including most recent postcode, a known date of the address, CLS proxy ID and NHS number.

The data was matched by NHS England on the following basis:

- Name, sex, date of birth and postcode
- Name, date of birth and sex
- NHS England then flagged cohort members in their system and matched their information to their NHS Number (NHSNO)

Cohort members are free to withdraw their consent to health linkages and to data sharing at any time. The numbers in the document are true at the time the data were shared via the UKDS but may change slightly over time. Researchers with access to this dataset, will be able to use the Consent information provided (see table 5).

To maintain confidentiality of cohort members the linked health records are made available for researchers in a pseudo-anonymised version using the MCS identifier variable (MCSID) which is the same ID used for the other research datasets available from the UKDS.

3.3 Matching rates

A total of 9,214 participants agreed to health linkage out of 10,757 who took part in the Age 17 sweep, corresponding to a consent rate of 85.7%. At the time of this linkage CLS identified 7,396 participants who were eligible for linkage. The remaining 1,814 consenting participants were not included in the matching file as their NHS number indicated Scottish or Northern Irish residence or no current link to England. Those who had withdrawn from the study were not included (four cases). Participants were selected based on the following criteria:

- Whether they had a valid English or Welsh NHS number (6,548 cases were identified)
- Without NHS number but originally sampled from England and Wales (817 cases)
- Other participants (Scottish or Northern Irish) with current address in England (31 cases)

CLS received HES data for a total of 6,027 MCS participants out of 7,396 sent for matching.

In 2024, CLS received HES data for a total of 3925 MCS Age 17 participants out of 7,396 originally sent to NHS England.

CLS received data for (6479) individuals across both linkages, and 6247 have been included in the data provided, after withdrawals are taken into account.

Table 2 below shows the number of successful matches to HES records.

Table 2: Consent and overall matching

Matching	2020	2024
Consent at time of linkage	9,214	9,214
Consent rate	85.7%	85.7%
Sent for matching	7,396	7,386
Data received at CLS	6,027	3925
Data received, minus withdrawals	6,014	3923
Coverage %	81.3%	53.1%

Data was available and matched for a total of 6247 participants in the different HES datasets. For each of the datasets, the matching was as shown in Table 3:

Table 3: Matching for each HES database

HES dataset	MCS participants			HES records			Variables
	2020	2024	All	2020	2023	All	Excluding ID vars
AE	4789	2269	5108	18950	4709	23659	167
APC	4163	1114	4446	14209	2884	17093	332
OP	5327	2553	5565	95032	20021	115053	146
CC	25	19	43	34	22	56	29
ECDS	1888	3002	3709	3884	7219	11103	281
Total	6014	3923	6247	132109	34855	166964	955

4. The research datasets

4.1 Licensing

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

- Establish the necessary agreement with the UKDS and abide by the terms and conditions of the UKDS Secure Access licence
- Specify the exact variables that they require for their project and will only be given access to a tailor-made subset of the HES data as specified in their application (note: any cohort members who have requested a deletion of their data will not be included in the tailor-made subset)

For details on how to apply for the data, please refer to section 5 of this document.

4.2 Datasets

Datasets are long in structure, apart from the consent data, which has one row per cohort member

Table 4: List of available datasets and contents

Name of the dataset	Content summary
mcs_eng_health_nhs_hes_ae_2007_to_2019.sav	Accident and Emergency episodes
mcs_eng_health_nhs_hes_apc_2000_to_2022.sav	Admitted Patient Care episodes
mcs_eng_health_nhs_hes_cc_2009_to_2022.sav	Critical Care episodes

Name of the dataset	Content summary
mcs_eng_health_nhs_hes_consent_linkage_info_2025_deposit.sav	Consent data
mcs_eng_health_nhs_hes_ecds_2019_to_2022.sav	Emergency Care dataset episodes
mcs_eng_health_nhs_hes_op_2003_to_2022.sav	Outpatient care episodes

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 5: List of documents

Documentation file	**File name**
User guide	MCS_HES_UserGuide_v2.pdf
CLS Data Dictionaries	MCS_HES_Variables_List_v2.xlsx
NHS Data Dictionaries	ECDS_ETOS_v3.1.1.xlsx HES+TOS+V2.03.xlsx
HES Analysis Guide	HES_analysis_guide_december_2019.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	OPCS-4.9 to OPCS-4.10 Summary of Core Changes Nov 2022 V1.0.pdf OPCS410 ToCE Analysis Nov 2022 V1.0.xlsx OPCS410 CodesAndTitles Nov 2022 V1.0.txt OPCS410 Metadata File Description V1.0.pdf OPCS410 MetaData Nov 2022 V1.0.txt OPCS410 ToCE Specification Nov 2022 V0.1.pdf
A&E Diagnosis and Treatments	A&E Diagnostic and treatment codes.xlsx

Acronyms

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

A&E: Accident and Emergency

EDCS: Emergency Care Data Set

APC: Admitted Patient Care dataset

CC: Critical Care

CCU: Coronary Care Unit

CLS: Centre for Longitudinal Studies

HCP: Health Care Provider

HDU: High Dependency Unit

HES: Hospital Episodes Statistics

ICU: Intensive Care Unit

OP: Outpatients

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

UKDS: UK Data Service

NHS Data Dictionaries

The data dictionaries from NHS England[1] are available in the supplementary documents. These dictionaries will help in interpreting the data. The NHS data dictionaries contain the full variable description and value labels, and when the variable came into use or was retired.

CLS Data Dictionaries

The data dictionaries generated by CLS provide detailed information for each of the four HES research datasets linked to MCS and curated by CLS. They include the variables names, format, labels or titles, positions in each dataset. They also provide information of the values included in each variable and a column to specify whether the variables will be requested as part of the data application.

These data dictionaries are based on NHS England documentation mentioned above.

HES Analysis Guide

To use the HES datasets, users are required to be familiar with the HES Analysis Guide provided by NHS England[2]. This document has been supplied as a supplementary documentation file.

International Classification of Disease v10 (ICD-10)

These supplementary files originate from the WHO website[3] and will only made available for approved projects:

- ICD-10: International statistical classification of diseases and related health problems-V1-eng.pdf

[1] NHS Data Dictionaries, NHS Digital: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary (accessed 14th May 2024)

[2] https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/users-uses-and-access-to-hospital-episode-statistics

[3] International statistical classification of diseases and related health problems, 10th revision, Fifth edition, 2016 https://apps.who.int/iris/handle/10665/246208, Accessed 14th May 2024

- 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

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[4]:

- 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.

Table 4: Operations Classifications and Standard Procedures (OPCS) versions

Version	Time period
OPCS4.10	1 April 2023 until further notice
OPCS4.9	1 April 2020 until March 2023
OPCS4.8	1 April 2017- 31 March 2020
OPCS4.7	1 April 2014- 31 March 2017
OPCS4.6	1 April 2011- 31 March 2014

[4] 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. The datasets can be downloaded here: https://isd.digital.nhs.uk/trud/users/authenticated/filters/0/categories/10/items/119/releases Accessed 14th May 2024

OPCS4.5	1 April 2009- 31 March 2011
OPCS4.4	1 April 2007- 31 March 2009
OPCS4.3	1 April 2006- 31 March 2007
OPCS4.2	Up to 31 March 2006

4.4 Identifiers

MCSID is the anonymised unique cohort member identifier which is used to maintain the confidentiality of cohort members in the linked health records. The MCSID can also be used to merge this data and other deposited MCS datasets.

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.

Variables that have been included in the dataset unchanged also have the same variable name as in the NHS data dictionaries.

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 diag_01 becomes D_diag_01 as it has been truncated to 3 characters.

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 loop-ups such as to the international coding ICD-10, OPCS4 and other diagnostic and treatment look ups. The APC and OP datasets use ICD-10 codes for recording diagnoses (D_diag_nn) and OPCS-4 to record operations and procedures (opertn_nn), 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.

The administrative variables are: Strategic Health Authority of Commissioning Office (PURSTHA), Strategic Health Authority of residence in the year of treatment (RESSTHA_HIS) and Regional Office of the GP practice (GPPRACRO). The health care providers (PROCODE, PROCODE3, PROCODE5, SENDER), Strategic Health Authority of GP practice (GPPRSTHA), Primary Care Trust if the GP practice (GPPRPCT), have been given in pseudonymised form.

Identification of HES episodes and spells

NHS administratively organises the data by Hospital Spells and Episodes which are recorded separately as single record (row of data) per patient. An **episode** is defined by NHS as a continuous period of admitted patient care administered under one consultant within healthcare providers. A **hospital spell** is defined by the total time spent by a patient in the same care provided by the hospital, from date of admission to date of discharge. Spells may contain a single episode or multiple episodes at the same health provider. If a patient is transferred to another consultant in the same healthcare provider, this new episode will be part of the same spell but recorded in a new row.

NHS administrative data only provides a date of discharge if the episode was the last service provided by a consultant/medical practitioner at that particular health provider. As a result, multiple episode spells may be identified by looking at records that have the same admission date (variable admidate). Only the last episode of the spells will have a discharge date (variable disdate). The previous episodes of the same spell do not have a discharge date.

Cohort members may have multiple episodes as part of the same spell recorded with the same admission date at a single healthcare provider or may have different episodes as part of different spells in the same hospital or in various health providers.

Within multi-episode spells, the last episode has all the diagnosis codes registered in that spell in variables D_diag_01 to D_diag_20. To avoid having duplicate diagnosis codes, researchers need to consider data rows which have a date in variable disdate.

Missing data

Some of the variables may only contain data for a few cases and mostly missing cases. For example, the OP dataset contains the variable LOCTYPE ‘Location Type’, which has missing values for 96% of records. The variable selection list contains a column ("counts") with counts of how many records contain non-missing data for each variable.

The missing cases have been recorded with the coded '-1' for most variables. A few variables requested by CLS did not contain information for the cohort members (i.e. VIND, 'V code indicator' OR WELL_BABY_IND 'Well baby indicator flag'). These variables did not contain any useful information and were removed.

Similarly, diagnostic codes in the OP dataset (variable diag_[01-12]) are mostly coded as "R69X" meaning "unknown and unspecified causes of morbidity" (98%). Note that these are included in the count of non-missing data previously mentioned.

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 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 to comply with the small numbers section of the data analysis guide from NHS England[5].These include all administrative variables relating to service providers, these have been pseudonymised.

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

Users who use fine grained geographical information, which is available as part of HES but also as part of the cohort data, must comply with the small numbers section of the data analysis guide from NHS England when publishing data.

[5] https://digital.nhs.uk/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hes_analysis_guide_december_2019-v1.0.pdf also available in the supplementary documentation. Accessed 14th May 2024

A detailed description of the de-identification of the variables can be found in **Appendices 1 to 3** of this document.

4.7 The Accident and Emergency (A&E) data

The A&E dataset details each attendance to an Accident and Emergency care facility in England, between 01-04-2007 and 31-03-2020 (inclusive). It includes major A&E departments, single specialty A&E departments, minor injury units and walk in centres in England. People can have more than one medical record in a single year or different years. If a patient arrives and is sent to a different clinic (i.e. walk-in clinic), this may appear as two records.

The A&E information is described in detail in the NHS A&E data dictionary, which is provided as a supplementary documentation file. The number of linked cohort members and records found in the data are detailed in **Table 3.**

A list of the available variables can be found in 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.

The file “A&E Diagnostic Treatment codes” can be used to interpret the diagnostic and treatment codes (DIAG_01 to DIAG_12, DIAGA_01 to DIAGA_05 and DIAGS_01 to DIAGS_05). The codes have been taken from the NHS England website, linked in the file. Not all of the values map to the supplied metadata: in some cases, an alternative coding schedule has been used for DIAG_01 to DIAG_12, this should be indicated in DIAGSCHEME. In other cases, this could be an input error.

4.8 The Admitted Patient Care (APC) data

The APC data summarises episodes of care for admitted patients, where the episode occurred between 01-04-1997 and 31-03-2023 (inclusive). An episode is a period of care under a single consultant at a single hospital – there can be more than one record for an admission period.

The number of linked cohort members and records found in the data are detailed in Table 3.

The APC dataset contains the majority of the available administrative information for the research participants. People may have multiple episodes to one admission, ordered by the episode order variable ‘epiorder’.

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

The APC information is described in detail in the NHS APC data dictionary, which is provided as a supplementary documentation file.

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

Not all values in the treatment specialty variable (TRETSPEF) are labelled in the data dictionary; it is likely these are due to data quality issues.

4.9 The Critical Care (CC) data

The CC dataset covers records of critical care activity between 01-04-2009 and 31-03-2023 (inclusive). This is the smallest of the five datasets, the number of linked cohort members and records found in the data are detailed in Table 3.

The variables are specified in detail in the NHS CC data dictionary, which is provided as a supplementary documentation file.

All critical care records have a parent APC record. The variable called D_susid was obtained by linking the record identifier “susrecid” in the critical care dataset to the corresponding “susrecid” in the APC dataset. The variable D_susid can be used to link the APC and CC datasets; researchers who wish to look at the critical care dataset should take care to select this variable in both datasets.

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

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

Researchers requesting data from the CC dataset should request D_SUSID from both the APC and CC datasets to link them together.

4.10 The Emergency Care Data

The Emergency Care dataset replaces the Accident and Emergency Data Set; it covers attendances to both urgent and emergency care facilities in England from 01-04-2020 to 31-03-2023.

The number of linked cohort members and records found in the data are detailed in Table 3. The data are described primarily using SNOMED-CT codes. The codes specific to the ECDS dataset can be found in the Enhanced Technical Output Specification (ETOS) [6], and the full UK specific vocabulary can be downloaded from the NHS England TRUD [7], however the values in the data have been labelled in this deposit.

4.11 The Outpatient Care (OP) data

The OP dataset lists the outpatient appointments between 01-04-2003 and 31-03-2023 (inclusive).

The number of linked cohort members and records found in the data are detailed in Table 3. The details of these variables are included in the NHS OP data dictionary, which is included as a supplementary documentation file.

Most of diagnostic codes (variable diag_[01-10]) are coded as “RX69X” unknown and unspecified causes of morbidity (98%). The Classification of Interventions and Procedures (variables, opertn [01-19]) and version of classification have just over one fifth (22%) of values coded as X997, which is not in the scope of the dictionary.

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

[6] https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0092-2062-commissioning-data-sets-emergency-care-data-set Accessed 30th August 2022

[7] https://isd.digital.nhs.uk/trud/user/guest/group/0/pack/26 Accessed 30th August 2022

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

The diagnosis variables (D_diag_01 to D_diag_05) use ICD-10 codes; these are included in the supplementary data files (see section 4.2). Similarly, the operation codes (opertn_01 to opertn_20) use OPCS-4 codes, please see section 4.2 in this document for advice on interpretation.

4.12 Consent Information dataset

The Consent Information Dataset contains a record for every cohort member with a record regarding whether they did or did not consent to health linkages at age 18. This table will be provided with the data and does not need to be requested in the HES Variables request form

Table 5: Consent Information Data Dictionary

Variable	Variable description	Values
mcsid	Research ID	
linkreq	Cases requested from NHS England	Data requested Apr 2001 - Mar 2023
		Consented, but not requested
		Data not requested, or consent withdrawn
datareturned	Data matched by NHS England	Data returned
		No data returned
		Data not requested, or consent withdrawn

5. Data access and variable selection

5.1 UKDS Secure Access application

Access to the HES linked data will only be provided via the UKDS Secure Lab Applicants wishing to access this data need to establish agreement with the UKDS and abide by the terms and conditions of the UKDS Secure Access licence. Before gaining access, researchers must make an application detailing the intended analysis and provide a justification as to why this data is requested.

5.2 Selection of variables

Researchers must specify the list of variables that they require for their project and will only be given access to a tailor-made subset of the HES data as specified in their application.

This should be done using the Excel spreadsheet

MCS_HES_Variables_List_v2.xlsx

Each data sheet has its own worksheet. Please type 'yes' next to each required.

Note that to link the Critical Care data to the relevant Admitted Patient Care record, researchers need to select D_SUSID in both datasets.

5.3 CLS Licence Agreement

In addition to registering and submitting an application to the UKDS, the organisation requesting to use the linked health data will also need to enter into a Licence Agreement with CLS.

Users should complete the '*CLS Licence Agreement – NHS England data'* document. This document will be provided by the UKDS at the point of applying to access the data.

We advise users to pay extra attention to Schedule 1 section of the *CLS Licence Agreement* and provide information on the expected measurable benefits to Health and/or Social Care of the research project. This is an NHS England requirement for accessing administrative health data and CLS will assess if the information provided in this section meets this requirement before approving applications.

Users will also need to ask their organisations to provide ‘Organisational Security Assurance’ in the relevant section of the CLS Licence Agreement. The organisation’s security assurance can be a Security Level Systems Policy (SLSP), or Data Security and Protection toolkit or International Organisation for Standardisation (ISO27001). Users should provide a copy of the associated documentation with their applications.

6. Disclosure control: requirements for data users

6.1. UKDS requirements

As the HES data linked to the longitudinal MCS data are only 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 will be applied 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

6.2. NHS England requirements

The NHS England have also have a number of specific requirements and these are specified below:

- 'Small numbers' in HES are the numbers 1 to 5. Low-level analyses are more likely to contain small numbers, which might facilitate identification of individual patients, especially at a local level. They might also allow identification of a hospital consultant, where local knowledge identifies a single consultant treating patients in a particular specialty.
- Small numbers are not necessarily a problem when they cover a broad geographical area, because the patient would not normally be identifiable (see

Table 1 of the Guide for analysis of HES, for the acceptable levels). However, data that are likely to be more sensitive, e.g. deaths (see 6.2.1 of the Guide for analysis of HES), should still be treated with care if they are likely to identify individuals. Small numbers within local authorities (LAs), wards, postcode districts, CCGs providers and trusts may allow identification of patients and should not be published/released.

- When publishing/releasing HES data, you must make sure that cell values from 1 to 5 are suppressed at a local level to prevent possible identification of individuals from small counts within the table. Zeros (0) do not need to be suppressed. If only one cell requires cell suppression, you must suppress at least one other component cell (the next smallest) to avoid calculation of suppressed values from the totals. You should replace these values with ’*' and add a note: ‘*’ in this table means a figure between 1 and 5.
- The rules on suppression of low cell counts should be considered wherever small numbers are encountered, irrespective of whether the count is directly a count of patients. The rules cover several types of analysis (e.g. episodes, admissions and deaths) and measures based on small numbers, such as bed days. While a bed day measure may not appear to be disclosive, a small number of bed days may imply a small number of cases so similar suppression is needed.
- Certain other measures, such as average times waited or length of stay, appear not to give any disclosive information on the number of cases, but at times they may do so, e.g. a mean of 5 days with up to 5 cases implies no case exceeded 25 days. In such cases, the averages might not be disclosive, but judgement still needs to be taken as to whether they imply something more about individual cases.
- An alternative to suppressing values from 1 and 5 is to consider a higher level of aggregation for one or more items, e.g. move from trust level to Area Team/Commissioning region of treatment, or from diagnosis at the 4-character level to the 3-character level, or group using wider age bands. A higher level of aggregation is the preferred option if several cells are affected by the suppression rule.

- Another option is to provide the data at the requested low level (if necessary for purpose), but anonymising the level of aggregation, i.e. replace identifying codes or labels with arbitrary reference numbers.
- In addition to this, as detailed in the small number table, there are a number of diagnosis and procedure codes which are covered by the small numbers guidance. This list is currently under review and may be subject to change once ratified. Advice should be sought from the HSCIC if there are any doubts around any potentially sensitive ICD10 or OPCS codes.

For further information on disclosure control please refer to the Guide for analysis of the Hospital Episode Statistics document. It is included in the supplementary documentation and can be downloaded here https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/users-uses-and-access-to-hospital-episode-statistics

Appendix 1. Modifications to the Accident and Emergency Data

Variable name	NHS original variable name	Variable description	Modification
D_GPPRAC	GPPRAC	Code of GP practice	Recoded to pseudonymised code
D_GPPRPCT	GPPRPCT	PCT of GP practice	Recoded to pseudonymised version
D_GPPRSTHA	GPPRSTHA	SHA of GP practice	Recoded to pseudonymised version
D_INVEST_[01-09]	INVEST_[01-09]	A&E Investigation	Truncated to 2 characters
D_ORGPPPID	ORGPPPID	Organisation for the patient pathway provider	Recoded to pseudonymised version
D_PCTTREAT	PCTTREAT	PCT of Treatment	Recoded to pseudonymised version
D_PCTCODE02	PCTCODE02	Historic PCT of responsibility	Recoded to pseudonymised version
D_PCTCODE06	PCTCODE06	Current PCT of responsibility	Recoded to pseudonymised version
D_PCTCODE_HIS	PCTCODE_HIS	PCT of responsibility (legacy vr)	Recoded to pseudonymised version
D_PROCODE	PROCODE	Organisation code (code of provider	Recoded to pseudonymised version
D_PROCODE3	PROCODE3	Provider code - 3 character	Recoded to pseudonymised version
D_PROCODE5	PROCODE5	Provider code - 5 character	Recoded to pseudonymised version
D_PURCODE	PURCODE	Commissioner Code 5 char	Recoded to pseudonymised version
D_PURSTHA	PURSTHA	Commissioner's Strategic Health Authority	Recoded to pseudonymised version

D_RESPCT_HIS	RESPCT_HIS	PCT of residence	Recoded to pseudonymised version
D_RESSTHA_HIS	RESSTHA_HIS	SHA of residence - mapped according to data year	Recoded to pseudonymised version

Appendix 2. Modifications to the Admitted Patient Care Data

Variable name	NHS original variable name	Variable description	Modifications
D_ALCDIAG	ALCDIAG	Principal alcohol related diagnosis	Truncated ICD-10 diagnosis code to 3 characters
D_ANAGEST	ANAGEST	Gestation period in weeks at first antenatal assessment	Top coded to 42
D_CAUSE	CAUSE	ICD-10 Cause code	Truncated to 3 characters
D_DIAG_[1-20]	DIAG_[1-20]	ICD-10 Diagnosis codes	Truncated to 3 characters
D_GESTAT	GESTAT	Length of gestation	Top coded to 42.
D_GPPRAC	GPPRAC	Code of GP practice	Recoded to pseudonymised version
D_GPPRPCT	GPPRPCT	PCT of GP practice (gpprpct)	Recoded to pseudonymised version
D_NUMPREG	NUMPREG	Number of pregnancies	Top coded to 5 (except 99)
D_ORGPPPID	ORGPPPID	Organisation for the patient pathway provider	Recoded to pseudonymised version
D_PCTCODE02	PCTCODE02	Historic PCT of responsibility	Recoded to pseudonymised version
D_PCTCODE06	PCTCODE06	Current PCT of responsibility	Recoded to pseudonymised version

Variable name	NHS original variable name	Variable description	Modifications
D_PCTCODE_HIS	PCTCODE_HIS	PCT of responsibility (legacy vr)	Recoded to pseudonymised version
D_PCTNHS	PCTNHS	Primary care trust of responsibility - NHS	Recoded to pseudonymised version
D_POSTDUR	POSTDUR	Postnatal days of stay	Recoded to: 0 or 1 days, 2 or more days to avoid certainty of DOB of infant
D_PROCODE	PROCODE	Organisation code (code of provider)	Recoded to pseudonymised version
D_PROCODE3	PROCODE3	Provider code - 3 character	Recoded: Derive from PROCODE recode
D_PURCODE	PURCODE	Commisioner Code 5 char	Recoded to pseudonymised version
D_PURRO	PURRO	Regional Office of Purchaser	Recoded to pseudonymised version
D_REFERORG	REFERORG	Referring Organisation code (referorg)	Recoded to pseudonymised version
D_RESSTHA_HIS	RESSTHA_HIS	SHA of residence - mapped according to data year	Recoded to pseudonymised version
D_SENDER	SENDER	CDS Sender Identity (sender)	Recoded to pseudonymised version
D_SUSID	SUSRECID	Secondary Uses ID	Recoded to pseudonymised version
D_WAITDAYS	WAITDAYS	Duration of elective wait	Top coded to 365

Appendix 3. Modifications to the Outpatient Care Data

Variable name	NHS original variable name	Variable description	Modifications
D_DIAG_[1-12]	DIAG_[1-12]	All Diagnosis codes	Truncated to 3 characters
D_GPPRAC	GPPRAC	Code of GP practice	Recoded to pseudonymised code.
D_GPPRACHA	GPPRACHA	Health Authority of GP practice	Recoded to pseudonymised version
D_GPPRACRO	GPPRACRO	Regional Office of GP Practice	Recoded to pseudonymised version
D_GPPRPCT	GPPRPCT	PCT of GP practice (gpprpct)	Recoded to pseudonymised version
D_GPPRSTHA	GPPRSTHA	SHA of GP practice	Recoded to pseudonymised version
D_OPERTN_[1-19]	OPERTN_[1-19]	Operative procedure	Truncated to chapter (1 character)
D_PCTTREAT02	PCTTREAT02	PCT of Treatment	Recoded to pseudonymised version
D_PCTCODE02	PCTCODE02	Historic PCT of responsibility	Recoded to pseudonymised version
D_PCTCODE06	PCTCODE06	Current PCT of responsibility	Recoded to pseudonymised version
D_PCTCODE_HIS	PCTCODE_HIS	PCT of responsibility (legacy vr)	Recoded to pseudonymised version
D_PROCODE	PROCODE	Organisation code (code of provider)	Recoded to pseudonymised version

D_PROCODE3	PROCODE3	Provider code - 3 character	Recoded to pseudonymised version
D_PROCODE5	PROCODE5	Provider code - 5 character	Recoded to pseudonymised version
D_PURCODE	PURCODE	Commissioner Code 5 char	Recoded to pseudonymised version
D_PURSTHA	PURSTHA	Commissioner’s Strategic Health Authority	Recoded to pseudonymised version
D_REFERORG	REFERORG	Referring Organisation code (referorg)	Recoded to pseudonymised version
D_SENDER	SENDER	CDS Sender Identity (sender)	Recoded to pseudonymised version

Appendix 4: Modifications to the Emergency Care Dataset

Variable name	NHS original variable name	Variable description	Modifications
attendance_source_organisation	d_attendance_source_organisation	Organisation site identifier for the site from which a patient arrived at the emergency department	Recoded to pseudonymised version
commissioner	d_commissioner	Organisation Identifier of the Organisation commissioning healthcare	Recoded to pseudonymised version
conveying_ambulance_trust	d_conveying_ambulance_trust	Organisation Identifier of the Conveying Ambulance Trust	Recoded to pseudonymised version
gpprac	d_gpprac	GP Practice	Recoded to pseudonymised version
interchange_sender	D_interchange_sender	CDS interchange sender identity	Recoded to pseudonymised version
lpi_organisation_code	d_lpi_organisation_code	Learning and Performance Initiative Organisation Code	Recoded to pseudonymised version
pds_general_practice	d_pds_general_practice	General Practice Medical code (PATIENT REGISTRATION)	Recoded to pseudonymised version
Practice_code_patient_registration	D_practice_code_patient_registration	General Practice	Recoded to pseudonymised version
Prime_recipient	D_prime_recipient	CDS Prime Recipient Site	Recoded to pseudonymised version
provider_code	d_provider_code	Organisation site identifier	Recoded to pseudonymised version
receiving_site	d_receiving_site	Organisation site identifier for site receiving the patient	Recoded to pseudonymised version
Residence_ccg	D_residence_ccg	CCG from residence	Recoded to pseudonymised version

Residence_ccg_from_patient_postcode	D_ Residence_ccg_from_patient_postcode	CCG from patient postcode	Recoded to pseudonymised version
Sha_commissioner	D_sha_commissioner	Strategic Health Authority, from Commissioner	Recoded to pseudonymised version
Sha_provider	D_sha_provider	Strategic Health Authority of provider	Recoded to pseudonymised version
site	d_site	Site Identifier	Recoded to pseudonymised version