Institute of Education

UCL

Next Steps

Linked health administrative datasets – Hospital Episode Statistics (HES)

User Guide (Version 1)

September 2020

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Questions and feedback about this user guide should be sent to clsfeedback@ucl.ac.uk.

How to cite this guide

Kerry- Barnard, S., Damiani, E., Gomes, D., Calderwood, L. (2020) *Next Steps: A guide to the linked health administrative datasets – Hospital Episode Statistics (HES).* London: UCL Centre for Longitudinal Studies.

Acknowledgements

CLS would like to thank NHS Digital for the provision of the linked data.

This guide was first published in September 2020 by the UCL Centre for Longitudinal Studies.

UCL Institute of Education

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 Institution of Education (IOE), University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies: tel: +44 (0)20 7612 6875 email: clsfeedback@uck.ac.uk

Contents

About Next Steps

Next Steps is a longitudinal cohort study, following a nationally representative group of nearly 16,000 people born in England in 1989-90. The study began when cohort members were 14 years old. With sweeps every year for the first seven years, it has captured incredibly rich information about their educational trajectories during adolescence and early adulthood.

Today, Next Steps continues to chart this cohort's experiences with a wider disciplinary scope, providing unparalleled insight into the many different aspects of their lives and transitions into adulthood.

A vital source of evidence, Next Steps has had a major influence on national education policy and cast light on a wide range of important social issues, including the effects of zero hours contracts and bullying.

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 Next Steps. 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. Next Steps (previously known as the Longitudinal Study of Young People in England (LSYPE1)) is a large and nationally representative longitudinal cohort study following the lives of around 16,000 people in England born in 1989-1990.

The study started in 2004 and was funded and managed by the Department for Education until it transferred to the UCL Centre for Longitudinal Studies (CLS) in 2013. Annual interviews with study members were carried out between 2004 and 2010 (ages 14 to 20), as well as interviews with parents for the first four sweeps. The study focused primarily on educational transitions and was supplemented by linked administrative data from the National Pupil Database. Following the initial survey at age 13-14, the cohort members were interviewed every year until 2010.

In 2013 CLS successfully re-started the study and collected another sweep of data (sweep 8) at age 25 (in 2015/2016). The scientific content of the study was broadened to provide a multi-disciplinary research resource, supplemented by a wide-range of administrative data linkage consents. Extensive efforts were made to maximise the size and representativeness of the sample and interviews were achieved with almost 8,000 study members.

2. Consent to health data linkage

In sweep 8 data linkage was a major part of the study, with 10 different linkage consents covering education, health, economics and criminal justice. One of these consents was linkage to health records.

Considerable effort was expended in developing an approach that would maximise consent rates, particularly as the survey was conducted using a mixed-mode approach. A data linkage leaflet was included in the advance mailing, sent at the

start of each batch of fieldwork. It gave information on the purpose, types, value and process of data linkage, and encouraged study members to contact the study team with any questions they might have.

During the interview, and following an introduction page, consents were recorded electronically directly into the survey instrument. All participants were then sent a confirmation of their consents (as part of their 'Thank you' letter) by post.

Detailed information on the fieldwork and consent collection can be found in the Next Steps Age 25 Technical report and Next Steps Age 25 User Guide. All documents can be found under 'documentation' at https://cls.ucl.ac.uk/cls-studies/next-steps/next-steps-age-25-sweep/

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. It is comprised of four datasets: Accident and Emergency episodes dataset (A&E), Admitted Patient Care episodes dataset (APC), Adult Critical Care episodes dataset (CC) and Outpatients episodes dataset (OP).

Table 1: List of datasets provided

Name of the dataset	Contents
A&E	Attendance to Accident and Emergency care facility years 2007-2017
APC	Attendance to Admitted Patient Care years 1997-2017
CC	Attendance to Critical Care years 2009-2017
OP	Attendance to Outpatient years 2003-2017

The years of data provided are the earliest years which data was available from NHS Digital. CLS will refresh this data in the future to get post 2017 data.

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 Digital 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[1]

In 2015 CLS made a request to NHS Digital, who are the data holder, to link all consenting Next Steps participants to their HES records. The data linkage was carried out by the NHS Digital team.

3.2 Matching strategy

A 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.

The matching exercise was carried out by the NHS data team in two stages:

a) Matching using the participant's personal information

CLS sent a matching file to NHS Digital containing the following information on the cohort member: 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, and CLS proxy ID. The data was matched on the following basis:

- Name, sex, date of birth and postcode,
- Name, date of birth and sex,

[1] https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

NHS Digital then flagged cohort members in their system and matched their information to their NHS Number (NHSNO). NHS Digital generated a ‘matching reference number’ which was provided to CLS with the CLS proxy ID.

b) Matching using ‘NHS matching reference number’

For this HES data linkage, CLS sent NHS Digital a file containing the ‘NHS matching reference’ previously provided to CLS, a previous CLS proxy ID and a new CLS proxy ID.

NHS Digital extracted the HES data for all cohort members previously flagged under the ‘NHS matching reference’ number provided.

NHS Digital sent the pseudo-anonymised linked HES data with the CLS proxy IDs to CLS.

To maintain the confidentiality of cohort members the linked health records are made available for researchers in a pseudo-anonymised version using the Next Steps identifier variable (NSID) used for the rest of the research available from the UKDS.

3.3 Matching rates

A total of 4895 participants agreed to health linkage out of 7707 who took part in the survey, corresponding to a consent rate of 63.5%.

A total of 4679 Next Steps participants were successfully matched out of 4895 corresponding to a successful linkage rate of 93.5%. This excludes a small number of cohort members who either withdrew their consent following sweep 8 both prior and subsequently to data linkage.

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

Table 2. Consent and overall linkage

Number in Next Steps Age 25 Survey	7707
Number with valid consent	4895
Consent rate	63.5%
Total number with matched HES data	4579
Linkage rate	93.5%

Data was available and matched for a total 4579 participants in the different databases. For each of the datasets, the matching was as shown in Table 3:

Table 3. Matching for each HES database

Database	Number of research participants
A&E	3746
APC	3036
OP	4099
CC	35
Total	4579

4. The research datasets

4.1 Licencing

The linked NHS Digital 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:

- 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, and

- enter into a Licence Agreement with CLS.

For details on how to apply for the data, please refer to section 6 of this document.

4.2 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:

Documentation file	**File name**
User guide	NextSteps_HES_UserGuide_v1.pdf
NHS Data Dictionaries	DD-AE-V11.pdf DD-APC-V11.pdf DD-OP-V11.pdf Hes_data_dictionary_-_adult_critical_care.pdf
CLS Data Dictionaries	NextSteps_HES_Variables_List_v1.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	OPCS48 Metadata File Description V1.0.pdf OPCS48 ToCE Analysis Nov 2016 V1.0.xlsx OPCS48 ToCE Specification 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

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 digital[2] are available in the supplementary documents.These dictionaries will help in interpreting the data. The NHS data dictionaries contains 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 Next Steps 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 Digital documentation mentioned above.

[2] NHS Data Dictionaries, NHS Digital:

A&E: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/3/l/hes_data_dictionary_-_accident_and_emergency.pdf

APC: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/1/c/hes_data_dictionary_-_admitted_patient_care.pdf

CC: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/7/9/hes_data_dictionary_-_adult_critical_care.pdf

OP: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/4/i/hes_data_dictionary_-_outpatients.pdf

HES Analysis Guide

To use the HES data datasets, users are required to be familiar with the HES Analysis Guide provided by NHS Digital [3]. 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[4] and will only made available for approved projects:

- 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

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[5]:

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

---

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

[4] 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

[5] 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.

- 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.

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

4.3 Identifiers

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

4.4 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 asto the international coding such as 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.

Variables referring to health administrative groups been described using the lookups available from the Office of National Statistics[6].

The variables that describe the heath 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 so there are some codes without labels.

---

[6] *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

The administrative variables are: Strategic Health Authority of GP practice (GPPRSTHA), Primary Care Trust if the GP practice (GPPRPCT), 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 data dictionary by NHS Digital provides the codes for the Regional Office, but not all values aligned with the data; where this occurs the label will say “not in dictionary”.

Identification of HES episodes and spells

NHS administratively organises the data by Hospital Spells and Episode 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 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_17. 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 it only has data for 5% of the total 68,043 medical records.

The missing cases have been recorded with the coded '-1' for most variables. A few variables requested by CLS did not contain information of 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 "RX69X" unknown and unspecified causes of morbidity.

4.5 Data de-identification

CLS is committed to protect research participant's 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 Digital[7].These include GP practice codes (gpprac), the providers codes (procode and procode3). The later variable is a truncation of procode which is at a broader level of granularity. Variables including specific GP and health providers were pseudonymised.

[7] 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.pdf, also available in the supplementary documentation.

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

Similarly, some variables had categories grouped into wider categories to avoid the possibility of data disclosure. They include maternity variables that provide length of gestation, age of baby in days or birthweights. In order to comply with the small numbers section of the data analysis guide from NHS Digital, fine grained geographical information is removed.

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

4.6 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 31-03-2017 (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 number of research participants included in this dataset is 3,746 (n=3,746). They have 15,877 available medical records that are included in 98 variables.

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

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.

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 digital 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.

4.7 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 31-03-2017 (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 dataset contains 11,915 medical episodes and has 99 variables. The number of research participants included in this dataset is 3,063 (n=3,063).

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 on 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.

4.8 The Critical Care (CC) data

The CC dataset covers records of critical care activity between 01-04-2009 and 31-03-2017 (inclusive). This is the smallest of the four datasets and only contains the medical records of 35 participants (n=35) and include 54 critical care episodes.

The dataset contains 33 variables that 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 on 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.

4.9 The Outpatient Care (OP) data

The OP dataset lists the outpatient appointments between 01-04-2003 and 31-03-2017 (inclusive).

68,043 outpatient episodes were available for the 4,099 research participants that provided consent for the data linkage. The episodes are recorded in 73 variables.

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_[01-12]) are coded as “RX69X” unknown and unspecified causes of morbidity (n=66,205). The Classification of Interventions and Procedures (variables, opertn_[01-24]) and version of classification have just under one third (29%) 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.

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, variable descriptions 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.

Variables describing administrative groups use look ups which can be found on the Office of National Statistics website (see section 4.2), however not every value in the dataset could be matched to a label. In the APC dataset, these 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 data dictionary by NHS Digital provides the codes for the Regional Office, but not all values aligned with the data; these have been labelled “not in dictionary”.
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.

5. Disclosure control: requirements for data users

5.1. UKDS requirements

As the HES data linked to the longitudinal Next Steps data are only available via the UDS 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.

5.2. NHS Digital requirements

The NHS Digital 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

6. Data access and variable selection

6.1 UKDS Secure Access application

Access to the HES linked data will only be provided via the UKDS Secure Lab, via the researcher's own institutional desktop PC or at the Safe Room at the UK Data Archive. 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.

6.2 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 NextSteps-HES_Variables-List.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.

6.3 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 Digital data'* document. This document will be provided by the UKDS at the point of registration.

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-Digital 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.

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_IMD04	IMD04	IMD Index of Multiple Deprivation	Rounded to nearest 1
D_IMD04C	IMD04C	IMD Crime Domain	Rounded to nearest 0.5
D_IMD04ED	IMD04ED	IMD Education, Skills and Training Domain	Rounded to nearest 1
D_IMD04EM	IMD04EM	IMD Employment Deprivation Domain	Rounded to nearest 0.05
D_IMD04HD	IMD04HD	IMD Health and Disability Domain	Rounded to nearest 0.05
D_IMD04HS	IMD04HS	IMD Barriers to Housing and Services Domain	Rounded to nearest 0.1
D_IMD04I	IMD04I	IMD Income Domain,	Rounded to nearest 1
D_IMD04IA	IMD04IA	IMD Income Affecting Older People Index	Rounded to nearest 0.1
D_IMD04IC	IMD04IC	IMD Income Affecting Children Index	Rounded to nearest 0.5
D_IMD04LE	IMD04LE	IMD Living Environment Domain	Rounded to nearest 1
D_INVEST_[01-09]	INVEST [01-09]	A&E investigation	Truncated to 2 characters
D_PROCODE	PROCODE	Organisation code (code of provider	Recoded to pseudonymised version
D_PROCODE3	PROCODE3	Provider code - 3 character	Recoded: derived from PROCODE recode

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_ANTEDUR	ANTEDUR	Antenatal days of stay	Recoded to: 0 or 1 days, 2 or more days to avoid certainty of DOB of infant
D_DIAG_[1-17]	DIAG_[1-17]	All Diagnosis codes	Truncated to 3 characters
GESTAT_M	GESTAT	Length of gestation	Top coded to 42.
GP_PRAC	GPPRAC	Code of GP practice	Recoded to, pseudonymised code.
D_IMD04	IMD04	IMD Index of Multiple Deprivation	Rounded to nearest 1
D_IMD04C	IMD04C	IMD Crime Domain	Rounded to nearest 0.5
D_IMD04ED	IMD04ED	IMD Education, Skills and Training Domain	Rounded to nearest 1
D_IMD04EM	IMD04EM	IMD Employment Deprivation Domain	Rounded to nearest 0.05
D_IMD04HD	IMD04HD	IMD Health and Disability Domain	Rounded to nearest 0.05
D_IMD04HS	IMD04HS	IMD Barriers to Housing and Services Domain	Rounded to nearest 0.1
D_IMD04I	IMD04I	IMD Income Domain,	Rounded to nearest 1

Variable name	NHS original variable name	Variable description	Modifications
D_IMD04IA	IMD04IA	IMD Income Affecting Older People Index	Rounded to nearest 0.1
D_IMD04IC	IMD04IC	IMD Income Affecting Children Index	Rounded to nearest 0.5
D_IMD04LE	IMD04LE	IMD Living Environment Domain	Rounded to nearest 1
D_NUMPREG	NUMPREG	Number of pregnancies	Top coded to 5 (except 99)
D_POSTOPDUR	POSOPDUR	Post-operative duration in days	Top coded to at 15 and above
D_PROCODE	PROCODE	Organisation code (code of provider)	Recoded to pseudonymised version
D_PROCODE3	PROCODE3	Provider code - 3 character	Recoded: Derive from PROCODE recode
P_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_IMD04	IMD04	IMD Index of Multiple Deprivation	Rounded to nearest 1
D_IMD04C	IMD04C	IMD Crime Domain	Rounded to nearest 0.5
D_IMD04ED	IMD04ED	IMD Education, Skills and Training Domain	Rounded to nearest 1
D_IMD04EM	IMD04EM	IMD Employment Deprivation Domain	Rounded to nearest 0.05
D_IMD04HD	IMD04HD	IMD Health and Disability Domain	Rounded to nearest 0.05
D_IMD04HS	IMD04HS	IMD Barriers to Housing and Services Domain	Rounded to nearest 0.1
D_IMD04I	IMD04I	IMD Income Domain,	Rounded to nearest 1
D_IMD04IA	IMD04IA	IMD Income Affecting Older People Index	Rounded to nearest 0.1
D_IMD04IC	IMD04IC	IMD Income Affecting Children Index	Rounded to nearest 0.5
D_IMD04LE	IMD04LE	IMD Living Environment Domain	Rounded to nearest 1
D_OPERTN_[1-19]	OPERTN_[1-19]	Operative procedure	Truncated to chapter (1 character)
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_WAITDAYS	WAITDAYS	Duration of elective wait	TOPCODE to 365