UK Biobank

Hospital inpatient data

Version 4.0

http://www.ukbiobank.ac.uk/

June 2023

This document provides an explanation of the hospital inpatient data available in UK Biobank.

Contents

1. Understanding Hospital Inpatient Data

This document provides information about the hospital inpatient data available in UK Biobank. Inpatients are patients who are admitted to the hospital and occupy a bed. This includes admissions where an overnight stay is planned and day cases.

Data on hospital inpatient admissions for **England** are provided to UK Biobank by the Data Access Request Service (DARS), managed by NHS England. The dataset is called Hospital Episode Statistics (HES) Admitted Patient Care (APC).

Data on hospital inpatient admissions for **Wales** are provided to UK Biobank by the Secure Anonymised Information Linkage (SAIL) Databank at the University of Swansea, managed by NHS Wales Informatics Service's Information Services Division (ISD). The dataset is called Patient Episode Database for Wales (PEDW) Admitted Patient Care (APC).

Data on hospital inpatient stays for **Scotland** are provided to UK Biobank by Public Health Scotland, part of NHS National Services Scotland (NSS). There are two datasets published: the General Acute Inpatient and Day Case - Scottish Morbidity Record (SMR01), and the Mental Health Inpatient and Day Case - Scottish Morbidity Record (SMR04). The published data does not contain data for maternity hospital admissions. These are provided in a separate file (Maternity Inpatient and Day Case - Scottish Morbidity Record (SMR02) data) which has not yet been made available for researchers to access via the UK Biobank Data Showcase.

Records date back to 1997 for England, 1998 for Wales and 1981 for Scotland and contain coded data on admissions, operations and procedures. UK Biobank receives hospital inpatient data periodically from each provider; for more information, please see data providers and dates of data availability.

Not every participant will have a hospital inpatient record, as not all have been admitted to hospital within the period covered.

All datasets currently available contain data on admissions and discharge, diagnostic and operation codes, with the exception outlined above about maternity admissions for Scotland.

There are differences between England, Wales and Scotland in how they collect and organise their data, and it is important to understand these differences in order to be able to interpret the data correctly. In all three countries admitted episodes of patient care are assigned one main (primary) code as well as a variable number of secondary codes:

- **England**: HES APC data can include up to 20 different diagnostic codes within one episode (1 main clinical code (indicating the reason for the admission) and 19 secondary (or underlying) clinical codes) and up to 24 operation/procedural codes. Episodes are coded at admission and then each time a patient moves between different hospital units. See the NHS England data dictionary for further information.
- **Wales**: PEDW APC data can include up to between 12 and 14 different diagnostic codes within one episode (1 main clinical code and 11-13 secondary codes) and up to 12 operation/procedural codes. Episodes are coded at admission and then each time a patient moves between different hospital units. See the data dictionary for Wales for further information.
- **Scotland**: SMR01 and SMR04 data can include up to 6 different diagnostic codes within one episode (1 main clinical code and 5 secondary codes) and up to 4 operation/procedural codes. Episodes are coded at discharge from hospital. See the data dictionary for Scotland for further information.

The variability in the maximum number of possible codes that can be recorded for an episode in each country may lead to variation in the likelihood of conditions being coded that are not the main (primary) reason for a particular episode of care. Researchers should also be aware that each nation has its own national coding guidelines, coding training manuals and quality standards, which might impact how hospital admissions are coded across the three nations.

2. Clinical classifications

All clinical data in the hospital inpatient data are coded according to the World Health Organization's ICD (International Classification of Diseases and Related Health Problems). All operations and procedures are coded according to the OPCS (Office of Population, Censuses and Surveys: Classification of Interventions and Procedures).

All of the current UK Biobank linked English and Welsh hospital data are coded in ICD-10 and OPCS-4. However, because the collection of Scottish data began earlier (in 1981), the earlier Scottish data (those collected prior to 1997) are coded in ICD-9 and OPCS-3.

Both ICD and OPCS classifications are revised periodically to account for newly emerging conditions and revised definitions. Amendments are made to these coding systems that are specific to the UK. More detail on the different versions can be found on the NHS England Technology Reference data Update Distribution website.

Please note that the clinically modified versions of the ICD used in the United States (ICD-9-CM and ICD-10-CM) are not used in the UK.

3. Data quality

The quality of hospital inpatient data has improved over time, with the implementation (by the data providers) of new cleaning and derivation rules. Most of these cleaning rules check the validity and format of the data-fields. As a result, the quality, coverage and integrity of the data may vary from year to year. All of this can vary by country.

Each data provider has a different approach to validating and checking the data:

- **England**: Information on the quality of HES data can be found here.
- **Wales:** Information on data quality is produced in annual reports and can be found here.
  Scotland: An audit process has been in place since 1990. More information can be found here.

When using the hospital inpatient data, it is worth considering that the following can change over time:

- Changes in the versions of the clinical classifications used to record diagnoses, procedures and interventions (i.e. ICD and OPCS coding; see Section 2);
- Certain data-fields may change, be added or become obsolete over the years;
- Changes to the range of values and/or admissible codes within data-fields may change over time, due to improvements in data cleaning and/or availability of more detailed codes.
- Relative numbers of records within the datasets may change as reporting standards change within the data providers (see below)

NHS England has notified data recipients that efforts to standardise the recording of Same Day Emergency Care (SDEC) activity will mean changes within HES data products, with a possible reduction in HES APC episodes as this activity is increasingly reported in the Emergency Care Data Set (ECDS). This change is not possible to fully quantify but should affect records starting in 2022, with a full standardisation of reporting within ECDS scheduled for July 2024. For more information about the changes and how they may affect interpretation of HES APC data, please see the NHS webpage detailing the possible impact of the recording changes.

4. Data cleaning performed by UK Biobank

Occasionally, invalid codes still occur within the dataset and we perform some verification and cleaning rules for each data-field. These rules check whether the ICD and OPCS codes correspond to the correct version of the coding system, and that values are within the given coding range and are in a standard format.

Where ICD or OPCS codes contain trailing characters (such as dashes and X's) or other additional characters that are not part of a valid code, UK Biobank applies cleaning rules to strip the trailing characters. Asterisks and daggers, which have a specific meaning in ICD-10, are made available in the addendum fields, however a portion of Scottish data is missing data in the addendum fields. We are looking into addressing this in due course.

It was observed in Welsh-source data that diagnosis codes appeared in some cases to be ICD-O-3 morphology codes rather than ICD-10 codes, or it was ambiguous whether the code was an ICD-10 code or the histology component of an ICD-O-3 morphology code. This resulted in an outsize number of (valid or invalid) ICD-10 codes beginning with "M." Consequently, any Welsh ICD-10 diagnosis code was removed that consisted of either a) an "M" followed by four digits also appearing as a valid ICD-O-3 histology code, or b) an "M" followed by five digits. This was to avoid the possibility of ICD-O-3 morphology codes being misinterpreted as ICD-10 diagnosis codes, although in practice we cannot guarantee that some valid Welsh-source ICD-10 diagnoses have not also been removed.

5. Data available in UK Biobank Showcase

As stated in Section 1, currently the inpatient data from England and Wales includes maternity and psychiatric episodes of care; the data from Scotland includes only general and psychiatric episodes.

Hospital inpatient data can be accessed in a variety of different formats. The most suitable format to use for any particular research project will depend on the nature of the research being undertaken (some projects may wish to use several of the formats available).

'Record-level' hospital inpatient data is provided in the form of six interrelated data tables, accessed separately from the main UK Biobank dataset. See Section 6 for further information of the structure of the record-level data, and Section 7 for details of how the record-level tables are accessed.

Due to the complexity and size of the record-level data, we have created summary fields that provide the first date of any given diagnostic or operation code, which may be sufficient for many researchers' needs. See Section 8 for further details on summary fields related to diagnoses and operations, and Section 9 for details on other summary fields (for example those related to administrative data).

Researchers interested in identifying participants with particular health conditions may find the "Algorithmically-defined outcomes" fields helpful. These combine carefully selected codes from self-report, hospital inpatient and death data together, and provide information on the date when the disease outcome of interest is thought most likely to have first occurred or been recorded, as well as information about the sources of data indicating the occurrence of the condition, for a number of selected conditions. These algorithmically defined outcomes may be expanded in the future to include primary care data and incorporate additional conditions. See Category 42 on Showcase for further information.

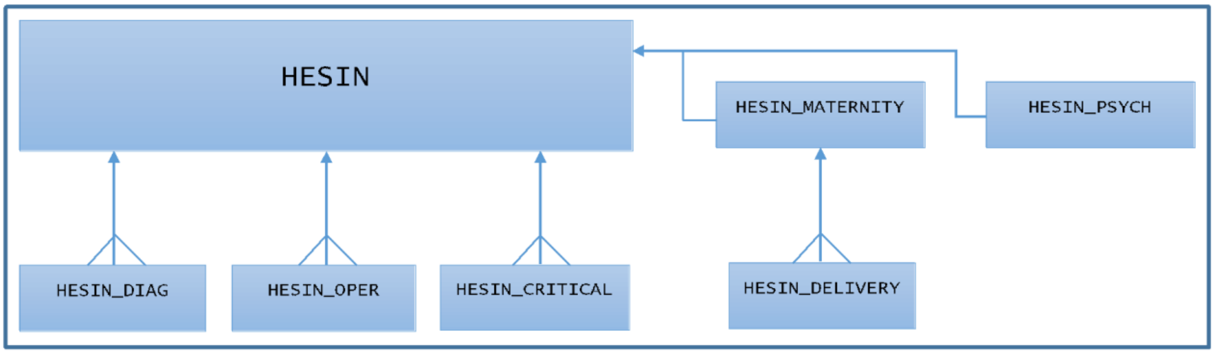
We have also made available a list of data-fields called "First occurrences" that map the clinical codes from self-report, hospital inpatient admissions, primary care (for participants where this is available) and death data to 3-character ICD-10 codes, using NHS ontology mapping tools to map to sets of Read and UK Biobank codes in the primary care and self-report data respectively. We also provide, for each participant, the date that code first occurred in any source. See Category 1712 in the Data Showcase and the accompanying documentation for further information.

6. Record-level hospital inpatient admissions data

Within the hospital inpatient dataset, each inpatient episode for a participant is stored as a single record, i.e. a row of data, or as multiple contiguous records. This is unlike the format of the UK Biobank main dataset, which provides a single row of data per participant.

In the inpatient data from England and Wales, a single admission is called a hospital inpatient "spell", and within each spell there can be one or more "episodes". See Appendix A for a more detailed explanation of these terms. The inpatient data from Scotland also contains "episodes" but they are not grouped into spells in this way.

Hospital inpatient data contains a large number of data-fields, some of which (such as maternity and administrative psychiatry fields), will only be of interest to a minority of researchers. Hence, we have provided the record-level data in the form of seven interrelated data tables, as shown below:



These tables contain the following data:

- **HESIN** – This is the overall master table, providing information on inpatient episodes of care for England, Wales and Scotland (but currently excluding maternity inpatient episodes for Scotland), including details on admissions and discharge, the type of episode. This also includes, where applicable, how an episode fits into a hospital spell.

- **HESIN_DIAG** – Diagnosis codes (ICD-9 or ICD-10) relating to the inpatient episode of care for England, Wales and Scotland (but currently excluding maternity episodes for Scotland).

- **HESIN_OPER** – Operations and procedural codes (OPCS-3 or OPCS-4) relating to the inpatient episode of care (but currently excluding maternity episodes for Scotland).

- **HESIN_CRITICAL** – Contains further information about those hospital episodes that required treatment in a critical care unit. For example, it gives the number of days of (basic & advanced) cardiac and respiratory support received by a patient.

- **HESIN_PSYCH** – A sibling table to HESIN containing information on administrative aspects of inpatient episodes relating to psychiatry, such as the history of psychiatric care and the legal status of the admission. These data are available for available for England and (in a subset of fields) Scotland only. Psychiatric diagnoses and procedures are contained in the HESIN_DIAG and HESIN_OPER tables.

- **HESIN_MATERNITY** – A sibling table to HESIN containing data-fields (currently for England and Wales only) specifically relating to maternity inpatient episodes of care, such as antenatal information and whether anaesthetic was administered, etc.

Diagnoses and procedures relating to maternity episodes are contained in the HESIN_DIAG and HESIN_OPER tables.

- **HESIN_DELIVERY** – A child table to HESIN_MATERNITY relating to birth episodes. Contains information such as birth weight, delivery method and place, and the status of the baby.

Note that only the HESIN and HESIN_CRITICAL tables contain information relating to the inpatient episode dates. Hence, for example, if you are interested in the dates of diagnosis for some particular condition, you will need access to the HESIN_DIAG table for the diagnosis codes, and the main HESIN table to use the episode start date or admission date as a proxy for the date of diagnosis.

Detailed information about the above tables and the data-fields they each contain can be found in the HES data dictionary (Resource 141140) in Category 2000.

7. Accessing record-level hospital inpatient admission data

Due to the format and complexity of the record-level data, access is not provided as part of a main UK Biobank dataset. Instead, you can access an approved record table via the "Data Portal" via your project's "Downloads" page. From here, you can either download some or all of the record-level data or analyse it in-situ via SQL.

Participants in the record-level data are labelled with the same identifiers used in your main dataset allowing the two datasets to be combined.

Access to each table is granted by adding the appropriate fields from Category 2006 to your basket, as detailed in the inpatient data dictionary.

For example: adding Showcase Field 41234 to a project basket grants access to all fields in the HESIN_DIAG table.

(Note that a field such as 41234 above will still appear in your main dataset, but it simply provides a count of the number of rows that exist for that participant in that table.)

Research projects performing detailed analyses of the hospital inpatient data will need record-level access. However, the record-level data is complex, amalgamated from multiple sources containing different data-fields and different data codings, and may contain numerous 'glitches' that will not have been dealt with by our current 'light-touch' data cleaning. Researchers using the record-level data should be aware of these issues.

Access to the Data Portal is described in the Data Access Guide on the Accessing UK Biobank data page on Showcase. Tables on the Data Portal can either be downloaded in full, or queries can be run through the Portal to download specific subsets of data. Some examples of SQL queries on the inpatient tables are given in Appendix B.

Due to the complex nature of the record-level data, we have provided summary fields that contain the first date for each diagnostic and operation code per participant, which are intended to obviate the need for many researchers to access the more complex record-level data. The summary inpatient fields are described in the next two sections.

8. Summary diagnosis and operation/procedural data-fields

As outlined in Section 5, the summary diagnosis and operation/procedural data-fields provide the first date that a particular ICD/OPCS code appears in a participant's inpatient records. These are accessed as part of a project's main dataset rather than the Data Portal.

For diagnoses, there are separate fields for each of ICD-9 and ICD-10 and for both the main diagnosis position and for either main (primary) or secondary position. There are similar fields for operations/procedures again for both OPCS-3 and OPCS-4, both for the main procedure position or any position in the data. Summary fields for secondary diagnoses or external causes do not include the accompanying date information.

Further details about the summary diagnosis and operation/procedure fields can be found in the Hospital Inpatient Data Dictionary (Resource 141140) in Category 2000.

9. Other summary fields

Categories 2001, 2003, and 2004 provide summary data-fields relating to administration (including admission and discharge details and treatment specialties), maternity care, and psychiatric admissions respectively.

Each summary data-field provides an overview of the amount and type of data stored across all inpatient records for that data-field for each participant. These summary fields are not, of themselves, particularly useful for research purposes, but are provided to illustrate the quantity and type of data available in the corresponding record-level tables via the HES portal.

For example: summary Field 41227 (Status of baby at birth) provides for each participant a list of each distinct value that appears in the birstat field in the HESIN_DELIVERY table in that participant's inpatient records.

Appendix A: Spells & Episodes

Within the inpatient data from England and Wales, a (hospital) “spell” is a total continuous stay of a patient in a single hospital from admission to discharge.

A spell is split into one or more (consultant) “episodes”, each of which is a continuous period of admitted patient care administered under one consultant within that one hospital provider. If a patient is transferred to another consultant during a spell, a new episode is generated.

For the inpatient data from England and Wales, each row in the HESIN table corresponds to a single episode, and there is a data-field “epiorder” which numbers (starting from 1) the position of an episode within a spell.

The data from England also has data-fields “spelbgin”, indicating whether an episode is the first in a spell, and “spelend” indicating whether an episode is the last in a spell. The data from Wales does not contain these data-fields. The values in these data-fields are not always consistent however, for example an episode that is not the end of a spell is sometimes labelled as if it is.

Where possible, we have attempted to group episodes into spells, with episodes in a single spell assigned a common “spell_index”, and a “spell_seq” data-field, giving the order of the episodes within that spell. However, the issues with the data-fields mentioned above means that the “spell_index” data-field cannot always be relied upon to fully group episodes into spells. Researchers particularly interested in hospital spells should treat these fields with caution, and we recommend you develop your own rules for assigning episodes to spells.

The inpatient data from Scotland includes episodes of care but is not arranged into spells in this way.

Appendix B: Sample SQL queries

The following gives some simple examples of how hospital inpatient data can be investigated using SQL directly in the Data Portal. These examples should be read in conjunction with the inpatient data dictionary (see the resources tab in Category 2000 on Showcase). Note that SQL generally ignores whitespace (including line breaks) so the spacing in the examples can be altered without having any effect. Also, it is not necessary for SQL statements such as SELECT to be written in uppercase. This is done simply for clarity, and a lower-case "select" will work just the same.

Example 1: To fetch all the fields from table **hesin**, enter:

```
SELECT * FROM hesin
```

To select just the first 100 records we would use:

```
SELECT FIRST 100 * FROM hesin
```

To select the next 100 rows (i.e. rows 101 to 200) we can use:

```
SELECT FIRST 100 * FROM hesin
OFFSET 100
```

Example 2: To select just a subset of fields from **hesin** and join these to the primary ICD-10 diagnosis from the **hesin_diag** table:

```
SELECT  hesin.eid,
        hesin.ins_index,
        dsource,
        epistart,
        epiend,
        admidate,
        disdate,
        diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1
```

The `level=1` picks only the primary/main diagnosis, and means that only one row in **hesin_diag** will be joined to each **hesin** row. If this condition were omitted the query would return multiple rows for each hesin row, one for each associated diagnosis code.
The eid and ins_index fields need to be prefixed by the table name because they appear in both tables in the join.

Example 3: To return all the OPCS4 operation codes, primary and secondary, and episode start dates for records starting from 1st July 2010, and link them to the appropriate participants (via the **hesin** table):

```
SELECT  hesin.eid,
        epistart,
        level,
        oper4
FROM hesin JOIN hesin_oper USING(eid, ins_index)
WHERE epistart >= '2010-07-01'
```

Example 4: Return a subset of fields from records where the ICD-10 code I21.1 (Acute transmural myocardial infarction of inferior wall) appears as the primary diagnosis:

```
SELECT hesin.eid,
       hesin.ins_index,
       dsource,
       epistart,
       epiend,
       admidate,
       disdate,
       diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1 and diag_icd10='I211'
```

Note that the decimal point in the code I21.1 is not present in the data, i.e. it appears as I211.

Example 5: Continuing from Example 4, we can search more generally for all codes starting I21 (i.e I21.0 to I21.4, and I21.9) by amending the above slightly to:

```
SELECT hesin.eid,
       hesin.ins_index,
       dsource,
       epistart,
       epiend,
       admidate,
       disdate,
       diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1 and diag_icd10 LIKE 'I21%'
```

The “like” searches for a pattern in the field, and the % functions as a “wildcard” matching any sequence of characters.

If we were interested in secondary diagnoses as well, but not in “external causes”, we could replace this with:

```
SELECT hesin.eid,
       hesin.ins_index,
       dsource,
       epistart,
       epiend,
       admidate,
       disdate,
       diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level in (1,2) and diag_icd10 LIKE 'I21%'
```

i.e allow level to be either 1 or 2 (primary or secondary).

Example 6: Following on from example 5, we could instead count the number of distinct participants having each of these codes in the inpatient data as follows:

```
SELECT diag_icd10, COUNT(DISTINCT hesin.eid) as number_of_pts
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1 and diag_icd10 LIKE 'I21%'
GROUP BY diag_icd10
```

Which will provide a short table giving the number of distinct participants for which each of the codes I210 – I219 appears as a primary diagnosis in their data. Note that if a participant had both (for example) I214 and I219 in their data they would be counted for both.