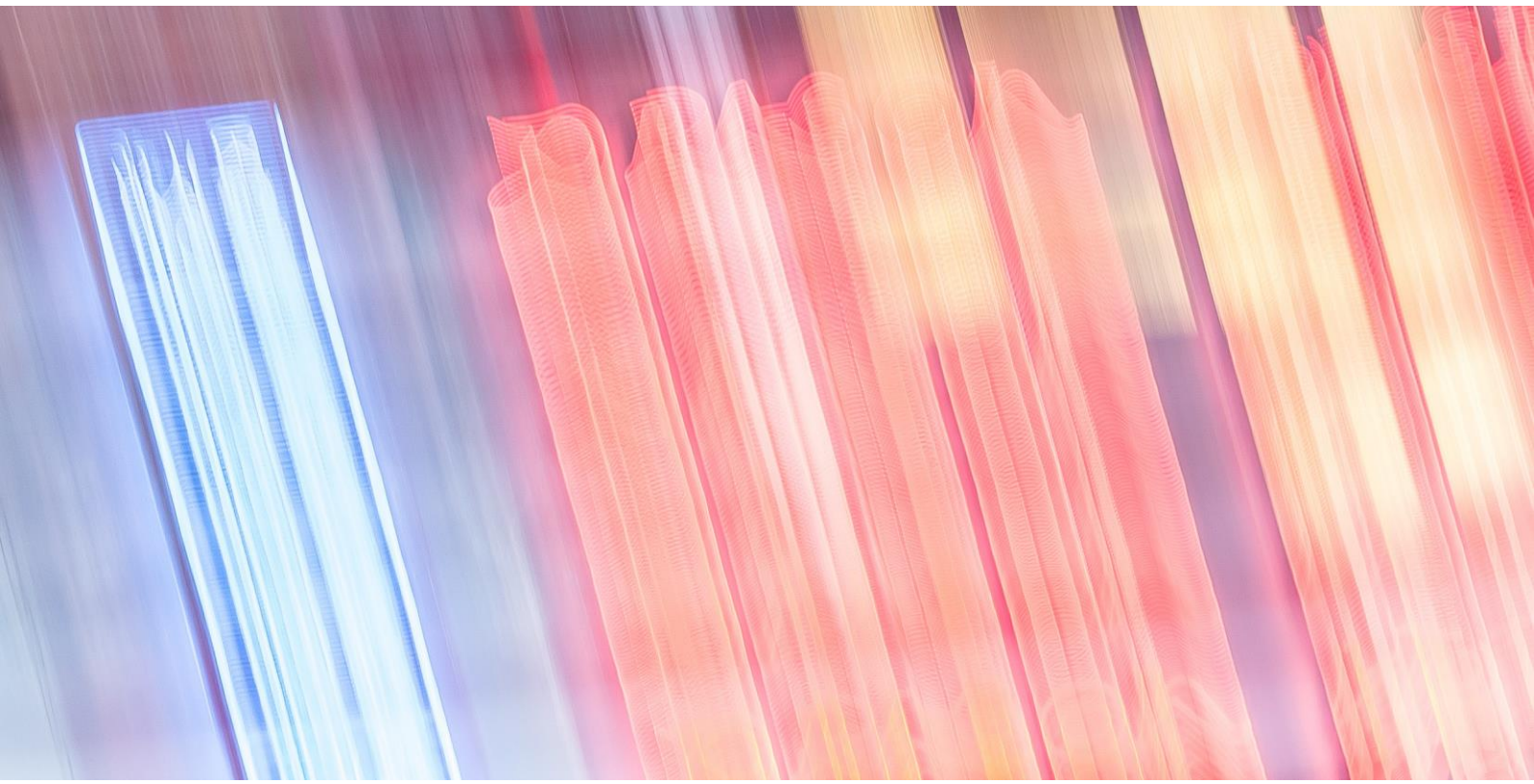
Announcement of methodological change

Impact of changes to Hospital Episode Statistics (HES) processing from April 2021

Version 3.0 - August 2022



Information and technology
for better health and care

Contents

Summary 4

What has changed? 4
Version Control 4
Feedback 5

Background 6

Improving our Data Processing Services (DPS) 6
Person identifier using the Master Person Service 6
Replacement of HESID with Person_ID 6
Update of cleaning rules and derivations 7
How we use our reference data 7

Description of changes 8

Faster data processing 8
Person identifier using the Master Person Service 8
Update of cleaning rules and derivations 10
How we use our reference data 10

Impact 11

Person identifier using the Master Person Service 11
Patient counts – by year 13
Patient counts – over multiple years 17
Consistency of HESID and Person_ID – by year 18
Consistency of HESID and Person_ID – over multiple years 20
Permutations of HESID and Person_ID 21
1 Person_ID linked to multiple HESIDs – by year 22
1 Person_ID linked to multiple HESIDs – over multiple years 23
Person_IDs linked to multiple HESIDs – contribution of algorithm differences 24
One HESID linked to multiple Person_IDs – by year 28
One HESID linked to multiple Person_IDs – over multiple years 29
1 HESID linked to 2 Person_IDs – record level 30
Unmatched patients 32
Assigning Person_ID to APC episodes at spell level 34
Reference data 36
Example – two organisations merging 36
Example – organisation transferring services to different organisations 37
Example – postcode allocation to geography changed during period 38
Example – new postcode opened during the reporting period 38

Appendices **39**

Appendix A - HESID methodology 39

Summary

This section briefly outlines key points about the changes that will be made to the Hospital Episode Statistics (HES) data set and the effects on uses of the HES data and associated statistical publication series.

What has changed?

NHS Digital are improving how we collect, process and access health and care data using new technologies and processes to do so in a smarter, more efficient way. This will enable faster access to better linked data giving commissioners and researchers a clearer picture of health and care.

HES data comes from the routine exchanges of information between providers and commissioners of healthcare for NHS patients in England. Healthcare providers collect administrative and clinical information locally to support the care of the patient. The data is submitted to the Secondary Uses Service (SUS), which, as well as making it available to the commissioners, also copies the information to a database. At regular intervals, an extract is taken from this database, cleaned and enhanced with additional derived data items. The resulting data is made available as the HES data set.

HES is an important national data asset used for Official Statistics, parliamentary questions, commissioning analysis and research projects amongst other things.

As part of this overarching program of work we are improving how we process the HES data set.

At the time of original publication, the changes described in this paper were planned to be effective from the processing of provisional July 2021 year-to-date data, released in September 2021. Any updates to operational information was notified on the ‘HES data changes in 2021’ webpage.

We are changing how we identify people within national data sets which use this new technology, including the HES data set. Using one method for doing this across data sets helps us increase the amount of usable, better quality, linkable data available to support research and planning.

As part of our work in adopting new technologies and processes with our HES data we are also updating the nature and extent of the cleaning performed and derivations added to the data, taking the opportunity to align methodologies with other data assets and sources where it is beneficial to do so.

This includes changes in how we handle reference data in HES, such as validating submitted organisation codes, to align to the way this is done for other data sets (such as SUS+ and the Mental Health Services Data Set) that are collected by NHS Digital.

Version Control

This document and associated data tables are subject to change as we are continuing to investigate the potential impacts that making these changes will have on the HES data and associated statistics.

The current version of this document is 2.0 and replaces version 1.0 published in December 2020.

The current version of document is 3.0 and replaces version 2.0 published in June 2021.

This document refers to the following versions of data and processes:

- HES Admitted Patient Care (APC) data – 1997-98 to 2019-20
- HES Outpatient (OP) data – 2003-04 to 2019-20
- HES Accident & Emergency (A&E) data – 2007-08 to 2019-20
- MPS algorithm as at July 2020.
- HES-specific MPS data preparation rules as at June 2021

Updates to content from v1.0 to v2.0 of this document and accompanying tables include:

- Person identifier changes:
    - Update of all data following implementation of further data preparation rules
    - Addition of data from 1997-98 to 2000-01 to some APC tables that previously began at 2001-02.
    - Addition of 2019-20 data for APC, OP and A&E
    - Extension of analysis to consider counts of records (episodes, appointments or attendances) as well as counts of patients
    - Analysis over most recent 10-year period as a common use case (2010-11 to 2019-20)
    - Investigation of some potential explanations for the difference in outputs between old and new methods.
    - Removal of section on patient counts subdivided by key variables
    - Removal of Person_ID methodology Appendix pending release of separate Person_ID handbook
- Reference data changes:
    - More detail including examples of impact on provider code validation and mapping, and geography-related derivations

Feedback

We would appreciate any feedback on the methodological change paper and associated data tables, particularly with respect to how these changes will affect users in any ways that have not been communicated as part of this document.

Feedback can be submitted via enquiries@nhsdigital.nhs.uk with the subject "**HES Methodological Changes 2021**".

Background

This section provides information on the work we are doing to transform how we process health and social care data and what we make available in the Hospital Episode Statistics data and publication series.

Improving our Data Processing Services (DPS)

We are implementing modern technologies and processes to enable us to perform our statutory role as the safe haven for health and care information. This requires us to collect and process data needed to run the health service.

These secure technologies and processes will enable us to collect, process and access data in a smarter, more efficient way. This will lead to faster access to better linked data giving commissioners and researchers a clearer picture of health and care.

Most importantly, our data processing services will improve patient care by empowering the health and care system to use information more effectively for research into the prevention and treatment of diseases and planning services essential to the sustainability of the NHS.

Person identifier using the Master Person Service

NHS Digital have developed a new standard person identifier using an algorithm that:

- utilises insight from our Master Person Service (MPS) that confirms the individual's identity using Patient Demographics Service (PDS) data on Spine
- assigns a common ID across all NHS Digital data sets, known as the Person_ID.

The Person_ID would be pseudonymised uniquely for each customer, and consistently across the data sets they have access to.

This enables direct linkage, on a patient basis, of records across data sets without the need to access patient identifiers or utilise different data set specific algorithmic matching methods.

Replacement of HESID with Person_ID

The data in HES consists of information about individual consultant episodes, outpatient attendances and A&E attendances, with no links between them. However, several such activity records may be related to a single patient.

To address this NHS Digital has historically created a derived field known as the HES Patient ID (HESID) to provide a way of tracking patients through the HES database. This was designed to be resilient to data quality issues and remove the need to access personal identifiable data to link records relating to the same patient together.

HESID has allowed users of HES data to:

- Measure counts of patients rather than attendances, appointments, episodes.
- Link together activity in the data relating to the same patient such as continuous inpatient spells.
- Identify and track specific patient cohorts across time.

The current version has been in place since March 2009.

Provisional HES data is processed and released as year-to-date data on a monthly basis. **When the provisional April to July 2021 data is processed and released in September 2021, the patient identifier in HES will change from the HESID to the Person_ID.**

All prior years of data will be re-processed to assign a Person_ID, and from this point only the Person_ID will be assigned.

Users of HES will be able to use Person_ID for all existing uses of HESID and will also be able to use it to easily link patient activity in HES to activity for the same person in other NHS Digital data sets that use this person identifier, such as the Mental Health Services Data Set (MHSDS) or Community Services Data Set (CSDS).

Update of cleaning rules and derivations

We are taking the opportunity to align methodologies with other data assets and sources where it is beneficial to do so. This primarily means aligning with current SUS outputs provided to commissioners for operational management processes.

This includes retirement of some legacy derivations, updates to a small number of derivations and the creation of a smaller number of new derivations.

These updates to cleaning rules and derivations will apply from the April to July 2021 provisional data release in September 2021 onwards. Prior years will not be reprocessed.

How we use our reference data

HES is a cleaned, standardised version of the data submitted to SUS.

In both the cleaning and the standardising of the data a range of different reference data sources are manually updated and utilised throughout the year. As part of our changes, we are moving towards aligning with most current methods and approaches utilised by other national data sets such as MHSDS & CSDS.

The changes described in this paper will apply from the April to July 2021 provisional data release in September 2021 onwards. Prior years will not be reprocessed.

We will continue to work towards utilising centrally managed corporate reference data which can be updated more frequently and will reduce dependency on manual processes. Future steps in this development are unlikely to affect users or outputs significantly, but if there are impacts we will communicate these at the relevant time.

Description of changes

This section describes the changes that will be made to the processing of HES data, which could change how you utilise the HES data and interpret statistics in the HES publication series.

Faster data processing

We are using modern data processing services to streamline and automate existing ways of collecting, processing, and accessing data. This will give us the power to process larger volumes of data, faster than ever before, whilst providing the tools to manage incoming data to ensure it is accurate, useful, and secure.

Person identifier using the Master Person Service

The patient identifier in HES will change from the HESID to the Master Person Service (MPS) Person Identifier (Person_ID). This change will take effect from the April to July 2021 year to date provisional HES data release in September 2021 and will be applied to all previous years of HES data.

This means that any analysis produced by NHS Digital using the patient identifier to count patients after that date will use Person_ID for all time periods.

Similarly, any new extracts provided under DARS agreements from September 2021 will only include Person_ID for all time periods. Further information on the provision of Person_ID for DARS customers is available here.

Using the Person_ID helps us increase the amount of usable, better quality data available to support research and planning.

Patients can visit multiple places where they register to receive care or treatment. They move, get married (or divorced) and go on holiday. At any given time, we store the health and care records of individuals as recorded in various systems around the country. The aim is to link together as much activity for the same patient as possible while avoiding false positives matches.

The Person_ID is derived by first assessing the data quality of submitted patient identifiers on each record, before passing the details to the MPS. MPS attempts to confirm the identity of each patient by matching each record to data held in the Personal Demographics Service (PDS), the national electronic database of NHS patient details, and finding a single best result.

MPS uses a four-stage algorithm to try and find a single best match to PDS.

For HES data, the data items used to match to PDS are NHS number, date of birth, gender (SEX), and postcode (HOMEADD).

Some data cleaning of HES data fields (such as consistent spacing of postcode values) is undertaken prior to submitting records to the matching process to ensure the best possible chance of finding a match while avoiding false matches. Documentation for these cleaning rules is available in the Person_ID handbook for HES users.

Where a single best result is not found, a further attempt is made using other logic to link together records relating to the same patient (using the cleaned values) and assign them the same Person_ID.

If there were missing or incorrect relevant data items on a record, it may not be possible to verify the patient's identity in PDS or link to any other records using the further process, so this Person_ID may be assigned to one record only and would never be assigned to another record. Such Person_IDs are described as 'unmatched'.

Further detail regarding the use of the four-stage algorithm and the further matching process for HES data may be found within the upcoming release of the Person_ID handbook for HES users.

To assign a Person_ID to all historic HES records, the finalised annual refresh HES data for prior years was processed a year at a time (initially working backwards from 2018-19, then processing 2019-20 when it became available).

Differences between the methods for deriving HESID and Person_ID mean that records that were linked together under a single HESID may now have different Person_IDs, or records that had different HESIDs may now be linked together under a single Person_ID.

The Impact section of this document includes analysis of how the change in patient identifier may affect counts of patients in HES data and their related records.

The explanation of the Person_ID process for historic and future data is more complex for APC data and includes an additional method step from 2021-22 data.

An admission to hospital can be a single APC episode, or a series of APC episodes which are referred to as a 'spell'. HESID processing assigned a HESID to each episode based on the submitted values for that episode, and similarly Person_ID has been assigned to historic APC records at an episode level for APC data (and will be for 2020-21).

From the April to July 2021 provisional data released in September 2021 onwards, in the explanation above the 'record' for HES APC data will be the first episode in the spell. The Person_ID will therefore be assigned to all episodes in a spell based on the submitted values on the first episode in the spell. This aligns the approach to generating person identifiers in HES with the approach in SUS+. The identification of the episodes comprising a spell, and the identification of the first episode also use the SUS+ methodology. (It would not have been possible to do this for all historic data as the current method for defining spells has only been in place since 2017-18.)

In particular, this means that unmatched Person_IDs will be unique in HES data up to 2020-21 (as each was generated for a single episode) but may be repeated in HES data from 2021-22 onwards (as an unmatched Person_ID generated for the first episode in the spell would be assigned to all episodes in the spell).

The Impact section of this document includes estimates of how many episodes in 2018-19 and 2019-20 may have been assigned a different Person_ID if the assignment had been at spell level.

The HESID is used in several other processes in HES, such as linkage between the A&E and APC data sets and the generation of continuous inpatient spells. Person_ID will also replace HESID in these processes.

Update of cleaning rules and derivations

Several legacy derivations will be retired, a small number of pre-existing derivations will be updated, and a smaller number of new derivations will be created. Details of these changes can be found within the HES Data Dictionary documentation available from Hospital Episode Statistics Data Dictionary - NHS Digital.

How we use our reference data

We are aligning our handling and updating of reference data to other data sets.

There will be two areas of change:

- The criteria for accepting values in some fields as valid
- The frequency with which derivations will be rerun with latest reference data

For example, the provider code field (PROCODE) in HES is checked by comparing to a list of valid codes.

In the current process, PROCODE is considered valid if:

- it is open at the end of the relevant data period
- it corresponds to an organisation that provides NHS-funded care

If the raw PROCODE value for a record is not considered valid, HES attempts to correct or 'map' an updated PROCODE value. If this is not successful, the record will be deleted.

Following the change, a PROCODE value will be considered valid if:

- it is open at the activity date (attendance, episode end date or appointment date)
- it corresponds to an organisation that provides NHS-funded care

If a PROCODE value is not considered valid, the same attempts to correct the value will be made.

This will affect the provider code values and related derivations that users see in the data for providers that open, merge or close during the year. Correctly submitted codes prior a merger will not be treated as invalid, and the mapping process will be triggered less often. The data will reflect more clearly the organisation providing care at the time of activity. Examples are provided in the Impact section.

The other main example is the use of postcode reference data in deriving geography codes.

In the current process, fields such as CCG of residence would be derived based on the postcode being a valid postcode at the period end and using the record from the postcode reference table that was applicable at the period end.

Following the change, derivations will use the record from the postcode reference table that is applicable at the activity date.

If the first 'open' date for a submitted postcode is after the activity date, then the postcode will be treated as invalid, and geography values will be 'Unknown'. This differs from current processing where geography values would be derived as long as the postcode was open at the period end.

In the current process for producing provisional year-to-date HES, all derivations are run each month with the current reference data.

Following the change, only a subset of derivations will be rerun for all year-to-date records each month. This subset has been selected as the derivations likely to change when reference data is updated, and includes provider code, related derivations and postcode-based derivations. Details of which derivations will be rerun each month will be made available shortly.

Impact

This section summarises the impact of these changes on how you can interpret the results of analysis using HES data and the publication series.

We quantify some impacts of the change from HES ID to Person_ID, considering the standard HES data sets. The impact on more complex HES processing such as the linkage between A&E and APC will be assessed and explained in separate documents where applicable to be published at a later date.

As the reference data processing changes are prospective and the impact will depend on particular circumstances from year to year we describe rather than quantify how the data will differ following these changes.

Person identifier using the Master Person Service

The switch from HESID to Person_ID does not change any of the submitted underlying data in HES. It only affects the patient identifier or derived fields that utilise the patient identifier.

If you query HES to generate counts of events such as admissions or procedures (for example number of hip replacements by hospital site) you will be unaffected by this change. If there are 15 million episodes within a current year's APC data set, then there will still be 15 million episodes following the change. This will be the same for the number of admissions and discharges.

However, if you use HES data to

- count patients rather than activity
- follow specific patients over time
- analyse linked A&E and APC records
- analyse continuous inpatient spells.

you may get different results using Person_ID.

Some examples of how the patient identifier for each record could be affected are:

Example 1 – records that had the same HESID also have the same Person_ID

Record number	HESID	Person_ID
1	123	ABC
2	123	ABC

Example 2 – records that had different HESIDs now have a single Person_ID

Record number	HESID	Person_ID
1	123	ABC
2	123	ABC
3	456	ABC
4	789	ABC

Example 3 – records that the same HESID now have different Person_IDs

Record number	HESID	Person_ID
1	123	ABC
2	123	ABC
3	123	DEF

Example 4 – more complex interactions

Record number	HESID	Person_ID
1	123	ABC
2	123	DEF
3	456	ABC
4	789	DEF

In this example, the records that were considered to relate to a single patient with HESID 123 now relate to two different Person_IDs (ABC and DEF), and activity formerly relating to two different patients according to HESID (123 and 789) now relates to the same patient (Person_ID DEF).

In this section we provide analysis of:

- headline differences in patient counts over various time intervals
- consistency between HESID and Person_ID
- one-to-many and many-to-one relationships between some HESIDs and Person_IDs
- some ways in which these relationships are explained by methodology differences, such as the use of gender in matching records.

There are some records with incomplete or invalid key data items that cannot be linked to a patient under the HESID process, the Person_ID process, or both. A new HESID or Person_ID would be generated for each such record. We also include analysis of the number of such ‘unmatched’ HESIDs and Person_IDs.

Patient counts – by year

Table 1a and Charts 1a to 1c show the difference between the total count of HESIDs and the total count of Person_IDs for each data set and year.

Please note this is a net position – there could be multiple Person_IDs within the set of records associated with 1 HESID, and one or more of those Person_IDs could also appear on records associated with other HESIDs.

For most years, there is less than 1% difference in the total count per year and per data set of patients by HESID and by Person_ID. The count of Person_IDs is usually lower than the count of HESIDs, suggesting the Person_ID algorithm is on average matching more records per patient than the HESID algorithm.

Table 1a: Count of unique patients by HESID and Person_ID with percentage change, by HES data set and year

	APC			Outpatients			A&E			All HES		
Year	HESID	Person_ID	Change	HESID	Person_ID	Change	HESID	Person_ID	Change	HESID	Person_ID	Change
1997-98	7,098,467	7,216,972	1.67%							7,098,467	7,216,972	1.67%
1998-99	7,230,413	7,338,458	1.49%							7,230,413	7,338,458	1.49%
1999-00	7,227,455	7,294,413	0.93%							7,227,455	7,294,413	0.93%
2000-01	7,212,972	7,253,228	0.56%							7,212,972	7,253,228	0.56%
2001-02	7,140,716	7,168,242	0.39%							7,140,716	7,168,242	0.39%
2002-03	7,274,255	7,294,661	0.28%							7,274,255	7,294,661	0.28%
2003-04	7,456,986	7,469,279	0.16%	15,586,744	15,609,878	0.15%				17,802,539	17,845,025	0.24%
2004-05	7,506,716	7,505,948	-0.01%	16,209,084	16,204,785	-0.03%				18,281,807	18,276,268	-0.03%
2005-06	7,764,257	7,764,354	0.00%	17,073,118	17,033,899	-0.23%				19,035,185	18,994,966	-0.21%
2006-07	7,848,061	7,850,551	0.03%	17,193,104	17,149,394	-0.25%				19,107,346	19,064,576	-0.22%
2007-08	8,136,911	8,123,604	-0.16%	17,398,261	17,321,948	-0.44%	8,486,309	8,463,273	-0.27%	23,071,141	22,875,497	-0.85%
2008-09	8,473,905	8,474,964	0.01%	18,161,882	18,082,032	-0.44%	9,349,040	9,339,367	-0.10%	23,919,644	23,767,611	-0.64%
2009-10	8,643,228	8,643,942	0.01%	18,975,771	18,932,763	-0.23%	10,451,519	10,425,994	-0.24%	25,014,248	24,863,800	-0.60%
2010-11	8,793,316	8,794,300	0.01%	19,393,375	19,267,126	-0.65%	10,762,169	10,743,179	-0.18%	25,395,260	25,219,550	-0.69%
2011-12	8,835,863	8,836,713	0.01%	19,612,172	19,542,330	-0.36%	11,489,265	11,478,801	-0.09%	25,994,166	25,901,704	-0.36%
2012-13	8,881,350	8,882,520	0.01%	20,287,852	19,886,145	-1.98%	11,918,667	11,876,673	-0.35%	26,885,543	26,376,995	-1.89%
2013-14	8,998,434	8,987,779	-0.12%	21,850,777	21,780,798	-0.32%	11,936,228	11,926,931	-0.08%	28,253,070	28,128,459	-0.44%
2014-15	9,177,783	9,175,171	-0.03%	21,858,272	21,355,644	-2.30%	12,622,161	12,571,677	-0.40%	28,590,859	28,003,130	-2.06%
2015-16	9,343,463	9,342,315	-0.01%	22,155,070	22,052,101	-0.46%	13,000,379	12,910,584	-0.69%	28,927,726	28,695,362	-0.80%
2016-17	9,533,174	9,533,563	0.00%	22,462,906	22,454,335	-0.04%	13,101,852	12,989,135	-0.86%	29,120,025	28,950,565	-0.58%
2017-18	9,573,301	9,573,895	0.01%	22,733,719	22,532,459	-0.89%	13,225,213	13,171,532	-0.41%	29,490,423	29,205,316	-0.97%
2018-19	9,691,466	9,689,669	-0.02%	23,329,798	22,916,413	-1.77%	13,846,501	13,766,559	-0.58%	30,315,781	29,790,813	-1.73%
2019-20	9,660,409	9,658,104	-0.02%	25,310,312	24,967,675	-1.35%	13,984,678	13,930,207	-0.39%	32,429,462	31,989,483	-1.36%

Chart 1a: Admitted Patient Care (APC) - Count of unique patients by HESID, Person_ID, percentage change, by year

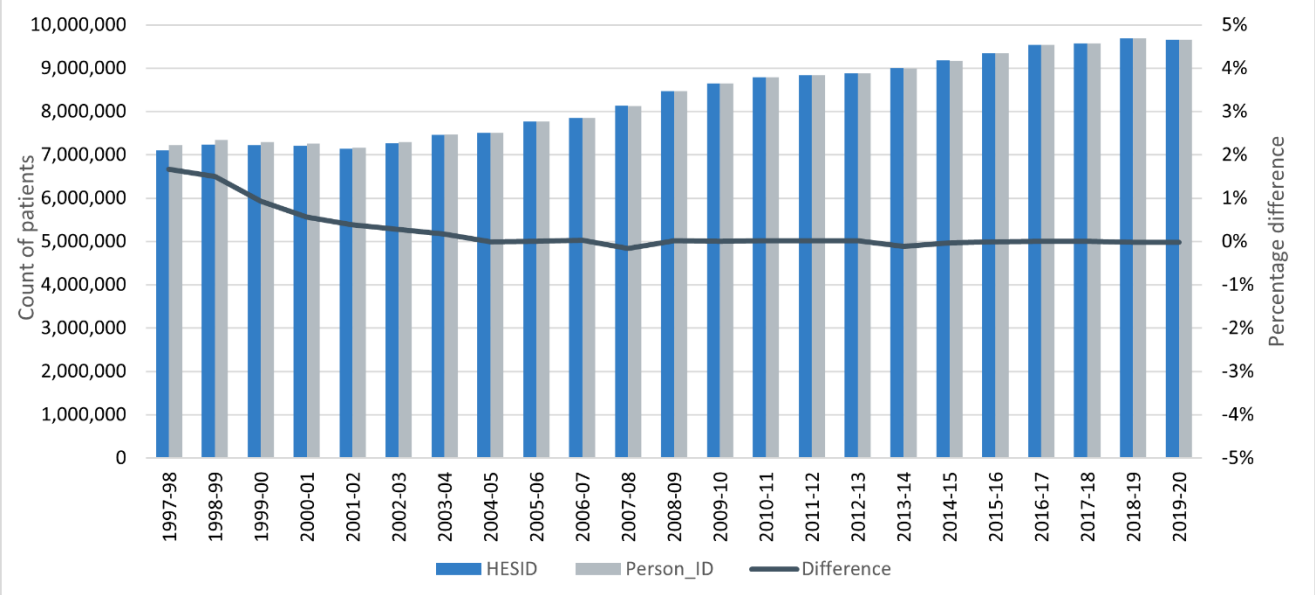


Chart 1a shows that in the earliest years of APC data (1997-98 to 2003-04) there are more patients according to the Person_ID algorithm than under the HESID algorithm. This is likely to be because completeness of fields such as NHS number and date of birth is lower in the earlier years, and the Person_ID algorithm relies more on these fields, while the use of local patient identifier at an earlier stage of the HESID algorithm has achieved more matches.

The difference in patient counts reduces to a consistent minimal level in the most recent years, being within 0.03% or less difference in 11 of the past 12 years since 2008-09.

This coincides with increased NHS number coverage over time, as shown in Table 8 in the accompanying data file.

Chart 1b: Outpatients - Count of unique patients by HESID, Person_ID, percentage change, by year

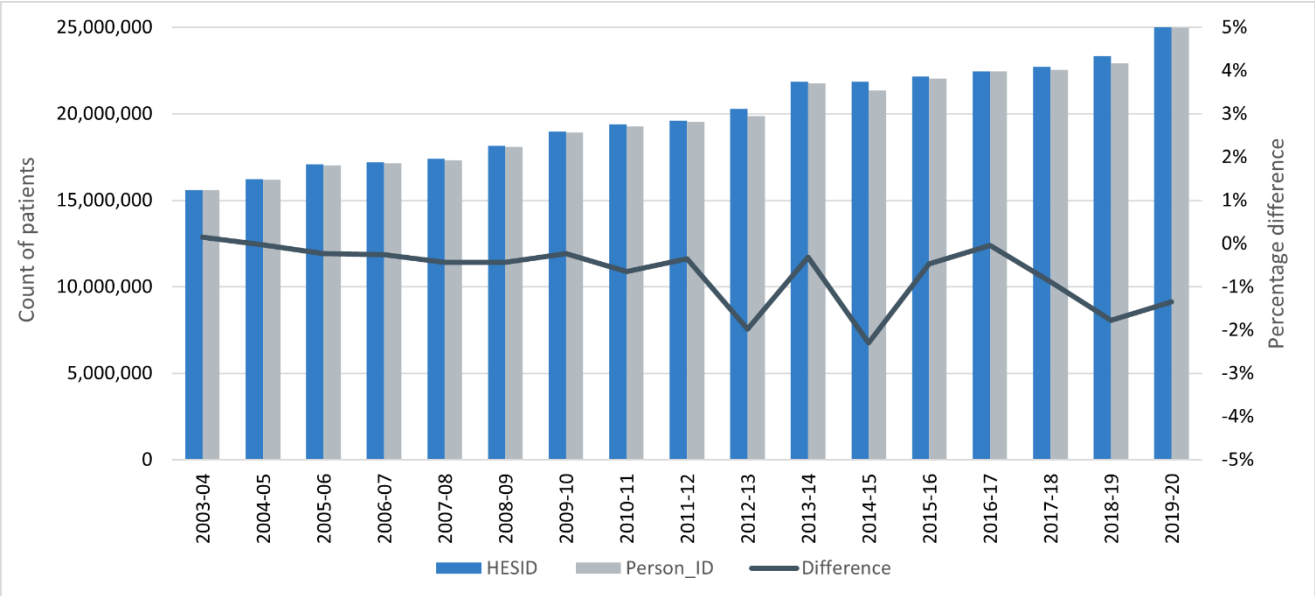


Chart 1b shows that for Outpatients, counting patients by Person_ID results in fewer patients than HESID in all years apart from the first recorded year in 2003-04. In 13 out of the 17 years the change is within 0% to -1% of the HESID count.

The difference in counts of HESIDs and Person_IDs is between -1% and- 2% for 2012-13, 2018-19 and 2019-20, and -2.3% for 2014-15.

The larger difference in these years is likely to be due to specific data quality issues in the SEX field for each of these years, as missing or unknown values of SEX are treated differently in the algorithms.

The HESID algorithm requires a matching and known value of SEX on each record to generate a match, so each record with a missing or unknown value of SEX will generate a new HESID even if the record would match to an existing patient on other details. The Person_ID algorithm does not use the SEX values in the same way so has potentially linked these records to other records for the same patient, resulting in a lower total patient count.

The known data quality issues where large numbers of records from a single trust had null or 'not known' SEX values in each of these years are:

- 2012-13 - Royal Berkshire NHS Foundation Trust, c.460,000 records affected
- 2014-15 - The Christie NHS Foundation Trust, c.460,000 records affected
- 2018-19 - Royal Brompton & Harefield NHS Foundation Trust, c.261,000 records affected
- 2019-20 - Royal Brompton & Harefield NHS Foundation Trust, c.261,000 records affected

Chart 1c: Accident & Emergency - Count of unique patients by HESID, Person_ID, percentage change

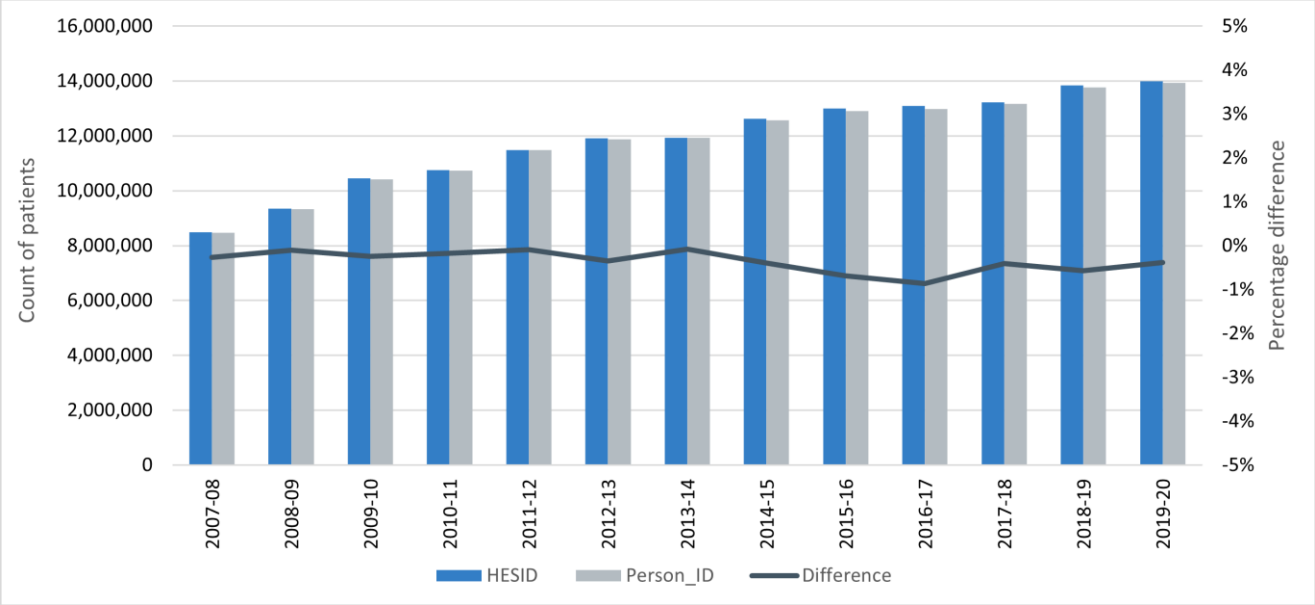


Chart 1c shows that for A&E, counting the Person_ID results in fewer unique patients than counting the HESID in all years of data. The change from HESID to Person_ID reduces the patient count by between -0.08% and -0.86%, suggesting that for A&E attendances Person_ID is enabling more matching of patient activity than HESID.

Patient counts – over multiple years

It is likely that some information about patients such as postcode may change over time, as they move house. Analysing over a longer time interval means patients are perhaps more likely to have been seen by multiple providers who may have recorded their details slightly differently.

Each algorithm aims to continue matching activity for the same patient through such changes and variation in values.

Table 1b shows the count of unique patients using both HESID and Person_ID over all available years and over the last 10 years.

Table 1b: Count of unique patients by HESID and Person_ID, by HES data set across multiple years

	APC	Outpatients	A&E	All
Years	**1997-98 to 2019-20**	**2003-04 to 2019-20**	**2007-08 to 2019-20**	
HESID	68,242,147	73,243,657	61,366,076	101,929,796
Person_ID	68,809,135	69,891,406	59,397,689	96,902,493
Change	0.8%	-4.6%	-3.2%	-4.9%
Years	**2010-11 to 2019-20**	**2010-11 to 2019-20**	**2010-11 to 2019-20**	**2010-11 to 2019-20**
HESID	43,153,277	62,281,361	55,017,857	79,369,656
Person_ID	43,091,352	59,707,971	53,702,740	75,284,955
Change	-0.1%	-4.1%	-2.4%	-5.1%

For APC, there are more unique Person_IDs than HESIDs considering 1997-98 to 2019-20 together. This is likely due to the lower data completeness for NHS number and date of birth in the earliest years, as noted for Table 1a, which limits the success of the Person_ID algorithm in matching to a patient record.

For Outpatients and A&E, we can observe an overall reduction in the number of unique person identifiers generated using the new algorithm.

Many users of HES data may limit their analysis to a more recent period – here we consider the last 10 years.

For APC, there is minimal difference between the two counts of unique patients over these 10 years, suggesting that the high levels of NHS number completeness in these years results in both algorithms consistently matching records to patients.

For Outpatients and A&E, the percentage difference in the counts over the last 10 years is smaller than over all years.

The results may differ for any specific subgroup of patients. For example, if considering patients seen by a particular provider, the impact of the change is likely to vary depending on the data quality of the patient identifiers submitted by that provider. Although the number of Person_IDs in Outpatients or A&E overall is less than the number of HESIDs, for a particular provider or other subgroup of patients it is possible that there could be more Person_IDs than HESIDs.

Consistency of HESID and Person_ID – by year

To understand the extent to which the change to Person_ID will affect patient level analysis, we need to understand how often the allocation of records to patients differs between HESID and Person_ID.

Where the set of records linked to a patient is the same under both algorithms, this means that the switch from HESID to Person_ID would not affect analysis involving that patient - all the activity that would be counted for them when counting by HESID would be included when counting by Person_ID.

We can assess this by counting the number of HESIDs that have the same Person_ID for all of their associated records, and that Person_ID does not appear on any records that have any other HESID. We will refer to this as a ‘1-1 mapping between HESID and Person_ID’.

Table 2a shows the count of unique patients with 1-1 mappings by HES data set and year.

Table 2a: Count of unique patients with a 1-1 mapping between HESID and Person_ID in both directions, by HES data set and year

	APC			Outpatients			A&E			All HES		
Year	1-1 patients	HESIDs	Percentage	1-1 patients	HESIDs	Percentage	1-1 patients	HESIDs	Percentage	1-1 patients	HESIDs	Percentage
1997-98	6,807,554	7,098,467	95.9%							6,807,554	7,098,467	95.9%
1998-99	7,074,321	7,230,413	97.8%							7,074,321	7,230,413	97.8%
1999-00	7,102,829	7,227,455	98.3%							7,102,829	7,227,455	98.3%
2000-01	7,105,409	7,212,972	98.5%							7,105,409	7,212,972	98.5%
2001-02	7,067,686	7,140,716	99.0%							7,067,686	7,140,716	99.0%
2002-03	7,220,482	7,274,255	99.3%							7,220,482	7,274,255	99.3%
2003-04	7,414,906	7,456,986	99.4%	15,372,461	15,586,744	98.6%				17,519,955	17,802,539	98.4%
2004-05	7,472,433	7,506,716	99.5%	16,043,559	16,209,084	99.0%				18,061,834	18,281,807	98.8%
2005-06	7,730,701	7,764,257	99.6%	16,867,488	17,073,118	98.8%				18,781,225	19,035,185	98.7%
2006-07	7,820,048	7,848,061	99.6%	16,987,879	17,193,104	98.8%				18,860,545	19,107,346	98.7%
2007-08	8,081,290	8,136,911	99.3%	17,196,867	17,398,261	98.8%	8,371,579	8,486,309	98.6%	22,479,267	23,071,141	97.4%
2008-09	8,451,085	8,473,905	99.7%	17,938,678	18,161,882	98.8%	9,260,418	9,349,040	99.1%	23,414,053	23,919,644	97.9%
2009-10	8,622,562	8,643,228	99.8%	18,814,452	18,975,771	99.1%	10,327,434	10,451,519	98.8%	24,506,612	25,014,248	98.0%
2010-11	8,777,096	8,793,316	99.8%	19,137,763	19,393,375	98.7%	10,662,785	10,762,169	99.1%	24,932,172	25,395,260	98.2%
2011-12	8,822,316	8,835,863	99.8%	19,443,167	19,612,172	99.1%	11,402,347	11,489,265	99.2%	25,672,780	25,994,166	98.8%
2012-13	8,870,200	8,881,350	99.9%	19,697,849	20,287,852	97.1%	11,788,832	11,918,667	98.9%	26,035,548	26,885,543	96.8%
2013-14	8,973,979	8,998,434	99.7%	21,679,957	21,850,777	99.2%	11,871,021	11,936,228	99.5%	27,912,564	28,253,070	98.8%
2014-15	9,162,645	9,177,783	99.8%	21,227,858	21,858,272	97.1%	12,491,286	12,622,161	99.0%	27,748,263	28,590,859	97.1%
2015-16	9,331,755	9,343,463	99.9%	21,965,057	22,155,070	99.1%	12,813,543	13,000,379	98.6%	28,463,887	28,927,726	98.4%
2016-17	9,524,847	9,533,174	99.9%	22,392,902	22,462,906	99.7%	12,879,304	13,101,852	98.3%	28,726,823	29,120,025	98.6%
2017-18	9,566,344	9,573,301	99.9%	22,385,633	22,733,719	98.5%	13,096,147	13,225,213	99.0%	28,945,292	29,490,423	98.2%
2018-19	9,680,133	9,691,466	99.9%	22,703,002	23,329,798	97.3%	13,680,632	13,846,501	98.8%	29,451,728	30,315,781	97.1%
2019-20	9,641,823	9,660,409	99.8%	24,750,770	25,310,312	97.8%	13,812,356	13,984,678	98.8%	31,577,806	32,429,462	97.4%

For each data set, a very high proportion of patients have a 1-1 mapping between HESID and Person_ID in each year – over 99% in APC for each year since 2002-03 and over 99.5% for each year since 2008-09, and consistently above 98% for both Outpatients and A&E (apart from the years where Outpatient data is affected by known data quality issues).

To reiterate, any annual counts including only these patients will be wholly unaffected by the change from HESID to Person_ID.

An extended version of Table 2a in the accompanying reference tables shows that the percentage of HES records within each year that have the 1-1 patient mapping is similar to the percentage of patients.

Consistency of HESID and Person_ID – over multiple years

As mentioned previously, there are likely to be more versions of a patient's details across longer time periods, as they have contact with different care providers or their details such as home postcode change.

Table 2b shows the percentage of patients for each HES data set that have a 1-1 mapping of HESID to Person_ID for all of their records across a number of years. The figures for 2010-11 to 2019-20 are patients that have a 1-1 mapping if we consider only the last 10 years of HES data. These patients may be linked to other HESIDs or Person_IDs across the full time period.

Table 2b: Count of unique patients with a 1-1 mapping between HESID and Person_ID in both directions, by HES data set across multiple years

	APC	Outpatients	A&E	All
Years	**1997-98 to 2019-20**	**2003-04 to 2019-20**	**2007-08 to 2019-20**	
Patients with 1-1 mappings	64,832,280	66,865,256	56,725,556	87,639,632
HESIDs	68,242,147	73,243,657	61,366,076	101,929,796
Percentage	95.0%	91.3%	92.4%	86.0%
Years	**2010-11 to 2019-20**	**2010-11 to 2019-20**	**2010-11 to 2019-20**	**2010-11 to 2019-20**
Patients with 1-1 mappings	42,839,854	58,273,452	52,010,506	72,042,357
HESIDs	43,153,277	62,281,361	55,017,857	79,369,656
Percentage	99.3%	93.6%	94.5%	90.8%

Counting across all years of HES data, 95% of patients will make the same contribution to a count of APC patients by HESID or by Person_ID, and similarly over 90% of patients for Outpatients or A&E.

For each data set, the proportion of patients with a 1-1 mapping between HESID and Person_ID is higher for the last 10 years – this is likely to be due to a combination of more consistent patient data recording such as NHS number, and fewer patient contacts so less opportunity for different values to be recorded for a specific patient. For the last 10 years of APC, over 99% of patients have a 1-1 mapping between HESID and Person_ID.

The percentage of 1-1 mappings will differ for any specific patient cohort, and users transitioning from using HESID to Person_ID in longitudinal patient studies are likely to want to assess individually the impact on their analysis.

Permutations of HESID and Person_ID

If we consider the 2 sets of patient identifiers as an index, not all patient IDs will have a 1-1 mapping between HESID and Person_ID.

As the algorithms have different rules, the Person_ID algorithm may assign the same Person_ID to records that had different HESIDs. We will describe this as 1 Person_ID linked to multiple HESIDs. A simple example of 1 Person_ID linked to 2 HESIDs is:

Record number	HESID	Person_ID
1	123	ABC
2	123	ABC
3	456	ABC

Similarly records that the HESID algorithm identified as relating to a single patient may be associated with more than 1 Person_ID. We will describe this as 1 HESID linked to multiple Person_IDs. A simple example of 1 HESID linked to 2 Person_IDs is:

Record number	HESID	Person_ID
1	123	ABC
2	123	ABC
3	123	DEF

Impact on counting patients

When cohorts have been created using the HESID as the patient identifier, this may have an impact on the number of patients who would now be included in each cohort – multiple HESIDs changing to a single Person_ID would reduce the number of patients, while a single HESID changing to multiple Person_IDs would increase the number of patients.

Impact on counting activity

This may also have an impact on activity counted as relating to a given patient in a cohort.

Where multiple HESIDs link to a single Person_ID, additional activity would now be counted against the single Person_ID. For example, if previous delivery episodes in HES are used to estimate the number of previous deliveries that a person has had, a HESID with 2 previous deliveries and another HESID with 1 previous delivery could change to a Person_ID with 3 previous deliveries.

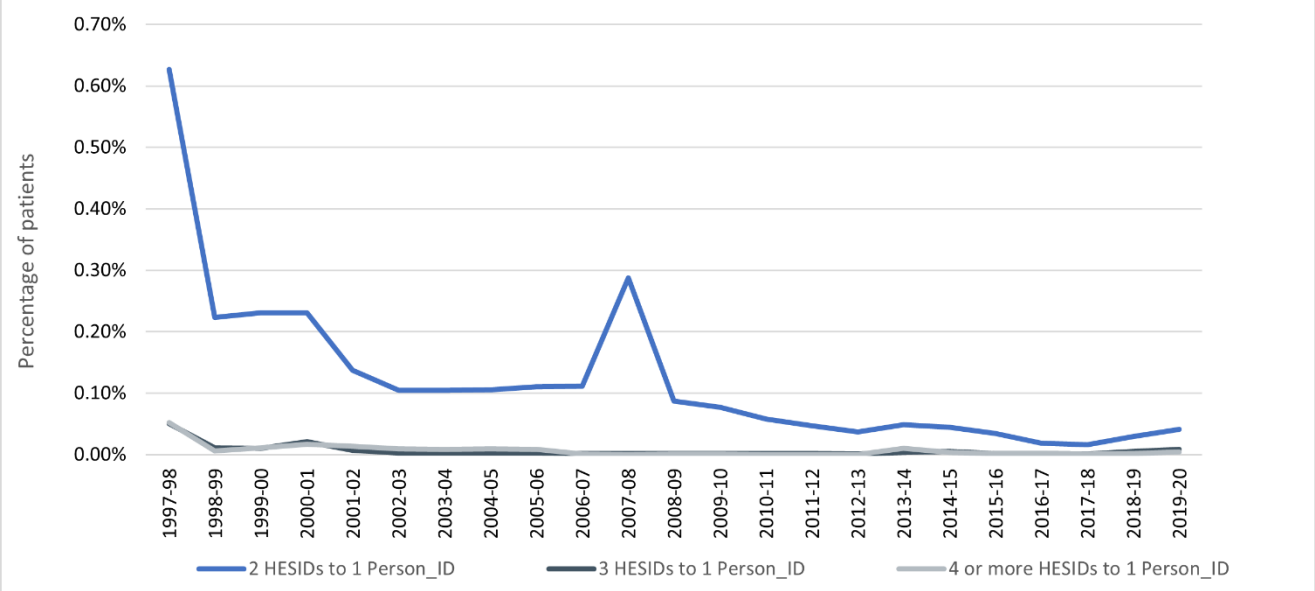
Where 1 HESID links to multiple Person_IDs, activity that could previously be counted as relating to the same individual may now be separated between multiple individuals. For example, if 3 delivery episodes had the same HESID, the separation of this activity between two Person_IDs could mean the analysis now counts a person with 1 previous delivery and another person with 2 previous deliveries.

The following sections of this report quantify the number of instances of each type of change discussed above by data set and by time period.

1 Person_ID linked to multiple HESIDs – by year

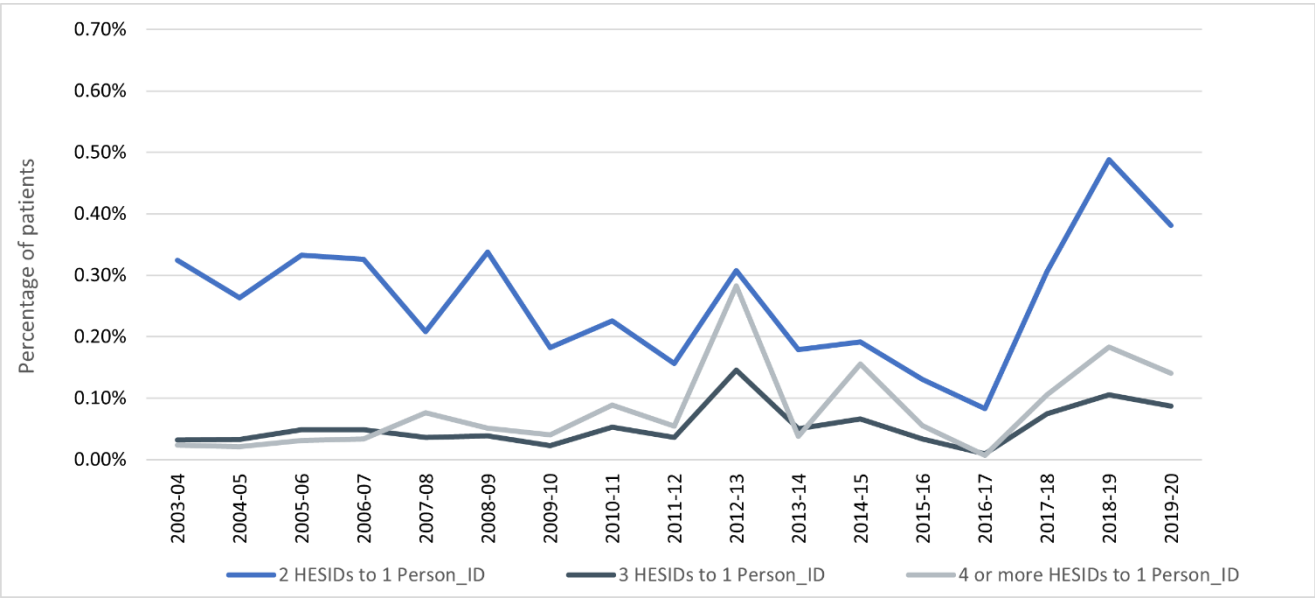
Charts 2a to 2c show the frequency of Person_IDs linked to multiple HESIDs for each data set and year. The number of linked HESIDs is counted within each data set – for example, 2 HESIDs in APC could be linked to 1 Person_ID in APC, and the same Person_ID is linked to another HESID in Outpatients. In these charts this would be counted as 2 HESIDs linked to 1 Person_ID in Chart 2a.

Chart 2a: Percentage of Person_IDs linked to multiple HESIDs, Admitted Patient Care, by year



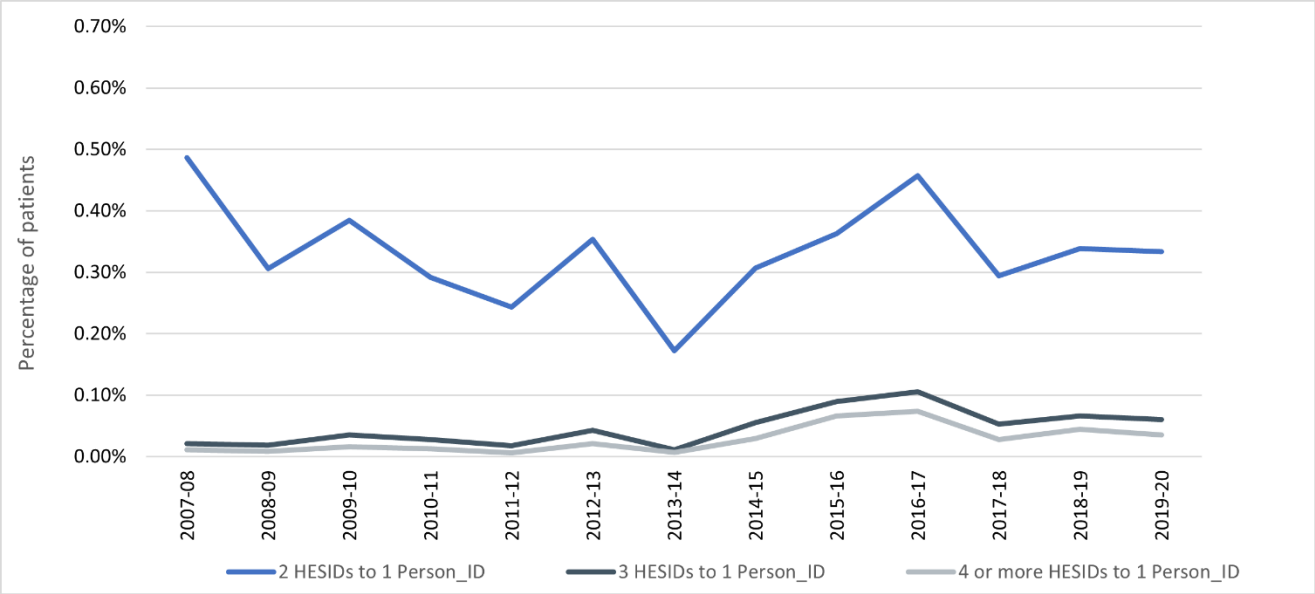
Within every year of APC data after the first (1997-98), fewer than 0.3% of Person_IDs are linked to 2 HESIDs. For every year since 2002-03, at most 0.01% of Person_IDs are linked to more than 2 HESIDs.

Chart 2b: Percentage of Person_IDs linked to multiple HESIDs, Outpatients, by year



The percentage of Person_IDs linked to multiple HESIDs within each year of Outpatient data is highest in the years affected by known data quality issues (2012-13, 2014-15, 2018-19 and 2019-20).

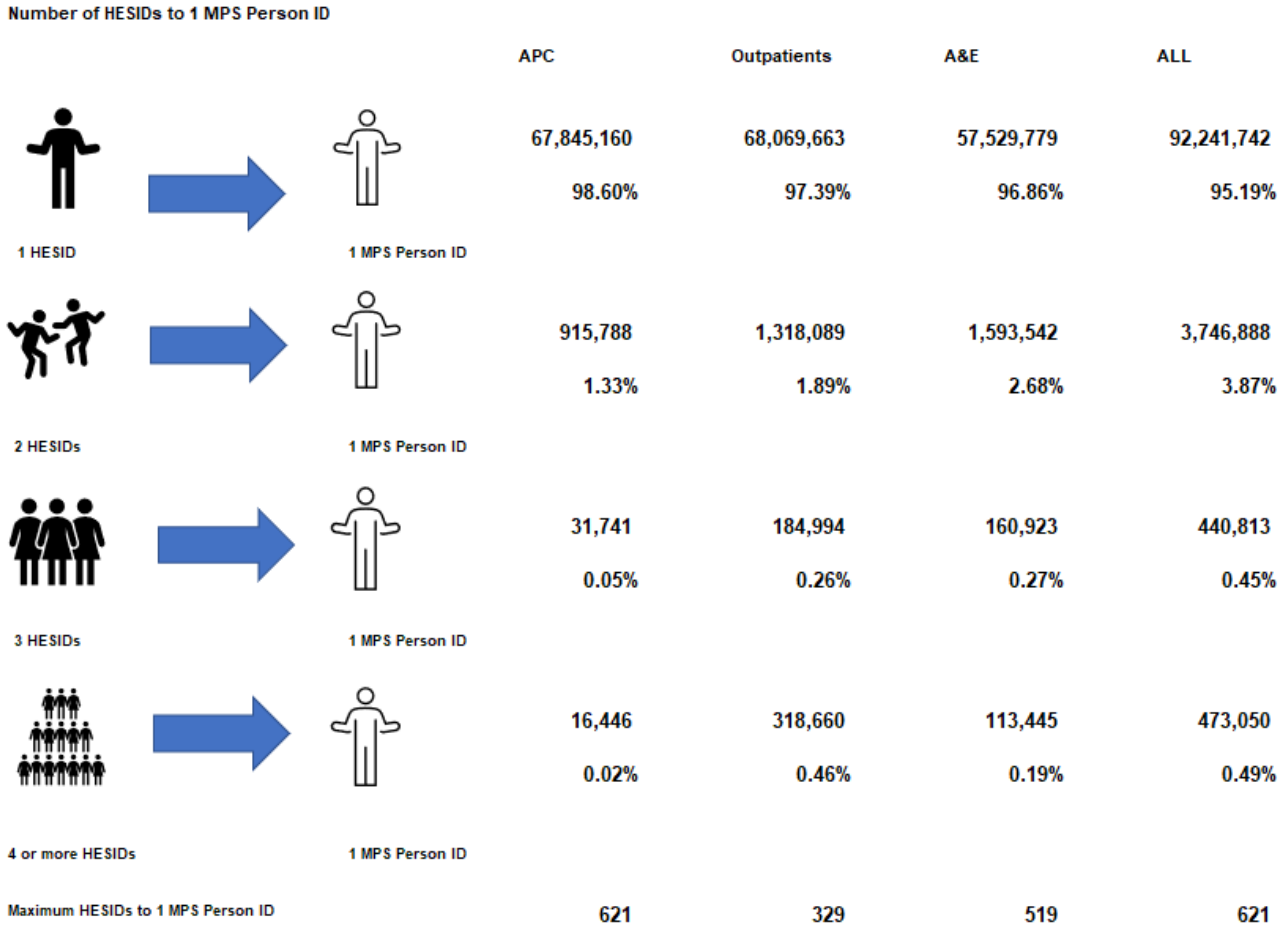
Chart 2c: Percentage of Person_IDs linked to multiple HESIDs, A&E, by year



1 Person_ID linked to multiple HESIDs – over multiple years

Chart 3 shows the number and proportion of Person_IDs linking to 1,2,3 or 4 or more HESIDs across all years (1997-8 onwards for APC, 2003-04 onwards for Outpatients and 2007-08 onwards for A&E).

Chart 3: Number and percentage of Person_IDs by number of linked HESIDs, by data set, all available years



The proportion of Person_IDs linked to multiple HESIDs counting across all available years is greater than the annual proportions in Charts 2a to 2c. The Person_ID algorithm has identified more records as belonging to the same patient in each instance than the HESID algorithm.

Tables 3b to 3e in the accompanying data file provide a more detailed breakdown of the '4 or more HESIDs' category.

Person_IDs linked to multiple HESIDs – contribution of algorithm differences

Gender

A key difference between the HESID and Person_ID algorithms is how gender (the SEX field in HES) is used to match records to a patient.

The HESID algorithm (Appendix A) uses a number of 'passes' to attempt matches on different combinations of personal identifiers, but all passes require a matching and known value of SEX i.e., a record with a missing or 'Unknown' value of SEX will not be assigned the same HESID as any other record.

In the Person_ID algorithm, gender (i.e., the SEX field in HES) is only used in some of the attempts to match records, i.e., records with missing or unknown gender can be matched on other data items such as NHS number and date of birth alone.

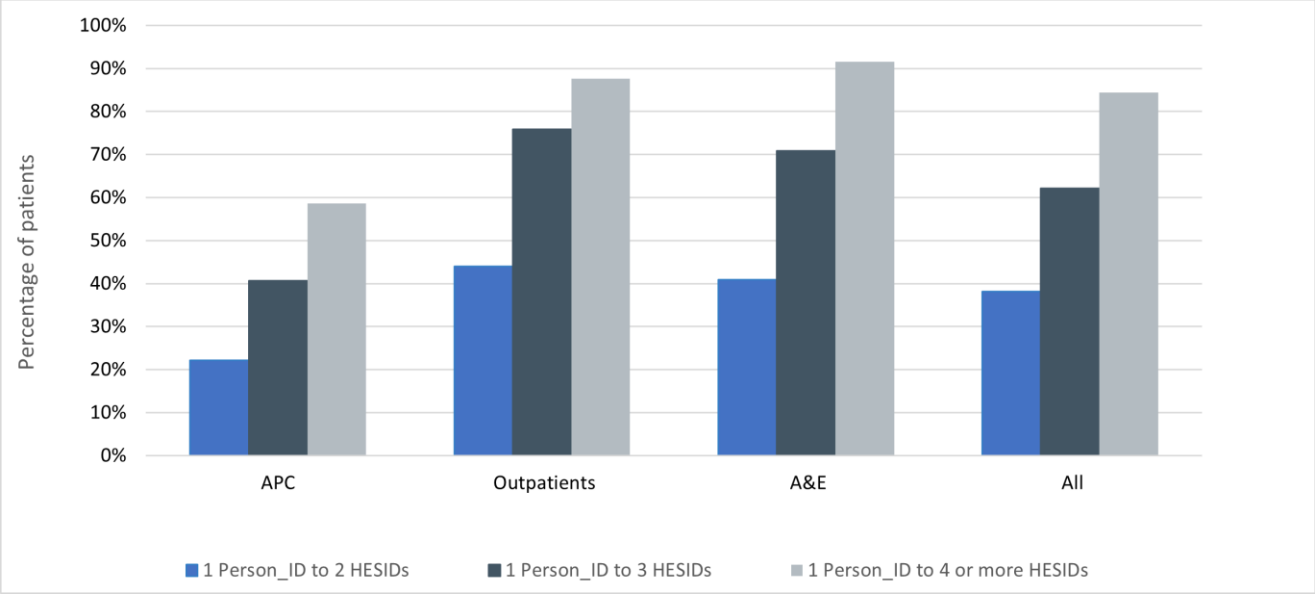
Where 1 Person_ID is linked to multiple HESIDs, we can quantify how often this can be explained by the SEX values on the HES records by counting Person_IDs where

Records with missing SEX	+	Records with SEX = 0 or 9	+	Number of distinct SEX values of 1 or 2	=	Number of HESIDs linked to the MPS Person_ID

Chart 4a below shows the percentage of Person_IDs linked to multiple HES IDs in Chart 3 where the number of HESIDs is wholly explained by the number of missing or different SEX values.

These percentages are based on looking only within the specified data set i.e. do the number of SEX values on APC records explain the number of HESIDs in APC linking to a Person_ID in APC data, irrespective of SEX values that may be associated with those HESIDs in other data sets.

Chart 4a: Percentage of Person_IDs linking to multiple HESIDs where gender values explain number of HESIDs, by HES data set, all available years



22% of the instances of 1 Person_ID linked to 2 HESIDs in APC are explained by the SEX values on the HES records. The same explanation accounts for a larger proportion of the instances of 1 Person_ID to 3 HESIDs (41%) and 1 Person_ID to 4 or more HESIDs (59%).

The corresponding proportions are higher for Outpatients and A&E, with 88% of instances of '1 to 4 or more' in Outpatients, and 92% of similar instances in A&E being explained by the SEX values.

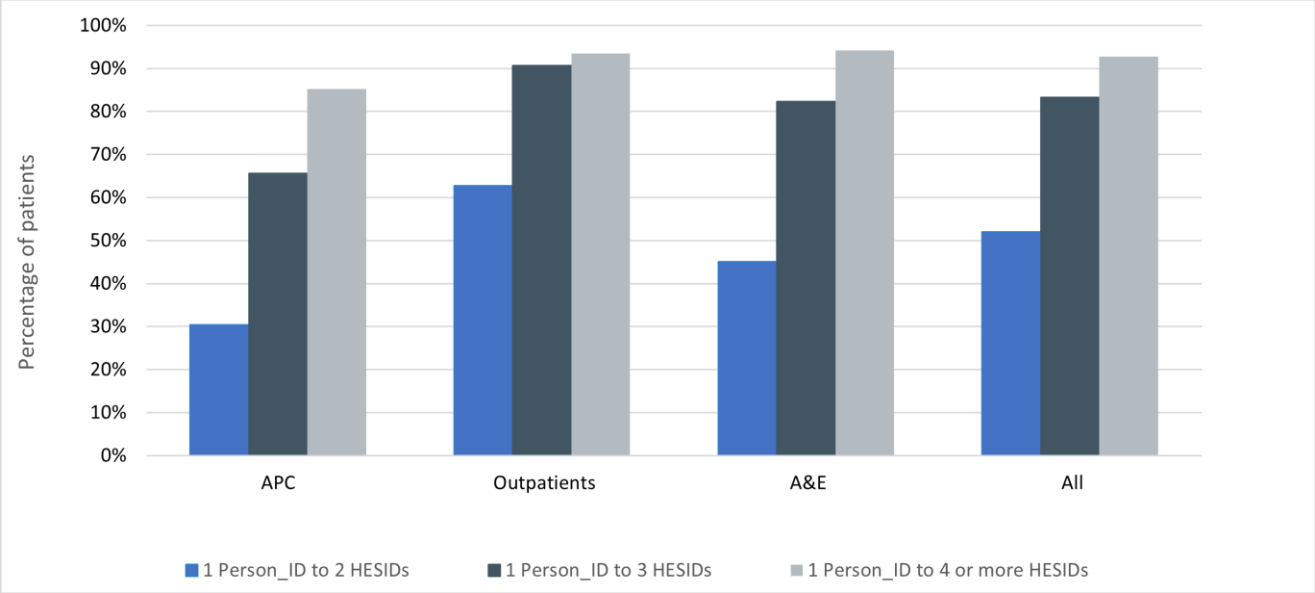
Please note that these percentages only include the instances where the number of HESIDs is explained *exactly* by the number of SEX values – there will be additional instances where the SEX values explain why there is more than one HESID linked to the Person_ID, but not the exact number of HESIDs that are linked.

There could of course be other features of these records that would also mean separate HESIDs but a single Person_ID. We are not claiming that these records would definitely have a single HESID if the HESID algorithm no longer had a dependency on consistent SEX values, just that the different SEX values are sufficient to explain why there are multiple HESIDs but a single Person_ID.

Chart 4b shows the same information just for the most recent 10 years (i.e., where 1 Person_ID links to the number of HESIDs within the last 10 years, and this can be explained by SEX values on records within the last 10 years).

These percentages are based on looking only within the specified data set at records in the specified years, i.e., do the number of SEX values on APC records in these 10 years explain the number of HESIDs in APC linking to a Person_ID in these 10 years of APC data, irrespective of SEX values that may be associated with those HESIDs in other years of APC data or in other data sets.

Chart 4b: Percentage of Person_IDs linking to multiple HESIDs where gender values explain number of HESIDs, by HES data set, 2010-11 to 2019-20



Over this shorter period, SEX values account for noticeably higher proportions of the instances of 1 Person_ID linking to multiple HESIDs in APC (up to 85% of the Person_IDs that link to 4 or more HESIDs) and Outpatients (94% of the equivalent group).

Date of birth (DOB)

The HESID and Person_ID algorithms also differ in their definitions of a partial DOB match, and in what is considered as a valid DOB to match on. Quantifying the precise extent to which these differences contribute to 1 Person_ID linking to multiple HESIDs would require lengthy analysis.

We can however estimate an upper limit on the proportion of instances of 1 Person_ID linked to 2 HESIDs that could be explained by differences in DOB handling logic.

If DOB is the same across a set of records, and meets the validity criteria for both algorithms, then we can assume that any differences in DOB partial matching methods or validity criteria do not explain why the records were assigned multiple HESIDs but 1 Person_ID.

We can count the number of Person_IDs linked to 2 HESIDs where DOB *could* explain this by:

- Discarding the Person_IDs linked to 2 HESIDs that are explained by missing or unknown SEX values.
- Counting the instances where:

- DOB differs i.e., there were two or more distinct values of DOB across all records for the Person_ID or
- At least one DOB was <01/01/1895 or
- At least one DOB was 01/01/1901 or
- At least one DOB was 31/12/1899

This does not mean that DOB handling differences *do* explain why there is 1 Person_ID linked to 2 HES IDs – a full DOB could differ but still meet the partial matching conditions for both algorithms.

However, if this number is small then it suggests there is limited value in further work to obtain a more precise estimate.

Table 3 shows the number and percentage of Person_IDs linked to 2 HESIDs where DOB *could* explain why there are 2 HESIDs but 1 Person_ID

Table 3: Number and percentage of Person_IDs linking to 2 HESIDs with consistent SEX values where DOB varies, across multiple years

	APC		Outpatients		A&E		All HES	
	Person_IDs linking to 2 HESIDs that could be explained	Percentage	Person_IDs linking to 2 HESIDs that could be explained	Percentage	Person_IDs linking to 2 HESIDs that could be explained	Percentage	Person_IDs linking to 2 HESIDs that could be explained	Percentage
All years	38,043	4.2%	49,556	3.8%	33,021	2.1%	133,908	3.6%
2010-11 to 2019-20	3,845	3.6%	18,232	2.9%	20,845	2.2%	45,543	2.8%

This shows that differences in the way the algorithms use DOB for matching will explain at most a small proportion of the instances of 1 Person_ID linking to 2 HESIDs – at most 4.2% of the relevant Person_IDs in APC, and smaller percentages over the last 10 years.

One HESID linked to multiple Person_IDs – by year

Charts 5a to 5c show the percentage of HESIDs linked to multiple Person_IDs for each data set and year.

Chart 5a: Percentage of HESIDs linked to multiple Person_IDs, Admitted Patient Care, by year

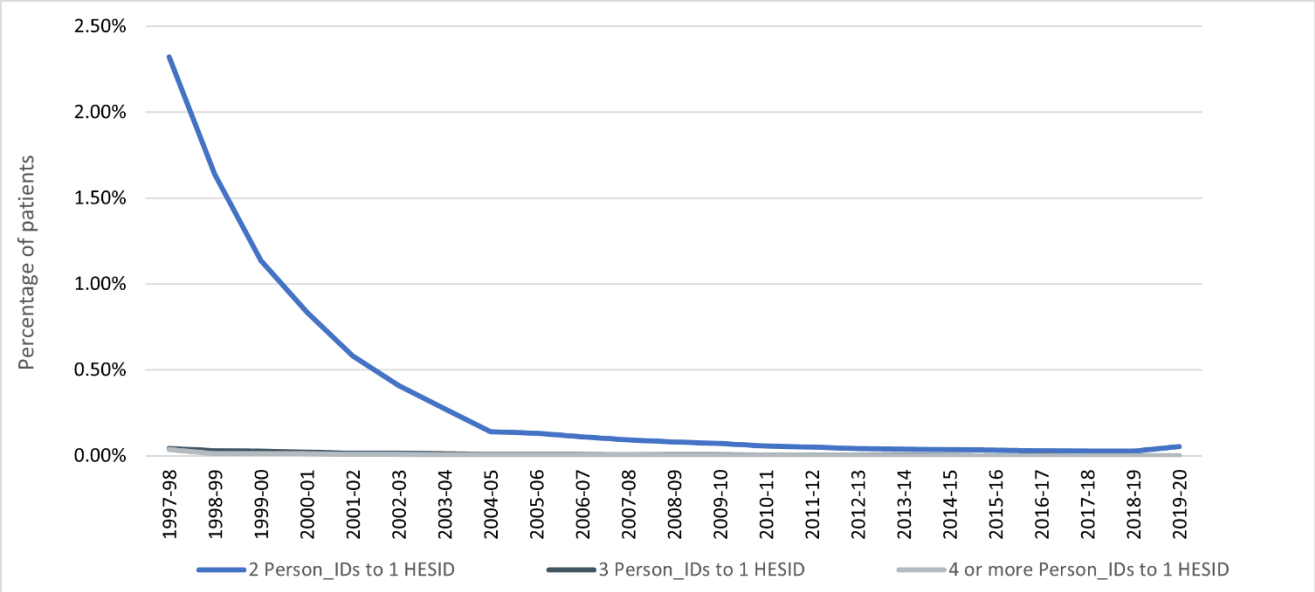


Chart 5a shows that for APC data in 1997-98, 2.3% of HESIDs were linked to 2 Person_IDs, but this proportion decreased quickly over time and has been less than 0.1% since 2007-08.

The higher proportions in earlier years are likely to have arisen because of differences between the ‘further passes’ of each algorithm that increase matching rates. Under the HESID algorithm, some records for a patient can be matched on NHS number, and other matches for the same patient can rely on local patient ID, with both types of match able to be assigned the same HESID. Under the Person_ID algorithm, the two types of match are kept separate – records matched using local patient ID cannot be assigned the same Person_ID as records assigned an Person_ID by confirmation against PDS.

Fewer than 0.1% of HESIDs were linked to more than 2 Person_IDs in any year of APC data, falling to fewer than 0.01% of HESIDs from 2008-09 onwards.

Chart 5b: Percentage of HESIDs linked to multiple Person_IDs, Outpatients, by year

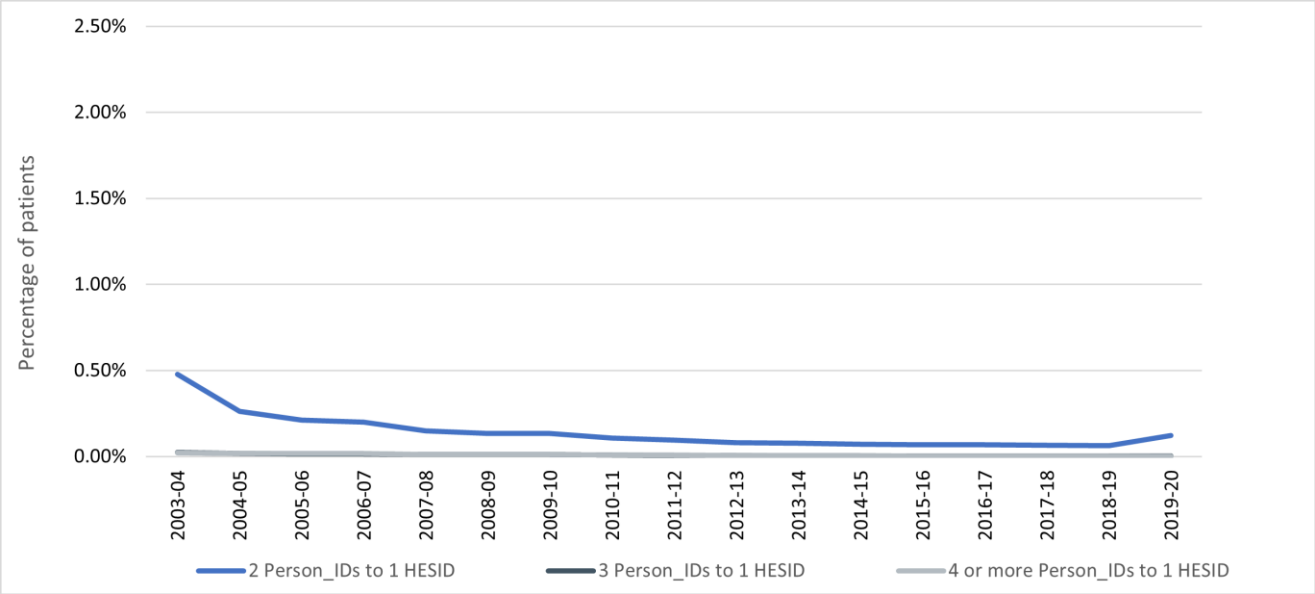
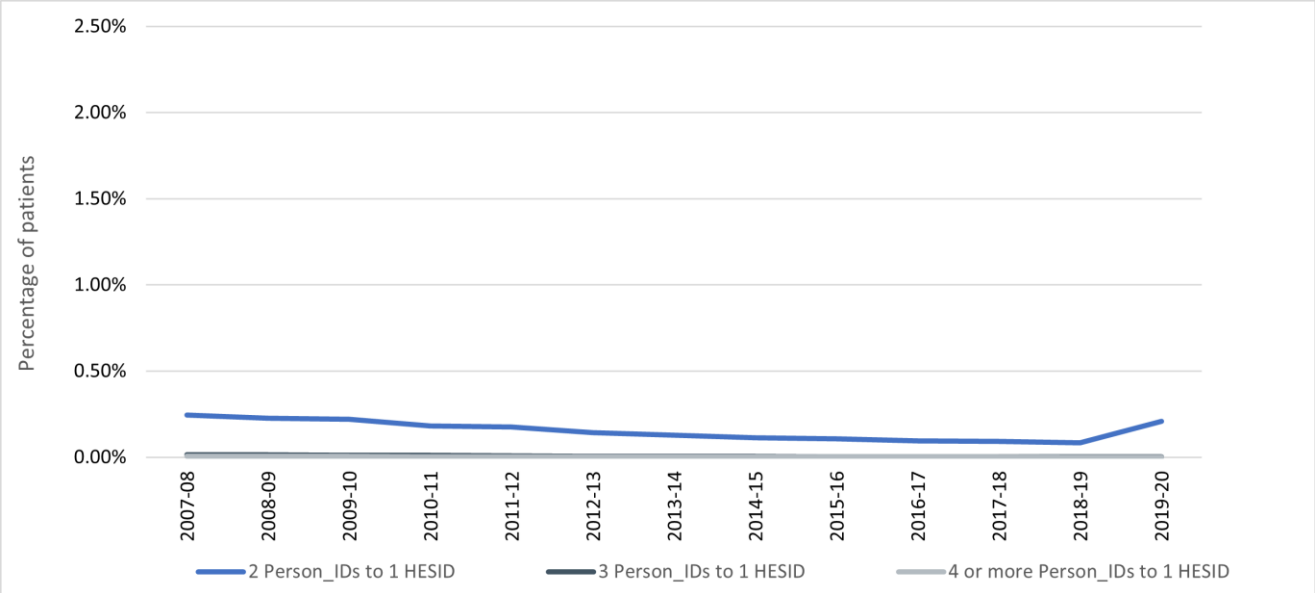


Chart 5c: Percentage of HESIDs linked to multiple Person_IDs, A&E, by year

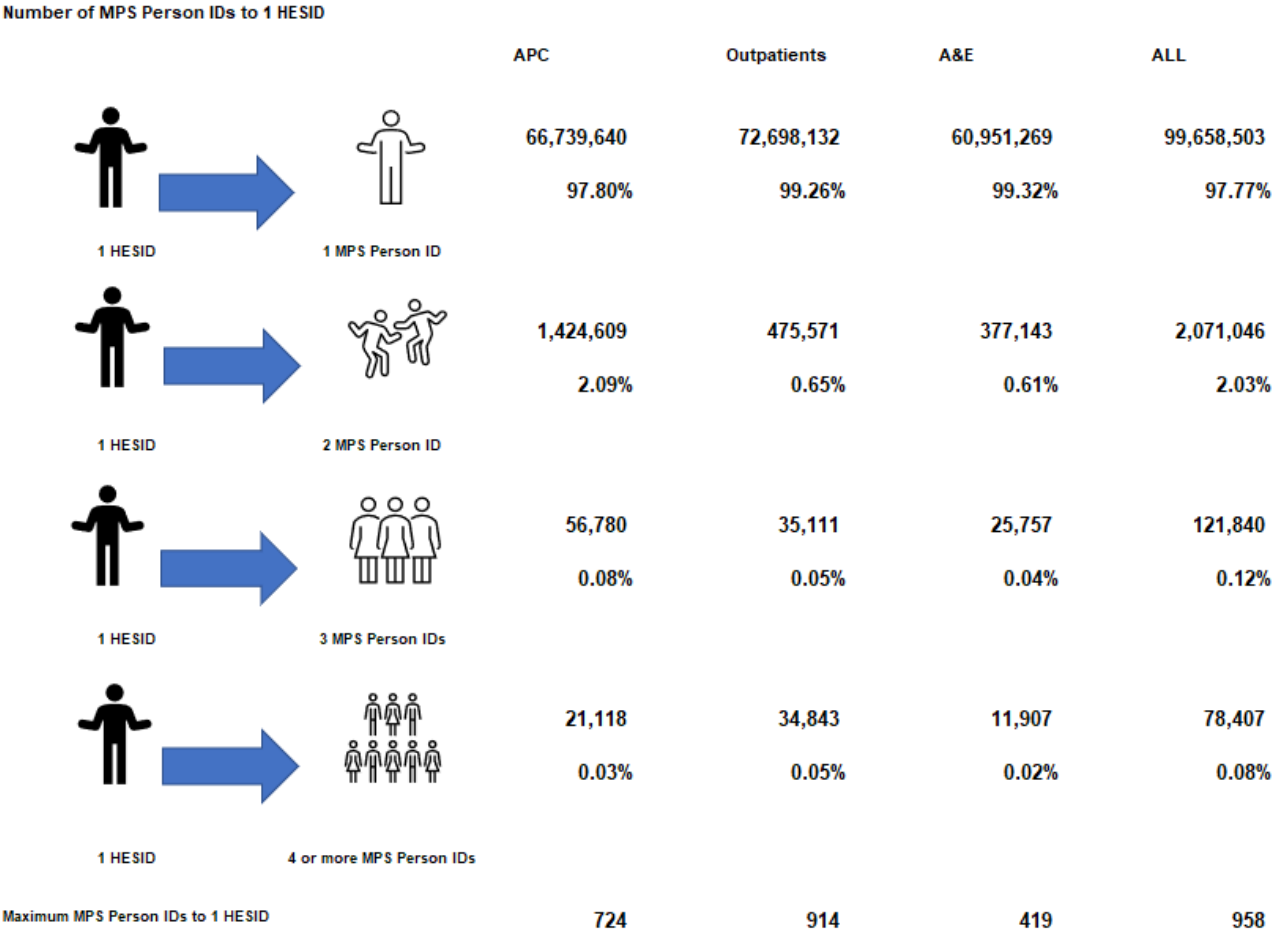


For both Outpatients and A&E, well under 0.5% of HESIDs are linked to 2 Person_IDs in a year in all years following the first year of Outpatients, and less than 0.05% of HESIDs are linked to more than 2 Person_IDs in any year.

One HESID linked to multiple Person_IDs – over multiple years

Chart 6 shows the number and proportion of HESIDs linking to 1,2, 3 or 4 or more Person_IDs across all years (1997-8 onwards for APC, 2003-04 onwards for Outpatients and 2007-08 onwards for A&E).

Chart 6: Number and percentage of HESIDs by number of linked Person_IDs, by data set, all available years



The 2.1% of HESIDs in the APC data which are linked to 2 Person_IDs are largely those seen already for 1997-98 in the analysis by year – if a HESID is linked to 2 Person_IDs in a year then it will also be counted as linked to 2 Person_IDs across multiple years. The percentage is much lower for Outpatients and A&E data.

For each data set, the proportion of HESIDs linking to more than 2 Person_IDs is very small.

Tables 5b to 5e in the accompanying data file provide a more detailed breakdown of the '4 or more HESIDs' category.

The small percentages throughout Chart 6 suggest that the Person_ID algorithm performs similarly to the HESID algorithm in handling patient detail changes over time.

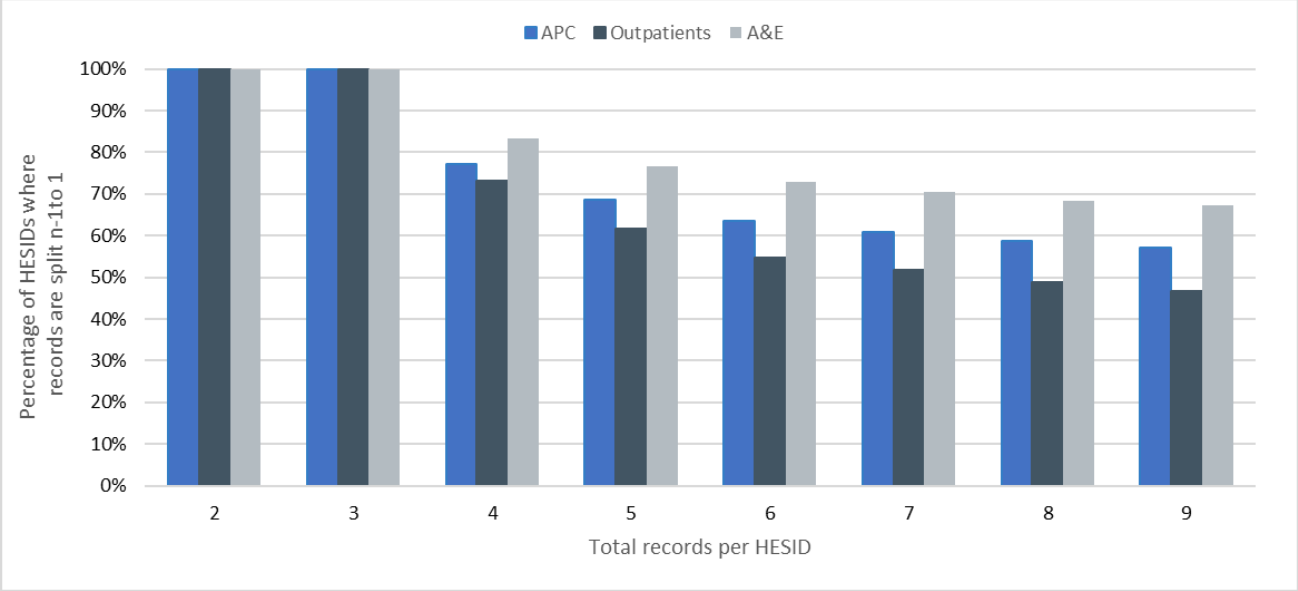
1 HESID linked to 2 Person_IDs – record level

From Chart 6, most instances of 1 HESID linked to multiple Person_IDs involve 2 Person_IDs. For these patients, the total number of records with the HESID will now be split between the 2 Person_IDs – for example if there were 6 records with the HESID, there could be 5 records with 1 of the Person_IDs and 1 with the other, or a 4:2 or 3:3 split.

This separation is likely to have least impact on analysis where the split is highly skewed, so that there is a 'majority' Person_ID associated with a high proportion of the records, keeping as much of the 'patient history' together as possible.

Chart 7a shows the percentage of 1 HESID – 2 Person_ID patients with fewer than 10 records where all but one record has one of the Person_IDs and only 1 record has the other.

Chart 7a: Percentage of HESIDs with fewer than 10 records and linked to 2 Person_IDs where the relevant records with each Person_ID are in the ratio n-1:1, where n is the total number of records per HESID, by HES data set, multiple years

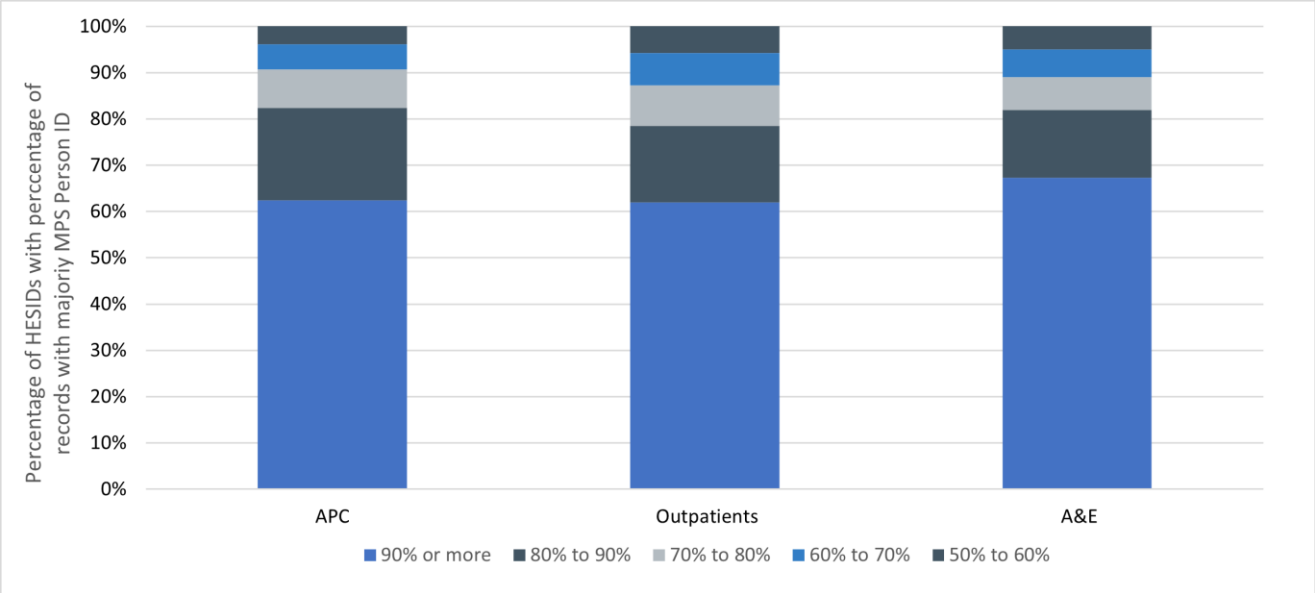


Where there are only 2 or 3 records associated with the HESID, the split is inevitably 1:1 or 2:1 respectively.

Where there are more records, an n-1:1 split is the most common pattern, particularly for APC and A&E. This split is still the most common for Outpatients, but as the number of records increases, the percentage of HESIDs with this pattern does drop below 50%.

Chart 7b shows the distribution of the percentage of records with the 'majority' Person_ID for HESIDs with more than 10 records.

Chart 7b: Percentage of HESIDs with 10 or more records linked to 2 Person_IDs, by frequency of the more common Person_ID among the HES records, by HES data set, multiple years



For each data set, over 60% of the relevant HESIDs that have 10 or more records have at least 90% of records with one of the Person_IDs and at most 10% of records with the other Person_ID. This means that for example, if the HESID is found on 50 records, at least 45 of those records have the same Person_ID and at most 5 have a different Person_ID.

This considers only the records with the original HESID and their Person_IDs. Those Person_IDs may also be associated with other HESIDs. In the following example, the 10

records with HESID 123 are split between the 2 Person_IDs in the ratio 3:7, but the Person_ID for the 7 records also appears on 4 records for another HESID, so has 11 activity records in total.

HESID	No of records	Person_ID	No of records
123	10	ABC	3
		DEF	11
456	4		

Unmatched patients

Records may be submitted with no patient identifiers, or poor quality values such as invalid or default values. (Not all records without patient identifiers reflect poor data quality - providers are required to remove patient identifiers before submitting records with certain 'legally restricted codes' or where a patient has withdrawn consent.)

Both the HESID algorithm and the Person_ID algorithm will assign such records a person ID, but this will be an 'unmatched' ID. The record will not be matched to any existing records, so this will be a new ID, and no subsequent records will be matched to this person ID. Up to 2020-21, all unmatched IDs will be seen once only across all years and all data sets. From 2021-22, any unmatched ID assigned to the first episode in an APC spell will be seen repeated on all episodes in that spell.

(Please note that the opposite of 'unmatched' in this context is 'matchable' rather than 'matched' – records not given unmatched IDs are of sufficient quality that a match would be possible if the relevant details were submitted on another record, but this 'matchable' ID may so far be seen only once throughout HES and may or may not be seen again in future.)

Because the algorithms differ in the data items used, some records will be assigned an unmatched HESID, but not an unmatched Person_ID, and vice versa.

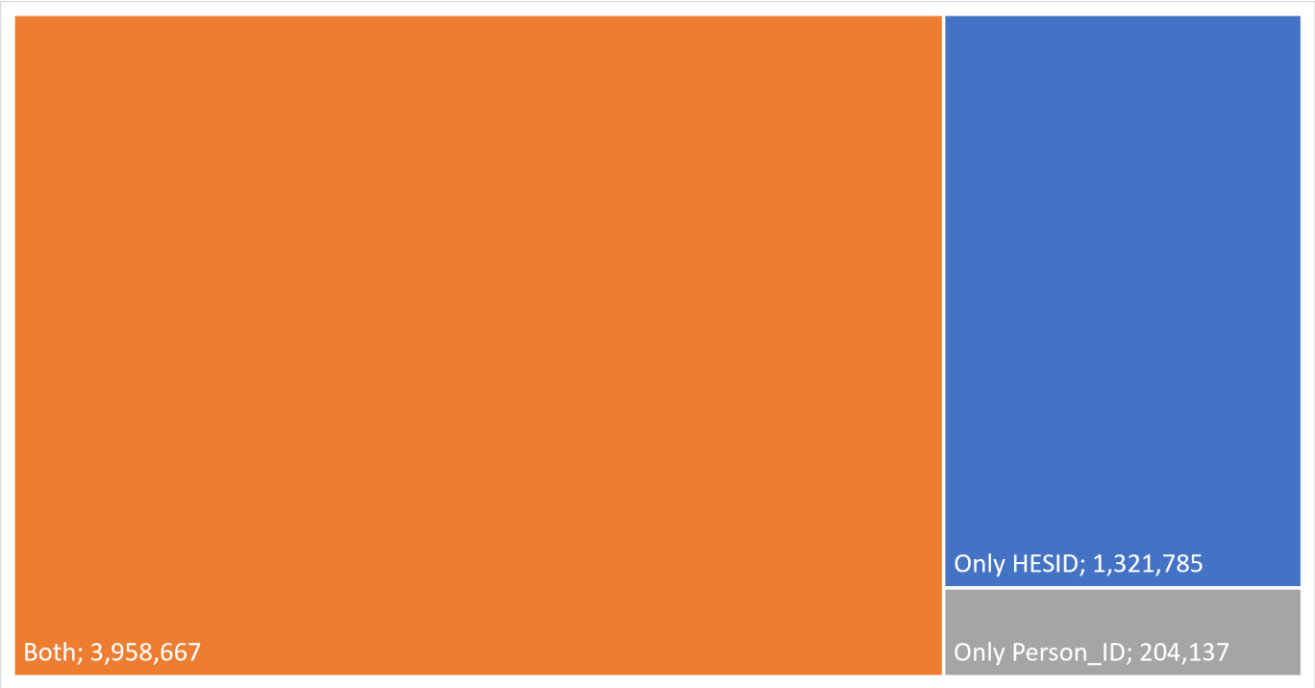
As mentioned earlier in this paper, the HESID algorithm includes gender in its matching criteria at every stage. Records with a missing or unknown SEX value will therefore always be unmatched by the HESID algorithm, and each such record will have an unmatched ID. The Person_ID algorithm does not have the same dependence on gender values, so SEX values alone would not result in unmatched Person_IDs for these records (they may be unmatched for other reasons).

Charts 8a to 8c show tree maps representing the number of unmatched IDs for each HES data set according to the algorithms.

Chart 8a: Admitted Patient Care - Comparison of unmatched patient count by HESID, Person_ID and both

Chart 8b: Outpatients - Comparison of unmatched patient count by HESID, Person_ID and both

Chart 8c: Accident & Emergency - Comparison of unmatched patient count by HESID, Person_ID and both



For APC, 95% of records with an unmatched HESID also have an unmatched Person_ID, suggesting the application of the 'unmatched' criteria produces similar results from each algorithm.

For Outpatients and A&E, 29% and 25% of the records with an unmatched HESID do not have an unmatched Person_ID i.e., the Person_ID algorithm considers the data quality to be sufficient for matching. This does not on its own mean that such records will be matched to other records under Person_ID, as such records with similar enough details would also need to be submitted. However, the other analysis earlier in this paper shows that more matching has taken place in HES to date, with the number of Person_IDs being smaller than the number of HESIDs in these settings.

Only 4% of Outpatient records with an unmatched Person_ID do not have an unmatched HESID, with an equivalent figure of 5% for A&E.

This provides evidence that transitioning to the use of Person_ID in all HES data sets will enable users to perform additional patient level analyses utilising more of the Hospital Episode Statistics data than was possible using HESID.

Assigning Person_ID to APC episodes at spell level

From 2021-22, the Person_ID will be derived for the first episode in each APC spell and assigned to all episodes with the same SUSSPELLID.

In some instances, this may mean that episodes are assigned the same Person_ID when the same submitted activity and data values would have resulted in being assigned different Person_IDs up to 2020-21.

This may result in (slightly) smaller patient counts for equivalent activity from 2021-22.

The first episode in each spell using SUS+ methodology is not flagged in HES, and it would be complex to derive this for periods prior to 2021-22 to assess exactly how often the Person_ID for the first episode differs from the Person_ID for other episodes in the same spell.

However, for 2018-19 and 2019-20 we can calculate the maximum number of episodes for which the Person_ID for the first episode *may* differ, as an indication of the potential impact on patient counts from 2021-22 onwards.

In a spell where each Person_ID derived at episode level was a (different) unmatched Person_ID, we know that the Person_ID for every episode except the first will be different to the Person_ID for the first episode.

In all other spells, we know that the largest possible number of episodes where the Person_ID could differ from the Person_ID for the first episode is the total number of episodes less the number of episodes with the least frequent Person_ID within the spell.

For example, in a spell with 6 episodes, 1 with one Person_ID and 5 with another, then at most 5 episodes could have a different Person_ID to the one for the first episode – this would be the maximum for this spell. However, if the first episode is among the 5 with the common Person_ID, then only 1 episode has a different Person_ID to the first episode.

Table 4 shows the number and percentage of such episodes in 2018-19 and 2019-20.

Table 4: Maximum number of APC episodes for which Person_ID could differ from first episode in spell, 2018-19 and 2019-20

	Spells where all episodes in spell have an unmatched Person_ID		All other spells		All spells	
Year	Number of episodes where Person_ID differs from first episode Person_ID	Percentage of all APC episodes	Maximum possible number of episodes where Person_ID differs from first episode Person_ID	Percentage of all APC episodes	Maximum possible number of episodes where Person_ID differs from first episode Person_ID	Percentage of all APC episodes
2018-19	33,133	0.15%	4,954	0.02%	38,087	0.17%
2019-20	32,423	0.14%	6,286	0.03%	38,709	0.17%

This shows that the number of episodes where the Person_ID for other episodes could differ from the Person_ID for the first episode in the spell was very small for the two most recent years.

Assigning a common Person_ID to all episodes in the spell for Type 1 episodes could be considered beneficial, as based on being in the same spell they would now be counted as a single patient rather than multiple patients.

Reference data

Using different rules to decide when a submitted organisation code is valid will affect the organisation code and related derivations for records submitted by organisations that undergo changes during a reporting period. There will therefore be changes to organisation level activity counts or other measures presented in national statistics and official statistics reports.

Example – two organisations merging

Organisation A and organisation B merge on 1 August 2021, forming a new organisation C, and the submitted outpatient appointments between April and August 2021 are as follows:

Organisation	Appointment dates	Count of appointments
A	1 Apr 2021 to 31 Jul 2021	300
B	1 Apr 2021 to 31 Jul 2021	500
C	1 Aug 2021 to 31 Aug 2021	200

Under current HES processing rules, for the August 2021 year-to-date data, validity of the submitted organisation codes would be assessed as at 31 August 2021. Provided the closure of A and B is recorded in the August quarterly release of ODS organisation data, organisation codes A and B would be identified as invalid, and the provider mapping process would begin.

Organisation C would be identified as the successor organisation to A and B, and the organisation level data would be reported as:

Organisation	Number of appointments 1 Apr 2021 to 31 Aug 2021
C	1000

Following the change, validity of the submitted organisation codes would be assessed as at the appointment date. Organisation codes A and B were valid at the appointment dates, so the provider mapping process would not be applied. The organisation level data would be reported as:

Organisation	Number of appointments 1 Apr 2021 to 31 Aug 2021
A	300
B	500
C	200

Example – organisation transferring services to different organisations

Organisation A exists until 1 August 2021, when parts of its operations transfer to organisation B, and other parts to Organisation C. The submitted outpatient appointments between April and August 2021 are as follows (for this illustration ignore any other activity for organisations B and C):

Organisation	Appointment dates	Count of appointments
A	1 Apr 2021 to 31 Jul 2021	600
B	1 Aug 2021 to 31 Aug 2021	100
C	1 Aug 2021 to 31 Aug 2021	50

Under current processing rules organisation code A would be identified as invalid at 31 August 2021, and the standard mapping process would be unable to allocate any of A's activity to organisation codes B or C. The records would therefore be deleted, and only the following would be reported:

Organisation	Number of appointments 1 Apr 2021 to 31 Aug 2021
B	100
C	50

The only way that A's records could remain in the data set would be via an additional 'manual clean' at the reporting period end to allocate them between organisations B and C.

Following the change, all records would be valid at the activity date, and no provider mapping or manual updates would be attempted. The activity would be reported as

Organisation	Number of appointments 1 Apr 2021 to 31 Aug 2021
A	600
B	100
C	50

Deriving geography-related fields based on the validity of the postcode at the activity date rather than the reporting period end will also affect counts at each geography.

Example – postcode allocation to geography changed during period

Although changes to create new CCGs or other geography codes mostly happen only at the start of a financial year, there can be updates to the postcode reference data during the reporting period.

Suppose the fictional postcode YY11 1YY moves from one CCG to another at 1 Aug 2021. This would be recorded in the reference data as:

Postcode	CCG	Start date	End date
YY11 1YY	00A	1 Apr 2021	31 Jul 2021
YY11 1YY	00B	1 Aug 2021	

Under current processing, in August 2021 year-to-date data, outpatient appointments would pick up the reference data values in force at the end of August 2021 as follows:

Appointment date	HOMEADD	CCG_RESIDENCE
25 May 2021	YY11 1YY	00B
10 Aug 2021	YY11 1YY	00B

Following the change, outpatient appointments would pick up the reference data values in force at the activity date:

Appointment date	HOMEADD	CCG_RESIDENCE
25 May 2021	YY11 1YY	00A
10 Aug 2021	YY11 1YY	00B

Example – new postcode opened during the reporting period

Suppose the fictional postcode XX11 1XX first appears in the postcode reference data with an effective date of 1 August 2021, and geography codes

Postcode	LSOA11	CCG
XX11 1XX	E01023201	11A

Under current processing, in August 2021 year-to-date data, an outpatient appointment on 25 May 2021 with this postcode would pick up the reference data values in force at the end of August 2021, which would be the following:

Appointment date	HOMEADD	LSOA11	CCG_RESIDENCE
25 May 2021	XX11 1XX	E01023201	11A

Following the change, this postcode would be assessed to be invalid at the activity date, and the same record would appear as

Appointment date	HOMEADD	LSOA11	CCG_RESIDENCE
25 May 2021	XX11 1XX	NULL	59999

Appendices

Appendix A - HESID methodology

The HESID is a pseudonymised number which uniquely identifies a patient and provides a way of tracking them in HES. It is useful in linking all activity records for the patient together without viewing patient identifiable or 'clear' fields such as the NHS number.

A combination of fields known as the patient key is created from HES data and is used in the HESID matching process. A patient can have different patient keys as values like home address, hospital provider and local patient ID can change. Each individual patient key can only be allocated to one HESID, but several different patient keys can be allocated to the same HESID.

The HESID to patient key mapping is stored in a table known as the HESID Index. The information from an activity record is only added to this index if there are sufficient valid data items to create a match. Otherwise, the activity record cannot be matched to the Index, and is assigned a new unique HESID value (an unmatched ID).

If an activity record includes enough information to attempt a match, but no match is found, a new HESID is created, and the record details are added to the Patient HESID Index, because another activity record may match it at some later date.

The matching process involves three main steps.

The first step is driven by NHS number, and attempts to perform a match using the following patient identifying information:

- Sex (Exact match)
- Date of Birth (Partial match)
- NHS number (Exact match)

The second step is driven by Local Patient Identifier within Provider and attempts to perform a match using:

- Sex (Exact match)
- Date of Birth (Partial match)
- Postcode (Exact match)
- Provider Code + Local Patient Identifier within Provider (Exact match)

The third step is driven by Date of Birth and attempts to perform a match using:

- Sex (Exact match)
- Date of Birth (Exact match)
- Postcode (Exact match)

It will not consider two records for matching if the NHS numbers differ, except when one record has a null NHS number. Where it can match to a record in the Index with a null NHS Number, the match would be disallowed if the HESID is associated with another NHS number in the Index. Step 3 will not match records if the match can happen on pass 3 and pass 3 only and if the postcode relates to a communal establishment such as a hospital, prison, army barrack etc. which cover a large range of potential patients. For a record to be able to match on pass 3 only NHS number should be blank and either Provider Code or Local Patient Identifier should be blank.

Notes on valid data values for HESID

To create the patient keys and the HESID, the algorithm compares a range of fields. To aid data quality, the data is subjected to a range of validation rules to ensure that they are valid and in the correct format.

To avoid matching together large numbers of records with missing, invalid or default values, patient keys which include these values are treated differently. Instead of being added to the main HESID index and processed, these patient keys are moved to a separate table where each instance of this poor-quality patient key that occurs in the source data is given a unique HESID (an unmatched ID) rather than matching to an existing HESID.

1) A Date of birth is valid if:
   - It is not null
   - It is a valid date
   - It is no earlier than 1895/01/01
   - It is not later than the end of the current data year

2) An NHS number is valid if:
   - It is not null
   - It consists of exactly 10 digits
   - The 10 digits are not all the same
   - It is not of the format “n00000000n” (where the first and last digits are the same)
   - It is not the dummy/default value “2333455667”
   - The modulus 11 check digit is correct

3) A Postcode is valid if:
   - It is not null
   - It is exactly 8 characters long
   - It is of the format AXXX�9AA, AXX��9AA, or AX���9AA, where A is any uppercase alphabetic character (A Z), X is any uppercase alphanumeric character (A – Z, 0 9), 9 is any digit (0 9), and is a space
   - It does not start with ‘ZZ’

In addition to the validation checks detailed above, several further criteria are applied to the data:

4) Postcode exclusions for pass 3
   A match cannot be created on pass 3 of the algorithm if the postcode is on the list of postcodes that cover a large range of potential patients in communal establishments such as hospitals, prisons, and military establishments.

5) Local Patient ID
   For matching purposes, all zeros and spaces are removed from local patient identifiers, which cover local PAS or case note numbers.

6) Date of birth partial matching
   Two DOBs partially match if:
   - Neither DOB is 1901/01/01
   - Neither DOB is 1899/12/31
   - The two DOB values are no more than 14 years apart
   - The two DOB values are the same or two components (i.e., YYYY, MM or DD) of the two DOB values match or two components of the two DOB values match when the MM and DD parts of one of them are swapped