HESonline™
HOSPITAL EPISODE STATISTICS

NHS
The Information Centre
for health and social care

Replacement of the HES Patient ID (HESID)

Contents

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

1 Executive summary

The HES Patient ID (HESID) provides a way of tracking patients through the HES database without identifying them. It is central to many HES outputs including spell construction, emergency readmissions and linkage to other datasets, such as mortality.

It has two main advantages over using patient identifiers, such as NHS Number:
1) It is derived from several different patient identifiers, so is more resilient to data quality or coverage problems affecting individual fields
2) It is a pseudonym field generated in HES processing so it minimises the risk of identification of patients

An improved methodology for generating the HESID was rolled out in March 2009 that further increases its resilience to data coverage shortfalls and minimises the risk of patient identification. This changes the HESID from an integer value to a 32 character combination of numbers and letters, which will be unique to each customer who requests an extract. This enables customers to track patients through the HES database, but minimises patient identification as it prevents them from joining their data to those of a different customer.

This document outlines the reasons for this improvement, the impact of the change and also provides a comprehensive guide to the new methodology.

To maintain consistency and to avoid confusion, the following names will be used:
- HESID – a generic term that can refer to either version.
- HESIDv1 – a term referring specifically to the previous version of the HESID.
- HESIDv2 – a term referring specifically to the new version of the HESID.

This document is intended as a guide to how the HESID is generated and the changes between the old and the new versions. The main body of this document is non technical for general information purposes. Detailed specifications of the algorithm have been provided in Annex C for more technical readers.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

2 Introduction to the HESID

The data supplied by NHS providers, which flows from the Secondary Uses Service (SUS) into HES, consists of information about individual consultant episodes, outpatient attendances and A&E attendances, with no links between them. However, in reality, several such activity records may be related to a single patient.

While, theoretically, the NHS Number is unique to each individual patient and could be used as a unique identifier to link activity records to each patient, there are two major problems with this approach. The first is that the NHS Number is a directly identifiable field and therefore potentially could be used to disclose personal and sensitive information about a patient. The second is that coverage of NHS Number is not high enough to allow full matching to take place. In 2007-08 around 97% of inpatient activity records had a valid NHS Number. However, in 2000-01 this figure was around 83%, which equates to around 2 million records without an NHS Number.

To link all activity records for the same patient together the HESID was developed. HESID is a pseudonymised number which uniquely identifies each patient without the necessity of viewing patient identifiable or 'clear' fields such as the NHS Number.

HESIDv1 was an integer value. HESIDv2 is a 32 character combination of numbers and letters (eg B8945C0BCB16DC6ED1E58C19749A44CC).

The HESID is derived using a matching algorithm which looks at various combinations of the following patient identifiable fields:

NHS Number	(fieldname NEWNHSNO)
Date of Birth	(fieldname DOB)
Sex	(fieldname SEX)
Postcode	(fieldname HOMEADD)
Provider code	(fieldname PROCODET)
Local patient ID	(fieldname LOPATID)

The way that these fields are combined varies between HESIDv1 and HESIDv2. These combinations of fields are known as patient keys and form the basis of the HESID allocation process. Each individual patient key can only be allocated to one HESID, but a unique HESID can be mapped to several different patient keys.

The table which stores the HESID -> patient key mapping is known as the PATIENT HESID Index. There is only one HESID index which contains distinct patient keys mapped to HESID for all years' activity for APC (inpatient), OP (outpatient) and A&E (accident and emergency).

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 can never be matched with any other record so it is assigned its own unique HESID value and stored in a separate table (PATIENT HESIDS UNMATCHED).

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.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

3 Reasons for the update

It was previously recognised that there was potential to improve the HESID algorithm (HESIDv1) to make it better at matching the same patient. Of particular concern was the fact that the coverage of NHS Number deteriorates further back through time. This makes it appear that there is an increase in activities that require record linkage (such as emergency readmissions) where in actual fact it is simply that the patient identifiable fields required for linkage are becoming more complete.

Table 1: The proportion of inpatient HES records with valid NHS Numbers

Year	Proportion of Finished episodes where NHS Number is valid (%)	Proportion of Finished episodes where NHS Number is not valid (%)	Proportion of Finished episodes where the NHS Number was not known (%)
2008-09 (M12)	96.8%	0.0%	3.2%
2007-08	96.7%	0.0%	3.2%
2006-07	96.4%	0.0%	3.6%
2005-06	95.7%	0.0%	4.2%
2004-05	95.4%	0.0%	4.6%
2003-04	93.6%	0.0%	6.3%
2002-03	90.5%	0.0%	9.5%
2001-02	86.5%	0.0%	13.4%
2000-01	83.2%	0.2%	16.5%

HESIDv1 was a simple integer value used to represent a single patient. It was unique to HES and because it was a pseudonymised value, it did not contain any personal information about a patient. It could therefore not be unscrambled or unencrypted to reveal any identifiable details about the patient, such as date of birth or postcode.

However, information governance concerns were raised by PIAG (the Patient Information Advisory Group – now ECC, the Ethics and Confidentiality Committee) that HESIDv1 had become a patient identifiable field, following an audit into the processes around deriving and storing the HESID. This audit found that the required security standard was not sufficiently met. In particular, all customers received the same HESID, so different customers could potentially link their extracts together using HESID to obtain more information around a given patient.

This was addressed by encrypting the HESID to 256 bit standard and creating a different group pseudonym for each customer of a HES extract, so that different customers cannot link their HESIDs together. Each customer who requests an extract of HES data will receive a unique version of the HESID called EXTRACT_HESID.

This enforced change provided an opportunity to implement the improvement to the matching algorithm from 2 passes to 3 passes at the same time, addressing data quality issues mentioned earlier in the section around NHS Number.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

4 Comparison of the old and new algorithm

The new algorithm utilises an additional step in order to link records for the same patient which may have been missed by the old algorithm. The old algorithm used two passes of the data to link patients together, the new algorithm uses three.

The old two-pass algorithm used to derive HESIDv1 matched records with exactly the same NHS Number, sex and a partial match of DOB on the first pass. The second pass links additional records if they have a partial DOB match and an exact match of postcode (PCODE), local patient ID (LOPATID) and provider code (PROCODET). The second pass helps to link additional records together where there are data quality issues, such as missing NHS Number, or input errors with the source data (eg month and day switched in the date of birth).

The new three-pass algorithm used to derive HESIDv2 adds an additional step to the first two which looks at an exact match of sex, date of birth and postcode. This extra step helps to link patients who have attended a number of providers and may have a multitude of provider codes and local patient IDs. Records that match on the third pass are assigned the same HESID, unless they contain different NHS Numbers.

In outline, 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	Sex	(Exact match)
• DOB	Date of Birth	(Partial match)
• NEWNHSNO	NHS Number	(Exact match)

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

• SEX	Sex	(Exact match)
• DOB	Date of Birth	(Partial match)
• HOMEADD	Postcode	(Exact match)
• PROCODET + LOPATID	Local Patient Identifier within Provider	(Exact match)

The third step is driven by Date of Birth and will not consider two records for matching if the NHS Numbers differ, except when one record has a null NHS Number and the other record has a non-null NHS Number. It requires matches on the following fields:

• SEX	Sex	(Exact match)
• DOB	Date of Birth	(Exact match)
• HOMEADD	Postcode*	(Exact match)

* step 3 will not match records if the postcode is on a list of excluded postcodes. These postcodes include those of communal establishments such as hospitals, prisons and army barracks which cover a large range of potential patients.

Note: See Annex C for the full technical specification for the matching algorithm and Annex D for further information on partial matching and other data quality constraints applied when matching data.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

5 Worked example

The following table of fabricated data illustrates a simple isolated example of how the matching process works using some randomly generated data. The coloured cells show where a match has taken place within that pass (columns labelled Pass 1 match, Pass 2 match and Pass 3 match) and the rows that have been matched together (columns labelled Patient No.). For example rows 2 and 3 match on the first pass, row 1 then matches these on the second pass, and row 4 matches on the third pass. So for these 4 rows of data, the first pass identifies 3 patients, the second 2 and the final pass recognises that this is a single patient.

Table 2: Example of the 3-pass matching process with fabricated data

Record Number	NHS Number	Sex	Date of Birth	Postcode	Provider code	Hospital Patient ID (or PAS Number)	Patient No. after Pass 1	Patient No. after Pass 2	Patient No. after Pass 3	Notes
1	4085292717	1	13/04/1932	AC2 9BD	PROV_1	PEMH	1	2	2	
2	4085292717	1	13/04/1932	ID6 6PP	PROV_2	BALK	1	2	2	
3		1	13/04/1932	AC2 9BD	PROV_1	PEMH	2	2	2	
4	4131182365	1	13/04/1932	AC2 9BD	PROV_3	RGRF	3	3	2	
5	4551472733	1	13/09/1951	FH12 5KQ	PROV_4	HAMK	4	4	4	
6	4551472733	1	13/09/1951	CG11 9PH	PROV_5	PLRQ	4	4	4	
7	4680003619	1	20/04/2000	GB13 0QI	PROV_6	AMED	5	5	5	
8	4680003916	1	20/04/2000	GB13 0QI	PROV_6	AMED	6	5	5	
9		1	20/04/2000	GB13 0QI	PROV_7	QOTM	7	7	5	
10	4963118151	2	30/09/1942	RI19 4GI	PROV_8	MSRJ	8	8	8	
11		2	30/09/1942	RI19 4GI	PROV_9		9	9	8	
12	4126598182	2	30/09/1942	RI19 4GI	PROV_10	IQMM	10	10	8	
13	4983293268	2	30/09/1942	RI19 4GI	PROV_10	IQMM	11	10	8	
14	4070622473	2	16/01/2026	DP1 1GC	PROV_11	LEND	12	12	12	Not matched on any pass
15	4248864930	2	29/08/1994	EI9 6IS	PROV_12	JIEF	13	13	13	Not matched on any pass
No. of distinct patients							13	11	6	

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

6 How does the introduction of the new HESID affect me?

The switch from HESIDv1 to HESIDv2 does not change any of the underlying data in HES apart from the HESID itself. For example, if there are 15 million episodes within a current year's dataset with HESID as the patient index, then there will still be 15 million episodes using the new index. This will be the same for the number of admissions and discharges. What will change however is the number of individual patients within a given dataset.

6.1 Users of the HES interrogation system (Business Objects)

If you mainly query HES using Business Objects to generate aggregate reports, for example hip replacements by hospital site, it is likely that you will be unaffected by this change. However, if you have used counts of patients in an analysis rather than counts of admissions, for example, you may need to re-query Business Objects. This is because the switch from HESIDv1 to HESIDv2 will decrease the overall number of individual patients, as more patients are matched together. Section 7 shows some key totals that demonstrate the magnitude of the impact.

HESIDv1 has been removed from all Business Objects universes and HESIDv2 has been applied to all current and historic data.

To avoid confusion, the new HESID has been called PSEUDO_HESID in Business Objects.

6.2 Extracts

Each customer of the extract service who requests HESID will receive a HESID pseudonymised in a key that is unique to them called the EXTRACT_HESID. This means that you will not be able to link your extract to another customer's extract via their EXTRACT_HESID. However, if you request multiple extracts of HES data including HESID at different times, you will be able to receive all of your extracts pseudonymised to the same key as long as you quote the previous extract's reference number in the relevant place on your application form.

If you query or request extracts of patient level information but do not use the HESID to link activity records together for the same patient, or if you look at a single year's worth of information in isolation, you probably will not be affected.

If you use patient level information to link patient records together and you locally hold and use historic data you will be affected as the HESID you hold will not be compatible with any new data you may request and you will not be able to link activity for the same patient from your historic data to the new data.

In this case, you may request an update to your existing data which will map the new EXTRACT_HESID to the unique record number held within your dataset. This will give you a consistent patient index across all years' data you hold. You will be asked to sign and return an Information Governance declaration form, stating that you will not attempt to cross match the HESID with the EXTRACT_HESID, before any new data will be issued.

A link to this form is available from the HESonline website [http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=1090].

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

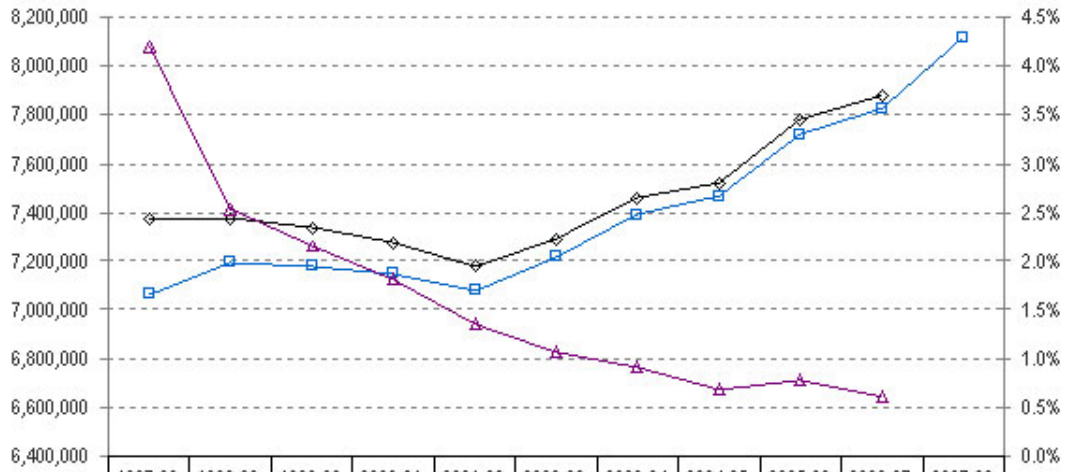
7 Impact of the change on patient numbers

7.1 Total patient counts

As has been stated earlier, the change from HESIDv1 to HESIDv2 has increased the level of matching, meaning more records can now be attributed to the same patient. This also has the effect of decreasing the total number of individual IDs, which means fewer unique patients.

Chart 1 shows the total count of patients admitted to hospital is higher under the two-pass algorithm than the three-pass algorithm. It also shows that the difference between the two total counts decreases through time, so the impact of the revised HESID is greater in earlier years. This can be explained by the improvement in NHS Number coverage through time, meaning that more legitimate matches are made under the first pass of the algorithm, so fewer patients are matched on the third pass.

Chart 1: Inpatient count with 2- and 3-pass HESID applied

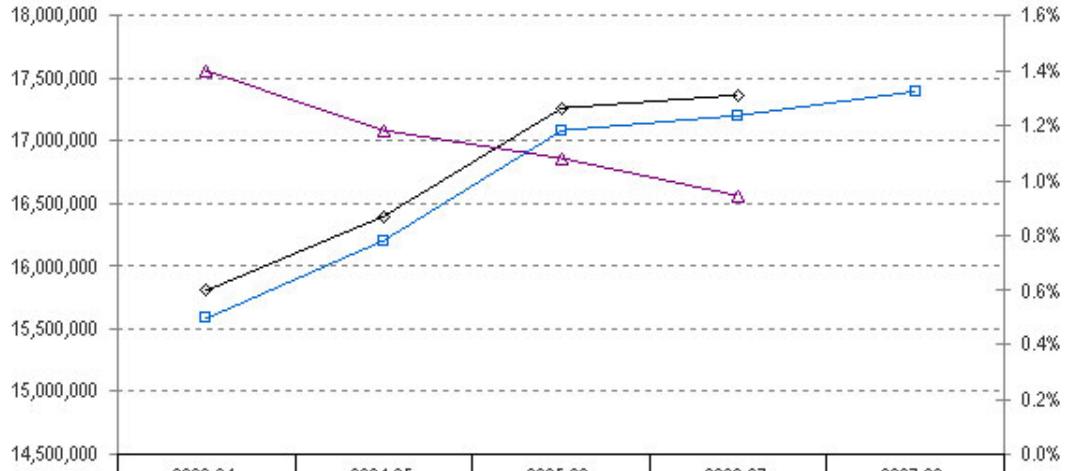


	1997-98	1998-99	1999-00	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08
Total 2 pass IP	7,375,732	7,378,776	7,336,977	7,278,336	7,176,568	7,290,151	7,459,391	7,519,545	7,777,691	7,876,083	
Total 3 pass IP	7,065,949	7,191,150	7,178,487	7,146,133	7,078,838	7,212,735	7,390,632	7,467,385	7,717,418	7,828,606	8,118,131
IP % reduction	4.2%	2.5%	2.2%	1.8%	1.4%	1.1%	0.9%	0.7%	0.8%	0.6%	

Chart 2 shows a similar pattern for the outpatient dataset. There are fewer patients counted under the three-pass than the two-pass algorithm, but with the gap between the numbers under the two algorithms narrowing in more recent years.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

Chart 2: Outpatient count with 2- and 3-pass HESID applied



	2003-04	2004-05	2005-06	2006-07	2007-08
Total 2 pass OP	15,807,606	16,402,450	17,258,826	17,356,788	
Total 3 pass OP	15,586,744	16,209,084	17,073,118	17,193,105	17,398,261
OP % reduction	1.4%	1.2%	1.1%	0.9%	

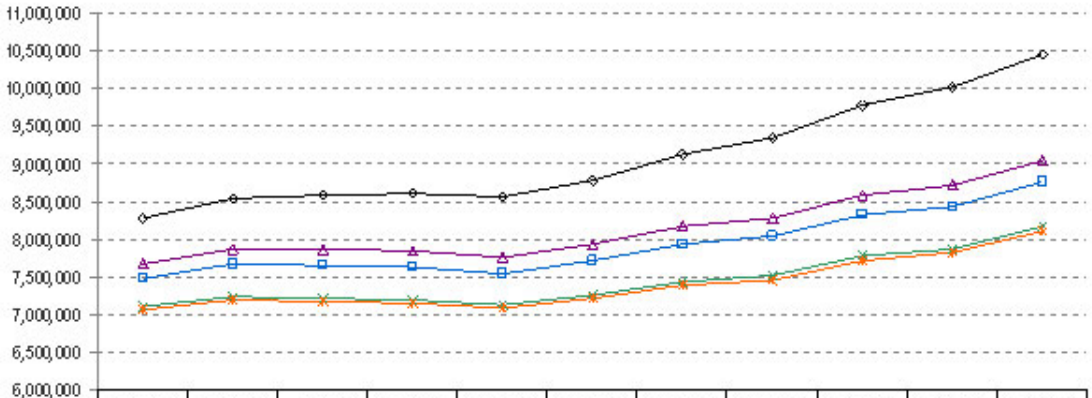
7.2 Patient counts subdivided by key variables

Further analysis has been performed that examines the total distinct combinations of HESID and another variable. This has been done separately for four variables: main specialty, provider, age and PCT. Therefore, each HESID is counted at least once, but may be counted multiple times if they have different values of the variable for the same HESID within the year (using the example of provider, where a patient who has been seen at several providers).

Chart 3 shows that there is a general year-on-year increase in the total number of admitted patients counted by the three-pass algorithm for each of the variables. The highest counts are for specialty. This shows that patients who have multiple episodes of treatment are often treated in different specialties. Age is the next highest, which is explained by patients having a birthday between hospital episodes. Provider is low still, which suggests that most patients who undergo several episodes of care are treated by the same provider. The total for PCT is just higher than the total number of patients without any other variables. This is to be expected as it will only be patients who moved house or changed GP practice across a PCT boundary during the course of the year and had a hospital episode either side of the boundary.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

Chart 3: Inpatient counts aggregated to key variables

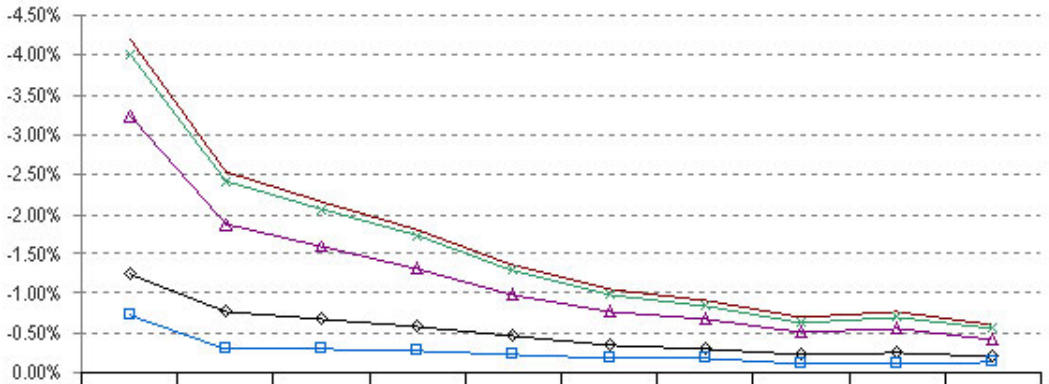


	1997-98	1998-99	1999-00	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08
Main Specialty	8,280,196	8,546,412	8,583,309	8,600,205	8,561,753	8,784,893	9,120,498	9,340,052	9,774,162	10,022,559	10,451,129
Provider	7,488,874	7,670,220	7,660,973	7,625,328	7,552,598	7,711,851	7,926,470	8,033,989	8,318,506	8,487,465	8,756,012
Age	7,672,542	7,867,898	7,866,176	7,836,990	7,764,807	7,934,616	8,178,688	8,278,967	8,580,681	8,709,374	9,042,169
PCT	7,102,705	7,232,593	7,216,483	7,186,153	7,120,112	7,260,147	7,444,246	7,527,717	7,776,578	7,878,756	8,177,284
Total	7,065,949	7,191,150	7,178,487	7,146,133	7,078,838	7,212,735	7,390,632	7,467,385	7,717,418	7,828,606	8,118,131

Chart 4 shows that for all variables the difference in the impact of the two- and three-pass HESID diminishes through the years. This is to be expected with improved NHS Number coverage meaning that more matches will be made on the first pass.

The lowest differences between the two- and three-pass HESID are for permutations with provider. This is to be expected, as where the patient is seen at the same provider, both pass one and pass two can be used to match the patient, so the two-pass algorithm has a greater chance of matching the same patient successfully. This shows that the three-pass algorithm is important for matching a patient who is seen at several providers.

Chart 4: Reduction of inpatient count following implementation of 3-pass HESID



	1997-98	1998-99	1999-00	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07
Main Specialty	-1.25%	-0.78%	-0.68%	-0.59%	-0.46%	-0.36%	-0.31%	-0.24%	-0.26%	-0.20%
Provider	-0.74%	-0.31%	-0.30%	-0.29%	-0.24%	-0.19%	-0.19%	-0.13%	-0.13%	-0.13%
Age	-3.22%	-1.87%	-1.59%	-1.32%	-1.00%	-0.77%	-0.68%	-0.51%	-0.56%	-0.43%
PCT	-4.01%	-2.42%	-2.06%	-1.73%	-1.29%	-1.00%	-0.85%	-0.63%	-0.71%	-0.56%
Total	-4.20%	-2.54%	-2.16%	-1.82%	-1.36%	-1.06%	-0.92%	-0.69%	-0.77%	-0.60%

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

8 How was the new algorithm determined?

The objective of the HESID is to make as many legitimate matches as possible while avoiding false positive matches. The nature of HES means that the data can be used in many different ways by different customers. These different uses are likely to attach different weights to the relative importance of making legitimate matches and avoiding false positives. However, a single solution is needed in HES[1]. For certain uses there are much greater risks in falsely matching two episodes to the same patient than failing to match two episodes to the same patient that should have been matched. Therefore, a conservative algorithm was employed to minimise the risk of false positives.

As outlined in section 4, HESIDv1 used a two-pass algorithm. It was decided not to alter the first two passes, but to add an additional pass to the algorithm.

The two core options that were considered were a third pass based on:

Option A:

• SEX	Sex	(Exact match)
• DOB	Date of Birth	(Exact match)
• HOMEADD	Postcode	(Partial match – on outward[2] portion of postcode only)

Option B:

• SEX	Sex	(Exact match)
• DOB	Date of Birth	(Exact match)
• HOMEADD	Postcode*	(Exact match)

8.1 Option A

Option A had been implemented in 2005 in HES, but had been quickly aborted due to the high number of false positive matches that were generated. A simple calculation shows that there are only around 200 million unique permutations of sex, date of birth and home address, which is only around four times higher than the English population of around 50 million. Assuming independence between the three variables, there would be an expected number of over ten million people in England who would share the same permutation of these three variables with someone else. This would create an unacceptably high number of false positive matches on HESID. In actual fact, the number would be even higher as the three variables would not be entirely independent (certain age groups would be more likely to live in certain postcode districts).

8.2 Option B

Option B was explored in more detail. Firstly the impact on the entire HES dataset up until 2006-07 of matching on the third pass under option B was ascertained. This showed that there were 9% fewer distinct HESIDs when matching under option B in comparison with the two-pass algorithm (47.6 million distinct HESIDs compared with 52.3 million). This is considerably higher than the impact of the third pass on individual years' data presented in section 7. This is largely because patients in different years who had different HESIDs under the two-pass algorithm could be given the same HESID under the three-pass algorithm.

The next stage was to estimate the number of these additional matches that are false positive matches. It is not possible to fully validate whether two records belong to the same patient or not within the HES dataset in order to determine the probability that matches on a particular subset of fields relate to the same patient. Some external validation would be required, which would be very costly and involve patient identification, so this approach was avoided. Therefore, a different approach is required to estimate the number of false positives that would be generated by various algorithms.

[1] Linkages between different datasets can have an additional field that indicates the quality of the match between the two records enabling customers to set their own threshold for quality of linkage (for example the HES ONS mortality linkage). However, such a solution cannot be employed for the HESID, which is linking a dataset to itself, as there would often be more than two records linked to the same patient and there may be differences in quality in the different pairs of linkage for the same patient that cannot be described by a single figure.
[2] An outward portion of postcode is the first part ie LS1 or W12.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

In most cases, two records in which the patient has the same gender, date of birth and postcode are caused either by the fact that the patient is the same or by the fact that due to coincidence two patients have the same date of birth, gender and postcode[3]. The probability of two records having the same postcode, date of birth and gender due to chance would be very similar to the probability of the two records having the same postcode and gender, but with the first record having a date of birth exactly 17 days higher than the other date of birth.

Therefore, the approach taken was to create a new field which added exactly 17 days to the date of birth. The number of matches between ([postcode]&[DOB]&[gender]) and ([postcode]&[DOB]+17&[gender]) was then assessed. This number was then halved as there are two ways that a pair of records can differ by 17 days (record A being 17 higher than record B or vice versa).

Under this methodology 186,000 records were returned from the index. Dividing this by two gives an estimate of 93,000 records that shared gender, postcode and date of birth due to coincidence. This is a very small proportion (around 2%) of the 4.7 million fewer distinct HESIDs under the three-pass algorithm. There will be other types of false positive matches, such as cohabiting twins, which will not be due entirely to chance and will make the actual proportion of false positive matches slightly higher.

While this is a fairly low number in comparison to the size of the dataset, the number will not be insignificant for all purposes. Therefore, it was decided to add two further safeguards to the algorithm:

1) Further analysis on the matching on date of birth plus 17 days showed that a relatively small number of postcodes were responsible for a large proportion of the false positive matches. These typically related to hospitals' own postcodes, army barracks or other communal establishments where a large number of residents of similar ages would be expected. It was decided to prevent any postcode that created more than ten matches on the date of birth plus 17 analysis from being able to create a match under the third pass.

2) The combination of postcode, date of birth and gender cannot override NHS Number, so in cases in which two records have a different NHS Number they will not be assigned the same HESID, even if they matched on postcode, date of birth and gender.

[3] There is also the case of twins and the use of a default date of birth such as January 1 where the patients would be different, but this difference could not be modelled using the date of birth + 17 day approach.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

9 User feedback

We would welcome feedback from users on this document, including suggested improvements and additions.

Please address all feedback to: HES.questions@ic.nhs.uk

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

Annex A: Glossary

EXTRACT_HESID - The new version of the HESID (ie HESIDv2) as viewed by HES users who request an extract of data using the extract service available from HESonline.

ECC - Ethics & Confidentiality Committee (formerly PIAG). An organisation accountable to the National Information Governance Board for Health and Social Care (NIGB) which was set up to consider ethical and information governance issues pertaining to the use and distribution of identifiable health and social care data.

HES - Hospital Episode Statistics. The central repository of NHS activity information including admitted patient care (APC), outpatient attendances (OP) and accident & emergency attendances (A&E).

HESIDv1 - The old version of the HESID based on a two-pass algorithm.

HESIDv2 - The new version of the HESID based on a three-pass algorithm.

HESID INDEX - The central reference table which maps individual patient keys to HESID.

PATIENT KEY - A combination of personally identifiable information used to derive the HESID. The patient key consists of a combination of NHS Number, sex, date of birth, postcode, provider code and local patient identifiers (PAS or case note numbers).

PIAG - Patient Information Advisory Group (superseded by ECC – see above).

PSEUDO_HESID - The new version of the HESID (ie HESIDv2) as viewed by HES users who query the data using the HES Interrogation System (Business Objects).

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

Annex B: Table details

Table 3: Details of tables referenced in the document	
Table name	**Table description**
DISTINCT_APP_HESIDS The primary key of this table is PATIENT_KEY	This is a working table which is populated with valid PATIENT_KEYS from source data (eg episodes for APC) for merging into the HESID Index.
HESID_POSTCODES_TO_IGNORE The primary key of this table is HOMEADD (ie postcode)	Postcodes of prisons and other institutions. These postcodes are ignored during third pass matching of source data to the HESID Index.
INVALID_APP_LADS The primary key of this table is source key, eg EPIKEY for APC data	This is a working table which is populated with invalid PATIENT_KEYS from source data (eg episodes for APC) for use when creating rows in PATIENT_HESIDS_UNMATCHED.
PATIENT_HESID_ACTIVITY The primary key of this table is HESID	Latest activity dates for HESIDs.
PATIENT_HESIDS The primary key of this table is PATIENT_KEY	The 'HESID Index' - the master index of PATIENT_KEYs mapped to HESIDs. One row per PATIENT_KEY. The same HESID can appear on one or more PATIENT_KEYs.
PATIENT_HESIDS_NEW The primary key of this table is HESID_FROM (ie the old HESID)	Any HESID which has been superseded by another HESID appears in this table, together with the most up-to-date HESID (see note below).
PATIENT_HESIDS_UNMATCHED The primary key of this table is HESID	All PATIENT_KEYs which are excluded from the standard HESID matching, due to poor quality. One row per instance of each poor quality PATIENT_KEY, of which each has a unique HESID taken from the stack.

Note: This does not map HESIDv2 to HESIDv1, rather it records changes in HESIDs after each run of HESID processing where new data (a new patient key) from a new HES data submission may act as a bridge record to cause two previously unmatched HESIDs to match together.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

Annex C: Algorithm technical specification

The following is a detailed breakdown of the algorithm used to match patients and generate the HESID.

Main procedure

```
BEGIN

      /*
      Extend the Patient HESID Index
      */

      FOR each activity record in the current data set
      DO
         IF    DOB (date of birth) is valid
         AND   Sex is valid and is known ('1' or '2')
         AND   NHS Number is valid (see Note 2)
         THEN
             Set Patient Match Key 1 to the combination of DOB, Sex and NHS
         Number (is this a direct concatenation?)
         ELSE
            Set Patient Match Key 1 to Null
         FI

         IF    DOB (date of birth) is valid
         AND   Sex is valid and is known ('1' or '2')
         AND   Postcode is valid
         AND   Local Patient Identifier is not null (see Note 4)
         THEN
            Set Patient Match Key 2 to the combination of DOB, Sex,
            Postcode, Provider Code (PROCODET), and Local Patient
            Identifier
         ELSE
            Set Patient Match Key 2 to Null
         FI

         IF    DOB (date of birth) is valid
         AND   Sex is valid and is known ('1' or '2')
         AND   Postcode is valid
         THEN
            Set Patient Match Key 3 to the combination of DOB, Sex, and
            Postcode
         ELSE
            Set Patient Match Key 3 to Null
         FI

         Set Patient Match Key for the activity record to the combination
         of Patient Match Key 1, Patient Match Key 2, and Patient Match Key
         3 (this is again a concatenation of the three variables above? So
         it may potentially match on all three, and the match on all three
         would become an identifier, not just the key with the highest
         place in the hierarchy?)

         IF    the Patient Match Key is not Null
         AND   the Patient Match Key is not already record in the Patient
         HESID Index
         THEN
            Record the Patient Match Key in the Patient HESID Index with a
            new HESID
         FI
      OD
```


Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

```
      /*
      Matching Step 1
      */

      FOR each unique non-null combination of Sex and NHS Number in the
          Patient HESID Index
      DO
         Retrieve from the Patient HESID Index a set of all records with
         this NHS number and Sex, ordered by HESID
         Apply Matching to the members of the retrieved set as described by
         Matching (Steps 1 and 2) below
      OD

      /*
      Matching Step 2
      */

      FOR each unique non-null combination of Sex, Postcode, Provider, and
          Local Patient Identifier in the Patient HESID Index
      DO
         Retrieve from the Patient HESID Index a set of all records with
         this Sex, Postcode, Provider, and Local Patient Identifier,
         ordered by HESID
         Apply Matching to the members of the retrieved set as described by
         Matching (Steps 1 and 2, a little confused as this is Step 2?)
         below
      OD

      /*
      Matching Step 3
      */

      FOR each unique non-null combination of Sex, Date of Birth, and
          Postcode (passing the additional step 3 postcode criteria) in
          the Patient HESID Index
      DO
         Retrieve from the Patient HESID Index a set of all records with
         this Sex, Date of Birth, and Postcode, ordered by HESID
         Apply Matching to the members of the retrieved set as described by
         Matching (Step 3) below
      OD

      /*
      Assign HESIDs to the Activity Records
      */

      FOR each activity record with a non-null Patient Match Key
      DO
         Use the Patient Match Key to assign a HESID to the activity record
         from the Patient HESID Index
      OD

      FOR each activity record with a null HESID
      DO
         Assign a new HESID to the activity record
      OD
END

Matching (Steps 1 and 2)

/*
Apply matching to a Set of Patient HESID Index records
*/
```


Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

```
BEGIN

      IF    the DOB for every record in the Set is 1901/01/01
      OR    the DOB for every record in the Set is 1899/12/31
      THEN
         All of the records in the Set match, and are assigned the same
         HESID value as the first record in the set

         FOR each replaced HESID
         DO
            Assign the HESID value from the first record to every record in
            the Patient HESID Index with the replaced HESID
         OD
      ELSE
         FOR each record in the Set (record 1)
         DO
            FOR each later record in the Set (record 2)
            DO
               IF the DOBs (exactly or partially) match (see Note 5)
               THEN
                  Assign the HESID value from record 1 to every record in
                  the set with the same HESID as record 2
                  Assign the HESID value from record 1 to every record in
                  the Patient HESID Index with the same HESID as record 2
               FI
            OD
         OD
      FI
END
```

Matching (Step 3)

```
/*
Apply matching to a Set of Patient HESID Index records
*/
BEGIN
      FOR each record in the Set (record 1)
      DO
         FOR each later record in the Set (record 2)
         DO
            IF    the Postcodes exactly match and fulfil the step 3
            postcode criteria
                  AND   the dates of birth exactly match
                  AND the sex exactly matches
                  AND match key 1 is null for either or both of record 1
                  and record 2
            THEN
               Assign the HESID value from record 1 to every record in the
               set with the same HESID as record 2
               Assign the HESID value from record 1 to every record in the
               Patient HESID Index with the same HESID as record 2
            FI
         OD
      OD
END
```


Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved

Annex D: Notes on valid data values

To create the patient keys and the HESID, the algorithm compares a range of fields. To aid data quality, the data are 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 partition area of the index called the 'unmatched table' where each instance of this poor quality patient key that occurs in the source data is given a unique HESID 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. See the NHS Number Check Digit Calculation file on the Connecting for Health website for more details.

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, there are a number of further criteria that 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. This list is reviewed regularly to ensure that it is kept up to date.

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 (ie YYYY, MM, 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.

Copyright © 2009, The Health and Social Care Information Centre. All Rights Reserved