The HES Processing Cycle and Data Quality
## Document Management

### Revision History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Summary of Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>24/02/2014</td>
<td>Reviewed &amp; updated to reflect 2013-2014 processing onwards</td>
</tr>
<tr>
<td>3.0</td>
<td>18/08/2015</td>
<td>The role of system integrator from BT has moved to the HSCIC.</td>
</tr>
<tr>
<td>4.0</td>
<td>26/09/2016</td>
<td>The corporate transformation from the Health and Social Care Information Centre to NHS Digital.</td>
</tr>
</tbody>
</table>

### Reviewers

This document must be reviewed by the following people:

<table>
<thead>
<tr>
<th>Reviewer name</th>
<th>Title / Responsibility</th>
<th>Date</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>HES Data Quality Team</td>
<td>Secondary Care</td>
<td>26/09/2016</td>
<td>4.0</td>
</tr>
</tbody>
</table>

### Approved by

This document must be approved by the following people:

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Title</th>
<th>Date</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julie Clough</td>
<td></td>
<td>Principal Information Analyst</td>
<td>26/09/2016</td>
<td>4.0</td>
</tr>
</tbody>
</table>

### Document Control:

The controlled copy of this document is maintained in the NHS Digital corporate network. Any copies of this document held outside of that area, in whatever format (e.g. paper, email attachment), are considered to have passed out of control and should be checked for currency and validity.
Contents

Part One: the HES Processing Cycle

Overview of Process
Overview Diagram of Process
Data collection
Monthly Provisional Data
Annual Refresh / Final Year Data

Part Two: HES Data Quality

Overview of the cleaning process
Provider organisation code mapping
Derivation rules
Duplicate detection and removal process
HES ID
Decodes
Feedback and Approval
Causes of Data Quality issues
HES Data Quality Notes
HES Data Quality team
Contacting the HES Data Quality Team
Part One: the HES Processing Cycle

Overview of Process

HES data for admitted patients, outpatients and A&E 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 pre-arranged dates during the year, SUS takes an extract from their database and sends it to HES. Data on SUS will continue to change, but HES data is fixed as it was when that particular extract was taken. This is why there are likely to be differences between analyses from the SUS and those from HES.

The HES data quality team then validates and cleans the extract, before deriving new items and making the information available in the data warehouse. Data quality reports and checks are completed at various stages in the cleaning and processing cycle.
Overview Diagram of Process

The diagram above is a simplified version of how HES is extracted. It can be broken down as follows:

**PAS** (Patient Administration System): system used by trusts to submit data to SUS.

**XML** (Extensible Markup Language): standard specification used for submitting and validating data to SUS.

The core elements of SUS and HES are founded on Commissioning Data Sets (CDS).

Go to the Commissioning Data Sets
SUS (Secondary Uses Service): a data warehouse used to store information.
Examples of extracts that are created from this data warehouse:

- **PbR** (Payment by Results)
- **SEM** (SUS Extract Mart)
- Read more about SUS

**SEM** (Standard Extract Mart) is a data mart that contains submitted patient activity. It is the source for **HES** data.

**Raw SUS Data**: This is raw HES data straight from SUS. The data is cleaned and additional fields are derived.

**Published HES Data**: Cleaned data is stored here, description fields are added and the data is made available to HES users.

The HES Data Quality team clean the data and create the Published HES Data from the Raw SUS Data.
**Data collection**

Initially data for HES publications was collected annually from Provider submissions. After a number of years the frequency of collections increased to quarterly to allow analysis and investigation (these were not published) and a final annual publication was released at the end of the year.

These collections within the year are extracted directly from SUS (Secondary Uses Service). They are cumulative and reflect any changes that have been made to the central data warehouse.

**Monthly Provisional Data**

Since April 2008, at pre-arranged dates during the year, on a monthly basis, HES extracts are taken from the SUS data warehouse. HES represents a series of fixed positions aligned to extracted data, whereas SUS is continuously updated whenever data is submitted. This is why there can be differences between SUS and HES even when looking at the same time period.

Each extract is cumulative and contains data submitted for the financial year so far, i.e. Month 1 will only contain the data submitted with an activity date in April, but Month 6 will contain data submitted with an activity date from April to September. One of the reasons for this is that additional data may be needed to update patient records from earlier in the year, e.g. an episode may potentially run for several months or an amendment may need to be made as clinical coding takes place on discharge.

**Annual Refresh / Final Year Data**

Each Final Year data publication of HES collects and holds data based on a financial year. At the end of each year, after the 12 monthly submissions, there is an additional submission date to support what is called the Annual Refresh. This gives providers the opportunity to revise and update their submissions for the year.

A provisional extract is generated from the Annual Refresh submission and is referred to as the “Month 13” extract. This allows HES users an early provisional view of the Final Year data before final publication.

The final publication is subject to consultation and providers are given the opportunity to amend provider organisation code mapping or the removal of any duplicate records.

While data relating to episodes and spells for a particular year can be amended and updated in SUS long after the year has passed, no further SUS updates are applied to HES which is fixed after Final Year data publication.

Submission timetable of HES for providers can be found here: [http://content.digital.nhs.uk/sus/pbrguidance](http://content.digital.nhs.uk/sus/pbrguidance)
Part Two: HES Data Quality

To make HES data usable and more accessible, we clean the data as detailed below.

Overview of the cleaning process

Provider organisation code mapping

During this stage any old or invalid provider organisation codes are changed or merged to the correct provider code using reference data based on information from the Organisation Data Service (ODS) website.

- Read more about ODS
- Provider code mapping methodology

Example

Provider A merged with Provider B to form Provider C in 2011.

Where any data submissions are showing A or B as a Provider, these are mapped to Provider C.

Where no appropriate Provider code can be found, this data is deleted during processing.
Derivation rules

Cleaning and derivation rules for HES have been developed over time and continue to evolve to enhance the data set further. The rules can be found at http://content.digital.nhs.uk/article/1825/The-processing-cycle-and-HES-data-quality and are also available in the HES Data Dictionary. http://content.digital.nhs.uk/hesdatadictionary

These rules have two main purposes:

- To clean common and obvious data quality errors (Correction and Validation).
- To derive additional data items to populate the HES dataset (Derivation).

Example of Correction rules

Rule #0150 looks for evidence where a Birth Episode (CDS type 120) has been incorrectly submitted to SUS as a General Episode (CDS type 130). If evidence is found then the Episode Type of the record is altered to reflect this.

**Rule 150: Epitype reset to 3**

When the episode type is not coded as an NHS hospital birth record, the admission method (admimeth), date of birth (dob), episode order (epiorder) and episode start date (epistart) are examined to see whether they indicate that the record is a birth record.

If so, the episode type (epitype) is changed to reflect this.

For all records:
If epitype = 1,2,4
and dob and epistart <> null and dob = epistart
and admimeth = 82
and epiorder = 01
Set epitype = 3

Without this clean, the number of birth records (Episode Type 3) in HES would tend to be lower than the actual number of births taking place. This improves the value of HES as a statistical data source. This is another example of why there can be differences between SUS and HES data.
Example of Validation rules

There are a number of validation rules to collect and classify field values that are not recognised against established values. Rule #090 checks the Main Speciality (mainspef) field for unrecognised values and changes to a value for not known (&).

Rule 90: Mainspef validation
If the main specialty is null or contains an invalid entry, it is overwritten with the appropriate value for not known.

For epitype = 1, 2, 3 or 4:
Overwrite invalid mainspef with ‘&’

The clean makes it easy to filter out values that are not known from any analysis for this field. Without the clean the user would need to either list all the recognised values or search for all the different incorrect values that were submitted.

Examples of Derivation

Rule #1200 uses the postcode (homeadd) from each submitted CDS record to derive additional geographical data items relating to the episode of care.

Rule 1200: Postcode Derived Items
homeadd (the full postcode of the patients home address) is used to populate the derived fields giving other area classifications for the patient’s place of residence

Look homeadd up in the Postcodes table
If the Postcode is found
Then
Set the following items from the Postcode entry:

gridlink            ward91
currward            oacode
rescty              oacode6
resha               soal
resstha              soam
resptct             rururb_ind
resladst            resstha02
resro               pcon
Set the PCFOUND flag to Y
Else

gridlink = ‘Y’
curnward = ‘Y’
rescty = ‘Y’
resladst = ‘Y’
resgor = ‘Y’
respct06= ‘59999’
respct02 = ‘59999’

Set the PCFOUND flag to N

End If

HES uses reference data from the ONS Postcode Directory to derive data items such as Parliamentary Constituency. This allows record level data to be easily aggregated to enable effective spatial analysis to be performed. The below table shows the available geographic fields that this derivation creates:
<table>
<thead>
<tr>
<th>Field from SUS</th>
<th>Fields available in HES Output Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMEADD</td>
<td>Postcode of Usual Address</td>
</tr>
<tr>
<td></td>
<td>Postcode District</td>
</tr>
<tr>
<td></td>
<td>Cancer network</td>
</tr>
<tr>
<td></td>
<td>Cancer registry</td>
</tr>
<tr>
<td></td>
<td>Current electoral ward</td>
</tr>
<tr>
<td></td>
<td>Gridlink (Restricted)</td>
</tr>
<tr>
<td></td>
<td>Ocode 2001 Census Output Area (Restricted)</td>
</tr>
<tr>
<td></td>
<td>Ocode06 2001 Census Output Area (6 chars)</td>
</tr>
<tr>
<td></td>
<td>Pfound PC Found</td>
</tr>
<tr>
<td></td>
<td>Pcon Parliamentary constituency</td>
</tr>
<tr>
<td></td>
<td>Rescty County of residence</td>
</tr>
<tr>
<td></td>
<td>Resgor Govt Office region of residence</td>
</tr>
<tr>
<td></td>
<td>Resha Health authority of residence</td>
</tr>
<tr>
<td></td>
<td>Resladist LA district of residence</td>
</tr>
<tr>
<td></td>
<td>Respct02 Historic PCT of Residence</td>
</tr>
<tr>
<td></td>
<td>Respct06 Current PCT of Residence</td>
</tr>
<tr>
<td></td>
<td>Resro Region of Residence</td>
</tr>
<tr>
<td></td>
<td>Respsth02 Historic Strategic HA of Residence</td>
</tr>
<tr>
<td></td>
<td>Respsth06 Current Strategic HA of residence</td>
</tr>
<tr>
<td></td>
<td>Rural Ind Rural/Urban Indicator</td>
</tr>
<tr>
<td></td>
<td>Soal Lower Super Output Area</td>
</tr>
<tr>
<td></td>
<td>Soarn Middle Super Output Area</td>
</tr>
<tr>
<td></td>
<td>Imd04 IMD Decile Group</td>
</tr>
<tr>
<td></td>
<td>Imd04c IMD Crime domain</td>
</tr>
<tr>
<td></td>
<td>Imd04ed IMD Education, Skills and Training</td>
</tr>
<tr>
<td></td>
<td>Imd04em IMD Employment domain</td>
</tr>
<tr>
<td></td>
<td>Imd04hd IMD Health and Disability domain</td>
</tr>
<tr>
<td></td>
<td>Imd04hs IMD Barriers to Housing and Service</td>
</tr>
<tr>
<td></td>
<td>Imd04i IMD Income domain</td>
</tr>
<tr>
<td></td>
<td>Imd04ia IMD Income Affecting Adults domain</td>
</tr>
<tr>
<td></td>
<td>Imd04ic IMD Income Affecting Children domain</td>
</tr>
<tr>
<td></td>
<td>Imd04le IMD Living Environment domain</td>
</tr>
<tr>
<td></td>
<td>Imd04rk IMD Overall Rank</td>
</tr>
<tr>
<td></td>
<td>Ward91 Electoral ward 1981/1991</td>
</tr>
</tbody>
</table>

NB. Other fields are derived from the postcode field in other cleaning rules.
Duplicate detection and removal process

Data is sent to SUS using one of two different protocols:

- Net Change Interchange allows the sender to add, identify and replace or delete individual records.
- Bulk Update Interchange allows the sender to add or replace a set of records covering a specified period.

If the user intended to replace a pre-existing record or a set of records, but does not identify them correctly, the replacement records are added to the dataset, creating duplicated ones. This is referred to as a failure to overwrite and is what the HES duplicate detection and deletion process focuses on.

Since there are field variations in the formation of each of the HES collections (inpatient, outpatient and A&E) the process to detect duplicates in each collection is slightly different and full details are provided in the HES Duplicate Identification and Removal Methodology document published on this link.

Feedback and improvement

During the processing of monthly HES data all duplicates are removed from HES without further consultation.

Details of all deletions undertaken are shared with trusts so that if necessary they can request further data deletions from SUS to help ensure SUS remains accurate.
HES ID

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:

- It is derived from several different patient identifiers, so is more resilient to data quality or coverage problems affecting individual fields.

- It is a pseudonym field generated in HES processing so it minimises the risk of identification of patients.

A document containing more information about HESID and its methodology can found at http://content.digital.nhs.uk/article/1825/The-processing-cycle-and-HES-data-quality
Decodes

Decodes are the description fields in HES that provide text explaining what a field value means. These fields are generated by matching fields and their values with the appropriate reference data tables. Some examples of this:

<table>
<thead>
<tr>
<th>Sex</th>
<th>Sex Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not known</td>
</tr>
<tr>
<td>1</td>
<td>Male</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
</tr>
<tr>
<td>9</td>
<td>Not specified</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient Classification</th>
<th>Classpat Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ordinary admission</td>
</tr>
<tr>
<td>2</td>
<td>Day case admission</td>
</tr>
<tr>
<td>3</td>
<td>Regular day patient excluded from HES</td>
</tr>
<tr>
<td>4</td>
<td>Regular night patient excluded from HES</td>
</tr>
<tr>
<td>5</td>
<td>Mothers and babies using only delivery facilities</td>
</tr>
<tr>
<td>8</td>
<td>Other maternity event</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 char main operative procedure</th>
<th>3 char main operative description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01</td>
<td>Major excision of tissue of brain</td>
</tr>
<tr>
<td>G35</td>
<td>Operations on ulcer of stomach</td>
</tr>
<tr>
<td>J21</td>
<td>Incision of gall bladder</td>
</tr>
<tr>
<td>O10</td>
<td>Complex reconstruction of shoulder</td>
</tr>
<tr>
<td>R36</td>
<td>Routine obstetric scan</td>
</tr>
</tbody>
</table>
Feedback and Approval

An important part of the cleaning process is feedback. The HES Data Quality team liaise with providers on a range of subjects, including:

- Incorrectly submitted provider organisation codes.
- Missing data / gaps in coverage.
- Duplicate data.
- Specific queries when issues are found.

They also:

- Encourage providers to clean (correct) their data and resubmit to SUS before the next extract is taken, where possible.
- Offer advice where issues occur in the submission process.
- Recommend to providers, where necessary, that they request deletions of data from within SUS itself.

When processing the Annual Refresh of HES at the end of a year, the team seek approval from providers before removal of duplicates and provider code mapping.
Causes of Data Quality issues

Common causes of Data Quality issues in HES include:

- Problems with XML translation process in CDS submission to SUS.
- Poor recording in either patient notes or coding onto Provider system.
- Failing to meet submission deadlines leading to gaps in coverage.
- The NHS is not “one size fits all”, numerous local systems and configurations which can vary greatly in quality and ability to upgrade.

Where these issues are known within an HES extract, they are recorded in the HES Data Quality Note (see below)
HES Data Quality Notes

When HES is published, additional data quality information and explanatory notes are also made available. For HES system users and those requesting extracts, Data Quality Notes are published highlighting any specific known issues with the data to be considered when analysing the data.

HES Data Quality Note content:
The HES Data Quality Notes are updated for each extract published and are available from 2007-08 data year onwards. A typical note will contain:

- Issues with specific fields.
- Issue with fields for a specific provider.
- Coverage issues.
- Processing errors:
  - Mapping issues / Duplicates missed / Autoclean error
  - Errors at SUS Ops stage / XML.
- Anything else that is considered useful for users to be made aware of.

Where possible each entry will highlight the impact on analysis of the area it relates to and give a timescale of when this issue will be resolved.

Each note records the issues that are found at the time of publication. If additional issues are found post-publication, the note will be updated to reflect this.

The notes can be found on the HES website on the link below:

HES Data Quality team

The HES Data Quality team aims to:

- Produce cleaned data for use by all HES users.
- Encourage providers to understand that the data is their responsibility and that poor quality data has consequences for them.
- Provide guidance, feedback and assistance to ensure data quality begins at the source.
- Record known data quality issues in the HES Data Quality Notes.
- Maintain the HES Data Dictionaries.

Contacting the HES Data Quality Team

To contact the HES Data Quality team about a Data Quality query or the content of this document, please send an email for the attention of the team to the NHS Digital mailbox at enquiries@nhsdigital.nhs.uk.