Resource report

closer
The home of longitudinal research

Understanding Hospital Episode Statistics (HES)

Andy Boyd[1,2], Rosie Cornish[1], Leigh Johnson[1], Shirley Simmonds[3], Holly Syddall[3], Leo Westbury[3], Cyrus Cooper[3], John Macleod[1]

[1] Population Health Sciences, Bristol Medical School, University of Bristol
[2] CLOSER, Institute of Education, University College London
[3] MRC Lifecourse Epidemiology Unit, University of Southampton

April 2018

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

MRC Medical Research Council

To cite this report, please use the following reference:

Boyd A., Cornish R., Johnson L., Simmonds S., Syddall H., Westbury L., Cooper C., Macleod J. Understanding Hospital Episode Statistics (HES). London, UK: CLOSER; 2017. Available from: https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf

CLOSER
UCL Institute of Education
20 Bedford Way
London
WC1H 0AL
United Kingdom

Tel: +44 (0)20 7331 5102
Email: closer@ucl.ac.uk
Web: www.closer.ac.uk
Twitter: @CLOSER_UK
YouTube: CLOSER

Acknowledgements & Background

This project was funded as part of the initial collaborative research programme entitled 'Cohorts and Longitudinal Studies Enhancement Resources' (CLOSER); ESRC grant reference: ES/K000357/1. CLOSER is a consortium including eight of the UKs major cohort and longitudinal studies. The CLOSER network brings these teams together to:

- stimulate interdisciplinary research across the major longitudinal studies
- provide shared resources for research
- assist with training and development for researchers in the use of longitudinal data at all career stages
- share information and expertise in longitudinal methodology

This project – CLOSER work package 7 – aims to understand the potential for the NHS Hospital Episode Statistics dataset to enrich study data sets and inform longitudinal research. The project draws on information collected by the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Hertfordshire Cohort Study (HCS). ALSPAC is funded by the Medical Research Council, the Wellcome Trust and the University of Bristol. The collection of Hospital Episode Statistics in ALSPAC was funded by the Wellcome Trust (WT grant reference: WT086118/Z/08/Z). HCS is funded primarily by the Medical Research Council, with additional support from the British Heart Foundation, the Arthritis Research Campaign, the National Osteoporosis Society, the Wellcome Trust, and the University of Southampton.

Much of the information presented within this report has been distilled from NHS Digital (and predecessor organizations) guidance documents. This includes our adaptations of some technical diagrams and graphics. The source documents have been acknowledged where appropriate.

Citation of the Report:
Boyd A, Cornish R, Johnson L, Simmonds S, Syddall H, Westbury L, Cooper C, Macleod J. (2018). Understanding Hospital Episode Statistics (HES). Bristol, UK: University of Bristol.

Contents

Summary 5
Aims 6
The Hospital Episode Statistics database 7
HES data coverage and collection 9
Data Collection 11
Data format 12
Hospital admissions records (aka Admitted Patient Care, APC) 15
Outpatient records (OP) 16
Accident and Emergency records (A&E) 17
Adult critical care records (ACC) 19
Data Quality 22
Quality assurance processes 25
The reliability and validity of self-reported hospital admissions. 28
Case study 3: The Hertfordshire Cohort Study 33
Conclusions 35

Summary

The Hospital Episode Statistics dataset offers a comprehensive resource for inpatient admissions, outpatient appointments and Accident & Emergency attendance records in England. Yet concerns regarding data quality remain and this resource is currently underutilised by the cohort and longitudinal study community.

Cohort and longitudinal studies – with support from their strategic funders – are investing considerable efforts in establishing linkages to routine health records. This will enable studies to link self-reported data, study assessed physiological, mental health and developmental data along with genomic data to objectively assessed and recorded clinical data. Linkage also has the appeal of being of low participant burden, relatively low cost and scalable to large study populations. Within England, the Hospital Episode Statistics (HES) data set is a priority target within linkage strategies given that it provides a centralised repository of secondary care admission and appointment records within NHS hospitals and independent sector health care providers (where the treatment was commissioned by the NHS). This includes records of hospital admissions (including births and also speciality care such as that provided by psychiatric hospitals), records describing appointments at outpatient services and attendance at Accident & Emergency units. As such, HES offers considerable value to the longitudinal research community given its breadth of clinical information, its near-universal (in England) coverage and the fact that it provides a longitudinal resource stretching back to 1989.

While HES has been used extensively by the research community, its use to date in cohort and longitudinal studies has been relatively limited and is currently hindered by issues relating to securing data access, a lack of visibility and familiarity within some research communities, and concerns regarding data quality. While CLOSER are considering access to health records elsewhere, this work package seeks to promote awareness and understanding of the HES dataset by summarising its history, content and structure and then discussing issues relating to data quality. This emphasis on data quality reflects wider concerns regarding the use of routine records in a secondary context and the potential within this for introducing misclassification errors. These errors could relate to the process of ‘diagnosis’ (e.g. doctors failing to accurately diagnose conditions) and ‘recording’ (e.g. errors in coding, errors in identifying patients accurately, loss of data during copying, transmission and processing) and errors in interpreting evidence drawn in part from secondary data. These errors are compounded in longitudinal observational research where we also need to account for changes in the data over time (resulting from change in policy or practice) and changes in participants over time (e.g. changing their identifiers, moving in and out of scope of routine datasets).

It is our intention that this report provides an overview describing HES and signposting readers to both technical documentation and also existing exemplar illustrations of linked study-HES research investigations. We will supplement our description of the HES dataset with a series of case study illustrations drawn from the ALSPAC birth cohort study and the Hertfordshire Cohort Study.

Aims

This report aims to summarise the Hospital Episode Statistics database and its use within cohort and longitudinal studies. Within this, we will:

1. Provide an overview of the Hospital Episode Statistics dataset, describing its history, coverage and content;
2. Describe the particular format of the HES datasets;
3. Summarise the process by which secondary care records are compiled into the HES datasets, the quality assurance processes deployed and academic assessments of HES quality;
4. Illustrate the reliability and validity of self-reported hospital admissions through two case studies – the first looking at broad reporting of admissions to hospital and the second looking specifically at self-reporting of self-harm in adolescents and young adults.
5. Illustrate the predictive potential of combined cohort and HES data using examples from HCS

The report is not based on a systematic review of HES technical specifications or the sum of research studies utilising HES in their research. Rather, it is based on the knowledge and experience of the authors gained from linking the ALSPAC and HCS studies to the HES data set.

Scope

This report describes the core HES data warehouse. It does not cover aligned NHS Digital databases such as the Mental Health Services Data Set, the Diagnostic Imaging Data Set or the Patient Reported Outcome Measures, nor does it cover linked Registers such as the Office for National Statistics Mortality Register. As a result, the report focuses on English inpatient admissions, outpatient appointments and attendance at Accident & Emergency units. However, researchers should be aware of the existence of the aligned databases and the fact that these can be linked within NHS Digital using NHS patient ID or the pseudonymised HESID patient ID that is generated across all databases using a consistent methodology.

The Hospital Episode Statistics database

History

The Hospital Episode Statistics (HES) database contains administrative data from English hospitals in the National Health Service (NHS). The impetus to develop such a database lies in the 1979 Royal Commission - a major review of information services within the NHS – which found that 'the information available to assist decision makers in the NHS leaves much to be desired'[1]. The Royal Commission recommended the establishment of a 'Steering Group on Health Services Information', which was subsequently convened under the chairmanship of Dame Edith Körner. The steering group reaffirmed that "improved data would help to improve the quality and efficiency of the NHS"[2] and in the early 1980s produced a series of six reports providing recommendations on improving data collation, processing, use and governance within the NHS. These recommendations provided the rationale and guidance for establishing a national centralised repository of hospital clinical information. The group's recommendations were accepted by the Secretary of State and enacted in full. It is worth noting that Körner describes their aim, based on pragmatic reasoning, as the need to collate "very spare minimum datasets" and that "Concern for the privacy of patients and staff and the exclusion from the group's remit of epidemiological concerns precluded the consideration of other data items"[3]. The group acknowledged the challenge in rolling out a standardised data collection and reporting mechanism across such a large organisation and the report documents were accompanied by training programmes and standardised glossaries and information classification tools.

Prior to 1987 hospital episode statistics were compiled on a 10% sample of admitted records. These data were collected in the 'Hospital In-Patient Enquiry', the 'Annual Hospital Returns' and the 'Hospital Activity Analysis' datasets[4]. These may have research value to some of the older CLOSER studies or other cohort and longitudinal studies. However, these records do not form part of the HES dataset and are therefore outside the scope of this report.

In 1989 the English HES database was established with the aim to record every 'episode' of admitted patient care delivered in England[5]. These episode data are generated at a local level and then centralised before being made available for secondary use in annual datasets. While HES initially only included records of inpatient care (including maternity services), the database has subsequently expanded to include four domains: i. inpatient

[1] Parris G. Towards a coordinated approach for management information in the NHS. Health Information & Libraries Journal. 1986 Jun 1;3(2):82-93.
[2] Havard J. Körner group urges changes in NHS statistics. British Medical Journal. 1982 Nov 27;285:1591.
[3] Körner E. Improved information for the NHS. British Medical Journal (Clinical research ed.). 1984 Dec 8;289(6458):1635.
[4] Ashley JS. Present state of statistics from hospital in-patient data and their uses. British journal of preventive & social medicine. 1972 Aug;26(3):135.
[5] http://content.digital.nhs.uk/hes

episodes (including maternity HES); ii. Outpatient episodes; iii. A&E attendance; and, iv. Critical Care. More recently, records of adult mental health community provided care have been added as an ancillary dataset. By 2017, NHS Digital estimated that the total annual HES extract contained over 125 million care records.

The primary purpose of the HES dataset is to facilitate hospital reimbursement from NHS England for the care they have provided (through the 'Payments by Results' (PbR) system); however, secondary uses – including research – are permitted and accommodated within the design of the HES system. While HES has been used extensively by the research community there are long-standing concerns regarding the quality, completeness and coverage of HES records[6].

Academic use of the HES dataset

NHS Digital do not maintain a register of academic publications informed through using the HES dataset. A recent review of publications using HES - and in particular the HES APC dataset - identified 264 publications using HES between 2011 and 2016[7], which follows on from a systematic review that identified 148 articles between 1989 and 2011[8]. Breakdown of these review results suggest that the annual publication numbers have risen from 2 in 1993 to 88 in 2015[9]. Use ranges from standalone analysis of HES extracts, through linking HES to registers and other sources of information, linking HES into longitudinal observational studies (see below) and to randomied controlled trial samples for long-term outcome assessments. Additionally, HES has been linked to research repositories – such as the Clinical Practice Research Datalink[10] – where it is available for onward sharing.

Within the cohort and longitudinal study community an increasing number of studies have linked participants to their routine HES records, including: ALSPAC[11], EPIC-Oxford[12], Hertfordshire Cohort Study[13], Millennium Cohort Study[14], Million Women Study[15], UK

[6] Spencer S. Hospital Episode Statistics (HES): Improving the quality and value of hospital data. A discussion document. 2011.
[7] Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). International journal of epidemiology. 2017 Mar 15:dyx015.
[8] Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes—systematic review and recommendations for reporting. The European Journal of Public Health. 2012 May 10;23(1):86-92.
[9] Ibid. 7.
[10] Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data resource profile: clinical practice research datalink (CPRD). International journal of epidemiology. 2015 Jun 1;44(3):827-36.
[11] Mars B, Cornish R, Heron J, Boyd A, Crane C, Hawton K, Lewis G, Tilling K, Macleod J, Gunnell D. Using data linkage to investigate inconsistent reporting of self-harm and questionnaire non-response. Archives of suicide research. 2016 Apr 2;20(2):113-41.
[12] Crowe FL, Appleby PN, Travis RC, Key TJ. Risk of hospitalization or death from ischemic heart disease among British vegetarians and nonvegetarians: results from the EPIC-Oxford cohort study. The American journal of clinical nutrition. 2013 Mar 1;97(3):597-603.
[13] Simmonds SJ, Syddall HE, Walsh B, Evandrou M, Dennison EM, Cooper C, Aihie Sayer A. Understanding NHS hospital admissions in England: linkage of Hospital episode statistics to the Hertfordshire cohort study. Age and ageing. 2014 Mar 4;43(5):653-60.

Biobank[16] and Whitehall II[17]. The Partridge Review[18] of data sharing by the NHS Information Centre has dominated recent efforts for studies to access and use HES records. Studies that had secured access to HES before the review have had to renegotiate access arrangements using a new framework. This new framework has taken several years to develop and marks a step change in approach to data sharing by NHS Digital (the NHS organisational unit now responsible for secondary use of HES information). Following considerable delay, the National Survey for Health and Development and Whitehall II have negotiated new permissions to extract and use HES, and studies such as Understanding Society have extracted HES equivalents in the other UK home countries.

HES data coverage and collection

Coverage

The HES database contains the records of inpatient admissions, outpatient appointments and Accident and Emergency attendances at NHS hospitals in England. This includes records of independently funded patients who are treated by an NHS provider and the records of non-English residents. The database also contains the records of admissions to independent (non-NHS) providers where that treatment is funded by the NHS[19]. Cross-border mechanisms exist for treatment of residents from other countries within the UK and foreign nationals. Cross-border treatment rates are relatively low, but should be accounted for when conducting linkage-based analysis. For example, the English HES database for 2013/14 includes ~57,000 inpatient episodes, ~277,000 outpatient appointments and ~49,000 Accident and Emergency attendances for patients resident in Wales (although these represent <0.4% of total recorded episodes). Similarly, ~10,500 English residents were admitted for care in Welsh hospitals[20].

Not every participant in a cohort or longitudinal study will have a HES record as some will live outside the catchment area of English NHS commissioned services, some will seek privately commissioned health treatment, and some will not have been admitted to hospital within the period covered by HES. It is also important to note that some participants may

[14] Millennium Cohort Study: Birth Registration and Hospital Episode Statistics Linkage. February 2007. Available From: http://www.cls.ioe.ac.uk/library-media/documents/MCS_Birth_Registration_and_Hospital_Records_User_Guide_v4.pdf

[15] Reeves GK, Balkwill A, Cairns BJ, Green J, Beral V, Million Women Study Collaborators. Hospital admissions in relation to body mass index in UK women: a prospective cohort study. BMC Med. 2014;12:45.

[16] UK Biobank: Hospital Episode Statistics data in Showcase. December 2013. Available from: http://biobank.ctsu.ox.ac.uk/showcase/docs/HospitalEpisodeStatistics.pdf

[17] Britton A, Milne B, Butler T, Sanchez-Galvez A, Shipley M, Rudd A, Wolfe CD, Bhalla A, Brunner EJ. Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): Extracting Information from hospital medical records versus the Hospital Episode Statistics database. BMC medical research methodology. 2012 Jun 21;12(1):83.

[18] Data Release Review. June 2014. Great Britain: Health and Social Care Information Centre. Available from: https://www.gov.uk/government/publications/review-of-data-releases-made-by-the-nhs-information-centre

[19] see www.content.digital.nhs.uk/hesdata

[20] The UK Government's Response to the House of Commons Welsh Affairs Committee Report: Cross-border Health arrangements between England and Wales. 10th September 2015. Great Britain: Department for Health.

have been admitted to hospital or received some other health treatment that should be recorded in HES, but whose record is not available due to record keeping error, coding error, linkage error or has been removed as a result of ethico-legal filtering (e.g. where selected patients records are removed from extracts as they have registered an objection to their records being used for this purpose[21]).

Equivalent data held across the UK

While the scope of this report is restricted to the English HES dataset, it is of interest to note that there are broadly comparable datasets collated across the four nations comprising the UK.

Each home nation of the UK records secondary care data within different datasets, each of which is available via a distinct mechanism from a different provider. This introduces both an ethico-governance challenge and a technical challenge for studies to address. Firstly, studies will need to develop infrastructure to accommodate the needs of these distinct providers. Secondly, once data have been acquired, there will be a subsequent challenge relating to integrating the data into a harmonised form. The latter interoperability challenge relates both to pooling information on study participants receiving treatment in the different home nations (i.e. where studies sample across the UK) and also to creating longitudinal participant records where participants move between the home nations (or seek treatment in different home nations). Cross-country analysis should be approached on the basis that policy and procedural differences may mean hospital admission and appointment records are not always measured in a comparable manner. UK Biobank have started to assess interoperability of secondary care records across England, Wales and Scotland[22].

The different systems operating within the home nations can be summarised as follows:

> **England:** The HES data warehouse, that has operated from 1990 to the present and access to which is managed by NHS Digital.
> **Wales:** The Patient Episode Database for Wales (PEDW)[23], that has operated from 1999 to the present and access to which - from a pragmatic research perspective - has often been negotiated via the Secure Anonymised Information Linkage (SAIL, University of Swansea) databank. However, the NHS Wales Informatics Service retain ownership of these data.
> **Scotland:** The Scottish Morbidity Record (SMR), that has operated from 1981 to present (although the system had significant change in 1997) and access to which is managed by the Information Services Division of NHS Scotland through their 'electronic Data Research and Innovation Service' (eDRIS) unit.

[21] See this site for further information: http://content.digital.nhs.uk/article/7092/Information-on-type-2-opt-outs
[22] Mapping inpatient hospital data across England, Scotland and Wales. July 2014. Great Britain: UK Biobank, University of Oxford. Available from: http://biobank.ctsu.ox.ac.uk/crystal/docs/inpatient_mapping.pdf
[23] Annual PEDW Data Tables -Notes and Definitions. 29th October 2008. Great Britain: Health Solutions Wales. Available from: http://www.infoandstats.wales.nhs.uk/Documents/869/NotesDefinitions_Oct08.pdf

Northern Ireland: The Northern Ireland Hospital Statistics Dataset (NIHSD), that has operated from 2007 and is collated by the 'Hospital Information Branch' of the Northern Ireland Department of Health. There is no current clear access mechanism, athough certain NI health datasets are available via the ESRC Administrative Data Research Centre – Northern Ireland in conjunction with the Northern Ireland Statistics and Research Agency (NISRA)[24].

Data Collection

Data are collected by the care provider while the patient receives treatment. The data are collated and then submitted to NHS Digital – the NHS unit responsible for collation and reporting of information within the English NHS – on a monthly basis[25] using the 'Commissioning Data Sets' (CDS)[26] standardised process. The mechanism by which information is collated and returned has changed over time. From 1989 until 1996 records were collated by the different regional health authorities. From 1996, a national service – known as the 'NHS-Wide Clearing Service' (NWCS) - was established. From 2006, this was replaced by a mechanism known as the Secondary Uses Service (SUS).

The SUS system takes monthly extracts from provider systems and returns these to NHS Digital. SUS provides this information to the PbR system, but also uses a copy of the information to populate a local SUS database that is subject to incremental change. In turn, snapshot extracts from the SUS database are used to populate the HES databases[27]. Once deposited in HES, the data are subject to data processing that maps provider codes, removes duplicate entries, cleans the data, derives new values and conducts disclosure control transformations. Data quality assurance checks are conducted at stages during these processes. Once finalised, the data are deposited in the HES data warehouse as a published annual dataset.

[24] ADRC-NI Data Prospectus. July 2017. Great Britain: ADRC-NI. Available from: https://www.adrn.ac.uk/media/174457/data-prospectus-v4-august-2017-pdf-version.pdf

[25] This practice has changed over time, initially submitted updates were provided annually, then quarterly before moving to monthly submissions.

[26] Commissioning Data Sets (CDS) v6.2 Standard Specification. 27th August 2012. Great Britain: Information Standards Board for Health and Social Care. Available from: http://content.digital.nhs.uk/media/17281/0092162010sspec/pdf/0092162010sspec.pdf

[27] The HES Processing Cycle and HES Data Quality v4.0. 26th September 2016. Great Britain: NHS Digital. Available from: http://content.digital.nhs.uk/media/1366/The-HES-processing-cycle-and-HES-data-quality/pdf/HESDQ_In_001_The_HES_Processing_Cycle_and_HES_Data_Quality.pdf

Data format

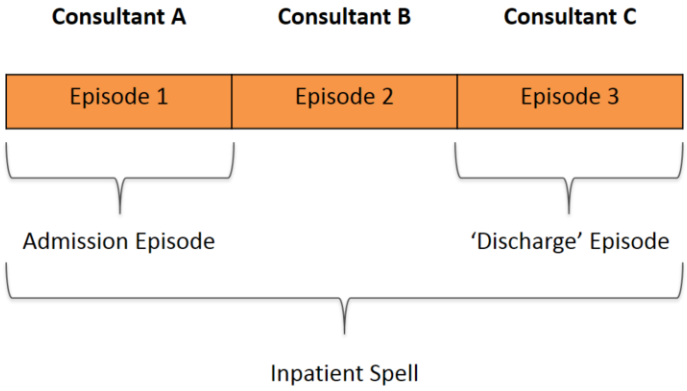
HES record level data are structured into *episodes* and *spells* and packaged into different data domains.

Defining Episodes and Spells

Patient admissions to hospital can be considered to comprise of *spells* – periods of continuous care in one provider institution[28] – which in turn can be sub-divided into *episodes* (figure 1) – periods of continuous care from a single consultant (although in practice the nominated consultant may be arbitrary given that consultant responsibility may be shared amongst a team).

Each row in the HES admitted patient care dataset contains the record of a single *episode*. These, in turn, can be grouped into admission *spells*. A *spell* commences when a patient is admitted for care and a consultant takes responsibility for that person's care. A *spell* ends when a patient is discharged, is transferred or dies.

Figure 1: Episodes and Spells in Hospital Episode Statistics patient admissions data.
Source: adapted from 'Methodology to create provider and CIP spells from HES APC data. 2014. England: NHS Digital.



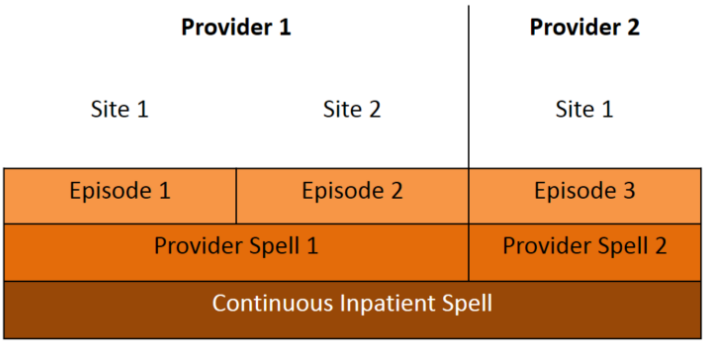
Most patient admissions spells comprise of a single episode. However, the patient record becomes more complicated where multiple consultants within the same provider institution treat a patient. This results in multiple episodes being generated within a single admitted inpatient spell (for example, Figure 1 illustrates an inpatient spell comprising three episodes). This could be a result of a patient being treated for different conditions, or where a patient is transferred to the care of a different consultant within the same provider institution.

Where a patient is transferred to a different NHS provider institution for continuing care, a new provider spell is created and, within this, new episode(s) are also created (Figure 2). This transfer signifies a change of organisation with responsibility for the patient and results in a patient being recorded as discharged from the first provider and admitted into the second provider. The Continuous Inpatient Spell (known by the CIP acronym) encompasses

[28] Note that an NHS provider institution can comprise multiple sites and multiple different hospitals.

the combined episodes and inpatient spells relating to the entire continuous sequence of care.

Figure 2: A Continuous Inpatient Spell in Hospital Episode Statistics patient admissions data.
Source: adapted from 'Methodology to create provider and CIP spells from HES APC data. 2014. England: NHS Digital.



This means that a CIP contains one or more provider spells, and that each provider spell contains one or more episodes. In practice, most CIPs contain one provider spell that contains one episode.

The HES annual dataset

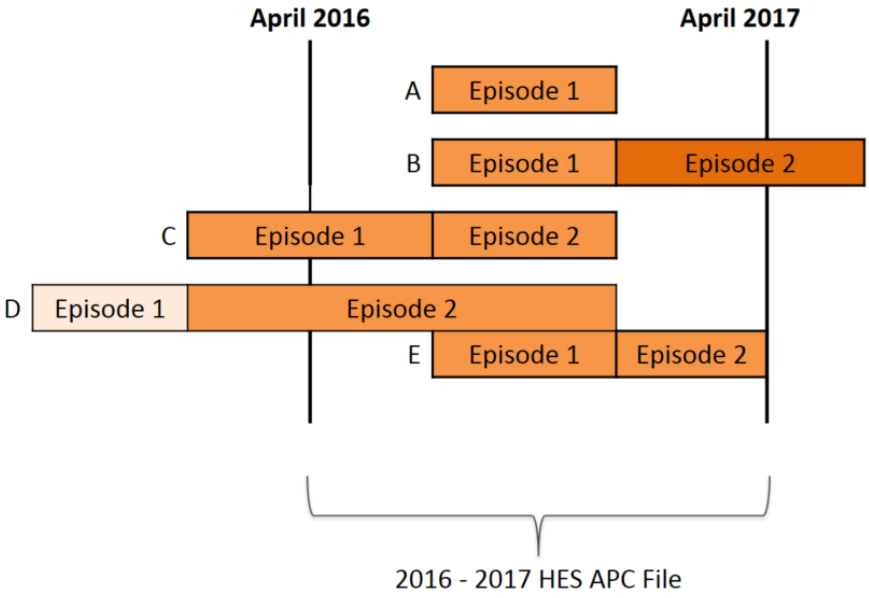
HES data are collated into in-year episodes defined by 'financial year'[29]. Once collated, processed and quality assessed, the yearly HES datasets are released for secondary use and are also used by NHS Digital statisticians to report national figures. The annual HES datasets are considered 'final' and theoretically will remain unchanged in perpetuity.

This means that CIPs, provider spells and episodes (or combinations of the three) can straddle financial year boundaries (Figure 3). Episodes that carry across the financial year boundary will remain 'unfinished' in the first year and be marked as finished in the year the patient was discharged. All the scenarios depicted in Figure 3 will result in records being added to the HES APC 2016/17 in-year dataset; only Episode 1 in Scenario D will not feature within the 16/17 dataset. Incomplete episodes (e.g. Episode 2 in Scenario B) should be identified in order to address double counting. In this scenario, where you had more than one year's worth of data (2016-17 and 2017-18) you would make sure you only counted Episode 2 once, even though it would be present in both datasets.

Figure 3: Episodes and spells across HES financial year data sets.

[29] The HES 'financial year' runs from the 1st April to the 31st March.

Source: adapted from 'Methodology to create provider and CIP spells from HES APC data. 2014. England: NHS Digital.



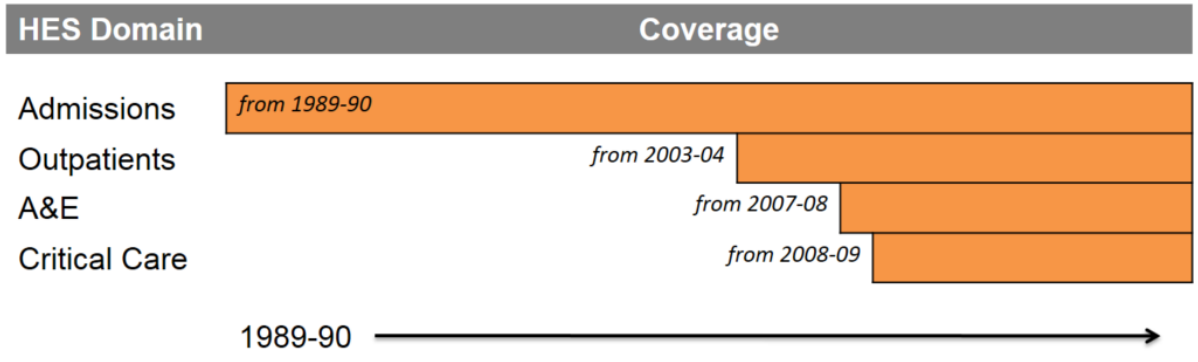
Content and data domains

The HES database comprises many data items that are defined within the national Commissioning Data Sets (CDS) standard. Additional data items are subsequently derived from the CDS values by NHS Digital as part of their processing activities.

The HES dataset is generally considered to fall into four domains, each of which is considered as a separate data product by NHS Digital, and each of which has a different history and coverage period (Figure 4). Each of these datasets comprises clinical information (typically primary and other diagnoses assigned; primary and other procedures carried out), socio-demographic information, administrative information (e.g. data on waiting times, admissions routes) and both person (including geographic residence information), event and provider (including geographic provider information) identifiers. The full content of the HES database is detailed in the online HES Data Dictionary[30]. The Data Dictionary includes both the CDS fields and the derived fields. It also contains notes on data cleaning processes.

Figure 4: Hospital Episode Statistics Data Domains.

[30] see www.content.digital.nhs.uk/hesdatadictionary The HES Data Dictionary is also available as PDF downloads (from the same link). It is also worth noting that the NHS are developing the NHS Data Model and Dictionary which are also available online: www.datadictionary.nhs.uk



Hospital admissions records (aka Admitted Patient Care, APC)

Hospital admissions include episodes of treatment that require the use of a hospital bed, although this does not necessarily indicate there has been an overnight stay. Admissions can be elective, or emergency. The key fields in the APC dataset are illustrated below (Table 1) and full documentation can be found in the HES APC data dictionary. From 1998 the APC dataset assigned episodes to a specific consultant using a specific consultant identifier.

Table 1: High level fields within the APC record

Identifiers	Clinical Information	Demographic information	Administrative	Maternity tail
HESID (specific to each data sharing agreement)	Diagnoses and procedures (up to 20 primary and secondary)	Age (years) at admission and discharge	Method of admission (e.g. elective or emergency, birth, transfer)	Gestational Age
Episode ID	Operation Dates	Gender	Episode start and end date	Parity
Date of admission	Consultant Speciality	Index of Multiple Deprivation (IMD)	Discharge method (e.g. self-discharge, died, transferred)	Birth Weight
A&E link ID	Augmented care location	Health, Electoral and census geographies	Discharge destination (e.g. home, other destination	Maternal age
Provider details (e.g. hospital code).		Ethnic group	Time waited	Mode of delivery (e.g. forceps or spontaneous)
Registered GP practice				Birth Order (for multiple births)
				Neonatal care

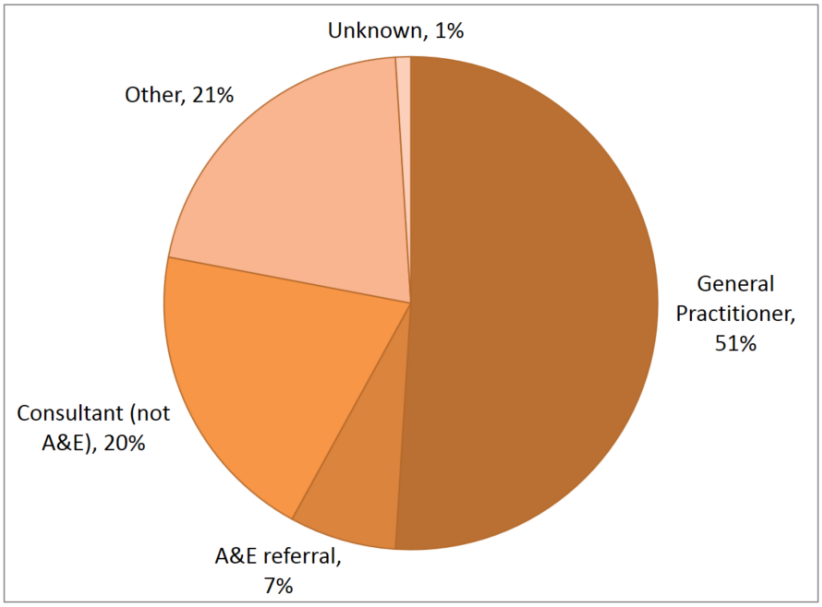
The APC file also includes the ‘Maternity HES’ records. Each birth generates at least two episodes, one recording details of the delivery (relating to the mother) and one episode per

child delivered (relating to a child). These episodes are supplemented with additional variables known as the 'maternity tail'. There is no NHS data service providing a systematic means of linking the mother's delivery episode with the baby's birth episode. However, researchers have demonstrated this is possible using probabilistic linkage algorithms on de-identified episode data[31].

Outpatient records (OP)

The HES OP dataset records outpatient appointments in English NHS hospitals and English NHS commissioned activity in the independent sector. Each appointment is represented by a distinct row of data (regardless of whether the patient attended the appointment or not). Treatment can take the form of a planned *series* of appointments, which can be identified in the dataset. A given patient may have multiple series of appointments in any given financial year. In 2015-16 there were 113.3 million outpatient appointments. Patients aged 60-79 years accounted for >30% of these and overall, women accounted for 58% of attended appointments[32]. Over half of all first appointments resulted from primary care referrals (see Figure 5).

Figure 5: OP first attendance by source of referral 2015-16
Source: NHS Digital. Adapted from: Hospital Outpatient Activity 2015-16. 1st December 2016. England: NHS Digital



Data completeness is an issue in some OP fields. While 'attendance type', 'source of referral' and 'main specialty' have high rates of completeness (>98%), the outcome variable is less complete (95%) and fields such as primary diagnosis (5%) and main procedure (26%) have low levels of completeness[29].

[31] Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking data for mothers and babies in de-identified electronic health data. PloS one. 2016 Oct 20;11(10):e0164667.
[32] Hospital Outpatient Activity 2015-16. 1st December 2016. Great Britain: NHS Digital. Available from: http://digital.nhs.uk/catalogue/PUB22596

The key fields in the OP dataset are illustrated below (Table 2) and full documentation can be found in the HES OP data dictionary[33].

Table 2: High level fields within the OP record

Identifiers	Clinical Information	Demographic information	Administrative
HESID (specific to each data sharing agreement)	Diagnosis (up to 12 primary and secondary diagnoses)	Age (years) at appointment	Attendance details
Appointment ID	Operative procedure(s)	Gender	Time waited
Appointment date	Consultant Speciality (e.g. Ophthalmology, Child and adolescent psychiatry)	Indices of Multiple Deprivation (IMD)	Appointment type (e.g. face-to-face, telephone)
Registered GP practice		Health, electoral and census geographies Ethnic group	

Accident and Emergency records (A&E)

The HES A&E dataset records attendance at Accident & Emergency departments. Within the NHS, A&E departments provide services for those seeking urgent care for injury and illness. Major A&E departments receive new patients on a continual basis and care is consultant led. The HES A&E dataset also includes attendance records for specialty A&E departments, walk-in centres and minor injury units. The key fields in the A&E dataset are illustrated below (Table 3) and full documentation can be found in the HES A&E data dictionary[34].

[33] Data Dictionary: Adult Critical Care. 2010. Great Britain: NHS Information Centre. Available from: http://content.digital.nhs.uk/media/1362/HES-Hospital-Episode-Statistics-Adult-Critical-Care-Data-Dictionary/pdf/CC_DataDictionary.pdf

[34] HES Data Dictionary: Accident and Emergency Version 2. 22nd September 2015. Great Britain: Health and Social Care Information Centre. Available from: http://content.digital.nhs.uk/media/18619/HES-AE-Data-Dictionary/pdf/DD_AE_v2.pdf

Table 3: High level fields within the AE record

Identifiers	Classification & clinical	Demographic information	Administrative
HESID (specific to each data sharing agreement)	Incident location (e.g. home, work, public place)	Age (years) at arrival	Arrival mode (ambulance or other)
Appointment ID	Patient group (e.g. road traffic accident, sports injury)	Gender	Attendance Category (first attendance or follow-up)
Arrival date and time	Diagnosis (up to 12 codes)	Indices of Multiple Deprivation (IMD)	Disposal (e.g. admitted, died, referred)
Registered GP practice	Anatomical area and side	Health, electoral and census geographies	Source for referral (e.g. self, GP, police, social services)
	A&E investigation (e.g. x-ray, toxicology)	Ethnic group	Visit duration

Each A&E entry – an attendance - provides a record of a single visit by an individual to A&E. Subsequent visits, such as for A&E provided follow-up care are recorded as a separate attendance (first and follow-up visits are distinguishable within the dataset). The majority of attendances at A&E are for a first visits and result in discharge with no further follow-up (Table 4).

Table 4: Headline A&E attendance figures 2015-16
Source: NHS Digital. Adapted from: Hospital Accident and Emergency Activity: 2015-16
10th January 2017. England: NHS Digital.

	Number	Percent
Total Attendances	20,457,805	100
Attendance Category		
First A&E attendance	19,249,491	94.1
Planned attendances	289,734	1.4
Unplanned attendances	457,415	2.2
Not known	461,165	2.3
Disposal Method		
Admitted to Hospital	4,123,765	20.2
Discharged with follow-up	4,048,970	19.8
Discharged no follow-up	7,627,315	37.3
Referred	2,593,890	12.7
Other	2,063,865	10.1
Patients' age		
0-4	2,054,092	10.0
5-14	2,096,128	10.2
15-44	8,066,288	39.4
45-64	3,922,683	19.2
65-84	3,080,507	15.1
85+	1,008,939	4.9
Not Known	229,168	1.1
Patients' Gender		
Female	10.2m	49.9
Male	10.1m	49.2

Adult critical care records (ACC)

The ACC dataset contains the records for critical care periods in adult designated wards (i.e. an Intensive Care or High Dependency Unit) where constant support and monitoring is required to maintain function in at least one organ. The ACC is a sub-set of the APC dataset that has extended data reporting requirements. These requirements are specified in the Critical Care Minimum Dataset (CCMD), which contains 34 separate fields (although

reporting is only mandatory for 14 of these)[35]. Key fields are illustrated below (Table 5). Full documentation can be found in the HES ACC data dictionary[36] and technical guide[37].

.

Table 5: High level fields within the ACC

Identifiers	Classification & clinical	Demographic information	Administrative
HESID (specific to each data sharing agreement)	Treatment function (e.g. transplantaion surgery, burns care)	Age (years) at arrival	Admission source (e.g. same hospital, transfer)
Provider code	Critical care level and duration of care at that level	Gender	ACC Unit function (e.g. renal, neuroscience)
Start date and time	Variables indicating duration of care in specific areas (e.g. renal support)	Indices of Multiple Deprivation (IMD)	Discharge location (e.g. ward details, home)
Registered GP practice	Maximum number of organs being supported	Health, electoral and census geographies	
		Ethnic group	

The majority (77%) of critical care patients are adults aged 50 or over (Figure 6)[38]. While this is billed as an 'adult' dataset, the scope is defined by the nature of the critical care ward function, therefore children (including neonates) may be included in the dataset if they were cared for in an adult ward. Patients can have multiple ACC stays. These stays may occur over the same or different time period, or relate to the same or different condition. It will not always be the case that a single ACC record entry will relate to a single APC episode[39].

[35] HES Data Dictionary: Adult Critical Care. 23rd February 2017. Great Britain: NHS Digital. Available from: http://content.digital.nhs.uk/media/23589/HESAdultCriticalCareDataDictionary/pdf/ACC_Data_Dictionary_Feb17.pdf

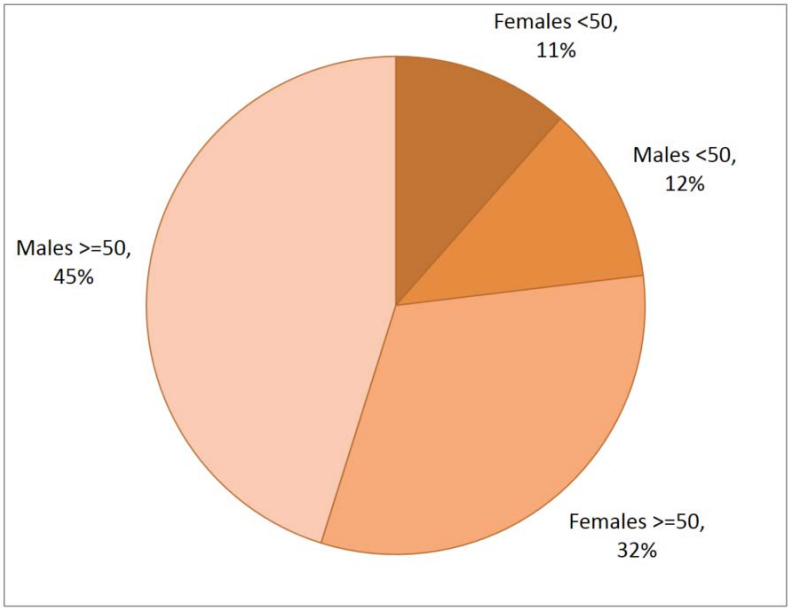
[36] See: http://www.datadictionary.nhs.uk/data_dictionary/messages/supporting_data_sets/data_sets/critical_care_minimum_data_set_fr.asp?shownav=1

[37] Hospital Adult Critical Care Activity 2016-17 Technical Guide. 3rd October 2017. Great Britain: NHS Digital. Available from: https://digital.nhs.uk/media/32942/Hospital-Admitted-Patient-Care-Activity-2016-17-Adult-Critical-Care-technical-guide/default/hosp-epis-stat-admi-acc-techguide-2016-17

[38] Hospital Adult Critical Care Activity 2015-16. 23rd February 2017. Great Britain: NHS Digital. Available at: https://digital.nhs.uk/catalogue/PUB23426

[39] NHS Digital have developed a 'best match' algorithm designed to link ACC records with APC episodes. This is described in: Hospital Adult Critical Care Activity: Technical Guide. 23rd February 2017. Available at: https://digital.nhs.uk/media/30567/Hospital-Adult-Critical-Care-Activity-2015-16-Technical-guide/Any/adul-crit-care-data-eng-apr-15-mar-16-tech

Figure 6: Percentage of Adult Critical Care records by age and gender, 2015-16
Source: NHS Digital. Adapted from: Hospital Adult Critical Care Activity. 23rd February 2016. England: NHS Digital. Unknown age and gender values have been excluded.



Coding

Different coding systems are in use through the HES datasets. Diagnoses in the APC and OP datasets are coded using the World Health Organisation's (WHO) International Classification of Disease (ICD) clinical classification system. Entries up until 1995 were coded using ICD version 9 and subsequent entries with ICD version 10. Establishing data interoperability between information coded to ICD 9 and ICD 10 can be challenging (although this challenge lies outside the scope of this report). Operative procedures are coded using the Office of Population Censuses and Survey's (OPCS) version 4 clinical classification. Both classification systems are updated periodically to accommodate new conditions. NHS coding and classification standards – and cross-mapping reference files - are managed by the 'Technology Reference data Update Distribution' (TRUD) unit, that is part of NHS Digital[40]. HES data can be linked to the NHS Healthcare Resource Group information on provision unit cost[41].

Within the A&E dataset, bespoke classifications are used within the AEPATGROUP (recording the reason for an A&E attendance) and DIAG2_NN fields (The A&E diagnosis comprises a six-character code made up of: diagnosis condition, sub-analysis, anatomical area and anatomical side[42]).

[40] See https://digital.nhs.uk/article/290/Terminology-and-Classifications
[41] NHS Reference Costs. 20th May 2015. Great Britain: Department for Health. Available from: https://www.gov.uk/government/collections/nhs-reference-costs.
[42] HES Data Dictionary: Accident and Emergency. 22nd September 2015. Available from: http://content.digital.nhs.uk/media/18619/HES-AE-Data-Dictionary/pdf/DD_AE_v2.pdf

Currently the NHS is moving to adopt the SNOMED CT nomenclature, and separately the WHO is working with SNOMED CT regarding the development of the ICD 11 standard. Both of these developments suggest the possibility of substantial changes in HES coding.

Data Quality

HES data quality has been an area of concern within health services and the academic community. Major reviews of HES data quality have identified substantial quality issues (see Table 6)[43] and a lack of engagement among consultants in data quality[44]:

> "coded data is still a poor reflection of clinical practice, and that many clinicians remain uninterested" Audit Commission 2009.

However, the change in the purpose of HES from being primarily a planning and management tool to being the main mechanism for reimbursement (after Payment by Results) has led to greater engagement regarding data quality and completeness[45]. Therefore, it is recognised that the NHS has invested considerable resource in improving both the quality of underlying record keeping and quality assurance mechanisms deployed when centralising these records and processing them into the released HES datasets[46]. It is important that research users recognise that quality is likely to vary on a temporal basis (where quality varies from year to year with general trends towards improved quality) and spatially (where local provider record-keeping and reporting practice is likely to vary). Researchers must also consider the separate issues - such as policy change – or changes in reporting specifications – that will introduce reporting variation (e.g. categories of treatment moving from one reporting domain to another or the impact of changing incentives to report certain treatments). This chapter of the report discusses the former category of quality assurance but does not discuss the latter category relating to change of policy or clinical practice.

[43] Hospital Episode Statistics (HES): Improving the quality and value of hospital data. 2011. Great Britain: NHS Information Centre. Available from: http://content.digital.nhs.uk/media/1593/Hospital-Episode-Statistics-Improving-the-quality-and-value-of-hospital-data/pdf/HES_-_Improving_the_quality_and_value_of_hospital_data.pdf

[44] Spencer SA, Davies MP. Hospital episode statistics: improving the quality and value of hospital data: a national internet e-survey of hospital consultants. BMJ open. 2012 Jan 1;2(6):e001651.

[45] Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, Faiz OD. Systematic review of discharge coding accuracy. J Public Health (Oxf). 2012;34(1):138-48

[46] Audit Commission for Local Authorities and the National Health Service in England and Wales. Improving data quality in the NHS: annual report on the PbR assurance programme. Audit Commission; 2010.

Table 6: Results of the Audit Commission's 2009 national clinical coding audit in selected specialties in a large NHS Trust.
Source: adapted from 'Hospital Episode Statistics (HES): Improving the quality and value of hospital data'.

Area audited	Primary Diagnoses Incorrect (%)	Secondary Diagnoses Incorrect (%)	Primary Procedures Incorrect (%)	Secondary Procedures Incorrect (%)
Theme – 110: Trauma & Orthopaedics	20	24.9	19.1	12.4
Speciality – 502: Gynaecology	8	19.7	12.4	12.5
HRG Chapter – L: Urinary Tract and Male Reproductive System	12.9	25.4	14.3	35.5
HRG – F36: Large Intestinal Disorders >69	3.3	28.3	27.3	7.1
Overall	**12.7**	**24.4**	**15.9**	**15.3**

NHS Data Quality Assurance

The accuracy of the original source data is the responsibility of the (approximately 700) provider institutions. However, NHS Digital has a legal duty[47] to assess data under their remit against defined standards. The link between the HES return and provider payments (the PbR system) provides a means to encourage complete and accurate reporting (i.e. hospital payments are calculated using the same source data that is then used to populate HES).

NHS Digital has embedded quality assurance checks within the design of the SUS. The SUS specification requires that returns are provided using a standardised XML schema with validation rules (Figure 7) that enforce data standardisation. Both the XML schema and the validation rules are developed by NHS Digital and rolled out nationally via a NHS standard[48]. Provider returns are audited (by NHS Digital appointed auditors) to check conformance and accuracy, although it is important to note that data quality auditing does not extend beyond the PbR subset of information, meaning that some HES information is not subject to audit checks[49] and may therefore be subject to less rigorous compilation at source.

It is not clear to these authors as to the extent to which NHS Digital has consulted the research community regarding incorporating research-orientated quality assessment checks

[47] Set out in the terms of the Health and Social Care Act (2012).
[48] Commissioning Data Sets (CDS) v6.2 Standard Specification. 27th August 2012. Great Britain: Information Standards Board for Health and Social Care. Available from: http://content.digital.nhs.uk/media/17281/0092162010sspec/pdf/0092162010sspec.pdf
[49] The Quality of Nationally Submitted Health and Social Care Data. 30th October 2014. Great Britain: NHS Digital. Available from: https://digital.nhs.uk/catalogue/PUB15783

or metrics, although we note that NHS Digital have made an explicit objective to consult on these issues[50].

Figure 7: Example SUS validation rules

Source: Adapted from the 'SUS Data Quality Dashboard Validation Rules v1.0 (August 2014)[51].

Example 1: NHS Number

1.1.2 NHS Number
The NHS Number is the unique identifier and is mandatory to record for each patient.

A valid NHS Number is populated with value in national standard format, with a valid check digit. For a number of sensitive diagnoses and procedures (e.g. IVF), SUS removes all patient identifiable data including the NHS Number, and derives an NHS number status indicator of 91. In these cases, a blank NHS Number will be classed as valid. A blank NHS number will also be accepted as valid when the treatment function code is 360 (Genitourinary Medicine).

Example 2: Primary Diagnosis

1.1.10 Primary Diagnosis
This is a clinical classification associated with the patient diagnosis.

The patient diagnosis is:
i. the main condition treated or investigated during the relevant episode of healthcare, and ii. where there is no definitive diagnosis, the main symptom, abnormal findings or problem.
A valid code at 4-digit level listed in the ICD-10 classification published by the World Health Organisation, excluding codes beginning with "R69" indicating an unknown diagnosis. The fifth character is checked to be either a numeric site code, or a '-'.
N.B. This data item is only analysed for episodes where the spell is finished, as it is normal for coding to occur once the spell has ended.
http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/p/pri/primary_(icd-10)_de.asp?shownav=1

Subsequent quality assurance processing is conducted by NHS Digital as the returned data are processed into the HES annual datasets. Before finalisation, the processed data are returned to provider institutions for final review. These processes are based on logical error and format assessments as there is no mechanism to refer to the underlying data (i.e. these checks are based around 'cleaning' processes rather than confirmation processes). Further checks compare returns by provider institution to assess patterns and outliers; through these

[50] The Quality of Nationally Submitted Health and Social Care Data. 30th October 2014. Great Britain: NHS Digital. Available from: https://digital.nhs.uk/catalogue/PUB15783
[51] The full SUS Data Quality Dashboard Validation Rules v1.0 document is available from: http://content.digital.nhs.uk/media/14973/SUS-Data-Quality-Dashboard-Validation-Rules-v10/pdf/SUS_Data_Quality_Dashboard_Validation_Rules_v1.0.pdf

comparisons NHS Digital identifies 'red flag' quality issues that require further investigation[52].

Since 2012, NHS Digital has published annual data quality guidance reports which include details of known issues[53]. These reports identify processes and practice areas that do not conform to the standard and monitor actions undertaken to address these non-conformities. For example, the 2014 assurance report describes improvements to reduce the number of blank fields included in Critical Care returns. One such improvement made adjustments to the data collection system by introducing a user prompt which is raised if users attempt to submit a blank field. The prompt asks the user why they are attempting to submit a blank field and encourages them to rectify this. The rate of blank fields has dropped substantially as a result[54]. This example illustrates that quality improvements are gradual, incremental and lead to progressive improvements over time (i.e. it is reasonable to expect that a recently introduced dataset may have inferior quality to a well-established data set). While improvements are likely to result from improving technology (e.g. natively digital data collection) as well as improved process and user training[55], improvements may also result from changing legislative requirements and high-profile reports such as Dame Fiona Caldicott's review of information governance[56] and the need for a robust data quality strategy[57].

Quality assurance processes

The following quality assurance processes – all conducted during the processing of the information received from SUS - are particularly noteworthy:

1. Provider organisation codes are assessed against a known master catalogue and are standardised to current values (or set to null);
2. Logical checks and data corrections are applied, for example where birth episodes are coded as general episodes (Figure 8);
3. Validation and correction rules are enforced (e.g. null values are set to the appropriate 'missing' value, a 'M' sex value is set to '1' as specified in the standard). As of September 2016 there were 120 rules applied to the APC dataset, 57 rules applied to the OP dataset and 43 rules applied to the A&E dataset;

[52] Quality assurance and audit arrangements for administrative data – exposure draft. July 2014. Great Britain: UK Statistics Authority. Available from: http://www.statisticsauthority.gov.uk/assessment/monitoring/administrative-data-and-official-statistics/quality-assurance- and-audit-arrangements-for-administrative-data---exposure-draft.pdf
[53] The Quality of Nationally Submitted Health and Social Care Data. 30th October 2014. Great Britain: NHS Digital. Available from: https://digital.nhs.uk/catalogue/PUB15783
[54] Ibid.
[55] While NHS Digital are not responsible for data quality in provider institutions, they do offer guidance, for example:http://content.digital.nhs.uk/media/21886/Performance-evidence-delivery-framework/pdf/Performance_evidence_delivery_framework__august_2016.pdf
[56] The Quality of Nationally Submitted Health and Social Care Data. 30th October 2014. Great Britain: NHS Digital. Available from: https://digital.nhs.uk/catalogue/PUB15783
[57] The NHS data quality strategy for 2015-2020 is summarized here: http://content.digital.nhs.uk/media/19015/Data-Quality-Assurance-Strategy-2015-2020/pdf/DQA-strategy-on-a-page.pdf

4. Derivations are conducted (e.g. postcode is assessed for validity against the ONS Postcode Directory and, where valid, the ONS Postcode Directory is then used for linkage to geographical indicators such as deprivation indices);
5. Algorithms use provider codes, patient identifiers and episode administrative values to identify and remove duplicate entries[58];
6. The pseudonymised HES ID is derived from NHS ID and other identifiers[59];
7. 'Decode' fields are added which provide metadata (e.g. value labels) to explain other fields.

Once the HES datasets have been finalised for the financial year, then no further attempt will be made to correct or clean the data (i.e. the dataset is final, and the records should remain identical in perpetuity). Researchers should note that extracts from this system could vary given that patients are being provided with a mechanism to object to the secondary use of their healthcare records and to block future data sharing.

Figure 8: Example HES logical check and transformation rule
Source: adapted from the 'HES Processing Cycle and HES Data Quality v4.0 (September 2016).

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

The full 'HES Processing Cycle and HES Data Quality v4.0' document is available from: http://content.digital.nhs.uk/media/1366/The-HES-processing-cycle-and-HES-data-quality/pdf/HESDQ_In_001_The_HES_Processing_Cycle_and_HES_Data_Quality.pdf

[58] Methodology for identifying and removing duplicate records from the HES dataset. 26th September 2016. Great Britain: NHS Digital. Available from: http://content.digital.nhs.uk/media/13656/HES-Duplicate-Identification-and-Removal-Methodology/pdf/HESDQ_In_006_HES_Duplicate_Identification_and_Removal_Methodology.pdf
[59] HESID Methodology. 21st May 2014. Great Britain: NHS Digital. Available from: http://content.digital.nhs.uk/media/1370/HES-Hospital-Episode-Statistics-Replacement-of-the-HES-patient-ID/pdf/HESID_Methodology.pdf

Observations on data quality

There is substantial evidence to suggest that there are widespread data quality issues in HES, particularly with older annual data extracts. The overview presented in this report is intended to highlight this fact and to illustrate the manner in which the NHS is addressing these shortcomings (and the changes in data process and quality over time resulting from this). Our report does not extend to a systematic assessment of error and we cannot contribute to the academic debate as to whether quality is improving or not. Systematic reviews of data accuracy in UK health records[60,61] have identified quality issues, but also improvements to data quality in discharge coding accuracy between 2001 and 2011. The 2011 review found that following the system improvements relating to PbR, the accuracy of the primary diagnoses improved from 74% ([IQR 59-92%] to 96% [89-96%], p=0.02). Having said this, other authors have noted that assessments of administration error such as these may not reflect on diagnostic accuracy in areas such as psychiatric disorders[62], and that administrative error (i.e. issues within the domain of the HES production team) was small in comparison to diagnostic error (i.e. issues outside the domain of the HES production team).

The newer datasets – for example the Critical Care admissions subset – are comprised of a mix of mandatory information and optional returns. In practice this results in substantial missingness and geographically clustered reporting practice. Certain classes of information, such as patient ethnic group, have been recognised as having poor levels of accuracy and completeness[63]. Researchers using the A&E dataset face challenges relating to its bespoke coding system, which can lack the coding granularity needed for research purposes[64].

Research users of HES should take note that changes in health service organisation (i.e. boundary change, merger or use of different provider) may result in change of quality or reporting practice, new coding schemes will result in changes, as will changing methodologies for record linkage (e.g. the introduction of NHS ID into HES returns in 1997), and changing incentives for reporting (e.g. the introduction of the PbR financial mechanisms). As a rule of thumb, data quality is higher in more recent datasets and higher in datasets which are more established (this does not mean there is not variation of quality within the different HES domains).

[60] Campbell SE, Campbell MK, Grimshaw JM, Walker AE. A systematic review of discharge coding accuracy. Journal of Public Health. 2001 Sep 1;23(3):205-11.
[61] Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, Faiz OD. Systematic review of discharge coding accuracy. Journal of public health. 2011 Jul 27;34(1):138-48.
[62] Davis KA, Sudlow CL, Hotopf M. Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. BMC psychiatry. 2016 Jul 26;16(1):263.
[63] Mathur R, Grundy E, Smeeth L. Availability and use of UK based ethnicity data for health research. Available from: http://eprints.ncrm.ac.uk/3040/1/Mathur-_Availability_and_use_of_UK_based_ethnicity_data_for_health_res_1.pdf
[64] Teyhan A, Cornish R, Boyd A, Joshi MS, Macleod J. The impact of cycle proficiency training on cycle-related behaviours and accidents in adolescence: findings from ALSPAC, a UK longitudinal cohort. BMC public health. 2016 Dec;16(1):469.

The reliability and validity of self-reported hospital admissions.

It is widely accepted that self-reported information provided by cohort and longitudinal study participants is subject to potential error and bias. Linkage to routine records has been identified as a means to collect information on participants that may be less susceptible to study or participant introduced biases[65,66], although it is important to stress that both record linkage as a process[67] and routine records as a source of information, are subject to error and bias. While some historical reporting errors[68] within the HES datasets are likely to have improved due to improved IT infrastructure, routine digitisation or digitally collected information, and quality improvement programs, reporting error is still cause for concern. Despite this, objectively recorded HES records provide a means to assess the reliability and validity of self-reported hospital admissions.

There are currently relatively few examples of cohort and longitudinal studies using HES in this way[69],[70]. In this report we will summarise two examples from the ALSPAC birth cohort. Firstly, we will summarise a sensitivity analysis of self-reported hospital admission (for any reasons). We then consider a focused example, relating to the potentially stigmatising practice of self-harm.

Case Study 1: Using linkage to Hospital Episode Statistics to investigate the accuracy of parent hospital admissions.

Introduction

This case study aimed to validate parental reported admissions to hospital for index participants in the ALSPAC cohort study against linked HES-recorded hospital admissions.

Methods

When ALSPAC participants reached age 18 the study sent a ‘fair processing’ information campaign seeking re-enrolment into the study and informing participants about ALSPAC’s proposed use of their routine health records. To test linkage methodologies, ALSPAC

[65] Calderwood L, Lessof C. Enhancing longitudinal surveys by linking to administrative data. Methodology of longitudinal surveys. 2009 Jan 26:55-72.
[66] Brett CE, Deary IJ. Realising health data linkage from a researcher’s perspective: following up the 6-Day Sample of the Scottish Mental Survey 1947. Longitudinal and Life Course Studies. 2014 Oct 30;5(3):283-98.
[67] Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, van der Meulen JH. A guide to evaluating linkage quality for the analysis of linked data. International journal of epidemiology. 2017 Sep 7;46(5):1699-710.
[68] Williams JG, Mann RY. Hospital episode statistics: time for clinicians to get involved?. Clinical Medicine. 2002 Jan 1;2(1):34-7.
[69] Britton A, Milne B, Butler T, Sanchez-Galvez A, Shipley M, Rudd A, Wolfe CD, Bhalla A, Brunner EJ. Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): Extracting information from hospital medical records versus the Hospital Episode Statistics database. BMC medical research methodology. 2012 Dec 1;12(1):83.
[70] Woodfield R, Grant I, Sudlow CL. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: a systematic review from the UK biobank stroke outcomes group. PloS one. 2015 Oct 23;10(10):e0140533.

selected a sub-sample of early responders to this fair processing campaign to establish a pilot linkage between responders and their HES records. Of those sent the fair processing materials (n=12,385), 3,195 (25.8%) responded and provided consent before the pilot selection cut-off date. These individuals comprised the sample included in this study.. In nine questionnaires completed when the children were aged between 6 months and 13 years old, parents were asked if their child had been admitted to hospital in the time since the issue of the previous questionnaire (the periods covered varied, ranging from 6 months to 4 years). We compared this information to data recorded in the HES database for the corresponding time period and calculated sensitivities, specificities and predictive values of the questionnaire-reported admissions using HES records as the reference standard.

Results

Between 4.5% and 10.5% of parents reported that their child had been admitted to hospital for each of the periods covered by the questionnaires (Table 7). Among those whose parent reported a hospital admission, at least 60% had one or more corresponding admission in the HES data. Where a hospital admission was not indicated on the questionnaire, an admission was found in the HES data for between 1.5% and 5.8% of the participants. We found that of those reporting an admission that was not seen in HES, between 6.8% and 25.8% were found to have an admission in the previous time period.

Conclusions

We found that the specificities and negative predictive values of parent-reported hospital admissions were high at all ages (i.e, parent-reported data correctly identified those *without* admission). The sensitivities and positive predictive values were lower (i.e, parent-reported data were less accurate in identifying those *with* admission). There are several possible explanations for this. Between 6.8% and 25.8% of reported admissions not seen in HES were seen in the previous time period, suggesting some degree of respondent error in recalling dates of admission. A proportion of respondents may have interpreted the questions about admission to hospital as including visits to A&E and/or outpatient appointments. The HES database only includes A&E data from April 2007 (when the ALSPAC children were aged 15-16 years old) and outpatient data from April 2003 (when they were 11-12) so it was not possible to examine whether this explained the low sensitivities. Further, some hospital admissions could be to non-NHS providers in England or non-English providers (for participants hospitalised while travelling or living abroad or in other UK countries) which are not recorded in HES.

Table 7: Agreement between self-reported hospital admission and admissions observed in linked HES records.

ALSPAC questionnaire and hospital admission data item	Timepoint	Mother reported response	HES admission found	HES admission not found
Has your child ever been admitted to hospital?	6 months	Yes	180 (72%)	89 (3%)
		No	69 (28%)	2,565 (97%)
Has your toddler been admitted since they were 6 months old?	18 months	Yes	153 (66%)	62 (2%)
		No	79 (34%)	2,593 (98%)
Has your toddler been admitted since they were 18 months old?	30 months	Yes	123 (62%)	44 (2%)
		No	75 (38%)	2,521 (98%)
Has your toddler been admitted in the past 12 months?	42 months	Yes	107 (67%)	77 (3%)
		No	51 (33%)	2,554 (97%)
Has your child been admitted since they were 3 years old?	57 months	Yes	172 (67%)	91 (4%)
		No	84 (33%)	2,383 (96%)
Has your child been admitted in the last 15 months?	69 months	Yes	93 (60%)	55 (2%)
		No	62 (40%)	2,446 (98%)
Has your child been admitted in the past year?	81 months	Yes	74 (62%)	38 (2%)
		No	44 (38%)	2,493 (98%)
Has your child been admitted in the last 2 years?	103 months	Yes	113 (60%)	52 (2%)
		No	74 (40%)	2,454 (98%)
Has your child been admitted in the past year?	157 months	Yes	188 (68%)	138 (6%)
		No	89 (32%)	2,229 (94%)

Case Study 2: Using linkage to Hospital Episode Statistics to assess inconsistent reporting of self-harm in the ALSPAC Cohort

Introduction

This case study was undertaken to understand whether prevalence estimates of self-harm derived from self-reported data from the ALSPAC index participants[71] may be affected by non-response/loss to follow-up or the misreporting of episodes by responding participants[72]. In this case study we summarise findings relating to misreporting of self-harm episodes, where we hypothesise that the accuracy of reporting may have been affected by issues such as denial, reinterpretation, problems with recall, current mood, social desirability, or by misinterpretation of the study question[73]. Previous studies have reported inconsistencies across different reporting modes[74,75,76,77], although it is not clear which mode or measure elicits more accurate response.

Methods

The case study compares self-harm status reported in postal questionnaires with linked data from participants' HES records.. As outlined in case study 1, 3,195 ALSPAC-enrolled individuals had provided explicit consent for linkage to their health records before the cut-off date for these case studies. Of these, 3,027 (24.4%) were considered eligible for this analysis (live born singleton and twin deliveries from mothers who had enrolled into ALSPAC during the initial 1990-92 recruitment campaign). Full details of the sampling methods are available elsewhere[72]. Of this sub-sample, 2,363 participants had responded to a self-harm questionnaire at age 16 and were subsequently linked to at least one HES record. We compared HES recorded self-harm with participant reported self-harm.

The measures of self-harm used were:

> ALSPAC self-reported: "Have you ever hurt yourself on purpose in any way (e.g. by taking an overdose of pills or by cutting yourself)?" Question L3a from the 'Life of a 16+ Teenager' questionnaire[78].
> HES, hospital admission for self-harm: ICD-10 codes Y10-Y34, X60-X84, X40-X49
> HES, A&E attendance for self-harm: A&E diagnostic codes 141/142 (poisoning) or reason for attendance codes as 'deliberate self-harm'.
>
> For a sub-analysis we included:
> HES, admission code for a mental health condition: ICD-10 codes F00-F99.

---

[71] Boyd, A., Golding, J., Macleod, J., et al. 2013. Cohort Profile: The 'Children of the 90s'-the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology,* 42**,** 111-127.

[72] Mars B, Cornish R, Heron J, Boyd A, Crane C, Hawton K, Lewis G, Tilling K, Macleod J, Gunnell D. Using data linkage to investigate inconsistent reporting of self-harm and questionnaire non-response. Archives of suicide research. 2016 Apr 2;20(2):113-41.

[73] Velting, D. M., Rathus, J. H. & Asnis, G. M. 1998. Asking Adolescents to Explain Discrepancies in Self-Reported Suicidality. *Suicide and Life-Threatening Behavior,* 28**,** 187-196.

[74] O'Sullivan, M. & Fitzgerald, M. 1998. Suicidal ideation and acts of self-harm among Dublin school children. *Journal of Adolescence,* 21**,** 427-433.

[75] Ross, S. & Heath, N. 2002. A study of the frequency of self-mutilation in a community sample of adolescents. *Journal of Youth and Adolescence,* 31**,** 67-77.

[76] Ougrin, D. & Boege, I. 2013. Brief report: The self harm questionnaire: A new tool designed to improve identification of self harm in adolescents. *Journal of Adolescence,* 36**,** 221-225

[77] Bjarehed, J., Pettersson, K., Wangby-lundh, M., et al. 2013. Examining the acceptability, attractiveness, and effects of a school-based validating interview for adolescents who self-injure. *The Journal of School Nursing,* 29**,** 225-234

[78] http://discovery.closer.ac.uk/item/uk.alspac/55acf0db-3528-4174-bf35-462274f973c1

Results

Of the 3,027 ALSPAC participants included in this study we found that 54 (1.8%) had at least one self-harm event recorded in HES, including 41 (1.4%) with one or more hospital admissions and 18 (0.6%) with one or more A&E only attendances. 82 (2.7%) had at least one hospital admission for a mental health condition. The prevalence of hospital admissions for self-harm recorded in HES was slightly higher amongst those who did not complete the self-harm questionnaire than amongst those who did; although this assessment was based on a very small sample size and there was no robust evidence for a difference.

Conclusions

The findings of this case study are based on small sample sizes. However, the results provide preliminary evidence to suggest that self-harm prevalence estimates derived from self-report may be underestimated. As such, this case study serves to illustrate the potential utility of combining self-reported self-harm data with clinical routine records.

Case study 3: The Hertfordshire Cohort Study

Introduction
The Hertfordshire Cohort Study (HCS) comprises 3000 men and women born between 1931 and 1939, whose social, behavioural and biological characteristics were assessed when they were aged 59-73 years. Baseline cohort data have been linked to[79] HES APC records of 8741 admissions and (2) ONS records of 275 deaths; these events were experienced by cohort members during the decade after their baseline investigations (median follow-up period 8.1 years).

Descriptive epidemiology
The linked data in HCS allowed hospital use to be explored at the individual level. Admissions were common: 75% of men and 69% of women were admitted to hospital at least once during the follow-up period; among them, median numbers of admissions were 3 in men (IQR 1,6) and 2 in women (IQR 1,5). 48% of those who were ever admitted experienced at least one emergency admission and 70% stayed overnight[80].

Prospective investigations: methodology
The linked data also presented an opportunity for prospective investigation of the predictors (among the cohort data) of hospital admission. Because analysis of these multiple-failure survival data was challenging, we first conducted a review of suitable statistical techniques[81]. This identified the Prentice, Williams and Peterson Total Time (PWP-TT) model as the method of choice because: it captures information from every admission a cohort member experiences rather than just the first; reflects increasing risk with accumulated admissions for an individual; excludes time spent in hospital from time at risk of admission; and recognises that times to admission are correlated within an individual's admission history. This technique was therefore used in two further investigations.

The outcomes examined in both studies were types of admission, including: any; elective (day case or overnight); emergency; long stay (>7days); and readmission within 30 days of discharge[82]. These measures were derived using three administrative fields from the APC record: date of admission, method of admission; and date of discharge. Doubts about the accuracy of clinical coding therefore do not apply. Death was considered an alternative failure event in each model, because we considered it to represent an outcome *worse* than admission.

Prospective investigations: examples
The first of the prospective studies investigated the relationship between baseline grip strength and subsequent admission[83]. In women, lower grip strength was strongly associated ($p<0.001$) with increased risk of each admission outcome, with or without adjustment for potential confounding variables [unadjusted hazard ratio per standard deviation (SD) decrease in grip strength for: any admission/death 1.10 (95% CI: 1.06,1.14),

[79] Simmonds SJ, Syddall HE, Walsh B, Evandrou M, Dennison EM, Cooper C, et al. Understanding NHS hospital admissions in England: linkage of Hospital Episode Statistics to the Hertfordshire Cohort Study. Age Ageing. 2014;43:653-60.
[80] Ibid.
[81] Westbury L, Syddall H, Simmonds S, Cooper C, Aihie Sayer A. Identification of risk factors for hospital admission using multiple-failure survival models: a toolkit for researchers. BMC Med Res Methodol. 2016;16:46.
[82] readmission within 30 days of discharge was a binary variable not suited to PWP analysis.
[83] Simmonds SJ, Syddall HE, Westbury LD, Dodds RM, Cooper C, Aihie Sayer A. Grip strength among community-dwelling older people predicts hospital admission during the following decade. Age Ageing. 2015;44:954-9.

elective admission/death 1.09 (95% CI: 1.05, 1.13), emergency admission/death 1.21 (95% CI: 1.13, 1.31), long-stay admission/death 1.22 (95% CI: 1.13, 1.32) and unadjusted relative risk per SD decrease in grip strength for 30-day readmission/death 1.30 (95% CI: 1.19, 1.43)]. In men, significant associations were seen only with emergency admission/death, long stay admission/death and readmission within 30 days/death.

The second study examined the relationship between admission and clustering, in individuals, of four poor health behaviours at baseline: smoking; high weekly alcohol intake; low customary physical activity; and poor diet[84]. Among men and women, increased number of poor health behaviours was strongly associated ($p<0.01$) with greater risk of subsequent long stay and emergency admissions, and 30-day emergency readmissions. Hazard ratios (HRs) for emergency admission for 3/4 poor health behaviours in comparison with none were: men, 1.37 (95% CI 1.11 to 1.69); women, 1.84 (95% CI 1.22 to 2.77). Associations were unaltered by adjustment for age, body mass index and comorbidity.

Conclusion
Studies in HCS demonstrate the potential of linked data, though it is underused due to ongoing access problems. Routinely collected data are a particularly valuable source of follow-up in an ageing cohort, amongst whom response rates may be impacted by declining health when participant involvement is required.

[84] Syddall HE, Westbury LD, Simmonds SJ, Robinson S, Cooper C, Sayer AA. Understanding poor health behaviours as predictors of different types of hospital admission in older people: findings from the Hertfordshire Cohort Study. J Epidemiol Community Health. 2016;70(3):292-8.

Conclusions

The HES dataset is a potentially powerful resource for longitudinal and cohort studies. Its strengths lie in the breadth of clinical and demographic data that are available, the fact that it has very good coverage of secondary care patient interactions with the health services (within England) and that routine centralisation should mean there is an efficient mechanism for studies to access the resource. The weaknesses of the resource relate to continuing concerns regarding data quality and data missingness, although there are encouraging signs of improvement. Recent delays in gaining access to the resource now seem to be abating, with ALSPAC, the National Survey of Health and Development and the Whitehall II cohort all having recently linked to HES for the first time or re-established existing linkages that were interrupted by process change following the Partridge Review of data releases by the then NHS Information Centre. Increasing enhancement of cohort and longitudinal study resources through linkage to HES will lead to greater utilisation of the HES resource. This is to be welcomed, although a challenge remains for studies to accurately communicate the nature of the dataset to research users. On a more positive note, in our case studies we have briefly demonstrated some of the benefits that triangulation to clinical records can bring to longitudinal observational studies. This is to be encouraged as is the reciprocal interrogation of linked self-reported – HES records to assess quality in order to improve understanding of HES and future data collection and processing practice. This potential will be enhanced when studies – as is the case with ALSPAC and UK Biobank – link to both primary care and secondary care can then triangulate across all sources.