Quality of routine hospital birth records and the feasibility of their use for creating birth cohorts

Joanna Murray¹, Sonia Saxena¹, Neena Modi², Azeem Majeed¹, Paul Aylin¹, Alex Bottle¹, on behalf of the Medicines for Neonates Investigator Group†

¹Department of Primary Care and Public Health, Imperial College London, 3rd Floor Reynolds Building, St Dunstan’s Road, London W6 8RP, UK
²Department of Neonatal Medicine, Imperial College London, London, UK

Address correspondence to Joanna Murray, E-mail: joanna.murray@imperial.ac.uk

Abstract

Background Linkage between routinely collected hospital birth and other records offers the potential for epidemiological and public health research by developing population-level birth cohorts with cradle-to-grave follow-up. Data from births in English National Health Service hospitals are collected in the Hospital Episode Statistics (HES) database but are of uncertain quality.

Methods We examined the range and completeness of birth information recorded in HES and tested an approach for minimizing the effect of hospital-level variations by selecting hospitals with high completeness of recording (≥90%) for key fields. We discuss important methodological considerations when using routine healthcare data to develop a birth cohort.

Results The proportion of missing data in key birth record fields has been decreasing annually, such as gestational age and birth weight (from 46.2 and 43.9% in 2005/06 to 18.1 and 16.9% in 2009/10, respectively). We compared the important characteristics such as size and access to specialist neonatal care between 71 high-coding and 85 low-coding hospitals and found no significant differences, suggesting hospitals with high birth record completeness may be generalizable and representative of all hospitals.

Conclusions The completeness of recording of hospital birth information varies greatly between hospitals in England but is improving. It may be preferable and valid to construct cohorts from only hospitals with high completeness of recording.

Keywords birth cohort, hospital admission records, linkage, longitudinal data, hospital episode statistics

Introduction

Potential uses of maternity data to monitor health outcomes

Information on birth characteristics, such as gestational age and birth weight, is needed for many health services and epidemiological research studies examining short- and longer term clinical outcomes.¹ Hospital birth records are routinely collected throughout England and can provide a rich electronic source of clinical information about the health status of individuals at birth, as well as for measuring the quality of maternity care. Life-long assessment of outcomes is difficult in practice and much infant research is therefore selective, poorly generalizable and focused on short-term outcomes. Secondary use of electronic health records cannot replace nor precisely replicate the value of data collected for bespoke prospective observational studies, but the potential benefits for research and patient care are well established²–⁴ and the use of these data involves comparatively minimal costs. These data can facilitate benchmarking and comparison of outcomes between different hospitals or geographical areas. Cohorts based on these data can provide representative population-level information that

¹Members of the Medicines for Neonates Investigator Group: Neena Modi, Peter Brocklehurst, Jane Abbott, Kate Conteloe, Elizabeth Draper, Azeem Majeed, Jacquie Kemp, Deborah Ashby, Alys Young, Stavros Petrou.
Joanna Murray, Doctoral Research Student
Sonia Saxena, Clinical Senior Lecturer in Primary Care
Neena Modi, Professor of Neonatal Medicine
Azeem Majeed, Professor of Primary Care
Paul Aylin, Clinical Reader in Epidemiology and Public Health
Alex Bottle, Senior Lecturer in Medical Statistics

© The Author 2012, Published by Oxford University Press on behalf of Faculty of Public Health. All rights reserved
is highly generalizable, has power to detect small effect sizes and relates directly to real clinical practice. Such cohorts have the potential to answer an extensive range of research questions that require longitudinal follow-up, from examining long-term health outcomes after preterm birth, to identifying risk factors for childhood hospital admission with influenza or bronchiolitis.

Linkage with other health records
In health systems with universal coverage, the value of clinical databases for research can be significantly enhanced by linking different data sources, creating the potential to develop cradle-to-grave data sets for the whole population. Probabilistic or deterministic linkage methods can be used to match records for the same individual from different databases, using unique personal identifiers such as National Health Service (NHS) number. Linkage between data sets can provide validation of recording, coding and completeness of data, help to develop more robust clinical case definitions and provide information on events that happen outside healthcare settings including deaths. Other possible benefits include adverse drug reaction monitoring, improved healthcare-associated infection surveillance with linkage between microbiological and clinical records and also identification and long-term follow-up of clinical trial participants. Area-level linkage to data on deprivation, pollution and other environmental data sources could provide valuable information for public health research studies. Large-scale population-based linkage presents its own complex ethical challenges of consent, confidentiality and secure data storage. These are best addressed by robust information governance and engaging with patients and the public about how their records are to be used.

How have administrative birth records been used so far?
Non-UK
The effectiveness of routine health record linkage in adults has been demonstrated in Australia and Canada, where it has improved both data quality and utility. Birth records have been successfully linked to hospital discharge data in Australia, with matching rates of 99%. In several regions of the USA, data from birth certificates have been linked to hospital discharge records to examine maternal outcomes. In Scandinavian countries, the assignment of unique personal identification numbers permits the development of population-based cohorts, facilitating a broad array of epidemiological studies, such as research in Denmark investigating the impact of place of birth and familial risk factors on risk of autism.

UK
In Wales, the Secure Anonymized Information Linkage (SAIL) databank collates anonymized person-based electronic health and social care data, which are now being combined to establish a Wales-wide Electronic Cohort for Children. This databank has already been used to identify potential clinical trial participants from primary care data. The Scottish Health Informatics Programme (SHIP) provides linked health and social care data from birth through to death. To date, SHIP has primarily been used for pharmacovigilance and diabetes epidemiology research. The Oxford Record Linkage Study contains over 10 million records for 5 million people since 1963. This has been used for longitudinal studies identifying maternal and perinatal risk factors for conditions such as inflammatory bowel disease, asthma and coeliac disease.

Administrative birth data are captured in England in both hospital discharge records and also birth registration information collected by the Office for National Statistics (ONS). The use of information in the English Hospital Episode Statistics (HES) delivery and birth records for research has been limited. A previous feasibility study showed high rates of linkage between HES maternity and ONS birth records. In addition to these national resources, the NHS Numbers for Babies (NN4B) Service, introduced in 2002, collects a small data set containing some key fields that are not recorded in birth registrations, such as gestational age. The NN4B service ensures every baby is allocated a unique NHS number shortly after birth. Linkage between NN4B and HES birth records indicated that improvements in the quality and completeness of HES maternity data are needed. HES maternity data have also been used to validate birth information recorded in the Millennium Cohort Study, a UK-wide longitudinal observational cohort of nearly 19 000 babies.

We aimed to examine the range and completeness of information recorded in hospital birth records in England and to identify the methodological issues when using such routine healthcare data to develop a birth cohort. First, we describe linkage issues and the HES data set. Second, we summarize the data quality and assess the feasibility of creating a birth cohort using records from hospitals with high data completeness.

Methods
Overview of HES data
HES is a complex administrative database containing details of all admissions to NHS (public) hospitals in England
electronically since 1989. HES contains a wide range of data including coded clinical diagnoses, procedures, geographical information on where patients are treated and demographic information such as patient age, sex and ethnicity. Further information on the HES database is presented in Appendix 1.

**HES maternity records and baby tails**

When a mother gives birth, her hospital admission record changes from a general inpatient admission record to a maternity record and is updated as such before it is submitted to HES. HES contains two types of maternity record, the delivery and the birth record, both of which contain a ‘baby tail’ of an additional 19 fields (http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=925). A list of some of the maternity data collected within HES can be found in Box 1 and HES maternity records are described in more detail in Appendix 1.

<table>
<thead>
<tr>
<th>Box 1 Key information collected in HES maternity records</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Method of onset of labour</td>
</tr>
<tr>
<td>• Place and date of delivery</td>
</tr>
<tr>
<td>• Miscarriages and ectopic pregnancies</td>
</tr>
<tr>
<td>• Live and still births</td>
</tr>
<tr>
<td>• Deliveries and births with complications</td>
</tr>
<tr>
<td>• Multiple deliveries</td>
</tr>
<tr>
<td>• Length of gestation</td>
</tr>
<tr>
<td>• Birth weight</td>
</tr>
<tr>
<td>• Neonatal unit admissions and level of care</td>
</tr>
<tr>
<td>• Antenatal and postnatal length of stay</td>
</tr>
</tbody>
</table>

**Linkage across time to produce a HES cohort**

The processing of HES birth records described for the first time here can be used to develop cohorts of children with the potential of long-term follow-up. Using a unique personal identifier, we can link individual birth records through time to subsequent admission records to provide longitudinal information on severe episodes of illness requiring hospitalization. The unique identifier used in HES (‘HESID’) is derived by matching records for the same patient using a combination of NHS number and local patient identifier, plus the patient’s date of birth, sex and postcode.

**Processing and methodological issues using HES maternity data**

HES data are divided into financial years from 1 April in a given year to 31 March in the following year. We outline the key data processing and methodological issues required to use these HES birth records for research purposes, described in detail in Appendix 1.

**Analysis**

The NHS Information Centre publish annual quality reports examining the coverage of the data submitted by NHS hospitals to HES, which have repeatedly found that for many hospitals completeness of birth fields is often inadequate. We examined the completeness of recording for baby tail fields over 5 years (2005/06–2009/10) and also compared the total number of births with ONS birth registrations. We compared the proportion of missing data for each baby tail field in 2005/06 to the most recent year available (2009/10), using chi-squared tests, with associated P-values presented in Table 1. Using a cut-off of 90% completeness of recording for key birth fields (gestational age and birth weight), we compared hospitals with highly complete data with the others in terms of hospital characteristics from available data to assess how representative of all of England a birth cohort formed from babies at high-coding hospitals would be. To test for significant differences between these hospitals with high versus low completeness of birth record fields, we used chi-squared tests to compare proportions and t-tests to compare mean values (associated P-values presented in Table 2).

**Results**

**Coverage and completeness of recording**

The completeness of recording of all birth fields has significantly improved over the 5 years between 2005/06 and 2009/10 (Table 1). The values presented differ from those published by HES online because of our extensive data cleaning processes, which resulted in the removal of some invalid birth records (such as those with an invalid date of birth). Among our 2009/10 data set, birth weight and gestational age were recorded for 83 and 82% of births, respectively.

**Comparison with data from ONS birth registrations**

Overall, the HES cohort captured 87% of all live births recorded by the ONS in England during the time period. There are various explanations for this difference. An estimated 97% of live births in England occur in NHS hospitals. Those occurring in private hospitals or at home may not be recorded in HES. In addition, because HES contains discharge records, some births which occurred within the
cohort year but were discharged from hospital after this period would not be captured in our data set. Most of the births that are not captured within our HES cohort may have been excluded as a result of the quality criteria we applied to remove records with invalid information reported in key fields, or due to the incomplete coverage issues described earlier.

Approach for minimizing the effect of variable data completion between hospitals
In our 2007/08 data set, 50% of records had no birth weight recorded, 48% were missing gestational age, 43% had no maternal age recorded and 17% had no indication of whether the baby received any specialist neonatal care, with considerable variation in the completeness of information recorded in birth records between NHS providers. Some hospitals with many births had recorded gestational age and birth weight in <10% of birth records, whilst other similarly large hospitals recorded this information for >90% of births. There appears to be no correlation between the number of births occurring in a hospital and the completeness of recording of these prematurity indicators. Therefore, depending on the study purpose and the exposure and outcome measures of interest, we suggest selecting birth records only from hospitals with high completeness of recording. We tested this by creating a 2007/08 birth cohort where we selected only birth records from hospitals where >90% of their birth records contained complete recording of key variables, birth weight and gestational age. The resulting cohort included 296,618 babies born at 71 hospitals across England. Table 2 shows a comparison of characteristics of included (n = 71) and excluded (n = 85) hospitals. The mean number of births, maternity beds and access to neonatal intensive care were mostly similar among hospitals with high and low completeness of recording.

<table>
<thead>
<tr>
<th>Baby tail fields in HES birth records (field name)</th>
<th>% Missing or unknown</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaesthetic given during labour or delivery (delpren)</td>
<td>41.9 41.8 44.8 29.6 16.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Anaesthetic given post-labour or delivery (delposn)</td>
<td>48.1 46.0 49.6 34.8 21.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Antenatal days of stay (antedur) (derived from other HES fields)</td>
<td>0.2 0.2 0.2 0.1 0.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Baby's age in days (neodur) (derived from other HES fields)</td>
<td>0.1 0.0 0.0 0.0 0.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Birth order (birorder)</td>
<td>33.9 36.9 39.4 24.9 13.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Birth weight (birweit)</td>
<td>43.9 47.1 50.1 31.3 16.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Delivery place change reason (delchang)</td>
<td>45.2 45.8 47.4 34.6 21.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Delivery method (delmeth)</td>
<td>35.1 35.8 44.3 30.6 14.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Delivery place (actual) (delplace)</td>
<td>44.0 46.8 57.0 41.3 17.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Delivery place (intended) (delinten)</td>
<td>41.3 42.7 43.6 30.2 14.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>First antenatal assessment date (anasdate)</td>
<td>41.7 44.3 44.6 34.7 20.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gestation in weeks at first antenatal assessment (anagast)</td>
<td>54.5 63.9 55.2 45.6 28.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Length of gestation (gestat)</td>
<td>46.2 54.2 48.0 34.6 18.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Birth status (birstat)</td>
<td>43.9 47.0 48.0 32.9 16.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Labour/delivery onset method (delonset)</td>
<td>36.2 37.7 41.1 25.5 11.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mother's age at delivery (matage)</td>
<td>42.4 43.3 43.0 34.5 30.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Neonatal level of care (neocare)</td>
<td>16.1 16.0 17.1 18.4 12.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of babies (numbaby)</td>
<td>31.8 33.3 36.1 23.6 11.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Resuscitation method (biresus)</td>
<td>44.2 45.3 48.0 34.2 21.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Status of person conducting delivery (delstat)</td>
<td>38.9 42.6 48.4 33.9 19.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total number of births(^a)</td>
<td>554,521 566,749 575,493 589,684 603,786</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^a\)Total number of births in a cohort after the removal of duplicate episodes and completion of data cleaning processes as described in Appendix 1.
Discussion

Main findings of this study

The completeness of recording of information in hospital birth records is highly variable between hospitals but is generally improving. The proportion of missing or unknown data in key birth record fields has decreased over the 5-year period, such as gestational age (from 46.2% in 2005/06 to 18.1% in 2009/10) and birth weight (from 43.9% in 2005/06 to 16.9% in 2009/10). We compared characteristics of hospitals with high (n = 71) and low (n = 85) completeness of recording of key birth fields and found no significant differences between them in terms of important hospital factors such as size or access to specialist Neonatal care. This suggests a birth cohort derived from these 71 hospitals would be representative of the whole country.

What is already known on this subject

Detailed clinical information is routinely captured in hospital birth and other records but their value for use in public health, epidemiological and health services research has yet to be fully exploited. The Healthcare Commission’s review of maternity services found that even among larger, well-respected maternity units, information technology was poor and data collection systems inadequate.36 Coverage of hospital deliveries was estimated to be around 73% between 2001/02 and 2005/06, but just 14% for home deliveries.32 As a result of these quality concerns, additional rigorous validation was carried out by the NHS Information Centre on HES maternity data, including intensive cleaning of systematic coding errors.32 Some of the quality and coverage issues specifically affecting HES maternity data are listed in Box 2.

Box 2 Quality and coverage issues affecting HES maternity data

- Stand-alone maternity systems in around 20 hospitals are not linked to their patient administration system, from which HES data are obtained (via the Secondary Uses Service).
- Some hospitals return data on the number of birth episodes or delivery episodes but not both.
- Transfer of maternity information between systems leaves scope for data errors and shortfalls.
- Stillbirths are not reliably recorded in every hospital and are not allocated an NHS number.28,31,34
- Lack of a priori definitions for data variables resulting in inconsistencies in data entry.
- Use of aggregate or categorized fields rather than the raw data.

What this study adds

Individual birth records linked to other healthcare records can provide a valuable source of longitudinal, population level data for developing cohorts from birth. We describe...
some of the methodological and processing issues that need to be considered when using birth records in this context. Completeness of birth information recording is highly variable between English NHS hospitals but has improved over the 5-year period from 2005/06 to 2009/10. Where key birth information such as gestational age and birth weight are missing, it may be necessary and valid to select data only from hospitals with high completeness of recording.

**The future of routine data collection for research**

With limited funding, restricted resources and improved recognition of the necessity for life-long follow-up for infant research, it is imperative that the research community, funders and the public maximize the value of information that can be gleaned from existing routine data sources and avoid duplication of effort. This was highlighted by the public inquiry report into the Bristol Royal Infirmary paediatric cardiac unit, which recommended that HES should be supported as a national resource to monitor healthcare outcomes. HES has been commonly used to examine disease time trends and standards of care at an aggregate level. Recording of birth information needs to improve in some NHS hospitals to provide robust baseline population estimates of birth outcomes. Encouragingly, the data we have presented from recent years indicates that completeness of recording of HES birth records is improving. Individual follow-up from birth is feasible using HES but the limitations and processing issues we have described are important methodological considerations to be taken into account.

Data capture is likely to be more complete and most accurate in records that are entered by clinicians at the point of care. In the UK for example, Neonatal Networks submit neonatal specialist care records to a National Neonatal Research Database managed by the Neonatal Data Analysis Unit (based at Imperial College London in collaboration with the Chelsea and Westminster NHS Foundation Trust). These records capture real-time operational data that are actively used in the processes of clinical care. The NHS Newborn Infant Physical Examination Programme and Newborn Hearing and Bloodspot Screening programmes provide additional sources of neonatal healthcare data in the UK. However, there are huge inconsistencies between data sets in the types of data captured and how they are defined. The NHS Information Centre has been commissioned to develop a maternity record. If this is to be truly useful it is essential that current uncertainties regarding the retention of identifiers, access for secondary use, clarity regarding data definitions and validation are addressed. A logical next step must also be to develop robust linkage between birth records and more detailed clinical data sets. For example, linkage between hospital and general practice records (belonging to both the mother and baby) would add substantial utility to the data, providing insight into the effects of exposures during pregnancy on subsequent infant health outcomes. Rather than existing ad hoc approaches to linkage, our aim should be routine database linkage on a national scale.

**Limitations of this study**

We were unable to determine how accurately the identifier we used to deterministically link records across time is allocated to unique individuals, although we know that an improved, complex algorithm is used to derive ‘HESID’, based on a combination of other identifiers as described earlier. Probabilistic approaches can facilitate linkage with records from other databases with differing availability and quality of unique identifiers, by developing algorithms to assign individual unique identifiers based on a combination of variables such as date of birth and postcode. Provided methods are tailored to the data sets used and the context of how data are collected is considered, linkage methods can correctly identify a high proportion of matched records even with limited quality or availability of identifiers.

Our comparison of maternity characteristics between hospitals with high and low completeness of recording was limited to specific fields for which we had access to information from all hospitals, such as their mean number of births, maternity beds and specialist neonatal care facilities. However, we are confident that our comparison of these key hospital factors between hospitals with high and low completeness of recording of birth information provides sufficient evidence that these do not significantly differ. We also compared the mean maternal age, the proportion of babies of non-white British ethnicity and the proportion of babies in the most deprived Carstairs quintile. These appear similar between the two groups, though data incompleteness means we cannot be certain of this.

**Conclusion**

Detailed clinical information including prematurity, birth weight, complications and comorbidities is routinely captured in hospital birth and other records. Despite some complex data-processing requirements, HES and other birth records can be used for population-scale epidemiological research. However, the completeness of birth information recorded in English hospitals is variable. Where key birth information such as gestational age and birth weight are missing, it may be preferable to select data only from hospitals with high levels of completeness of recording; our
analysis suggests that the results may be generalizable to all hospitals. HES data can be used to develop longitudinal cohorts by linking individual birth records to subsequent hospital admission and other healthcare, microbiological or treatment records for many vital purposes.

Details of contributors and guarantor

J.M., A.B. and S.S. conceived the study. J.M. performed all analyses and wrote the first draft. All authors contributed in the revision of the manuscript. J.M. is the guarantor.

Ethical approval

We have permission from the NIGB under Section 251 of the NHS Act 2006 (formerly Section 60 approval from the Patient Information Advisory Group) to hold confidential data. We have ethical approval to use them for research and measuring quality of delivery of health care from the South East Ethics Research Committee.

Funding

J.M. is funded through a National Institute for Health Research (NIHR) Programme Grant for Applied Research (RP-PG-0707-10010) held by N.M.; J.M., A.M. and N.M. declare financial support from this award for the submitted work. Support for data analysis was received from the Neonatal Data Analysis Unit led by N.M. Unrestricted support for the Neonatal Data Analysis Unit from Abbott International and Danone UK is acknowledged. A.B. and P.A. are part of The Dr Foster Unit at Imperial, which is principally funded via a research grant by Dr Foster Intelligence, an independent healthcare information company and joint venture with the Information Centre of the NHS. The Unit is affiliated with the NIHR Imperial Centre for Patient Safety and Service Quality at Imperial College Healthcare NHS Trust. The Department of Primary Care & Public Health at Imperial College is grateful for support from the NIHR Imperial Bio-venture with the Information Centre of the NHS. The Unit is

Conflict of interest

All authors have completed the Unified Competing Interest form at www.icmje.org/coiDisclosure.pdf (available on request from the corresponding author). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or Department of Health.

References


Appendix 1

Processing and methodological issues using HES maternity data

Overview of HES data

The basic unit recorded in HES is the finished consultant episode (the period of time during which a patient is under the care of one consultant). A 'spell' or admission is defined...
as the continuous period of time spent as a patient within one hospital, from admission to discharge or transfer to another provider and may therefore include more than one consultant episode.

**HES maternity records and baby tails**

The delivery record is for the mother and contains the same information as a general record plus an additional baby tail with information about the delivery. The birth record is for the baby and also contains general record information plus the baby tail, which contains identical information to the corresponding baby tail in the mother’s delivery record. Diagnoses and procedures recorded in the birth record refer to the baby and, conversely, diagnoses and procedures in the delivery record refer to the mother. For multiple births, separate tails for each baby will appear in the mother’s delivery record, but each birth record will contain only the individual baby’s relevant tail. HES data are collected from births in NHS and non-NHS hospitals and at home, although information from births outside NHS hospitals is often incomplete.

**Identifying birth episodes**

The first step in developing a birth cohort study using HES was identifying all individual birth episodes within a given year. To identify birth episodes, we used the ‘admmeth’ variable which contains a code recording how the patient was admitted to hospital. This field was used to select all records with an admission method coded as 82 (other: babies born in healthcare provider) or 83 (other: babies born outside the healthcare provider, except when born at home as intended).

**Removing duplicate birth episodes**

The next step was identifying any duplicate birth episodes (a typical but limited problem with administrative data) using the unique personal identifier (“HESID”). We looked specifically at the first episode within the given spell (i.e. the initial birth episode itself) using the ‘epiorder’ field which contains the number of the episode within the current spell. In our data set from financial year 2007/08, we identified 5600 individuals with more than one birth episode recorded, of which 5545 had two and 55 had more than two birth episodes recorded. We excluded individuals with more than two birth episodes from the cohort because it would be almost impossible to decipher which birth record ought to be retained when comparing more than two duplicates. Many records from individuals with two birth episodes were identical matches (2750 individuals), having two exact duplicate birth episodes with identical information recorded in all fields. Where this was the case, one of the identical records was deleted.

Among the remaining individuals with two birth episodes, the records did not contain matching information. It is very difficult to determine which of two birth records with the same unique identifier is likely to be the true record of that individual’s birth. In previous studies using HES and similar databases, the common approach has been to retain only the first observation recorded. We observed that the basic demographic information for each pair of birth episodes tended to be common to both and for most it was only the diagnostic fields that differed. Consequently, records for individuals with duplicate birth episodes were compared using recording of diagnostic information and only the birth episode with the most diagnostic information recorded (number of non-empty diagnosis fields) was retained in the cohort. It was not possible to determine whether this approach is better than randomly retaining one birth episode. However, the record with the most diagnostic information might be the more accurate of the two because it is common for hospitals to resubmit data after they have carried out all their diagnosis coding more thoroughly (their first submissions may contain only very basic diagnostic information). Therefore, records can appear as duplicates if hospitals fail to report them as re-submissions to the secondary uses service (the NHS data repository). Consequently, for an individual with two birth episodes, the one with more diagnostic information may be the more accurate record. Handling duplicate records is a common, challenging problem when using routine data such as HES and requires careful consideration of the study context.

**Summarizing information from additional episodes in a birth spell**

Babies can have more than one episode of care within their birth admission, for example if a baby receives specialist care from a different consultant, is transferred between hospitals or is admitted to a neonatal unit. These additional episodes occur within the same birth spell but, where the initial birth event would have an ‘epiorder’ value of 1, subsequent episodes in the birth spell have an ‘epiorder’ value >1. To facilitate one-to-many linkage to subsequent hospital admission records, it is easiest to develop a data set consisting of one observation (or row) per individual. Seven thousand four hundred and ten individuals in our 2007/08 birth cohort had more than one episode in their birth spell. To simplify the data set to consist of only one birth record per patient, we summarized key information such as any
premature or congenital anomaly diagnoses using flags. These were then linked back into the original birth episode and subsequent episodes in the birth spell were dropped.

**Cleaning variables**

A range of exclusion criteria were developed to clean key variable fields and examine the quality of coding. The Care Quality Commission (CQC) conducted a review exploring quality indicator specifications, used to assess the quality of HES maternity data from 2009/10.\(^{36,40}\) We combined the criteria identified within the CQC review\(^{36}\) and HES in-patient cleaning rules\(^ {41}\) and applied these to the HES birth fields to ensure suspicious data and invalid records were removed. In addition, we identified any stillbirths that were recorded (using the birth status and discharge method fields) and removed these from the final cohort of live births.

Identical baby tail information recorded for each baby can be found in their mother’s delivery record. Maternity systems can record up to nine birth tails for each delivery, allowing information from multiple births to appear in the mother’s delivery record. In some instances, we found a baby’s information was not recorded in the first field of a given variable. For example, if a baby was the second twin, their gestational age at birth (‘gestat’) may have appeared in the second field (‘gestat_2’) with the first field blank (‘gestat_1’) because in the mother’s delivery record this would contain the first twin’s gestation. To simplify analyses, we condensed individual birth records so that only one field for each variable existed. So for the above example, we transferred the relevant information for the gestation variable from ‘gestat_2’ into ‘gestat_1’ and then removed all additional fields for that variable (i.e. ‘gestat_2’ to ‘gestat_9’).

**Contextual issues to consider**

As well as the completeness and quality of recording, attrition is also important to consider because although an individual may be born in a English NHS hospital and therefore included in a HES birth cohort, migration, deaths and admissions outside of English NHS hospitals are difficult to identify without individual patient consent and follow-up.