HSR Health Services Research

© Health Research and Educational Trust
DOI: 10.1111/1475-6773.12272
RESEARCH ARTICLE

Identifying Possible False Matches in Anonymized Hospital Administrative Data without Patient Identifiers

Gareth Hagger-Johnson, Katie Harron, Arturo Gonzalez-Izquierdo, Mario Cortina-Borja, Nirupa Dattani, Berit Muller-Pebody, Roger Parslow, Ruth Gilbert, and Harvey Goldstein

Objective. To identify data linkage errors in the form of possible false matches, where two patients appear to share the same unique identification number.

Data Source. Hospital Episode Statistics (HES) in England, United Kingdom.

Study Design. Data on births and re-admissions for infants (April 1, 2011 to March 31, 2012; age 0–1 year) and adolescents (April 1, 2004 to March 31, 2011; age 10–19 years).

Data Collection/Extraction Methods. Hospital records pseudo-anonymized using an algorithm designed to link multiple records belonging to the same person. Six implausible clinical scenarios were considered possible false matches: multiple births sharing HESID, re-admission after death, two birth episodes sharing HESID, simultaneous admission at different hospitals, infant episodes coded as deliveries, and adolescent episodes coded as births.

Principal Findings. Among 507,778 infants, possible false matches were relatively rare (n = 433, 0.1 percent). The most common scenario (simultaneous admission at two hospitals, n = 324) was more likely for infants with missing data, those born preterm, and for Asian infants. Among adolescents, this scenario (n = 320) was more common for males, younger patients, the Mixed ethnic group, and those re-admitted more frequently.

Conclusions. Researchers can identify clinically implausible scenarios and patients affected, at the data cleaning stage, to mitigate the impact of possible linkage errors.

Key Words. Computerized patient medical records, data linkage, data quality, medical errors

Hospital Episode Statistics (HES) are used to capture all admissions to National Health Service (NHS) hospitals in England. An algorithm is used to create a unique and anonymous identifier (the HESID) which links episodes of care belonging to the same patient over time (internal data linkage). The algorithm is designed to minimize false matches (maximize specificity), a situation in which two patients might share the same HESID (Figure 1). False matches can lead to clinical harm (Joffe et al. 2012; McCoy et al. 2013; Middleton et al. 2013) and breaches of patient confidentiality (Connecting for Health 2005; HSCIC 2013b). It is possible that false matches can be identified even in anonymized datasets, using implausible clinical scenarios (e.g., a patient dies and then is apparently re-admitted). False matches can lead to incorrect estimates of incidence (Schmidlin et al. 2013) and biases estimates of relative risk. Given increasing emphasis on data linkage (Dunn 1946; Bohensky et al. 2010), understanding linkage errors is important for researchers and operational users of HES and other administrative datasets.

Internal data linkage algorithms that are deterministic use a set of rules for deciding whether two or more pairs of records should be deemed a match or nonmatch, according to patient identifiers (Christen 2012). They need to be sufficiently sensitive to link patient records together. Data linkage errors can occur for various reasons, including errors in patient identifiers, patient identifiers that match by coincidence, or errors in data submitted by hospitals (HSCIC 2009b, 2012a). Relatively little is known about the extent of linkage errors in HES, or indeed other administrative data sources. Even small amounts of linkage error disproportionately affecting certain groups can bias analyses (Lariscy 2011). Our aims were therefore to determine whether possible false matches could be detected in nonidentifiable HES data. This would provide a minimal estimate of the size of the problem, identify which groups of the population are more likely to have records that could have

Address correspondence to Gareth Hagger-Johnson, Ph.D., Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health, Centre of Paediatric Epidemiology and Biostatistics, 30 Guilford Street, London WC1N 1EH, UK; email: g.hagger-johnson@ucl.ac.uk. Katie Harron, Ph.D., is with the Institute of Health Informatics, Faculty of Pop Health Sciences, Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health, Centre of Paediatric Epidemiology and Biostatistics, London, UK. Arturo Gonzalez-Izquierdo, Ph.D., is with the Institute of Child Health, Faculty of Pop Health Sciences, Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health, Centre of Paediatric Epidemiology and Biostatistics, London, UK. Mario Cortina-Borja, Ph.D., is with the Centre for Maternal and Child Health Research, School of Health Sciences, Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health, Centre of Paediatric Epidemiology and Biostatistics, London, UK. Harvey Goldstein, Ph.D., is with the Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health, Centre of Paediatric Epidemiology and Biostatistics, London, UK. Nirupa Dattani, Ph.D., is with the City University London, UK. Berit Muller-Pebody, Ph.D., and Ruth Gilbert, M.D., are with the Public Health England, London, UK. Roger Parslow, Ph.D., is with the University of Leeds, Leeds, UK. Harvey Goldstein, Ph.D., is also with the University of Bristol, Bristol, UK.

Figure 1: Four Scenarios Following Data Linkage of Two Records

	Same individual	**Different individuals**
Same identification number	✓ True match	**False match** (specificity)
Different identification number	**Missed match** (sensitivity)	✓ True non-match

falsely matched, and evaluate whether possible false match scenarios can be found in two different clinical settings comprising different kinds of patients (e.g., infants in maternity units vs adolescents admitted as an emergency).

METHODS

In England, HES data are released by the Health and Social Care Information Centre (HSCIC) annually. Hospitals submit local data to the Secondary Use Service (SUS), which is regularly updated (HSCIC 2009b). Once a year, the HSCIC release a fixed extract from SUS, after cleaning the data (HSCIC 2012a) and assigning the HESID (HSCIC 2009b) to each episode of care. An episode is a continuous period of care provided by a consultant at a hospital (Health and Social Care Information Centre 2013). A patient's care pathway might involve several episodes at the same hospital, for example, if the patient receives care from several different physicians. To link episodes of care belonging to the same patient, an internal data linkage algorithm is used to assign the HESID.

The HES algorithm involves three passes, assigning the same HESID to records that match at any pass (HSCIC 2009b): (1) records with the same sex, date of birth, AND NHS number; (2) records with the same sex, date of birth, AND local patient identifier within each hospital; (3) records with the same sex, date of birth, AND postcode except those from communal establishments (Health and Social Care Information Centre 2013), or those having NHS numbers. The NHS number is a 10-digit identifier assigned to each UK citizen (HSCIC 2012b); 98.9 percent of HES records in 2011/2012 had a valid (although not necessarily correct) NHS number (HSCIC 2013a). In 2002, the NHS Numbers for Babies service was set up to issue an NHS number soon after birth (NHS Information Centre 2007). This reduced the potential for multiple births to be falsely matched at this step.

Data

Inpatient admissions for infants: HES inpatient admissions for infants (age 0–1 year, including stillbirths) from April 1, 2011 to March 31, 2012. Twenty hospitals do not submit maternity data to HES (Dattani, Datta-Nemdharry, and Macfarlane 2007) and are not included in our study population.

Inpatient admissions for adolescents: HES inpatient admissions for adolescents (age 10, 19) from 590 hospital codes (including hospitals with codes that changed over the study period) with at least one unplanned injury between 10–19 years of age from April 1, 2005 to March 31, 2011.

Outcome Variable

Six implausible clinical scenarios were used to identify possible false matches. These scenarios were chosen on the basis that they can be established even in nonidentifiable data and can be used to provide a minimal estimate of possible false matching. Scenarios 1 to 5 apply to infants; scenarios 2, 4, and 6 apply to adolescents:

1. *Multiple births with the same HESID.* Multiple births indicated by birth order or baby number sharing the same HESID.
2. *Re-admission after death.* Infants/adolescents who died in hospital (discharge destination coded as death or discharge method died/stillborn) and were subsequently re-admitted according to HESID.
3. *Birth, followed by a subsequent birth episode.* More than one birth episode with the same HESID.
4. *Simultaneous admission at different hospitals.* Admissions on the same day at different hospitals, with different discharge dates except where the method of admission/discharge was from/to another hospital (i.e., transfers).

5. *Infant episodes coded as deliveries.* Infant episodes coded as a delivery which could indicate a mother and baby sharing the same HESID. Currently, HES do not currently allow mothers and children to be linked, and so information pertaining to the mother should not appear on the infant's record.
6. *Adolescent episodes coded as births.* Adolescents having episodes coded as birth episodes (i.e., being born, not deliveries) were flagged as possible false matches.

Predictors of Simultaneous Admissions. The analytic sample for both extracts comprised those with available data on episode start/end dates and on variables hypothesized to be predictive of possible false matches: age, sex, ethnic group; and for infants, multiple births and gestational age. Ethnic minority status was recorded as one of 16 categories which we grouped into White, Mixed, Asian, Black, Chinese, and Other. We used information from the birth to identify multiple births using information on birth order. Gestational age <37 weeks were classified as "preterm." The Index of Multiple Deprivation 2004 (IMD2004) was used to provide an area-based, aggregated measure of socio-economic status (SES) for infants who were re-admitted (as this information was not available on the birth record) and for adolescents at any admission (HES Data Quality Team 2014). IMD2004 score was divided into thirds for each analytic sample in order to capture linear trends by ensuring sufficient numbers of patients from ethnic groups were included in each third.

Statistical Analysis

Descriptive analyses involved testing for differences across study variables for the most common possible false match scenario, using chi-square tests. For infants and adolescents, mixed effects logistic regression was used to identify predictors of the most common false match scenario (simultaneous admission) due to low statistical power on less common scenarios. A two-level model was fitted where hospital was the level 2 unit, allowing for clustering of infants within hospitals. Models were fitted using *Stata v12.1* (Stata Corp, College Station TX, USA).

In sensitivity analyses for the infant sample, we repeated the model after excluding hospitals considered to have poor quality data returns. Poor quality data hospitals were defined as those having >30 percent of delivery records missing information about the onset of labor, fewer than 500 admissions, <10 percent or >50 percent of labors induced, no elective cesareans, an unadjusted rate of spontaneous labors <50 percent or >90 percent, or >30 percent missing data on gestational age at delivery. These indicators were developed by panels of clinical and academic experts, taking information on validity, fairness, statistical power, and technical specification into account (Knight et al. 2013).

RESULTS

Infants

After excluding 150 (0.02 percent) infants with missing data on sex, the analytic sample comprised 733,770 unique HESIDS (infants, Figure S1) from 166 hospitals, with available data on study variables (April 1, 2011 to March 31, 2012). Variables with larger proportions of missing data were included in the analysis, by assigning a missing value indicator (Table 1). Of the unique HESIDS, 131,466 (17.9 percent) were re-admissions.

There were 433 (0.1 percent) possible false matches: multiple births sharing the same HESID ($n = 2$), died then re-admitted ($n = 18$), birth, followed by a subsequent birth episode ($n = 17$), simultaneous admission ($n = 324$), infants with delivery episodes ($n = 69$). Due to low statistical power for relatively rare scenarios, we focused on simultaneous admissions for the main analysis. The characteristics of infants apparently being admitted to two different hospitals simultaneously are shown in Table 1. They were more likely to be male, preterm, Asian, or missing data on ethnic group and gestational age. Ten hospitals accounted for 5.9 percent of apparent simultaneous admissions, 30 accounted for 9.6 percent, and 50 accounted for 15.7 percent.

The fitted logistic models showed that after adjusting for predictors (Table 2), apparent simultaneous admissions were more common for infants missing gestational age (OR = 2.35, 95 percent CI 1.42, 3.89) and those born preterm (OR = 2.12, 95 percent CI 1.42, 3.17), compared to infants born to term. The Asian ethnic group were also at increased risk (OR = 2.12, 95 percent CI 1.54, 2.91). Infants without data on birth order were less likely to have possible false matches (OR = 0.52, 95 percent CI 0.30, 0.92). There was significant variation across hospitals in the likelihood of a false match, comparable in size on the logistic scale to the effects for preterm and the Asian group.

Table 1: Descriptive Statistics for Study Variables According to Most Common Possible Match Scenario: Apparent Simultaneous Admission, Two Hospitals

	Infants (n = 733,770)					*Adolescents (n = 1,678,623)*			
	Not (n = 773,446)	*Simultaneous Admission (N = 324)*	*p*	*Total Sample*		*Not (n = 1678,303)*	*Simultaneous Admission (n = 320)*	*p*	*Total Sample*
Male	51.7%	56.8%	.07	51.7%	Male	42.4%	51.9%	.001	42.4%
Preterm*	7.9%	15.1%	<.001	7.9%	Age (M, SD)	15.1 (2.9)	14.0 (2.7)	<.001	15.1 (2.9)
White*	75.8%	66.8%	(ref)	75.8%	White*	85.3%	81.5%	(ref)	85.4%
Mixed*	4.6%	6.0%	.09	4.6%	Mixed*	1.8%	4.1%	.01	1.8%
Asian*	11.1%	18.4%	<.001	11.1%	Asian*	5.5%	6.3%	.50	5.5%
Black*	5.3%	4.4%	.83	5.3%	Black*	4.3%	4.5%	.76	4.3%
Chinese*	0.6%	1.0%	.26	0.6%	Chinese*	0.3%	0.5%	.70	0.3%
Other*	2.7%	3.5%	.22	2.7%	Other*	2.7%	3.2%	.62	2.7%
Multiple birth*	3.5%	3.8%	.75	3.5%	High socio-economic deprivation†	33.3%	36.8%	.19	33.3%
Missing data					*Missing data*				
Ethnic group	7.6%	2.5%	.001	7.6%	Ethnic group	24.9%	30.9%	.01	25.9%
Birth order	18.1%	19.4%	.53	18.1%	Socio-economic deprivation	2.8%	1.6%	.18	2.8%
Gestational age	25.0%	20.1%	.05	25.0%					

*Percentage calculated excluding patients with missing data (many singletons will have missing data on birth order, inflating the multiple birth rate).

†Highest tertile of area-based socio-economic deprivation (IMD2004 score) among infants re-admitted (IMD score not available for birth episodes).

Table 2: Odds Ratios (95% Confidence Intervals) for False Matches Defined as a Simultaneous Admission According to Study Variables

	All Infants *(n = 733,770)* *OR (95% CI)*	*Infants Re-admitted* *(n = 131,466)* *OR (95% CI)*		*Adolescents* *(n = 1,678,623)* *OR (95% CI)*
Male (vs female)	1.19 (0.96, 1.49)	0.90 (0.59, 1.36)	Male (vs. female)	1.37 (1.10,1.72)
Missing gestational age	2.35 (1.42, 3.89)	2.56 (1.21, 5.42)		
Born preterm	2.12 (1.42, 3.17)	2.46 (1.34, 4.53)	Age	0.91 (0.87,0.94)
Missing ethnic group (vs. white)	0.27 (0.12, 0.57)	0.78 (0.24, 2.57)	Missing ethnic group (vs. white)	1.30 (1.01,1.68)
Mixed ethnic group (vs. white)	1.55 (0.97, 2.50)	1.03 (0.37, 2.85)	Mixed ethnic group (vs. white)	2.61 (1.34,5.12)
Asian ethnic group (vs. white)	2.12 (1.54, 2.91)	1.06 (0.54, 2.08)	Asian ethnic group (vs. white)	1.15 (0.66,2.03)
Black ethnic group (vs. white)	1.20 (0.68, 2.11)	0.76 (0.23, 2.48)	Black ethnic group (vs. white)	0.83 (0.40, 1.71)
Chinese ethnic group (vs. white)	2.01 (0.64, 6.32)	2.04 (0.28, 14.84)	Chinese ethnic group (vs. white)	1.64 (0.23, 11.74)
Any other ethnic group (vs. white)	1.49 (0.80, 2.77)	0.74 (0.18, 3.07)	Any other ethnic group (vs. white)	1.01 (0.44, 2.33)
Missing birth order	0.52 (0.30, 0.92)	0.71(0.28, 1.77)		
Multiple birth (vs singleton)	0.76 (0.38, 1.41)	1.51(0.68, 3.37)	Number of re-admissions	1.01 (1.01, 1.01)
Number of re-admissions		1.05 (1.01, 1.09)	*Socio-economic deprivation*	
Socio-economic deprivation			Missing	0.61 (0.24, 1.54)
Medium (vs. low)*		0.86 (0.51, 1.45)	Medium (vs. low)	1.03 (0.77, 1.37)
High (vs low)*		0.81 (0.47, 1.40)	High (vs. low)	1.24 (0.93, 1.64)
	B (95% CI)	B (95% CI)		
Between-hospital variance	1.27 (1.04, 1.54)	1.09 (0.78, 1.52)	†	

*IMD score not available for birth episodes.

†The adolescent model was analyzed as a single level model.

Among the nested sample of infants re-admitted, missing gestational age and being born preterm were both associated with possible false matches (vs. born to term). Infants with more re-admissions were more likely to experience possible false matches (OR 1.05, 95 percent CI: 1.01, 1.09).

Sensitivity analyses restricted to 127 of 166 hospitals considered to have good quality data produced similar results, suggesting that results were not driven by hospitals producing poor quality data (Table S1).

We undertook further analyses of the infant delivery episodes to evaluate if age or diagnosis indicated linkage error or coding error. Of the 69 apparent infant delivery episodes, the age recorded was always <1 year consistent with the expected age for the study population. Primary diagnostic codes referred to the mother (*n* = 25, 36.2 percent; suggesting linkage errors), mother or infant (*n* = 17, 24.6 percent), or infant (*n* = 27, 39.1 percent; suggesting coding error).

Of the 324 apparent simultaneous admissions, age differed in 5.3 percent of record pairs, ethnic group in 16.1 percent of pairs, and primary diagnosis in 35.8 percent of pairs. There were 0.6 percent differing on all three of the variables, 6.8 percent on two variables, and 41.7 percent on one.

These results show that relying on demographic variables or diagnostic codes alone cannot determine if a genuine false match has occurred, but a large proportion of these simultaneous admissions differ on just three key variables.

Adolescents

The analytic sample comprised 1,678,623 adolescents with available data on study variables from 590 hospital IDs (Figure S2). Descriptive statistics are shown in Table 1. From 2004 to 2011, 71 died in hospital and were then apparently re-admitted, 65 adolescent episodes were coded as births, and 320 adolescents were apparently admitted to two different hospitals on the same day. Of the apparent simultaneous admissions, 6.3 percent were accounted for by 10 hospitals, 14.7 percent by 20 hospitals, and 21.3 percent by 30 of 320 hospital IDs that produced the simultaneous episodes. Descriptive statistics are shown in Table 1. Apparent simultaneous admissions were more likely to occur for males, younger patients, the Mixed ethnic group, or those missing data on ethnic group. Logistic modeling controlling for covariates showed, similarly, that simultaneous admissions were more common for males (OR = 1.37, 95 percent CI: 1.10, 1.72), younger patients (OR = 0.91, 95 percent CI: 0.87, 0.94), the Mixed ethnic group (OR = 1.30, 95 percent CI: 1.01, 1.68), and for those re-admitted more frequently (OR = 2.61, 95 percent CI: 1.34, 5.12). Missing data on ethnic group increased the odds of a simultaneous admission (OR = 1.30, 95 percent CI: 1.01, 1.68). Non-significant trends suggested that higher socio-economic deprivation was associated with possible false matching.

Discussion

Main Findings

In nonidentifiable hospital administrative data, possible false matches can be detected. They are relatively rare but have an impact on some patient groups more than others. Infants who were born preterm or those from Asian ethnic groups were more likely to experience the most common scenario (two admissions on the same day apparently at different hospitals). In adolescents, these scenarios were more likely to occur in the Mixed ethnic group. Results for infants were not different when restricted to good quality hospitals. There was significant variation across hospitals in the likelihood of possible false matches occurring; with a standard deviation on the logistic scale of about 1.2, which is comparable in size to the gender and ethnic group effects. Our findings suggest that errors may be reduced but are not eradicated by the algorithm used to anonymize patient records and link records belonging to the same person.

Strengths and Limitations

One strength of our analysis is that we used data from a whole country accessible to other researchers, describing scenarios that can be easily applied by other users of the data. We show that researchers can identity possible false matches and the groups of the population most likely to experience them, even in nonidentifiable hospital data. Data from two patient groups demonstrated that possible linkage error is not restricted to a particular age range or to maternity units, for example. Although we focused on six implausible clinical scenarios, others scenarios may also indicate possible false matches. One weakness is that our results provide a minimal estimate of linkage error as other possible false matches may also have been present. The scenarios could have been caused by data coding errors, as well as linkage errors.

Researchers can use clinical scenarios relevant to their own analyses, to assess data quality and modify data cleaning algorithms to detect further manifestations of possible linkage error. For example, although not done here, diagnostic codes could be used to detect implausible combinations of variables suggesting a possible false match has occurred (e.g., alcoholism in infants, males undergoing cesareans). We illustrated how to detect implausible scenarios in two patient groups, but the principles can be generalized more widely to other patients and other datasets. Researchers can devise clinical scenarios specific to their own study population and research question. Where available, data on patient identifiers can be used to clarify whether possible linkage error has occurred (Baker et al. 2012; Royal College of Paediatrics and Child Health 2013). A further weakness is that missing data are common in HES even for variables that are mandatory (e.g., ethnic group). This may have led us to underestimate the effect of these variables, although we did include the records in our model to avoid loss of data. Dialogue with hospitals and clinical coders is needed to improve data quality, identify coding errors, and prevent linkage errors.

Characteristics of excluded patients differed from the analytic sample and it is likely that records excluded from our analysis would be more prone to possible linkage error due to poorer data quality and missing data. However, the fact that results were similar when restricted to good quality data hospitals suggests that linkage errors are common across hospitals and are not caused by a small number of hospitals. Patient identifiers were not available, meaning that we could not identify the reason for the possible false match or identify possible missed matches (two different HESIDs for the same patient).

Context/Mechanisms

Relatively little is known about the mechanisms that create data linkage errors in HES and other administrative data sources. Table 3 provides some examples of the kinds of situations that might lead to errors in identifiers and false matching. True simultaneous admission to two hospitals on the same day (with no transfer recorded) is unlikely, so we hypothesize that data linkage errors following errors in patient identifiers, or coding errors, are the cause. Our additional analysis of the infant episodes coded as deliveries shows that there is no straightforward way to disentangle linkage errors from coding errors, given that mothers and infants are not currently linked in HES. We argue, however, that any information referring to the mother (including that the episode was a "delivery" rather than a birth) is still a form of linkage error—this

Table 3: Examples of How False Matches Might Be Produced by Errors in Identifiers or in Data Submitted by Hospitals

Situation	*Example*
Characteristics of patients	
Unconscious patients, frail patients with dementia, patients under the influence of alcohol or drugs, abandoned babies	Unknown dates of birth can result in missing data or default values (see below) (HSCIC 2010)
Unconventional surname	Naming conventions can contribute to linkage errors, because names may be stored differently on different databases, be presented by patients in different ways to different hospitals, and may be misunderstood by frontline staff. In an analysis of 100 records which failed to match, and did not have an NHS number, 37 failed to match because of the name (HSCIC 2009a)
Address out of date	If 40% of patients have not registered with a GP 6 months after moving, the address provided may not match that shown on the PDS (Millett et al. 2005)
"Complex case"	Examples include having no NHS number, invalid PDS record, demographic details out of date, demographic details not supplied by patient, Scottish patient presenting in England for the first time (NHS Connecting for Health 2012)
Misleading information given	A drug user may presents at two hospitals with different names
Visitor	A visitor to the United Kingdom may have no NHS number or postcode
Match by co-incidence	93,000 records shared sex, postcode, and date of birth by coincidence in 2006/2007 (excluding multiple births) (HSCIC 2009b). This could potentially explain records apparently showing two birth episodes separated in time
Multiple births	Multiple births share date of birth and postcode. Prior to 2002, they would have not received individual NHS numbers until they were registered. Prior research has shown that data about the first baby in a multiple delivery are more complete than subsequent babies (Dattani, Datta-Nemdharry, and Macfarlane 2007). For example, an infant death followed by apparent re-admissions might refer to twins
Communal establishments	Shared housing and communal living establishments can lead to false matches, although HES exclude a regularly updated list of communal postcodes from stage 3 of the algorithm. Previously, postcodes creating 10 matches were excluded (HSCIC 2009b)
Errors in data submitted by hospitals	
Default dates of birth entered	Default date of birth values (e.g., January 1, 1900) or estimated dates of birth may increase false matching (HSCIC 2010)

Continued

Table 3. *Continued*

Situation	*Example*
Default postcode entered	Default postcodes may be entered, which may or may not follow national guidelines (e.g., recognized default postcodes for homeless people and travelers may be entered as the hospital or embassy postcode by frontline staff)
Multiple births given same identifier	If hospitals give multiple births the same ID number (Dattani, Datta-Nemdharry, and Macfarlane 2007), or leave the field missing, this would increase the chance of false matching
Sharing NHS number	An NHS number can be unverified (not checked against the PDS) or invalid (fail a number check digit calculation) (HSCIC 2012b). Even valid NHS numbers can refer to the wrong patient. A check digit will fail 10% of the time due to typographical errors. Patients can therefore end up sharing NHS numbers (e.g., mother and infant sharing an NHS number might explain infant deliveries coded as births, or apparent simultaneous admissions to different hospitals on the same day)

information should have been recorded only on the mother's record, not the infant's. Errors in the recording of patient identifiers are made before the data is received at HSCIC. HSCIC apply some data cleaning algorithms and manual checks but rely largely on the quality of data submitted by trusts (HSCIC 2012a). Whereas clinical data are constantly updated and corrected on live systems by frontline staff and clinical coders, HES data released to researchers are a fixed extract (Health and Social Care Information Centre 2013). Improvements in data quality at source could reduce the risk of linkage error. However, NHS staff report that they do not feel confident about asking the right questions to obtain detailed information from patients (HSCIC 2009a). This could lead to selecting the wrong patient on a database or creating a new record after failing to identify the existing patient. Given that ethnic minorities were at greater risk of possible false matches in our analysis, better understanding of why patient records from these groups might have become confused could help improve recording, and improve patient safety/confidentiality. Socio-economic gradients exist in the quality of data held for other types of routinely collected data (Adams, White, and Forman 2004). Identifying the reasons why apparent simultaneous admissions are more likely in areas with high deprivation is a priority for further study.

Implications

Our results show that even when overall error rates are low, they may disproportionately affect certain patient groups, thereby potentially biasing results. We recommend that researchers use implausible clinical scenarios such as those demonstrated here to identify possible false matches prior to any subsequent analysis at the data cleaning stage, so that the potential impact on results can be mitigated. Researchers should routinely acknowledge that results based on linked data may be biased by linkage, particularly for analyses involving ethnic minority groups (Lariscy 2011). Overall linkage success rates are not adequate, because some groups of the population experience more linkage error than others (Lariscy 2011).

We have shown methods for identifying some manifestations of possible linkage error (apparent false matching) by using clinical scenarios and internal validation algorithms. Additional information on clinical scenarios, perhaps by using diagnostic codes, could identify further cases. This would only capture possible false matches, but not missed matches (two records belonging to the same patient with different HESIDs). If patient identifiers were available, actual false matches and also missed matches could be detected and verified. External validation with a "gold standard" database is needed to check whether two records genuinely belong to the same person or not. The Personal Demographic Service (PDS) is an example of a database that theoretically contains the most up-to-date patient demographics (Connecting for Health 2005; HSCIC 2013b). To protect patient confidentiality, however, the PDS is not currently available for researchers to use (Connecting for Health 2005; HSCIC 2013b). Longitudinal data on how patient identifiers change over time (e.g., address, surname) combined with information on how errors in identifiers occur in clinical settings could be used to determine why linkage errors occur, which groups of the population are more affected, and improve linkage success in the future. Data from multiple sources increase the chance of identifying true false and missed matches, and ensuring that the anonymization process has worked correctly.

Our results illustrate that the quality of linked data available to researchers depends on the underlying quality of data collected and submitted by hospitals combined with the algorithm used to link patient episodes over time. These results are relevant to current debates about whether patient identifiers should be pseudo-anonymized at source by scrambling ("hashing") the identifiers before they leave the hospital (ONS 2013; Hagger-Johnson et al. 2014). This would inevitably increase linkage error, because any identifier errors would be impossible to detect and it would not be possible to query the information or provide feedback about how to improve data quality, across multiple databases using different standards. The HESID algorithm is designed to perform linkage as best as possible with imperfect data from service providers, but only the HSCIC can make changes to this algorithm. No algorithm of this kind can remove linkage errors entirely.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported by a grant from the National Centre for Research Methods (NCRM); Economic and Social Research Council (ESRC).

Disclosures: The author has an honorary contract with the Health and Social Care Information Centre (HSCIC) as part of an NCRM funded project on data linkage errors. The HSCIC had no involvement in the preparation of this manuscript, nor the decision to publish. The views stated are those of the authors.

REFERENCES

Adams, J., M. White, and D. Forman. 2004. "Are There Socioeconomic Gradients in the Quality of Data Held by UK Cancer Registries?" *Journal of Epidemiology and Community Health* 58 (12): 1052–4.

Baker, M. G., L. T. Barnard, A. Kvalsvig, A. Verrall, J. Zhang, M. Keall, N. Wilson, T. Wall, and P. Howden-Chapman. 2012. "Increasing Incidence of Serious Infectious Diseases and Inequalities in New Zealand: A National Epidemiological Study." *Lancet* 379 (9821): 1112–9.

Bohensky, M., D. Jolley, V. Sundararajan, S. Evans, D. Pilcher, I. Scott, and C. Brand. 2010. "Data Linkage: A Powerful Research Tool with Potential Problems." *BMC Health Services Research* 10 (1): 346.

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* London: Springer.

Connecting for Health. 2005. *PDS Tracing Guidance for Front Line NHS Staff.* Leeds, UK: Health and Social Care Information Centre.

Dattani, N., P. Datta-Nemdharry, and A. Macfarlane. 2007. *Linking Maternity Data for England 2007: Methods and Data Quality (ONS Report).* London: Office of National Statistics.

Dunn, H. L. 1946. "Record Linkage." *American Journal of Public Health and the Nation's Health* 36 (12): 1412–6.

Hagger-Johnson, G. E., K. Harron, H. Goldstein, R. Parslow, N. Dattani, M. C. Borja, L. Wijlaars, and R. Gilbert. 2014. "Making a Hash of Data: What Risks to Privacy Does the NHS's Care Data Scheme Pose?" *British Medical Journal* 348: 348.

Health and Social Care Information Centre. 2013. *NHS Communal Establishment File Record Specification.* Leeds, UK: Health and Social Care Information Centre [accessed on October 29, 2013]. Available at http://systems.hscic.gov.uk/data/ods/datadownloads/onsdata

HES Data Quality Team. 2014. *HES 2013–14 Month 8 Inpatient Data Quality Note.* Leeds, UK: Health & Social Care Information Centre.

HSCIC. 2009a. *IQAP Guidance on Ethnic Naming Conventions.* Leeds, UK: Health and Social Care Information Centre.

HSCIC. 2009b. *Replacement of the HES Patient ID (HESID).* Leeds, UK: Health and Social Care Information Centre.

HSCIC. 2010. *IQAP Guidance on Unknown, Estimated and Default Birth Dates.* Leeds, UK: Health and Social Care Information Centre.

HSCIC. 2012a. *The HES Processing Cycle and HES Data Quality.* Leeds, UK: Health and Social Care Information Centre.

HSCIC. 2012b. *NHS Number Standard Specification.* Leeds, UK: Health and Social Care Information Centre.

HSCIC. 2013a. *The Quality of Nationally Submitted Health and Social Care Data.* Leeds, UK: Health and Social Care Information Centre.

HSCIC. 2013b. "Systems? Demographics? Personal Demographics Service (PDS)? Training and Guidance" [accessed on June 2013]. Available at http://systems.hscic.gov.uk/demographics/pds/training

Joffe, E., C. F. Bearden, M. J. Byrne, and E. V. Bernstam. 2012. "Duplicate Patient Records–Implication for Missed Laboratory Results." *AMIA Annual Symposium Proceedings Archive* 2012: 1269–75.

Knight, H., D. Cromwell, J. van der Meulen, I. Gurol-Urganci, D. Richmond, T. Mahmood, D. Richmond, A. Templeton, A. Dougall, and S. Johnson. 2013. *Patterns of Maternity Care in English NHS Hospitals.* London: Royal College of Obstetricians and Gynaecologists.

Lariscy, J. T. 2011. "Differential Record Linkage by Hispanic Ethnicity and Age in Linked Mortality Studies: Implications for the Epidemiologic Paradox." *Journal of Aging and Health* 23 (8): 1263–84.

McCoy, A., A. Wright, M. Kahn, J. Shapiro, E. Bernstam, and D. Sittig. 2013. "Matching Identifiers in Electronic Health Records: Implications for Duplicate Records and Patient Safety." *BMJ Quality and Safety* 22 (3): 219–24.

Middleton, B., M. Bloomrosen, M. Dente, B. Hashmat, R. Koppel, M. Overhage, T. Payne, T. Rosenbloom, C. Weaver, and J. Zhang. 2013. "Enhancing Patient Safety and Quality of Care by Improving the Usability of Electronic Health Record Systems: Recommendations from AMIA." *Journal of the American Medical Informatics Association* 20 (e1): e2–8.

Millett, C., C. Zelenyanszki, K. Binysh, J. Lancaster, and A. Majeed. 2005. "Population Mobility: Characteristics of People Registering with General Practices." *Public Health* 119 (7): 632–8.

NHS Connecting for Health. 2012. *NHS Number: Your Unique Patient Identifier. Guidance on Categories of Complex Case.* Leeds, UK: NHS Connecting for Health.

NHS Information Centre. 2007. *NHS Numbers for Babies Central Issue Service Guide for NHAIIS Users.* Exeter, UK: NHS Information Centre.

ONS. 2013. *Beyond 2011: Matching Anonymous Data.* London: ONS.

Royal College of Paediatrics and Child Health. 2013. *Overview of Child Deaths in the Four UK Countries: Overview of Child Deaths in the Four UK Countries.* London: Royal College of Paediatrics and Child Health.

Schmidlin, K., K. Clough-Gorr, A. Spoerri, M. Egger, M. Zwahlen, and for the Swiss National Cohort. 2013. "Impact of Unlinked Deaths and Coding Changes on Mortality Trends in the Swiss National Cohort." *BMC Medical Informatics and Decision Making* 13 (1): 1.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Figure S1: Flow Diagrams Showing How the Analytic Sample Was Determined for Infants.

Figure S2: Flow Diagrams Showing How the Analytic Sample Was Determined for Adolescents.

Table S1: Odds Ratios (95% Confidence Intervals) for Apparent Simultaneous Admissions According to Study Variables, Restricted to 127 Good Quality Data Hospitals (n = 645,507).