A guide to evaluating linkage quality for the analysis of linked data

Katie L Harron,1* James C Doidge,2,3 Hannah E Knight,1,4 Ruth E Gilbert,2 Harvey Goldstein,2,5 David A Cromwell1 and Jan H van der Meulen1

1Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK, 2Administrative Data Research Centre for England, UCL Great Ormond Street Institute of Child Health, UCL, London, UK, 3Centre for Population Health Research, University of South Australia, Adelaide, Australia, 4Lindsay Stewart Centre for Audit and Clinical Informatics, Royal College of Obstetricians and Gynaecologists, London, UK and 5Graduate School of Education, University of Bristol, Bristol, UK

*Corresponding author. Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, 15–17 Tavistock Place, London WC1 H 9SH, UK. E-mail: Katie.harron@lshtm.ac.uk

Editorial decision 26 July 2017; Accepted 8 August 2017

Abstract

Linked datasets are an important resource for epidemiological and clinical studies, but linkage error can lead to biased results. For data security reasons, linkage of personal identifiers is often performed by a third party, making it difficult for researchers to assess the quality of the linked dataset in the context of specific research questions. This is compounded by a lack of guidance on how to determine the potential impact of linkage error. We describe how linkage quality can be evaluated and provide widely applicable guidance for both data providers and researchers. Using an illustrative example of a linked dataset of maternal and baby hospital records, we demonstrate three approaches for evaluating linkage quality: applying the linkage algorithm to a subset of gold standard data to quantify linkage error; comparing characteristics of linked and unlinked data to identify potential sources of bias; and evaluating the sensitivity of results to changes in the linkage procedure. These approaches can inform our understanding of the potential impact of linkage error and provide an opportunity to select the most appropriate linkage procedure for a specific analysis. Evaluating linkage quality in this way will improve the quality and transparency of epidemiological and clinical research using linked data.

Key words: Record linkage, linkage error, bias, hospital records, data accuracy, sensitivity and specificity, selection bias, data linkage, administrative data
Introduction

Epidemiological and clinical research is increasingly based on datasets created by linking data from different sources such as administrative hospital datasets, clinical databases and national death registers.1,2 Due to the nature of many of these data sources, which are typically collected for financial or clinical management, a unique person identifier (e.g. National Health Service number in the UK) may not be available for linkage. Therefore, linkage is often based on a series of identifiers that are not unique, are prone to errors or missing values, or are dynamic (i.e. may change over time, such as postcode or name). Errors in the linkage process that arise from imperfect identifiers are increasingly being recognized as a potential source of bias in results from studies using linked data.3,4 However, assessing the extent to which linkage error affects results can be difficult for users of linked data who do not have access to the identifiable data used for linkage, which is often the case if the data linkage is performed by a third party. Separation of linkage and analysis processes in this way is recommended to preserve data security and personal confidentiality.5

Studies that evaluate linkage quality are therefore often restricted to estimates of the match rate (the proportion of records that were linked), sensitivity (the proportion of true links that were detected), or positive predictive value (the proportion of detected links that were true), which can be obtained, for example, by comparing a linked dataset with a ‘gold standard’ or reference dataset where true match status is known.6 However, these metrics are limited in their ability to tell us the degree to which linkage error might produce bias in outcomes of interest. In some instances, we can assume that effect sizes will be underestimated but, in most scenarios, it is not straightforward to predict the direction of bias that may result from linkage error. There are several factors that determine how linkage error affects an estimate, but one key factor is the distribution of errors with respect to variables of interest, and this is usually unknown. In complex analyses incorporating multiple variables, different variables can be affected by linkage error in different ways.

There is a lack of guidance on how to explore the extent to which error impacts upon analysis, and this area has been identified as a priority for research.7,8 This Education Corner article describes three simple approaches for evaluating quality of linkage, using an illustrative example of a linked dataset of maternal and baby hospital records. We aim to provide guidance that is applicable to both data providers and researchers, and to encourage the application of these methods among researchers using linked data.9

Why is linkage error important?

Linkage error can occur in two ways: false matches and missed matches. False matches occur when records belonging to different individuals are erroneously linked together. False matches typically (but not always) add noise to estimates, diluting the association between variables captured in different datasets and biasing effect estimates towards zero.3 Missed matches occur when records belonging to the same individual are not linked. When unlinked records are excluded from analyses, one consequence is reduced sample size and statistical power. If linkage is ‘informative’ (e.g. linkage to a disease register indicating the presence of a particular condition), a consequence of missed matches can be under-ascertainment of exposures or outcomes.10,11

An important further issue is that linkage errors do not always occur randomly, meaning that particular subgroups of individuals are often over- or under-represented amongst records affected by linkage error. Systematic reviews of studies comparing the characteristics of linked and unlinked records have identified that more vulnerable or hard to
reach populations are often missed, with the probability of a missed match being associated with a range of characteristics including gender, age, ethnicity, deprivation and health status. Consequently, the linked data may not be representative of the population of interest, which can reduce the study’s external validity, or may not capture subgroups that are of particular interest. As these demographic variables are often associated with exposures or outcomes of interest, differential rates of linkage error may also introduce bias. For example, unlinked mortality records in one particular ethnic group could lead to a distorted comparison of mortality rates by ethnicity.

If unlinked records are to be excluded from analysis, selection bias (or collider bias) can occur if selection into the linked dataset is related to both an exposure and an outcome of interest. For example, suppose it is more difficult to link records for low birthweight babies and also more difficult to link records from mothers who smoke. In this case, records for low birthweight babies that are successfully linked are more likely to be from mothers who do not smoke (since, in this example, records from mothers who smoke are more difficult to link). Conditioning on linked records could therefore induce a protective relationship between maternal smoking and low birthweight, analogous to the birthweight paradox described in epidemiological literature.

Evaluating the impact of data linkage error

The following sections describe three approaches to evaluating linkage quality (see Box 1 for a summary). The use of these methods can help researchers using linked data to understand the potential impact of linkage error on results, and comprise:

i. using a gold standard dataset to quantify false matches and missed matches;
ii. comparing characteristics of linked and unlinked data to identify potential sources of bias;
iii. using sensitivity analyses to evaluate how sensitive results are to changes in linkage procedure.

We use an illustrative example of linkage between hospital records for mothers and babies to demonstrate how these approaches can be implemented by researchers using linked data.

Illustrative example: linking hospital records for mothers and babies

The two most popular approaches to linkage have previously been described in an Education Corner article: deterministic (or rule-based) methods and probabilistic (or score-based) methods. Alternative methods also exist. In a previous study, we used a combination of these techniques to create a mother-baby cohort of records from Hospital Episode Statistics (HES), an administrative data resource that holds detailed information of all admissions to National Health Service (NHS) hospitals in England. The methods are described in full elsewhere, but comprised deterministic and probabilistic linkage of de-identified information in data items contained in both the mother’s delivery record and the baby’s birth record. Using a deterministic algorithm based on exact matching of hospital, maternal age, gestational age, baby’s sex, birth

Box 1. Summary of approaches to evaluating linkage quality

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Using a gold standard dataset to quantify false matches and missed matches</th>
<th>Comparing characteristics of linked and unlinked data to identify potential sources of bias</th>
<th>Sensitivity analyses to evaluate how sensitive results are to changes in linkage procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strengths</td>
<td>To quantify errors (missed matches and false matches)</td>
<td>To identify subgroups of records that are more prone to linkage error and are potential sources of bias</td>
<td>Assesses the extent to which results of interest may vary depending on different levels of error, and the direction of likely bias</td>
</tr>
<tr>
<td>Limitations</td>
<td>Easily interpretable; allows linkage error to be fully measured</td>
<td>Straightforward to implement and easily interpretable</td>
<td>Straightforward to implement</td>
</tr>
<tr>
<td>Technical requirements</td>
<td>Representative gold standard data are rarely available</td>
<td>Cannot be applied if systematic differences are expected between linked unlinked records (e.g. if linking to death register)</td>
<td>Results may be difficult to interpret as false matches and missed matches may impact on results in opposing or compounding ways</td>
</tr>
<tr>
<td></td>
<td>A representative group of records for which true match status is known; data linker capacity to perform evaluation (researchers rarely have access to gold standard data)</td>
<td>A linkage design where all records in at least one file are expected to link; provision of record-level or aggregate characteristics of unlinked records to researchers</td>
<td>Provision of information on the strength of the match (e.g. deterministic rule or probabilistic match weight)</td>
</tr>
</tbody>
</table>
order and GP practice code, 42% of baby records were linked to a maternal record. The match rate increased to 98% through the use of probabilistic linkage and additional variables (e.g. admission dates and mode of delivery). However, the extent to which these missed matches (or any false matches) affect analyses has not yet been explored.

Evaluating linkage error using ‘gold standard’ data

If data are available where the true match status of each pair of records is known, these ‘gold standard’ data can be used to test linkage algorithms and estimate rates of linkage error. There are various ways in which gold standard datasets can be derived, for example from an additional data source with complete identifiers, from a subsample of records that have been manually reviewed or otherwise determined to be matches (or non-matches), or from a representative synthetic dataset (e.g. generated through simulating data). Gold standard datasets allow us to identify: where errors have occurred in our linkage; where we have failed to link records that should have been linked (missed matches); or where we have linked together records belonging to different entities (false matches). Since gold standard data should be linked in the same way as the study data (for comparison), involvement of data linkers is required. Unlike the second two approaches, gold standard comparisons cannot be readily implemented by researchers who do not have access to the identifying data (e.g. in a ‘trusted third party’ model).

Creating the gold standard dataset

In order to create a gold standard for assessing the quality of linkage within the HES mother-baby cohort, we needed a dataset where the true match status of HES maternal and baby records was known. This was possible due to another study, which collected electronic records from maternity information systems (MIS) within 15 English obstetric units for births between April 2012 and March 2013. The MIS data captured NHS numbers for both mothers and babies together on the same record (this is not the case for HES). The MIS records were linked by NHS Digital to corresponding maternal and baby records in HES, using a deterministic approach based on NHS number, date of birth, sex and postcode. After excluding a number of uncertain links (see Supplementary Figure 1, available as Supplementary data at IJE online), the MIS-HES links provided a gold standard dataset that could be used to validate the same subset of births in the linked HES mother-baby cohort (Figure 1).

Evaluating linkage using the gold standard dataset

We compared linked records in the gold standard dataset (linked using direct patient identifiers including NHS number) with those in our HES mother-baby cohort (linked using indirect identifiers captured for birth and delivery records). To enable this comparison, we applied our original linkage algorithm to the same set of HES records captured in the MIS data. This allowed us to quantify the number of false matches or missed matches that our linkage algorithm...
produced, and to derive standard measures of linkage quality (sensitivity/recall, false match rate and positive predictive value/precision). These metrics are thought to be more useful than specificity (due to the imbalance between the number of non-matches versus matches in a linkage) or the F-measure (due to issues when comparing linked datasets of different sizes).

The gold standard dataset comprised records for 72,817 babies (Supplementary Figure 1). Of these, 72,520 (99.6%) were linked using the original linkage algorithm and 297 (0.4%) were unlinked. Of the 72,520 linked records, 71,884 were true matches, giving a positive predictive value of 99.1%, a false match rate of 0.9% (636 false matches/72,520 linked records) and a sensitivity of 98.7% (71,884/72,817 true matches).

The low error rates observed in this evaluation demonstrate that the original linkage algorithm was highly accurate. We might therefore assume that the impact of linkage error will be negligible, since, for example, excluding such a small proportion of the target population is unlikely to affect the generalizability of results or to dramatically reduce precision. However, further evaluation is needed to assess whether selection bias could be present, e.g. if records from a particular subgroup were more likely to be missed, or whether the 0.9% of false matches could have introduced enough noise to bias results.

Comparing characteristics of linked and unlinked data

In order to identify particular subgroups of records that are differentially missed during linkage, we can compare the characteristics of linked and unlinked records. This method of quality appraisal can only be implemented when all of the records in at least one of the files (or within the target sample within one of the files) are expected to link; it would not be useful, for example, when linking to a register of deaths to determine mortality, as there would be expected systematic differences between those who link and those who do not. This approach can be implemented if researchers have access to record-level or aggregate information on the characteristics of unlinked records.

Since data linkage studies are often characterized by large sample sizes, standardized differences can be more informative than P-values for comparing unlinked and linked records. Standardized differences are calculated as the mean difference divided by the standard deviation, and can be easily calculated in statistical software packages (e.g. using the ‘stddiff’ command in Stata). Standardized differences of 0.2, 0.5 and 0.8 represent small, moderate and large standardized differences, respectively. This helps us to identify variables that may have been more affected by linkage error and are therefore potential sources of bias.

Evaluating linkage using comparisons of linked and unlinked records

Compared with true matches in the gold standard, the 297 records that failed to link (missed matches) and 636 baby records that linked to the wrong mother (false matches) were more likely to be: multiple births; or babies with lower gestational age, lower birthweight or more neonatal medical conditions; or babies born by caesarean section; or those of non-White ethnic background (Table 1). Records with linkage errors were also more likely to be from babies born to nulliparous mothers or mothers without pregnancy risk factors. Linkage errors were strongly driven by data quality, since records with one or more missing values were less likely to link.

The results in this example indicate that although linkage error rates were low, there was still some potential for bias, as particular subgroups of records were more often affected than others. Whether these differences were large enough to introduce bias into results depends on the relationship between these variables and the parameters of interest. It is therefore helpful to explore how results of interest might change according to different levels of error.

Sensitivity analyses

In order to assess the sensitivity of results to different linkage procedures, we can perform sensitivity analyses, aiming to assess the extent to which results vary and the direction of likely bias. This can involve changing the linkage algorithm or varying the match weight threshold for probabilistic linkage, and re-running analyses to evaluate any impact on results. The aim of this approach is to determine whether decisions about the design of the linkage procedure could have had a substantial impact on inferences drawn from the linked data. These types of sensitivity analysis can be implemented by researchers without access to identifying data, if they are provided with match weights (in probabilistic linkage) or decision steps (in deterministic linkage). As these are not sensitive data, data providers and linkers are usually able to share these with researchers.

Evaluating linkage using sensitivity analyses

We conducted a sensitivity analysis to evaluate the mother-baby linkage by changing the threshold used in our linkage algorithm, and comparing results across different sets of linked records. We compared linkage results from our original probabilistic algorithm using a threshold weight of 20 for classifying records as links, with results from an
Table 1. Characteristics of records in the HES mother-baby cohort according to linkage status derived from gold standard data

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>True matches (N = 71884)</th>
<th>False matches (N = 636)</th>
<th>St. diff.</th>
<th>Missed matches (N = 297)</th>
<th>St. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
</tr>
<tr>
<td>Stillbirth</td>
<td>325 0.5</td>
<td>6 0.9</td>
<td>0.1</td>
<td>9 3.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Survival to postnatal discharge</td>
<td>71384 99.1</td>
<td>627 98.1</td>
<td>0.1</td>
<td>286 96.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Delivery risk factor(^a)</td>
<td>6738 9.4</td>
<td>105 16.4</td>
<td>0.2</td>
<td>49 16.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Female infant</td>
<td>34967 48.7</td>
<td>321 50.2</td>
<td>0.0</td>
<td>140 47.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Multiple birth</td>
<td>1961 2.7</td>
<td>126 19.7</td>
<td>0.6</td>
<td>31 10.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Caesarean section</td>
<td>18034 25.1</td>
<td>63 9.9</td>
<td>0.4</td>
<td>10 3.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Pregnancy risk factor(^b)</td>
<td>7388 10.3</td>
<td>103 16.4</td>
<td>0.6</td>
<td>1 0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Neonatal medical condition(^c)</td>
<td>6281 8.7</td>
<td>91 14.2</td>
<td>0.2</td>
<td>90 30.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Neonatal ICU</td>
<td>8461 11.8</td>
<td>32 5.0</td>
<td>0.2</td>
<td>33 11.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Parity: nulliparous</td>
<td>27125 37.7</td>
<td>335 52.4</td>
<td>0.3</td>
<td>192 64.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Gestational age group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full term (39+ wks)</td>
<td>45611 72.3</td>
<td>102 44.4</td>
<td>0.7</td>
<td>27 44.3</td>
<td>0.7</td>
</tr>
<tr>
<td>Early term (37–38 wks)</td>
<td>12721 20.2</td>
<td>66 28.7</td>
<td>17</td>
<td>27.9</td>
<td></td>
</tr>
<tr>
<td>Late preterm (34–36 wks)</td>
<td>3280 5.2</td>
<td>39 17.0</td>
<td>6</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>Moderate/very preterm (≤ 34 wks)</td>
<td>1494 2.4</td>
<td>23 10.0</td>
<td>11</td>
<td>18.0</td>
<td></td>
</tr>
<tr>
<td>Missing(^*)</td>
<td>8775 12.2</td>
<td>409 64.0</td>
<td>236</td>
<td>79.5</td>
<td></td>
</tr>
<tr>
<td>Birthweight (g)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1500</td>
<td>909 1.4</td>
<td>14 6.1</td>
<td>0.7</td>
<td>7 10.9</td>
<td>0.7</td>
</tr>
<tr>
<td>1500–2500</td>
<td>1798 6.0</td>
<td>45 19.7</td>
<td>12</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>2500–4000</td>
<td>51718 82.0</td>
<td>160 69.9</td>
<td>42</td>
<td>65.6</td>
<td></td>
</tr>
<tr>
<td>4000+</td>
<td>6687 10.6</td>
<td>10 4.4</td>
<td>3</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Missing(^*)</td>
<td>8769 12.2</td>
<td>410 64.2</td>
<td>233</td>
<td>75.8</td>
<td></td>
</tr>
<tr>
<td>Size for gestation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small (&lt; 10th percentile)</td>
<td>5274 8.4</td>
<td>25 11.1</td>
<td>0.2</td>
<td>5 8.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Normal</td>
<td>54367 81.6</td>
<td>187 93.1</td>
<td>51</td>
<td>85.0</td>
<td></td>
</tr>
<tr>
<td>Large (&gt; 10th percentile)</td>
<td>6344 10.1</td>
<td>13 5.8</td>
<td>4</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>Missing(^*)</td>
<td>8896 12.4</td>
<td>414 64.8</td>
<td>237</td>
<td>79.8</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>48896 68.0</td>
<td>408 63.9</td>
<td>0.3</td>
<td>165 55.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Mixed</td>
<td>3410 4.7</td>
<td>24 3.8</td>
<td>14</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>7367 10.3</td>
<td>49 7.7</td>
<td>20</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>4866 6.8</td>
<td>32 5.0</td>
<td>25</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4508 6.3</td>
<td>77 12.1</td>
<td>38</td>
<td>12.8</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>2834 3.9</td>
<td>49 7.7</td>
<td>35</td>
<td>11.8</td>
<td></td>
</tr>
<tr>
<td>Newborn length of stay (days)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 2</td>
<td>38329 53.3</td>
<td>315 49.3</td>
<td>0.2</td>
<td>131 44.1</td>
<td>0.7</td>
</tr>
<tr>
<td>2–6</td>
<td>28946 40.3</td>
<td>244 38.2</td>
<td>74</td>
<td>24.9</td>
<td></td>
</tr>
<tr>
<td>7+</td>
<td>4599 6.4</td>
<td>80 12.5</td>
<td>92</td>
<td>31.0</td>
<td></td>
</tr>
<tr>
<td>Maternal age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 20</td>
<td>2859 4.0</td>
<td>21 3.3</td>
<td>0.1</td>
<td>13 4.4</td>
<td>0.2</td>
</tr>
<tr>
<td>20–24</td>
<td>11752 16.4</td>
<td>88 13.8</td>
<td>42</td>
<td>14.1</td>
<td></td>
</tr>
<tr>
<td>25–29</td>
<td>19226 26.8</td>
<td>155 24.3</td>
<td>55</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>30–34</td>
<td>22377 31.1</td>
<td>220 34.4</td>
<td>101</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>35–39</td>
<td>12433 17.3</td>
<td>125 19.6</td>
<td>64</td>
<td>21.6</td>
<td></td>
</tr>
<tr>
<td>40+</td>
<td>3234 4.5</td>
<td>30 4.7</td>
<td>22</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>Income/deprivation quintile(^d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most deprived</td>
<td>27042 37.7</td>
<td>206 32.3</td>
<td>0.1</td>
<td>97 32.9</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>16394 22.9</td>
<td>170 26.7</td>
<td>86</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>13104 18.3</td>
<td>129 20.3</td>
<td>58</td>
<td>19.7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9040 12.6</td>
<td>77 12.1</td>
<td>37</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>Most affluent</td>
<td>6146 8.5</td>
<td>56 8.8</td>
<td>17</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Missing(^*)</td>
<td>155 0.2</td>
<td>2 0.3</td>
<td>2</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

\(^0.2, 0.5, \text{and } 0.8\) can be considered as small, medium and large effect sizes respectively. 
St. diff, standardized differences; ICU, intensive care unit; wks, weeks. 
\(^a\)Hypoxia, amniotic fluid embolism, placental-transfusion syndrome, umbilical cord prolapse, chorioamnionitis, fetal haemorrhage, birth trauma, complications of delivery, umbilical cord problem.  
\(^b\)Eclampsia, gestational hypertension, diabetes, placental abruption or infarction.  
\(^c\)Congenital anomaly, perinatal infection, neonatal abstinence syndrome, respiratory distress syndrome.  
\(^d\)Quintiles of deprivation were derived from the Index of Multiple Deprivation (IMD) score based on patient postcode in HES.  
\(^*\)Percentage of records with missing data (excluded from other category percentages).
algorithm that minimized false links by using a considerably higher threshold of 45. These thresholds were selected based on examination of the observed distribution of weights in our analysis; this distribution can differ substantially depending on the number and quality of matching variables, so thresholds are generally selected in the context of a specific linkage or analysis. The aim of this type of sensitivity analysis is to select thresholds that are likely to reflect plausible limits for the trade-off between false matches and missed matches. We also compared results with those from the initial deterministic linkage only, i.e. where records agreed exactly on hospital, maternal age, gestational age, baby’s sex, birth order and GP practice code.

As expected, increasing the match weight threshold in probabilistic linkage, or using deterministic linkage only, produced linkages that introduced fewer false matches but more missed matches (Table 2). This is because stricter linkage criteria make it less likely that records belonging to different entities will link by chance, but more likely that records with missing or incorrect linking variables will remain unlinked.

We expected that impact of linkage errors in each of these linkage scenarios would depend on the research question, and therefore assessed four different outcomes: proportion of stillbirths; the proportion of preterm births (< 37 weeks of gestation); the association between neonatal survival to discharge and delivery risk factors; and the association between delivery risk factors and ethnic group. Odds ratios were estimated from logistic regression models, adjusting for a number of maternal and neonatal risk factors (listed in Table 1, based on ICD-10 diagnosis codes listed in Supplementary Table 1, available as Supplementary data at IJE online). Analysis was performed in Stata 14.33

Proportions of stillbirths and preterm births
We expected that the generalizability (i.e. external validity) of the data would be affected by missed matches. In particular, we expected that records of preterm births or stillbirths would be less likely to link than those of later gestations or live births, and that the ascertainment of these outcomes would therefore be lower in datasets more affected by missed matches.

By comparing results across different linkage algorithms (Table 3), we observed that for preterm birth (7.65% of records in gold standard), ascertainment was similar for the

---

**Table 2. Linkage success for a range of linkage criteria**

<table>
<thead>
<tr>
<th></th>
<th>Original probabilistic linkage (threshold weight = 20)</th>
<th>High-threshold probabilistic linkage (threshold weight = 45)</th>
<th>Deterministic linkage only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked records</td>
<td>72520/72817</td>
<td>65020/72817</td>
<td>35324/72817</td>
</tr>
<tr>
<td>Missed match rate</td>
<td>297/72817</td>
<td>7797/72817</td>
<td>37493/72817</td>
</tr>
<tr>
<td>False match rate</td>
<td>636/72520</td>
<td>212/65020</td>
<td>22/35324</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>71884/72520</td>
<td>64808/65020</td>
<td>35302/35324</td>
</tr>
</tbody>
</table>

**Table 3. Comparison of outcome measures for a range of linkage criteria**

<table>
<thead>
<tr>
<th></th>
<th>Gold standard</th>
<th>Original probabilistic linkage</th>
<th>High-threshold probabilistic linkage</th>
<th>Deterministic linkage only</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Preterm births (95% CI)</td>
<td>7.65 (7.45–7.86)</td>
<td>7.64 (7.43–7.85)</td>
<td>7.31 (7.11–7.53)</td>
<td>7.43 (7.16–7.71)</td>
</tr>
<tr>
<td>% Stillbirths (95% CI)</td>
<td>0.47 (0.42–0.52)</td>
<td>0.46 (0.41–0.51)</td>
<td>0.44 (0.39–0.49)</td>
<td>0.45 (0.40–0.50)</td>
</tr>
<tr>
<td>Odds ratio (95% CI) for neonatal survival to discharge: mothers with delivery risk factors vs those without</td>
<td>0.42 (0.18–0.98)</td>
<td>0.35 (0.15–0.79)</td>
<td>P = 0.011</td>
<td>P = 0.143</td>
</tr>
<tr>
<td>P = 0.039</td>
<td>P = 0.044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds ratio (95% CI) for delivery risk factors: Black ethnicity vs White ethnicity</td>
<td>0.98 (0.88–1.09)</td>
<td>0.97 (0.87–1.08)</td>
<td>0.89 (0.79–1.01)</td>
<td>0.80 (0.66–0.96)</td>
</tr>
<tr>
<td>P = 0.700</td>
<td>P = 0.593</td>
<td>P = 0.067</td>
<td>P = 0.017</td>
<td></td>
</tr>
</tbody>
</table>
original linkage algorithm (7.64%), but ascertainment was a little lower when using a higher match weight threshold of 45 (7.31%) or deterministic linkage only (7.43%). All confidence intervals for proportions of stillbirths and preterm births estimated in the linked datasets overlapped with those in the gold standard, indicating that in this example, linkage error is unlikely to have resulted in substantial bias for these outcomes.

Association between neonatal survival to discharge and recording of delivery risk factors
We expected that statistical power would be affected either through missed matches (due to a reduction in the size of the study population) or a lack of precision introduced by false matches (leading to increased noise in the association between variables). Given the large sample size, we assumed that power implications would be most important for identifying associations with rare outcomes (e.g. mortality).

As expected, we found that as the number of linked records decreased due to more missed matches at higher thresholds (Figure 2), there was a reduction in the precision of estimates of association between delivery risk factors and survival (Table 3). Although an association between delivery risk factors and survival was observed in the gold standard data [adjusted odds ratio = 0.40; 95% confidence interval (CI) 0.17–0.95], this association was no longer observed in the deterministic linkage where less than half of records were linked (adjusted odds ratio = 0.52; 95% CI 0.22–1.25). These results demonstrate that there can be power implications when large numbers of unlinked records are excluded from analysis: a problem analogous to that of complete case analysis in the presence of missing data.

Association between delivery risk factors and ethnicity
We expected that selection bias could be introduced if selection into the linked dataset were associated with both the outcome and exposure. For example, earlier comparisons between linked and unlinked data had indicated that that records for mothers with delivery risk factors were less likely to link, and also that mothers in the Black ethnic group were less likely to link. If this was the case, mothers with delivery risk factors who were successfully linked would be more likely to be from other ethnic groups (since, in this example, those from the Black ethnic group were harder to link). Conditioning on linked records could therefore induce a spurious protective relationship between Black ethnicity and delivery risk factors.

In the gold standard data, we observed that 6.5% of mothers with delivery risk factors were from the Black ethnic group, whereas in the deterministic linkage only 4.7% of mothers with delivery risk factors were from the Black ethnic group. There was no true association between ethnicity and delivery risk factors. However, within the deterministically linked data there appeared to be a protective effect (odds ratio = 0.80; 95% CI 0.66–0.96, Table 3). These results indicate that in this example, some inferences from linked data could be incorrect due to selection bias in the presence of missed matches.

Discussion
We describe three approaches for evaluating linkage quality, and demonstrate how these methods can be used determine the extent to which linkage error may introduce bias for a specific research question. Our illustrative example showed that even with high linkage rates, particular subgroups of individuals are disproportionately affected by linkage error, as has been observed in previous literature. However, we demonstrate that sensitivity analyses can help us to understand the direction of any bias, and to assess whether linkage errors may influence inferences from the linked dataset.

Access to information required to evaluate linkage quality
In many jurisdictions, linkage is carried out by an independent body and information about the linkage processes is not readily available to researchers. Comprehensive guidelines on information that should be shared between data providers, linkers and researchers are available elsewhere. All of the approaches described in this article require some level of collaboration between data linkers and the researchers aiming to evaluate linkage quality, but only approach (i) (gold standard data) requires direct involvement of the linkers; approaches (ii) and (iii) can be

Figure 2. Number of linked records and percentage of missed matches and false matches for a range of linkage criteria. W = threshold used to classify links in probabilistic linkage.
implemented by researchers provided that certain non-
sensitive information is shared.

Our example of HES mother-baby linkage was sup-
ported by the availability of a subset of gold standard data,
which is one of the most useful tools for quantifying linkage
error. Consideration should always be given to the represen-
tativeness of the gold standard dataset. In our example,
the proportion of unlinked records in the gold standard
data was lower than in the HES mother-baby cohort over-
all, indicating that data quality in the hospitals contribut-
ing MIS data may have been slightly better than average. The
same applies to gold standard datasets created using subsets
of records that have complete data either on a single unique
identifier or on a sufficient set of partially identifying vari-
ables; records with high quality data may differ systematic-
ally from those of poorer quality data.

In practice, gold standard datasets are rarely available.
Even those that are generated (e.g. through manually re-
viewing a sample of linked and unlinked records, or by cre-
ating a synthetic dataset with the same characteristics as
the original data) are often only available to the data link-
ers and not to researchers. This means that data linkers
would need to evaluate linkage quality using a gold stand-
ard (as researchers generally would not have access to gold
standard data).

However, when gold standard datasets are not avail-
able, researchers can consider alternative approaches: com-
parisons of characteristics of linked and unlinked data, and
sensitivity analyses. These methods can be easily imple-
mented but require data linkers to provide information on
the characteristics of unlinked records and/or on the qual-
ity of each potential link. Providing record-level or aggre-
gate characteristics of unlinked records allows researchers
to compare linked and unlinked records, to identify any
potential sources of bias where particular subgroups of rec-
ords were missed from the linkage.

Sensitivity analyses can be performed if measures of
linkage certainty (e.g. match weights in probabilistic
matching or matching ranks/criteria in hierarchical deter-
ministic matching) are provided by data linkers alongside a
linked dataset. This makes any uncertainty or subjectivity
in the linkage process more transparent and allows re-
searchers to run analysis on different sets of linked records
in turn. It should be noted that caution is needed when in-
terpreting results of this type of sensitivity analysis, as in
most cases, none of the linkage algorithms compared will
be 100% accurate. The trade-off between false matches
and missed matches will vary depending on the linkage al-
gorithm, and these errors may impact on results in oppos-
ing or compounding ways. In our example, inferences
about the associations between delivery risk factors and
survival to discharge, and between ethnicity and delivery
risk factors, differed between the methods. Exploring dif-
f erences in results over a range of linkage algorithms in this
way can help researchers consider the pattern of any bias,
and to identify scenarios which are particularly likely to
produce substantial bias.

Further methods for evaluating linkage quality
In our example, we expected all babies to link with a
mother, which made comparisons between linked and un-
linked records easily interpretable and allowed us to di-
rectly estimate the proportion of missed matches. However,
careful consideration needs to be given to appropriate ref-
ence populations when all records are not expected to
link. For example, we would not expect all hospital records
to link with a mortality record and vice versa; rather, a suc-
cessful link indicates that an individual has died (‘infor-
mativ e linkage’). If that is the case, comparing the
characteristics of the individuals whose records were and
were not linked would also be affected by differences in
the groups for whom no linked records were available (i.e.
the difference between those who died and those who sur-
vived). In such situations, external reference data (e.g. age-
specific mortality rates) can allow us to assess how linkage
rates might differ for different subgroups.

Further methods not covered in this article can also be
used to evaluate linkage quality in the absence of a gold
standard. For example, estimates of false match rates can
be derived by applying linkage algorithms to records
known to have no match (e.g. attempting to link with mor-
tality records for individuals known to be alive, or attempt-
ing to link male patient records with maternity records).3
Alternatively, inconsistency checks, such as checks for ad-
missions following a patient’s death, linkage between a
male patient and a caesarean section, or linkage of one
mortality record to two different individuals, can be per-
formed post-linkage in de-identified data to identify false
matches. Linkage error rates estimated in this way
should be interpreted with caution, as not all errors may
have been identified and distinguishing between linkage
errors and data coding errors can be difficult. Nevertheless,
these methods can reveal useful information about the relative
distribution of errors across subgroups or with respect to variables of interest (i.e. whereas the ab-
solute error rate may remain unknown, higher rates of
inconsistencies may be observed with respect to some vari-
able of interest, implicating likely bias).35

Handling bias in the analysis of linked data
Evaluation of linkage quality can guide decisions about ap-
propriate study design. For example, if linkage is used to

identify individuals with a particular condition or disease (informative linkage), high levels of missed matches will lead to under-ascertainment, meaning that cohort study designs may be unsuitable (particularly for deriving estimates of prevalence or incidence). Where linkage rates are too low, researchers may conclude that linked data are not fit for these purposes. On the other hand, a case-control study may still be valid, whereby a high threshold is used to identify cases and a low threshold is used to identify controls (assuming no other biases are present). 37 In this scenario, records for which there is uncertainty about linkage would not be included in analysis.

An alternative, which still makes use of all available records, is to use multiple imputation to handle missing values due to unlinked or equivocal records. 38 Furthermore, information from match weights can be incorporated into imputation procedures, making use of variable distributions in candidate links (known as ‘prior-informed imputation’). 39,40 This method incorporates information from ‘auxiliary’ variables, such as individual characteristics associated with linkage quality (e.g. birthweight or ethnicity) to help correct for selection biases.

In situations in which we have information about how linkage error affects the distribution of outcomes and exposures in our data, it may be possible to use established techniques for quantitative bias analysis, to adjust for these errors. 41,42 This is particularly relevant for simple analyses, but becomes more complex with complicated designs involving more than two data sources and/or a number of covariates. Developing appropriate methods to handle bias arising from linkage error is a priority for methodological research. 43

Studies of linked data are often based on administrative data that have not been collected primarily for research. In addition to linkage error, researchers should also consider other issues specifically relevant to these types of data (e.g. missing data, coding changes, service changes etc.), and explore methods to handle any potential bias that is identified. 43,44

Summary

We describe three methods for evaluating linkage quality: applying the linkage algorithm to a subset of gold standard data to quantify linkage error; comparing characteristics of linked and unlinked data to identify potential sources of bias; and evaluating the sensitivity of results to changes in the linkage procedure. These methods are generalizable to many other linkage situations and can be used as a guide for evaluating the quality of linkage for population-based analyses of linked data. Researchers using linked data should collaborate with data providers to understand the data linkage process, including data extraction and cleaning, linkage methods and resulting data quality. Ultimately, this will improve transparency and enhance the value of linked data for epidemiological and clinical research. 44

Supplementary Data

Supplementary data are available at IJE online.

Funding

This work was supported by the Wellcome Trust [grant number 103975/Z/14/Z]. JvM is supported by the NIHR CLAHRC North Thames. The work was also supported by the Economic and Social Research Council through the Administrative Data Research Centre for England (ES/L007617/1).

Acknowledgements

We would like to thank Lynn Copley for managing the required data. We would also like to thank the NHS Trusts that participated in the RCOG MIS pilot project and have agreed to be named in resulting academic outputs: Birmingham Women’s NHS Foundation Trust, Bradford Teaching Hospitals NHS Foundation Trust, Cambridge University Hospitals NHS Foundation Trust, Central Manchester University Hospitals NHS Foundation Trust, Chelsea and Westminster Hospital NHS Foundation Trust, Homerton University Hospital NHS Foundation Trust, Liverpool Women’s NHS Foundation Trust, Mid Cheshire Hospitals NHS Foundation Trust, Norfolk & Norwich University Hospitals NHS Foundation Trust, Nottingham University Hospitals* NHS Trust, Southern Health and Social Care Trust, St George’s Healthcare NHS Trust, Newcastle Upon Tyne Hospitals NHS Foundation Trust, University College London Hospitals NHS Foundation Trust and Western Sussex Hospitals* NHS Foundation Trust. *These trusts contributed data for more than one hospital. Approvals for the use of HES data were obtained as part of the standard Hospitals Episode Statistics approval process. Hospital Episode Statistics were made available by NHS Digital, all rights reserved.

Conflict of interest: The authors state no relevant conflicts of interest.

References


