Administrative Data Research Network

An ESRC Data Investment

Better Knowledge Better Society

An Introduction to Data Linkage

Author: Katie Harron
Series editors: Elaine Mackey and Mark Elliot

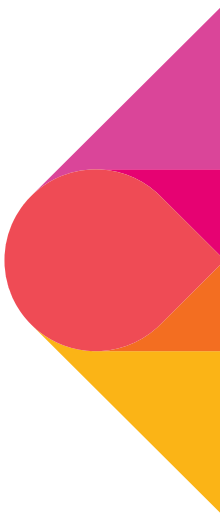
ADRN Publication - June 2016

Published in the UK in 2016 by
ADRN Publications
University of Essex
Wivenhoe Park
Colchester
CO4 3SQ

Introduction to Data Linkage

Written by Katie Harron

Edited by Elaine Mackay and Mark Elliot



Contents

1. Introduction 3
1.1. Description of ADRN 3
1.2. Purpose of this guide and who it is aimed at 3
1.3. Overview – how the intended audience should use this guide 3
2. An introduction to data linkage 4
2.1. Background 4
2.2. Record linkage methods 5
2.2.1. Deterministic linkage 5
2.2.2. Probabilistic linkage 5
2.2.2.1. Match weights 6
2.2.2.2. Thresholds 6
2.2.3. Designing an algorithm 6
2.2.4. Data preparation 7
2.2.4.1. Phonetic coding, string comparators and address standardisation 7
2.2.4.2. Capacity for linkage 7
2.2.5. Privacy preservation 8
2.2.5.1. Encryption 9
2.2.6. Linkage error 9
2.2.6.1. Impact of linkage error 11
2.2.7. Evaluating linkage quality 12
2.2.7.1. Measuring linkage error using 'gold-standard' data 12
2.2.7.2. Sensitivity analyses 12
2.2.7.3. Comparison of characteristics of linked and unlinked data 12
2.2.7.4. Statistical methods 13
2.3. Discussion of ADRN-specific issues 13
2.3.1. Future ADRN research 13
3. Signposting to other resources 15
4. Summary and conclusions 15
5. References 15
6. Acknowledgements 21
7. Glossary 21

List of Figures

Figure 1: Data linkage timeline and number of PubMed search results by year of publication with search term "record linkage" 4
Figure 2: Example classification of links using thresholds in probabilistic linkage 7
Figure 3: Separation of identifiers and attribute data. *Anonymous match key provides link between pseudo IDs from each data provider. 8

List of Tables

Table 1: Classification of record pairs in linkage. **Total non-matches can either be defined as all pairwise comparisons of records relating to different people, or all records from the primary dataset with a corresponding record in the linking dataset. 10
Table 2: Measures of linkage quality (A, B, C and D refer to Table 1.1) 11

1. Introduction

1.1. Description of ADRN

The Administrative Data Research Network (ADRN) is a UK-wide partnership between academia, government departments and agencies, national statistical authorities, funders and the wider research community to facilitate new economic and social research based on routinely collected government administrative data.

The Network is establishing a new legal, secure and efficient pathway for the research community to access de-identified linked administrative datasets. This will potentially benefit our society by providing a greater evidence base to inform policy.

The Network consists of:

- four data research centres (ADRCs):
  - ADRC England: led by the University of Southampton
  - ADRC Northern Ireland: led by Queen's University Belfast
  - ADRC Scotland: led by the University of Edinburgh
  - ADRC Wales: led by the Swansea University
- an overarching Administrative Data Service, which is the information and co-ordinating body of the Network
- data owners
- The Economic and Social Research Council (the funding body)
- The UK Statistics Authority

The ADRN has commissioned this guide on data linkage to support the development of knowledge and skills in the subject topic area.

1.2. Purpose of this guide and who it is aimed at

This guide is designed to give readers a practical introduction to data linkage and is aimed at researchers who would like to gain an understanding of data linkage techniques, either for the creation or analysis of linked data. It covers data preparation, deterministic and probabilistic linkage methods, and analysis of linked data, with examples relevant to health and other administrative data sources. This guide is relevant for academic researchers in the social and health sciences or those who work for government, survey agencies, official statistics, charities or the private sector.

1.3. Overview – how the intended audience should use this guide

Readers should use this guide as an introductory text on data linkage; to gain knowledge of the background and theory of data linkage methods; to understand how to perform basic deterministic and probabilistic linkage; and to consider how to evaluate and report data linkage quality.

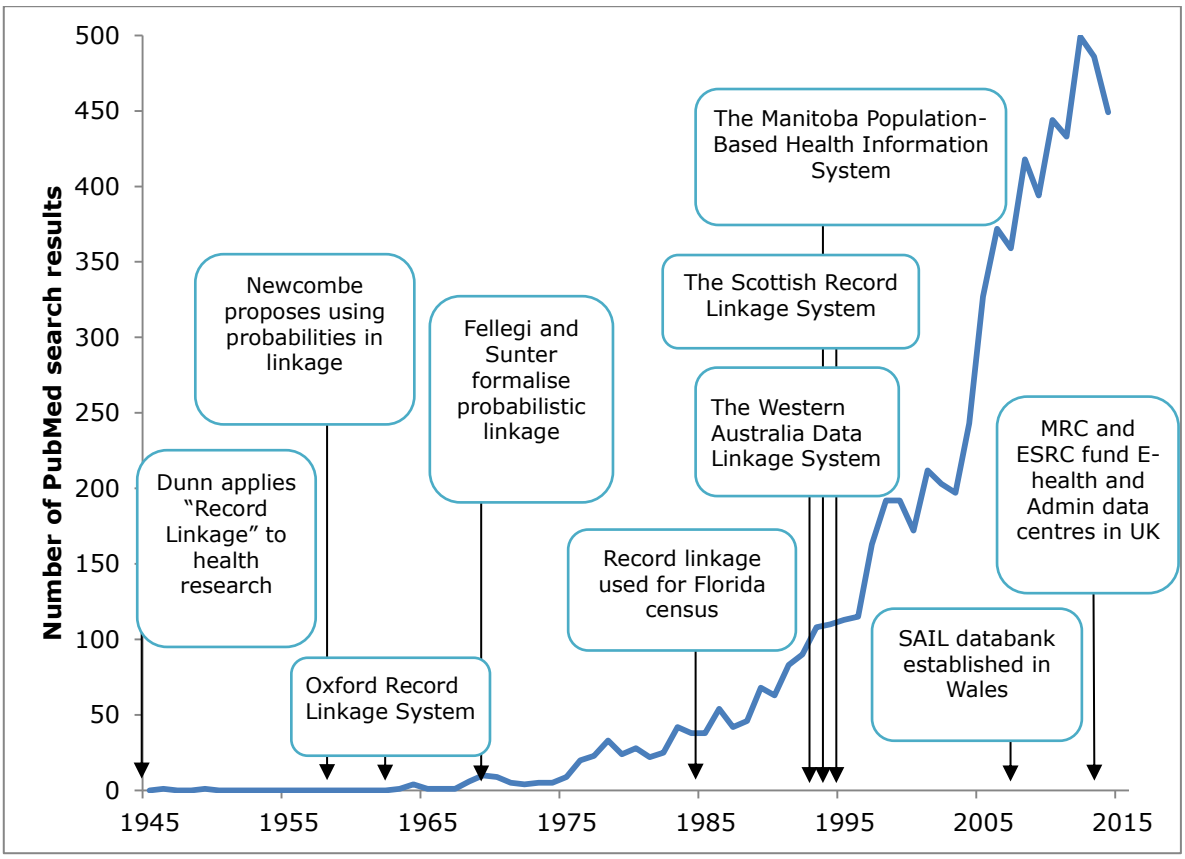


Figure 1: Data linkage timeline and number of PubMed search results by year of publication with search term "record linkage"

2. An introduction to data linkage

2.1. Background

The OECD defines record linkage as "a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record" [1]. In fact the term 'record linkage' was coined in 1946, when Dunn described linkage of vital records from the same individual (birth and death registrations) and referred to the process as "assembling the book of life" [2].

In recent decades, record linkage has become an increasingly used tool for service evaluation and research (see Figure 1). The development of computerised record linkage means that existing information relating to the same individual can be combined efficiently and cost-effectively, avoiding the high cost, time and effort associated with setting up new data collection systems [3]. Population-based linkage systems have been established in countries around the world, including in Australia, Canada, the UK and the Nordic countries among others [4-8]. As record linkage has become an established part of research relating to health and society, there has been an increasing interest in methodological issues associated with creating and analysing linked datasets [9-11].

There are a number of synonyms for record linkage depending on the field of application, including 'record matching', 'entity resolution' and 'merge-purge'. The term 'data linkage' technically covers other topics such as statistical matching and data wrangling, but is often used to describe record linkage, and the two terms will be used interchangeably in this text.

2.2. Record linkage methods

Record linkage is the process of bringing together information relating to the same individual from different sources, through comparing records and applying a set of linkage criteria or rules to determine whether or not records belong to the same individual.

The aim of linkage is to determine the true **match status** of each record pair:

- **Match**: records belong to the same individual
- **Non-match**: records belong to different individuals

Through the linkage methods that we use, a **link status** is assigned to each pair:

- **Link**: records classified as belonging to the same individual
- **Non-link**: records classified as belonging to different individuals

In a perfect linkage, all matches are classified as links, and all non-matches are classified as non-links. If record pairs are mis-classified, error is introduced:

- **False match:** records from different individuals link erroneously
- **Missed match:** records from the same individual fail to link

2.2.1. Deterministic linkage

Deterministic linkage is a relatively straightforward linkage method, typically requiring exact agreement on a unique identifier (such as a national insurance number) or on a specified set of partial identifiers (e.g. surname, sex and postcode) [12-18]. Deterministic methods are useful when records have unique (or at least highly discriminative) identifiers that are well completed and accurate. For example, the community health index (CHI) is used for much of the linkage in the Scottish Record Linkage System [7]. Modifications of strict deterministic linkage allow for small differences in identifiers, by using a succession of rules. For example, the deterministic algorithm used to link hospital admission records for the same individual in Hospital Episode Statistics is based on a sequential set of rules looking for agreement on a combination of identifiers [19]:

1. NHS number, date of birth and sex
2. Local patient identifier, hospital provider, date of birth, sex and postcode
3. Date of birth, sex and postcode

Deterministic methods are designed to avoid **false matches**, since it is unlikely that different individuals will share the same set of identifiers, although this can occur where there are identifier errors. On the other hand, deterministic methods requiring exact agreement on identifiers are prone to **missed matches**, as any recording errors or missing values can prevent identifier agreement [20, 21].

2.2.2. Probabilistic linkage

Probabilistic methods were proposed as a means to overcome some of the limitations of deterministic linkage, and to allow linkage in the presence of recording errors and/or without using a unique identifier. Newcombe was the first to propose probabilistic methods, suggesting that a match weight could be created to represent the likelihood that two records are a true match, given agreement or disagreement on a set of partial identifiers [3]. Fellegi and Sunter later formalised Newcombe's proposals into the statistical theory underpinning most probabilistic linkage today [22]. An alternative form of probabilistic linkage is the Copas-Hilton method, which uses statistical models to measure the evidence that records belong to the same rather than different individuals [23].

2.2.2.1. Match weights

In the Fellegi-Sunter approach, the contribution of each identifier to the overall match weight reflects its discriminative value, so that, for example, agreement on date of birth contributes more evidence of a match than agreement on sex [24-26]. Disagreement on an identifier contributes a penalty to the overall match weight. Match weights are calculated from two conditional probabilities:

- **M-probability**: the probability that an identifier agrees given records belong to the same individual
- **U-probability**: the probability that an identifier agrees given records belong to different individuals

The **u-probability** is calculated based on the frequency of values for each identifier. For example, the probability of chance agreement on sex would be ½. The probability of chance agreement on month of birth would be 1/12, and so on. The **m-probability** represents the error rate in a particular identifier, and is typically estimated during the linkage process and updated as more links are made. For example, if sex was miscoded in 5% of record pairs, the m-probability would be 0.95. The overall match weight is derived by calculating the ratio $\log_2(m/u)$ for each identifier, and summing across all identifiers.[a]

The summing of weights across identifiers relies on an independence assumption, i.e. that agreement on one variable is independent of agreement on another. This independence assumption does not always hold, and dependence between highly correlated variables such as dates (e.g. due date and actual delivery date) has been shown to have a negative impact on match weights, resulting in incorrect ranking of record pairs [27, 28]. Although dependence between identifiers is commonly ignored [29, 30], methods that account for identifier dependence have been shown to improve the quality of linkage [31]. One approach is to derive match weights jointly over a set of identifiers thus avoiding the need for independence between identifiers, which is an area of ongoing research.

2.2.2.2. Thresholds

To classify records as links, match weights are compared with a threshold or cut-off value. Choice of threshold values is important, since adjusting the thresholds alters the balance between the number of false matches and missed matches [32]. However, choosing optimal thresholds is not straightforward, and is often a subjective process based on manual review of record pairs. Typically, two thresholds are chosen: pairs with weights above the upper threshold are classified as links; pairs with weights below the lower threshold are classified as non-links; those in-between are subjected to manual review (Figure 2) [33]. Manual review involves deciding where the optimal thresholds lie, usually by inspecting pairs of records. An alternative method for choosing thresholds is to estimate error rates for a range of threshold values, and choose a threshold based on criteria for a particular study (e.g. a maximum allowable false match rate). Error rate estimation at different thresholds can be performed using a subset of data where the true match status is known ('training' or 'gold-standard' data), or using simulated data [34].

2.2.3. Designing an algorithm

Probabilistic linkage is more computationally intensive than deterministic linkage, but can lead to fewer missed matches, as it is more tolerant to missing values and recording errors [35-37]. For these reasons, probabilistic methods are often required for linkage of administrative data, which may be lacking in completeness and accuracy, and can be subject to changes over time (e.g. addresses) [3, 22]. In practice, linkage studies often use a combination of deterministic and probabilistic methods, using initial deterministic steps to reduce the number of comparison pairs for subsequent probabilistic linkage [38]. Linkage algorithms are often developed iteratively, through trial and error, manual review, linkage error rate estimation and evaluation of linkage quality.

[a] For a detailed description of methods for estimating conditional probabilities, see Winkler 2015 (Chapter 2: Probabilistic linkage. In: Methodological Developments in Data Linkage. Chichester: Wiley (10)).

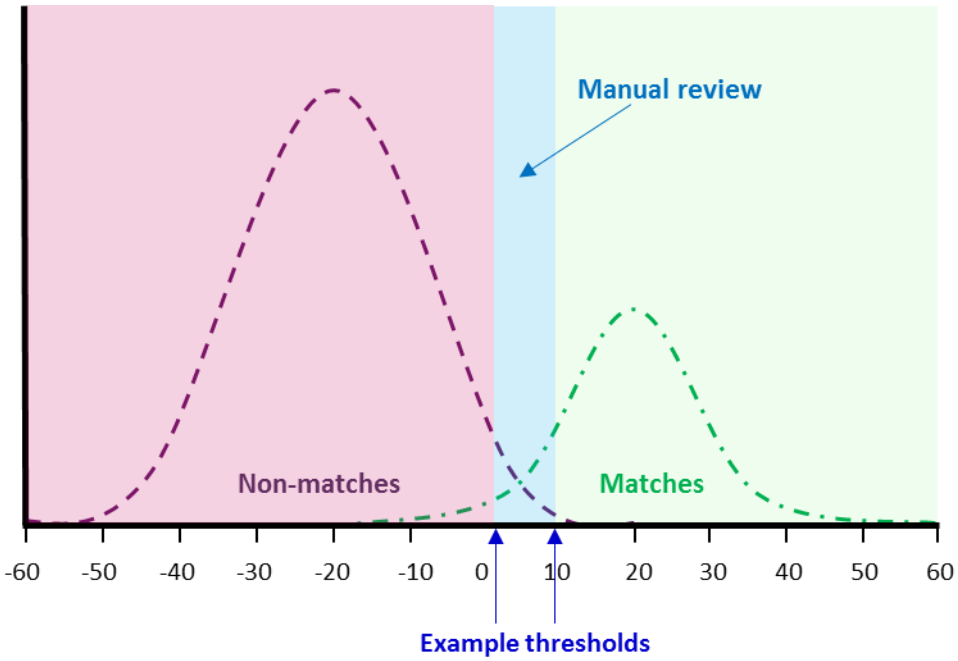


Figure 2: Example classification of links using thresholds in probabilistic linkage

2.2.4. Data preparation

Quality of linkage ultimately depends on the quality of the underlying data. If datasets contained sufficiently accurate, complete and discriminative information, data linkage would be a straightforward database-merging process. Since administrative datasets are generally created without linkage in mind, data are often messy, inconsistent, prone to missing values, and vary in structure, format and content.

For linkage to be successful, it is important that data from different datasets are cleaned and standardised in the same way. The data preparation stage of linkage converts raw data into a consistent format, resolving any consistencies. Data cleaning requires a level of balance: too much can lower the discriminative value of an identifier [39]. For example, if known nicknames are removed from the data, a smaller variety of names will be found, reducing the ability to distinguish between records belonging to different individuals.

2.2.4.1. Phonetic coding, string comparators and address standardisation

String variables, such as names or addresses, are particularly subject to typographical and data entry errors (e.g. Jonathan vs. Jonathon, Katie vs. Katy). Various string comparators and phonetic coding systems have been developed in order to overcome such small typographical errors [40, 41]. Soundex is a phonetic algorithm for indexing names in the English language, transforming names into four-character codes (e.g. Robert to R163). Other phonetic systems codes exist for different languages [42-45].

String comparators provide means for comparing strings that contain errors [46]. For example, *Edit Distance (or Levenshtein distance)* measures the 'distance' between two strings (based on the number of operations, deletions and insertions needed for two strings to be equal) [47]. The *Jaro-Winkler Comparator* is an extension of the Edit Distance which gives higher weightings to strings that match on prefixes, under the assumption that mistakes are more likely to occur towards the end of a string [48]. Methods for standardising addresses can also be used, to help overcome differences in format and abbreviations (e.g. St. for Street) [49].

2.2.4.2. Capacity for linkage

If every record in a first dataset is compared with every record in a second dataset, the total

number of pairwise comparisons is the product of file sizes: for two datasets of 100,000 records each, the total number of comparison pairs would be 10,000,000,000. As file sizes increase, this quickly becomes unmanageable. Therefore, blocking strategies are often used, which restrict the comparison pairs to those likely to match – for example, blocking on a particular geographical region or location would only consider pairs of records as potential matches if they agreed on that location. To account for errors in blocking variables, multiple blocks (e.g. year, location) can be used [50-53].

Although with modern computing power it is often possible to manage ad-hoc linkage of reasonably-sized datasets, establishing a large-scale linkage system involving the linkage of many large datasets over time requires a dedicated IT infrastructure and support. One example of such an infrastructure in the UK is the Secure Anonymised Information Linkage (SAIL), which maintains a scalable infrastructure to accommodate growing datasets and a growing user base [54].

2.2.5. Privacy preservation

Linkage can either be carried out in-house by a single data provider (on data that they own or have permission to access) or can be outsourced to another body, often known as trusted third party. For example in England, NHS Digital (formally the Health and Social Care Information Centre) acts as a trusted third party for many linkage projects involving health data. Trusted third parties are typically used in the situation where identifiable data cannot be released to analysts, and there is a need to keep identifiers separate from attribute data (known as the 'separation principle' [55]). This means that data linkers (the trusted third party) only have access to identifiers and serial record IDs (but no attribute data), and data users only have access to de-identified attribute data required for analysis (Figure 3). The trusted third party creates an anonymous match key to map together serial record IDs from each dataset, before stripping off identifiers and releasing to the research site. The research site then uses the match key to merge together attribute data from each provider (never accessing the original identifiers). The separation principle is recognised as good practice for protecting confidentiality [56]. The downside to this approach is that researchers often lack sufficient information on linkage methods and linkage quality, and may find it difficult to evaluate the impact of any linkage error on results [57-60].

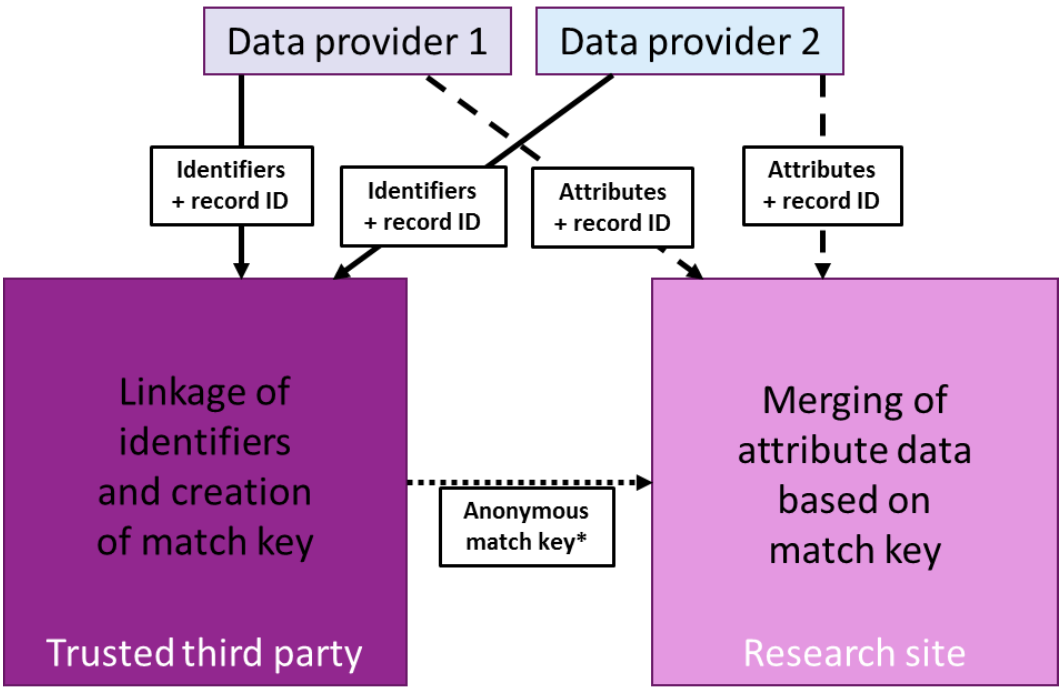


Figure 3: Separation of identifiers and attribute data. *Anonymous match key provides link between pseudo IDs from each data provider.

2.2.5.1. Encryption

Encryption ensures that data are secure while being transferred from one data holder to another (or from data holders to the trusted third party) and avoids the release of original identifiers. Encryption transforms identifiers into hashed values and hence prevents re-identification of individuals. Algorithms can be reversible (i.e. the original identifier can be obtained using an encryption 'key') or completely non-reversible (i.e. it is never possible to return to the original identifier).

One example of linkage using encryption is the Office for National Statistics (ONS) Beyond 2011 Programme, which explored different methods for linkage of administrative data to support the Census. Since the programme involved linking large quantities of information from different government departments on all individuals in England and Wales, the ONS made the decision to handle only non-identifiable data in order to maintain high levels of data security[b].

One of the limitations of linkage of encrypted identifiers is that, by design, similar identifiers look very different once encryption has taken place. For example, a hash function may transform the name "John" to the string "8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61" and the name "Jon" to "86 1A 42 1C 1A 05 E0 E8 FA 24 A1 53 41 59 69 1F". Although there is only one character difference in the original identifiers, the hashed values are completely different. This complicates the process of assessing similarity between identifiers.

One solution to this problem is the creation of 'match-keys', which take elements from each identifier, e.g. first letter of first name, first letter of second name, day of birth and postcode prefix, relying on the assumption that errors in these elements are minimal. Another proposed solution is the use of Bloom filters, which decompose a string into bigrams (2-character strings) and map these bigrams to a specific position in a binary array. Bloom filters are more complex data structures than standard hashing functions, and although they are anonymous, it is possible to compare two Bloom filters using a similarity index such as the Dice coefficient [61].[c]

The disadvantage of linkage methods using encrypted identifiers is that evaluation of linkage is problematic, since it is not possible to return to original identifiers to perform manual review, identify limitations in algorithms and assess performance. Implications for achieving adequate linkage quality should therefore be taken into account when considering the need for encryption.

2.2.6. Linkage error

Linkage error occurs when record pairs are misclassified (Table 1). Errors occur when identifiers are not sufficiently discriminative, or when available identifiers are prone to missing values, recording errors, or changes over time [62].

False matches, (also called false positives) where records from different individuals link erroneously, occur when different individuals have similar identifiers. These errors are more typical in large files (e.g. different people sharing the same sex, date of birth and postcode etc).

Missed matches, (also called false negatives) where records from the same individual fail to link, occur where there are errors in identifiers. This could be due to misreporting (e.g. typographical errors), changes over time (e.g. married womens' surnames) or missing/invalid data that prevent records from agreeing.

Although there is no universal measure of linkage quality, measures of linkage error typically

[b] (for further details, see Abbott et al 2015. Chapter 8: Large scale linkage for total populations in Official Statistics. In: Methodological Developments in Data Linkage. Chichester: Wiley (10)).
[c] For more details on Bloom filters, see Schnell 2015. Chapter 9: Privacy preserving record linkage. In: Methodological Developments in Data Linkage. Chichester: Wiley.

reported in the literature include sensitivity, specificity, match rate and false match rate (Table 2) [63, 64]. Interpreting these measures is not always straightforward. For example, match rate is only relevant if we expect all records to match; in practice, the target number of links (and true match rate) could be unknown. For example, linkage between the Millennium Cohort Study (MCS) and the National Pupil Database (NPD) achieved an 81% match rate for England, meaning that 81% of the children in the MCS were successfully linked to the NPD [65]. The remaining 19% could either be children never present in NPD (e.g. those in private / home schools) or missed matches (i.e. present in NPD but not possible to link). Furthermore, the match-rate for NPD would take a different value, as only a small proportion of children included in NPD would be present in MCS.

Although quantification of linkage error is important, it is also important to understand the impact of any errors on results [66]. The most appropriate linkage quality measures depend on the purpose of the linkage and the end use of the linked data: avoiding false matches is important for some studies, whereas for others, a high match rate may be more desirable.

For example, consider linkage between a cohort dataset and a cancer registry. A highly specific linkage (i.e. one where there were few false matches) would mean that all participants identified as having cancer really did have the disease. However, a strict linkage strategy may prevent some links from being identified, meaning that some of the controls also had cancer, but had not been identified. This could lead to dilution of any true associations, and would mean that the linked data may not be useful for providing estimates of cancer incidence. Conversely, if a more sensitive linkage were achieved, incidence estimates would be more accurate, as more cancer cases have been identified. However, some of the records may be falsely linked, meaning that a number of controls are misclassified as cases. It is important to understand the implications of linkage errors when considering study design and analyses.

		Match status	
		Match (same individual)	**Non-match (different individuals)**
Link status	**Link**	A: Identified match	B: False match
	Non-link	C: Missed match	D: Identified non-match
		Total matches	Total non-matches**

Table 1: Classification of record pairs in linkage. **Total non-matches can be defined either as all pairwise comparisons of records relating to different people, or all records from the primary dataset with a corresponding record in the linking dataset.

Measure	Calculation	Description
Match rate	$\frac{A+B}{N}$	The proportion of total records in the primary dataset (N) that are successfully linked
Sensitivity (recall)	P(identified match)= $\frac{A}{A+C}$	The proportion of matches that are correctly identified as links
Specificity	P(identified non-match)= $\frac{D}{B+D}$	The proportion of non-matches that are correctly identified as non-links
Positive predictive value (precision)	$\frac{A}{A+B}$	The proportion of links that are true matches
Negative predictive value	$\frac{D}{C+D}$	The proportion of non-links that are true non-matches

Table 2: Measures of linkage quality (A, B, C and D refer to Table 1.1)

2.2.6.1. Impact of linkage error

As data linkage becomes a more established methodology for observational research, there has been increasing interest in the impact of linkage error on analysis of linked data. The separation of processes for linkage and analysis (to help preserve privacy) can make it difficult for data users to interpret the level of uncertainty in linked data prepared for analysis [67]. The impact of linkage error on analysis of linked data depends on the structure of the data, the distribution of error, and the proposed analysis [32, 68].[d]

With health data, the number of false matches and missed matches can directly affect the estimation of prevalence or incidence rates [69-72]. False matches (low specificity) lead to overestimates of prevalence whilst missed matches (low sensitivity) lead to underestimates prevalence, [73, 74] although simultaneous errors can result in a fairly small net effect despite reasonably large error rates of both kinds [75, 76]. The impact of linkage error depends on the underlying prevalence of the target condition: analyses of rare conditions are more severely affected by linkage error compared with more common conditions, as overestimation is inversely related to the underlying prevalence.

In terms of measuring associations between variables from different datasets, false matches can increase the variability of estimates, dilute true relationships, and tend to lead to bias towards the null hypothesis[32] – i.e. they can increase the likelihood of a type 2 error. Missed matches can reduce the number of records available for analysis and so can result in a loss of statistical power [77, 78]. However, with the large sample sizes available in administrative data, a more serious problem associated with missed matches is selection bias, which occurs when particular groups are systematically less likely to link and hence are excluded from

[d] For a detailed review of the impact of linkage error on analyses, see Bohensky, 2015 (Chapter 4: Bias in data linkage studies. In: Methodological Developments in Data Linkage. Chichester: Wiley (10)).

analysis [79]. Such differential linkage is usually a result of differing data quality between subgroups of records [80-85]. A simple example of differential linkage is time-varying data quality: data quality in registries often improves over time as they become more established, and so more recent records may be more easily linked, which could impact on outcome trends [86].

2.2.7. Evaluating linkage quality

Assessing the quality of linkage is vital and allows any limitations of the linked data to be considered within analysis. Evaluation of linkage quality is typically done either through systematic quality assessment within large-scale linkage systems or on a project-specific basis, and can be done by the data linker, the data-user, or a combination of the two. For large-scale linkage systems, systematic quality assessment might include regular consistency checks and manual review of linked and unlinked records. For project-specific linkages, the nature of evaluation of linkage will depend on the nature of the planned analyses and the information available. For example, a particular study question might require high specificity, in which case evaluation would focus on the false match rate.

The most common methods for evaluating quality of linkage are as follows, and are expanded on in the following section:

- Comparing linked data with reference or 'gold-standard' datasets where the true match status is known;
- Structured sensitivity analyses where a number of linked datasets are produced using different linkage criteria;
- Comparisons of characteristics of linked and unlinked data to identify any potential sources of bias;
- Statistical methods accounting for linkage uncertainty within analysis (e.g. using missing data methods)

2.2.7.1. Measuring linkage error using 'gold-standard' data

Gold-standard (or reference) data for measuring linkage error can be obtained through detailed manual review of a sample of record pairs, the use of additional identifiers not available in all records, or well-validated external datasets [87-95]. Although convenient, creating a gold-standard dataset through manual review can take a substantial amount of time, particularly for large files [96, 97]. It is also important to consider the representativeness of reference datasets so that valid inferences about the linked dataset can be made. However, gold-standard data can be used to obtain linkage error rates, and if appropriate, measures of linkage sensitivity and specificity. Since linkage error often varies over time and for different groups of records, linkage error rates are most useful when presented across subgroups of records (i.e. to identify groups in which linkage is more or less successful).

2.2.7.2. Sensitivity analyses

Structured sensitivity analyses, e.g. comparing results from different linkage criteria or software, are useful for identifying the extent to which results depend on the linkage strategy used and for informing the most appropriate linkage strategy [79]. For example, Lariscy et al compared results from sensitive and specific linkage strategies, and observed a reverse in the direction of effect due to selection bias: Hispanic groups with more complicated name structures were less likely to link with strict linkage strategies, resulting in a biased sample [98]. Sensitivity analyses are useful for providing a range of plausible results, representing the uncertainty associated with linkage [12, 50, 74, 98-100].

2.2.7.3. Comparison of characteristics of linked and unlinked data

Comparing the characteristics of linked and unlinked data is useful for identifying potential sources of bias due to linkage error. For example, Ford et al. (2006) compared the characteristics of linked and unlinked maternal and birth records and found that particular groups of records – specifically those relating to still births or preterm births – were less likely to link [101]. Such differential linkage by subgroup is not uncommon and has been observed for more vulnerable groups, more severely ill patients, by age, institution, ethnicity and gender

[83, 102-105]. Comparisons of linked and unlinked records can be useful to identifying where modified linkage strategies may be required for specific groups of records [101]. It is also helpful to understand reasons why particular records have not linked (e.g. missing information or poor data quality).

Where not all records are expected to link (e.g. linkage between a study population and a disease registry), comparisons may need to be performed on a higher level. For example, age- and sex-distributions of linked records could be compared with distributions in population data, to establish how representative the linked data are of the target population.

2.2.7.4. Statistical methods

Statistical methods of adjusting analysis for linkage bias are an area of on-going research [106-109].[e] Until now, these methods have been limited to the context of regression analysis and rely on several assumptions.

Alternatively, viewing data linkage as a missing data problem has the potential to account for linkage error and uncertainty within analysis. Specifically, an extension to standard multiple imputation methods, able to handle 'partially observed' (or partially linked) data, was proposed by Goldstein et al. (2009) in the context of multi-level generalised linear regression [110]. These methods have been extended to the data linkage context, motivated by the idea that the ultimate purpose of linkage is not to combine *records*, but to combine *information* from records belonging to the same individual [111]. The goal in evaluating linkage quality then shifts from quantifying match rates and linkage error, to obtaining correct estimates for the outcomes of interest.[f]

Finally, there is ongoing research into the usefulness of statistical survey methodologies for addressing the problem of selection bias in unlinked data, i.e. by using population weights to account for groups or people who are more or less likely to be linked.

2.3. Discussion of ADRN-specific issues

An important role of the ADRN is to facilitate research using linked, de-identified administrative data. There are several ways in which the ADRN aims to protect data privacy, including only supporting accredited researchers who have had appropriate training on information governance, providing a secure environment in which researchers can work, and assessing proposals by an approvals panel. The ADRN ensures that research is only conducted on non-disclosive data, i.e. data from which information that directly identifies a person (such as name or address) has been removed. The current model for linkage via the ADRN is therefore to use a trusted third party with the facility for secure data linkage. Researchers then only have access to resulting linked data in a de-identified form.

Another very important aspect of ADRN's work is to develop high standards for sharing, linking and matching records securely and consistently. Understanding the quality of linked administrative data used in analyses is vital for appropriate interpretation of results, so that evidence can be used to better understand trends and patterns in the population and to inform strategies and policies that promote social wellbeing. In the context of both upholding data privacy and providing high quality research, the ADRN supports researchers to communicate on a project-specific basis with trusted third parties about linkage methodologies, data quality, and the potential impact of linkage errors on results.

2.3.1. Future ADRN research

There are several ongoing areas of research in data linkage methodology:

[e] For an example, see Chambers and Kim, 2015 (Chapter 5: Secondary analysis of linked data. In: Methodological Developments in Data Linkage. Chichester: Wiley (10)).
[f] For a detailed description of imputation methods for linkage, see Harron et al, 2015 (Chapter 6: Record linkage: a missing data problem. In: Methodological Developments in Data Linkage. Chichester: Wiley.(10)).

Software

Many commercial and open-source linkage software packages exist, including the following [9, 112, 113]:

- AUTOMATCH –Matchware Technologies [114]
- EpiLink –the Lombardy Cancer Registry [115]
- Link Plus –the US Centers for Disease Control (CDC) [116]
- Febrl – Australian National University [117]
- SALI – Software for Automated Linkage in Italy [14]

Data linkage commands also exist in a number of existing statistical software packages, including 'reclink' in STATA, the Link King in SAS, and 'RecordLinkage' in R [118]. These software packages have various limitations, e.g. a lack of flexibility in match weight calculation and parameter estimation, lack of capacity for linkage of large datasets, and limited user-friendliness [115, 119, 120]. Off-the-shelf software packages can be useful but do not always provide enough flexibility for specific linkage projects, resulting in duplication of work as different research groups develop new software or algorithms for particular purposes. Available methods for comparing linkage software are typically limited to assessing sensitivity and specificity of linkage [121]. Co-ordination and sharing of linkage algorithms could help improve and refine software development and reduce the burden on ad hoc linkage studies.

As the size of datasets to be linked increases, manual review will become infeasible, even where sufficient identifying information is available. In these cases, alternative approaches will become even more important. The practicalities of storing multiple candidate links and associated match weights or match probabilities need to be further explored. Graph databases (as opposed to traditional relational databases), could provide a technical solution to this problem, by storing records and links in the form of edges and nodes. This is an area of ongoing research [122-124].[9]

Evaluating linkage quality

The evaluation of linkage quality is vital to producing reliable results from studies using linked data, and is becoming increasingly important as linkage of administrative data underpins more research and service evaluation in the UK and internationally [58]. Due to restrictions on access to identifiable data for research, and the separation of linkage and analysis of linked data, research in this area is lacking. Development is needed of processes that allow linkage quality to be assessed while preserving confidentiality, particularly in complex linkage situations involving more than two data sources.

Reporting of studies using linked data

The importance of transparency in reporting of studies using linked administrative data is well-recognised [125]. Further research is needed to establish whether standards for reporting studies based on linked data improve communication between IT teams, data providers, and researchers using linked data, and ultimately improve the quality of studies themselves.

Unconsented linkage

Linked administrative data has the potential to support primary studies such as surveys or randomised controlled trials, from study design and recruitment, capturing outcomes, assessing generalisability, to monitoring implementation of effective interventions [126]. The potential for linkage of administrative data to assess long-term safety outcomes in existing, past, or 'dormant' clinical trials has recently been recognised in the UK and other countries [127]. However, current regulations on unconsented linkage of administrative data may prevent these opportunities from being realised. More evidence on the acceptability of and optimal approaches for unconsented linkage is required.

[9] For a detailed description on graph databases for linkage, see Farrow et al, 2015 (Chapter 7: Using graph databases to manage linked data. In: Methodological Developments in Data Linkage. Chichester: Wiley.(10)).

3. Signposting to other resources

In addition to the references cited in this guide, there are a number of useful textbooks on data linkage:

- Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*, Springer: Data-centric systems and applications.
- Harron, K., Goldstein, H., and C. Dibben (2015). *Methodological developments in data linkage*, Wiley.
- Herzog, T., F. Scheuren and W. Winkler (2007). *Data quality and record linkage techniques*, New York, Springer Verlag.

The ADRN also runs several courses on data linkage.

4. Summary and conclusions

The rich, informative datasets created by linking administrative data provide an invaluable source of information for research relating to health and society, allowing new insights into research questions that could not otherwise have been addressed [128-130]. Although the value of linked administrative datasets to research studies in general is well accepted, the dynamic, error-prone or incomplete nature of administrative data can make linkage less than straightforward. These complications can be compounded when data are anonymised or pseudonymised before linkage. Methods for data linkage have evolved over the years to accommodate imperfect data, but current techniques cannot eliminate linkage error entirely. With human involvement in the creation of these increasingly large and complex data sources, recording errors will always be an issue and lead to uncertainty in linkage. In addition, as more opportunities for linkage of cross-sectoral data arise, reliance on deterministic linkage of unique identifiers such as NHS number is unhelpful. There is an ongoing need for methodological research into the most effective ways of combining information relating to the same individual in different data sources.

Communication between data linkers and data users is vital for understanding the consequences of underlying data quality and errors in linkage, and for appropriately taking this into account within study design and analysis [131]. Researchers should be proactive in evaluating quality of linkage: data users need to know what to ask for, in terms of information required to evaluate the quality of linkage; data linkers (including trusted third parties) need to be willing to provide details of the linkage processes used to create linked datasets.

5. References

1. Organisation for Economic Co-operation and Development (OECD). *Glossary of Statistical Terms*. 02/09/14]; Available from: http://stats.oecd.org/glossary/.
2. Dunn, H., *Record linkage.* Am J Public Health, 1946. **36**(12): p. 1412-16.
3. Newcombe, H., et al., *Automatic linkage of vital records.* Science, 1959. **130**(3381): p. 954-959.
4. Roos, N.P., et al., *A population-based health information system.* Med Care, 1995. **33**(12): p. DS13-DS20.
5. Holman, C., et al., *Population based linkage of health records in Western Australia: development of a health services research linked database.* Aust N Z J Public Health, 1999. **23**(5): p. 453-459.
6. Lyons, R., et al., *The SAIL databank: linking multiple health and social care datasets.* BMC Med Inform Decis Mak, 2009. **9**: p. 3.
7. Kendrick, S. and J. Clarke, *The Scottish Record Linkage System.* Health Bull (Edinb), 1993. **51**(2): p. 72-9.
8. Furu, K., et al., *The Nordic countries as a cohort for pharmacoepidemiological research.*

Basic Clin Pharmacol Toxicol, 2010. **106**(2): p. 86-94.
9. Maggi, F., *A survey of probabilistic record matching models, techniques and tools*, in *Advanced Topics in Information Systems B*. 2008: Politecnico di Milano.
10. Harron, K., C. Dibben, and H. Goldstein, *Methodological developments in data linkage*. 2015: Wiley.
11. Christen, P., *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. 2012: Springer: Data-centric systems and applications.
12. Abrahams, C. and K. Davy, *Linking HES maternity records with ONS birth records.* Health Stat Q, 2002. **13**: p. 22-30.
13. Mears, G.D., et al., *A link to improve stroke patient care: a successful linkage between a statewide emergency medical services data system and a stroke registry.* Acad Emerg Med, 2010. **17**(12): p. 1398-1404.
14. Maso, L., C. Braga, and S. Franceschi, *Methodology used for software for automated linkage in Italy (SALI).* J Biomed Inform, 2001. **34**(6): p. 387-95.
15. Muse, A., J. Mikl, and P. Smith, *Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file.* Stat Med, 1995. **14**(5-7): p. 499-509.
16. Tromp, M., et al., *Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage.* J Clin Epidemiol, 2011. **64**(5): p. 565-572.
17. Yu, N., P.T. Donnan, and G.P. Leese, *A record linkage study of outcomes in patients with mild primary hyperparathyroidism: The Parathyroid Epidemiology and Audit Research Study (PEARS).* Clin Endocrinol, 2011. **75**(2): p. 169-76.
18. Poluzzi, E., et al., *Cardiovascular events in statin recipients: impact of adherence to treatment in a 3-year record linkage study.* Eur J Clin Pharmacol, 2011. **67**(4): p. 407-14.
19. Health and Social Care Information Centre, *Methodology for creation of the HES Patient ID (HESID)*. 2014.
20. Grannis, S., J. Overhage, and C. McDonald, *Analysis of identifier performance using a deterministic linkage algorithm.* Proc. AMIA Symp, 2002: p. 305–309.
21. Hagger-Johnson, G., et al., *Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records.* BMJ Open, 2015. **5**(8).
22. Fellegi IP and Sunter AB, *A theory for record linkage.* J Am Stat Assoc, 1969. **64**(328): p. 1183-1210.
23. Copas, J. and F. Hilton, *Record linkage: statistical models for matching computer records.* J R Stat Soc Ser A Stat Soc, 1990. **153**(3): p. 287-320.
24. Zhu, V., et al., *An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling.* J Am Med Inform Assn, 2009. **16**(5): p. 738-745.
25. Blakely, T. and C. Salmond, *Probabilistic record linkage and a method to calculate the positive predictive value.* Int J Epidemiol, 2002. **31**(6): p. 1246-52.
26. Sayers, A., et al., *Probabilistic record linkage.* Int J Epidemiol, 2015.
27. Tromp, M., et al., *Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies.* J Am Med Inform Assn, 2008. **15**(5): p. 654-660.
28. Herzog, T.H., F. Scheuren, and W.E. Winkler, *Record linkage.* WIREs Comp Stat, 2010. **2**(5): p. 535-543.
29. Herzog, T., F. Scheuren, and W. Winkler, *Data quality and record linkage techniques* 2007, New York: Springer Verlag.
30. Winkler, W.E., *The state of record linkage and current research problems*. 1999.
31. Daggy, J., et al., *A practical approach for incorporating dependence among fields in probabilistic record linkage.* BMC Med Res Methodol, 2013. **13**(1): p. 97.
32. Krewski, D., et al., *The effect of record linkage errors on risk estimates in cohort mortality studies.* Surv Methodol, 2005. **31**(1): p. 13-21.
33. Roos, L.L., A. Wajda, and J.P. Nicol, *The art and science of record linkage: Methods that work with few identifiers.* Comput Biol Med, 1986. **16**(1): p. 45-57.
34. Winglee, M., R. Valliant, and F. Scheuren, *A case study in record linkage.* Surv Methodol, 2005. **31**(1): p. 3-11.

35. Grannis, S., et al., *Analysis of a probabilistic record linkage technique without human review.* AMIA Annu Symp Proc., 2003. **2003**: p. 259-263.
36. Méray, N., et al., *Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number.* J Clin Epidemiol, 2007. **60**(9): p. 883-91.
37. Gomatam, S., et al., *An empirical comparison of record linkage procedures.* Stat Med, 2002. **21**(10): p. 1485-1496.
38. Jamieson, E., J. Roberts, and G. Browne, *The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies.* Method Inform Med, 1995. **34**(4): p. 371-7.
39. Randall, S.M., et al., *The effect of data cleaning on record linkage quality.* BMC Med Res Methodol, 2013. **13**(64).
40. Newcombe, H., M. Fair, and P. Lalonde, *Discriminating powers of partial agreements of names for linking personal records. Part I: The logical basis.* Method Inform Med, 1989. **28**(2): p. 86-91.
41. Newcombe, H., M. Fair, and P. Lalonde, *Discriminating powers of partial agreements of names for linking personal records. Part II: The empirical test.* Method Inform Med, 1989. **28**(2): p. 92-96.
42. Russell, R., *Soundex*, *available at: http://v3.espacenet.com/publicationDetails/originalDocument?CC=US&NR=1261167A&KC=A&FT=D&date=&DB=&locale= (accessed December 2013)*, United States Patent Office, Editor. 1918.
43. Russell, R., *Soundex*, *available at: http://v3.espacenet.com/publicationDetails/biblio?CC=US&NR=1435663&KC=&FT=E (accessed Deceber 2013)*, United States Patent Office, Editor. 1922.
44. Mortimer, J. and J. Salathiel, *'Soundex' codes of surnames provide confidentiality and accuracy in a national HIV database.* Commun Dis Rep CDR Rev, 1995. **5**(12): p. R183-6.
45. Zahoranský, D. and I. Polášek. *Rule based phonetic search approaches for central Europe*. in *International Symposium on Intelligent Systems and Informatics (SISY), 8th International Symposium on*. 2010. Subotica: IEEE.
46. Grannis, S., J. Overhage, and C. McDonald, *Real world performance of approximate string comparators for use in patient matching.* Stud Health Technol Inform, 2004. **107**(Pt 1): p. 43-7.
47. DuVall SL, Kerber RA, and Thomas A, *Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators.* J Biomed Inform, 2010. **43**(1): p. 24-30.
48. Winkler, W., *Chapter 11: Matching and Record Linkage*, in *Business Survey Methods*, B. Cox, Editor. 1995, Wiley: New York. p. 374-403.
49. Churches, T., et al., *Preparation of name and address data for record linkage using hidden Markov models.* BMC Med Inform Decis Mak, 2002. **2**(9).
50. Webster, A., et al., *Validity of registry data: Agreement between cancer records in an end-stage kidney disease registry (voluntary reporting) and a cancer register (statutory reporting).* Nephrology, 2010. **15**(4): p. 491-501.
51. Michelson, M. and C.A. Knoblock. *Learning blocking schemes for record linkage*. in *21st national conference on Artificial intelligence - Volume 1*. 2006. Menlo Park, CA.
52. Newcombe, H. and J. Kennedy. *Record linkage: making maximum use of the discriminating power of identifying information*. in *Communications of the ACM*. 1962. ACM.
53. Newcombe, H., *Record linking: the design of efficient systems for linking records into individual and family histories.* Am J Hum Genet, 1967. **19**(3 Pt 1): p. 335-359.
54. Jones, K.H., et al., *A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation.* Jf Biomed Inform, 2014. **50**(100): p. 196-204.
55. Kelman, C., A. Bass, and C. Holman, *Research use of linked health data - a best practice protocol.* Aust N Z J Public Health, 2002. **26**(3): p. 251-255.
56. Goeken, R., et al., *New methods of census record linking.* Hist Methods, 2011. **44**(1): p. 7-14.
57. Kelman CW, Bass AJ, and Holman CDJ, *Research use of linked health data—a best*

practice protocol. Aust N Z J Public Health, 2002. **26**(3): p. 251-255.
58. Harron K, et al., *Opening the black box of record linkage.* J Epidemiol Commun H, 2012. **66**(12): p. 1198.
59. Baldi, I., et al., *The impact of record linkage bias in the Cox model.* J Eval Clin Pract, 2010. **16**(1): p. 92-96.
60. Herman, A., et al., *Data linkage methods used in maternally-linked birth and infant death surveillance data sets from the United States (Georgia, Missouri, Utah and Washington State), Israel, Norway, Scotland and Western Australia.* Paediatr Perinat Epidemiol, 1997. **11**(S1): p. 5-22.
61. Schnell, R., T. Bachteler, and J. Reiher, *Privacy-preserving record linkage using Bloom filters.* BMC Med Inform Decis Mak, 2009. **9**(1): p. 41.
62. Sariyar, M., A. Borg, and K. Pommerening, *Missing values in deduplication of electronic patient data.* J Am Med Inform Assn, 2012. **19**(e1): p. e76-e82.
63. Pinto da Silveira, D. and E. Artmann, *Accuracy of probabilistic record linkage applied to health databases: systematic review.* Rev Saúde Públ, 2009. **43**(5): p. 875-82.
64. Christen, P. and K. Goiser. *Assessing deduplication and data linkage quality: what to measure?* in *Proceedings of the fourth Australasian Data Mining Conference*. 2005. Sydney.
65. Johnson, J. and R. Rosenberg, *Millennium Cohort Study: A guide to the linked education administrative datasets*. 2013.
66. Leiss, J.K., *A new method for measuring misclassification of maternal sets in maternally linked birth records: true and false linkage proportions.* Matern Child Health J, 2007. **11**(3): p. 293-300.
67. Chambers, R., *Regression analysis of probability-linked data*, in *Official Statistics Research Series, Volume 4*. 2009: Wellington.
68. Fett, M.J., *The development of matching criteria for epidemiological studies using record linkage techniques.* Int J Epidemiol, 1984. **13**(3): p. 351-55.
69. Anthony, S. and K. Bruin, *The reliability of perinatal and neonatal mortality rates: differential under reporting in linked professional registers vs. Dutch civil registers.* Paediatr Perinat Epidemiol, 2001. **15**(3): p. 306-314.
70. Brenner, H. and I. Schmidtmann, *Determinants of homonym and synonym rates of record linkage in disease registration.* Method Inform Med, 1996. **35**(1): p. 19-24.
71. Oberaigner, W., *Errors in survival rates caused by routinely used deterministic record linkage methods.* Method Inform Med, 2007. **46**(4): p. 420-424.
72. Clough-Gorr, K. *The challenge of unlinked deaths in health research: an example from the Swiss National Cohort Study*. in *Exploiting Existing Data for Health Research, Scottish Health Informatics Programme* 2011. St Andrews
73. Brenner, H., I. Schmidtmann, and C. Stegmaier, *Effects of record linkage errors on registry-based follow-up studies.* Stat Med, 1997. **16**(23): p. 2633-2643.
74. O'Reilly, D., M. Rosato, and S. Connolly, *Unlinked vital events in census-based longitudinal studies can bias subsequent analysis.* J Clin Epidemiol, 2008. **61**(4): p. 380-385.
75. Brenner, H. and I. Schmidtmann, *Effects of record linkage errors on disease registration studies.* Method Inform Med, 1998. **37**(1): p. 69-74.
76. Clark, D., *Practical introduction to record linkage for injury research.* Injury Prev, 2004. **10**(3): p. 186-191.
77. McGeechan, K., et al., *Evaluation of linked cancer registry and hospital records of breast cancer.* Aust N Z J Public Health, 1998. **22**(7): p. 765-770.
78. Haas, J., et al., *Creating a comprehensive database to evaluate health coverage for pregnant women: the completeness and validity of a computerized linkage algorithm.* Med Care, 1994. **32**(10): p. 1053-1057.
79. Lariscy, J.T., *Differential record linkage by hispanic ethnicity and age in linked mortality studies.* J Aging Health, 2011. **23**(8): p. 1263-1284.
80. Lawrence, D., et al., *Adjusting for under-identification of aboriginal and/or Torres strait islander births in time series produced from birth records: Using record linkage of survey data and administrative data sources.* BMC Med Res Methodol, 2012. **12**(1): p. 90-102.
81. DuVall, S.L., et al., *Evaluation of record linkage between a large healthcare provider and the Utah Population Database.* J Am Med Inform Assn, 2011. **19**(e1): p. e54-e59.

82. Coeli, C.M., et al., *Estimated parameters in linkage between mortality and hospitalization databases according to quality of records on underlying cause of death.* Cad Saude Publica, 2011. **27**(8): p. 1654-8.
83. Adams, M.M., et al., *Constructing reproductive histories by linking vital records.* Am J Epidemiol, 1997. **145**(4): p. 339-348.
84. Leiss, J.K., et al., *US maternally linked birth records may be biased for Hispanics and other population groups.* Ann Epidemiol, 2010. **20**(1): p. 23-31.
85. Boyle, D. and S. Cunningham, *Resolving fundamental quality issues in linked datasets for clinical care.* Health Informatics J, 2002. **8**(2): p. 73-77.
86. Newcombe, H.B., *Age-related bias in probabilistic death searches due to neglect of the "prior likelihoods".* Comput Biomed Res, 1995. **28**(2): p. 87-99.
87. Fonseca, M., et al., *Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database.* Cad Saúde Pública, 2010. **26**: p. 1431-1438.
88. Monga, H. and T. Patrick, *Error estimation in linking heterogeneous data sources.* Health Inform J, 2001. **7**(3-4): p. 135-137.
89. Newgard, C., *Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry.* Acad Emerg Med, 2006. **13**(1): p. 69-75.
90. Wiklund, K. and G. Eklund, *Reliability of record linkage in the Swedish Cancer-Environment Register.* Acta Oncologica, 1986. **25**(1): p. 11-14.
91. Zingmond, D., et al., *Linking hospital discharge and death records--accuracy and sources of bias.* J Clin Microbiol, 2004. **57**(1): p. 21-29.
92. Gill, L. *OX-LINK: The Oxford Medical Record Linkage System*. in *Record Linkage Techniques: Proceedings of an International Workshop and Exposition*. 1997. Washington DC: Federal Committee on Statistical Methodology, Office of Management and Budget.
93. Morris, A., et al., *The diabetes audit and research in Tayside Scotland (DARTS) study: electronic record linkage to create a diabetes register.* BMJ, 1997. **315**(7107): p. 524-8.
94. Potz, N., et al., *Probabilistic record linkage of infection records and death registrations: a tool to strengthen surveillance.* Stat Comm Infect Dis, 2010. **2**(1): p. 6.
95. Belin, T. and D. Rubin, *A method for calibrating false match rates in record linkage.* JAMA, 1995. **90**(430): p. 694-707
96. Qayad, M. and H. Zhang, *Accuracy of public health data linkages.* Matern Child Health J, 2009. **13**(4): p. 531-538.
97. Sauleau, E., J. Paumier, and A. Buemi, *Medical record linkage in health information systems by approximate string matching and clustering.* BMC Med Inform Decis Mak, 2005. **5**(1): p. 32.
98. Nitsch, D., et al., *Linkage bias in estimating the association between childhood exposures and propensity to become a mother: an example of simple sensitivity analyses.* J R Stat Soc Ser A Stat Soc, 2006. **169**(3): p. 493-505.
99. Churches, T. and K. Lim, *Using record linkage to measure trends in breast cancer surgery.* N S W Public Health Bull, 2001. **12**(4): p. 105-110.
100. Campbell, K., *Impact of record-linkage methodology on performance indicators and multivariate relationships.* J Subst Abuse Treat, 2009. **36**(1): p. 110-117.
101. Ford JB, Roberts CL, and Taylor LK, *Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data.* Paediatr Perinat Epidemiol, 2006. **20**(4): p. 329-337.
102. Bohensky, M., et al., *Empirical aspects of linking intensive care registry data to hospital discharge data without the use of direct patient identifiers.* Anaesth Intens Care, 2011. **39**(2): p. 202-8.
103. Bentley, J., et al., *Investigating linkage rates among probabilistically linked birth and hospitalization records.* BMC Med Res Methodol, 2012. **12**(1): p. 149.
104. Duvall, S., et al., *The impact of a growing minority population on identification of duplicate records in an enterprise data warehouse.* Stud Health Technol Inform, 2010. **160**(Pt 2): p. 1122-6.
105. Fournel, I., et al., *Contribution of record linkage to vital status determination in cancer patients.* Stud Health Technol Inform, 2009. **150**: p. 91-5.
106. Chambers, R., et al., *Inference based on estimating equations and probability-linked*

data, Centre for Statistical & Survey Methodology Working Paper Series, Editor. 2009: University of Wollongong. p. 38.
107. Kim, G. and R. Chambers, *Regression analysis under probabilistic multi-linkage.* Stat Neerl, 2011. **66**(1): p. 64-79.
108. Scheuren, F. and W. Winkler, *Regression analysis of data files that are computer matched - Part II.* Surv Methodol, 1997. **23**(2): p. 126-138.
109. Hof, M.H.P. and A.H. Zwinderman, *Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables.* Stat Med, 2012. **31**(30): p. 4231-4242.
110. Goldstein, H., et al., *Multilevel models with multivariate mixed response types.* Stat Model, 2009. **9**(3): p. 173-197.
111. Goldstein, H., K. Harron, and A. Wade, *The analysis of record-linked data using multiple imputation with data value priors.* Stat Med, 2012. **31**(28): p. 3481-93.
112. Cochinwala, M., et al., *Record matching: Past, present and future*. 2001, Computer Sciences, Purdue University
113. Gu, L., et al., *Record linkage: Current practice and future directions*, in *CSIRO Mathematical and Information Sciences Technical Report*. 2003, Citeseer. p. 83.
114. Jaro, M., *Probabilistic linkage of large public health data files.* Stat Med, 1995. **14**(5-7): p. 491-498.
115. Contiero, P., et al., *The EpiLink record linkage software.* Methods Inf Med, 2005. **44**(1): p. 66-71.
116. Campbell, K., D. Deck, and A. Krupski, *Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a basic deterministic algorithm.* Health Informatics J, 2008. **14**(1): p. 5-15.
117. Christen, P., *Febrl: a freely available record linkage system with a graphical user interface*, in *Proceedings of the second Australasian workshop on Health data and knowledge management - Volume 80*. 2008, Australian Computer Society, Inc.: Wollongong, NSW, Australia. p. 17-25.
118. Campbell, K. *Rule Your Data with The Link King©(a SAS/AF® application for record linkage and unduplication)*. 2005.
119. Newman, T. and A. Brown, *Use of commercial record linkage software and vital statistics to identify patient deaths.* J Am Med Inform Assn, 1997. **4**(3): p. 233-237.
120. Wajda, A., et al., *Record linkage strategies: Part II. Portable software and deterministic matching.* Method Inform Med, 1991. **30**(3): p. 210-4.
121. Ferrante, A. and J. Boyd, *A transparent and transportable methodology for evaluating Data Linkage software.* J Biomed Inform, 2012. **45**(1): p. 165 - 172.
122. Finney, J., et al., *An efficient record linkage scheme using graphical analysis for identifier error detection.* BMC Med Inform Decis Mak, 2011. **11**(1): p. 7.
123. Kirby, G., et al. *Comparing relational and graph databases for pedigree datasets* in *Exploiting Existing Data for Health Research (SHIP)*. 2013. University of St Andrews.
124. Farrow, J. *Improving linked data quality, research outcomes and reducing costs using graph theory and graph databases* in *Exploiting Existing Data for Health Research*. 2013. University of St Andrews.
125. Benchimol, E.I., et al., *The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement.* PLoS Med, 2015. **12**(10): p. e1001885.
126. Harron K, Gamble C, and Gilbert R, *E-health data to support and enhance randomised controlled trials in the UK.* Clinical Trials, 2014. **(online first)**.
127. Henry, D. and T. Fitzpatrick, *Liberating the data from clinical trials.* BMJ, 2015. **351**.
128. Kelman, C.W., et al., *Deep vein thrombosis and air travel: record linkage study.* BMJ, 2003. **327**(7423): p. 1072-1075.
129. Merrall, E.L., S.M. Bird, and S.J. Hutchinson, *Mortality of those who attended drug services in Scotland 1996–2006: Record-linkage study.* Int J Drug Policy, 2012. **23**(1): p. 24-32.
130. Holman, C., et al., *A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system.* Aust Health Rev, 2008. **32**(4): p. 766-777.
131. Bohensky, M.A., et al., *Development and validation of reporting guidelines for studies involving data linkage.* Aust N Z J Public Health, 2011. **35**(5): p. 486-489.

6. Acknowledgements

Many thanks to Ruth Gilbert and Harvey Goldstein for their contributions, advice and support for this work on data linkage.

7. Glossary

Blocking: Process of indexing datasets to reduce the number of record pairs for comparison. For example, in a linkage that uses blocking by year of birth, only records with the same year of birth would be considered as potential matches.

Bloom filter: Data structures used to allow comparisons of encrypted identifiers.

Deterministic linkage: Rule-based linkage typically requiring exact agreement on a unique identifier or set of partial identifiers.

Encryption: Method of de-identifying data values, for example through hashing. Encryption can be reversible or non-reversible.

False match: Records belonging to different subjects that have been linked together (otherwise known as a false positive).

Gold-standard data: A sample of data where the true-match status of record pairs is known, often created through manual review; used as training data to optimise linkage algorithms or to evaluate linkage quality.

Identifier: Data field compared between records to identify the same subject. Identifiers can be unique (e.g. NHS number, National Insurance number) or partial (e.g. sex, date of birth, postcode).

Linkage algorithm: Set of rules or criteria used to classify pairs of records as belonging to the same or different individuals, typically computerised.

Match key: Anonymous code allowing two or more sets of attribute data to be brought together.

Match weight: Numerical value representing the likelihood of two records belonging to the same subject given agreement on a set of partial identifiers.

Missed match: Records belonging to the same subject that have not been linked together (otherwise known as a false negative).

m- and u-probabilities: Conditional probabilities used to derive match weights in probabilistic linkage.

Probabilistic linkage: Linkage using match weights representing the likelihood of two records belonging to the same subject given agreement on a set of partial identifiers.

Sensitivity analysis: Process of systematically repeating linkage under different assumptions and exploring the effect on results.

Sensitivity: The proportion of true-matches that are correctly classified.

Specificity: The proportion of true non-matches that are correctly classified.

String comparator: Algorithm to provide a numerical value representing agreement or distance between two strings

Threshold: Cut-off value for classifying record pairs as links or non-links, based on a match weight.

Trusted third party: Body used to perform linkage, with separation from data providers.

Version history

Version	Changes made	By	Date
01.00	First published		2016-06-08

ADRN PUBLICATION
University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ
T. +44 (0) 1206 87 2976 E. help@adrn.ac.uk W. adrn.ac.uk

Funded by

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