Government Statistical Service

Best Practice and Impact

GSS Methodology Advisory Committee

Privacy, confidentiality and practicalities in data linkage

Professor Kerina Jones
Swansea University

Professor David Ford
Swansea University

A contributing article to the National Statistician's Quality Review into Privacy and Data Confidentiality Methods

December 2018

Contents

1. Focus

This article explores data linkage methodologies with particular focus on privacy and confidentiality. These two terms are often used interchangeably, but for clarification: privacy refers to a person's freedom from intrusion in their activities or information; confidentiality relates to information where there is an expectation that it will be held in confidence and not divulged without authorisation. The article introduces terminologies and concepts used in data linkage, and the main types of data linkage method in practical use. It briefly outlines the UK legislative and regulatory backdrop to emphasise the importance of privacy and confidentiality in context. Having outlined linkage methods, it discusses practicalities and the respective measures to evaluate linkage efficacy. From this, linkage methods in practice are illustrated via a selection of case studies. The choice of data linkage approach cannot be made in isolation, and so the material is drawn together and contextualised to propose a set of high-level questions and options to inform decision-making. Finally, it emphasises the importance of privacy and confidentiality in data linkage, and recommends a robust, proportionate, data governance framework with privacy by design to enable both the safe and optimum use of data.

Data linkage is a large subject and, naturally, this article can only cover a limited scope. Further information is provided in an annotated bibliography.

2. Introduction

2.1 Basic principles

Data linkage is defined as the processes involved in connecting records that relate to the same person, family, event, organisation or location (i.e. entities) within or between datasets. The term may also be referred to as entity resolution, de-duplication or record linkage. Within datasets it is used to clean the data by removing duplicate records and verify entity identities, and between datasets it enables data integration for combined analysis [1]. Methods of combining datasets that do not bring together specific entities (such as overlaying area-based deprivation indices and pollution levels) are outside the scope of this article. The growing availability of administrative datasets, and emerging data types in the increasingly connected digital world, is leading to vast expansions in data linkage across all sectors.

The process relies on the presence of linkage variables i.e. data items that are present in both records of interest, and it is successful when these variables match in a record pair. However, there are various methodologies for determining whether two records form an acceptable match, with the main categories being deterministic or probabilistic linkage. In seeking to link individual-level data, both of these rely on the use of person identifiable data (PID). As such, privacy and confidentiality are of utmost importance, but they are significant factors whatever and however entities are being linked [1, 2]. In an ideal situation, each entity represented in the dataset would have a consistent, unique identifier(s) on which to base the linkage, so that all true matches and all false matches would be designated as such, but data linkage is not that simple because administrative datasets are not perfect. Consequently, there is a trade-off across four possible results:

Match status	Actual	Assigned
True positive	Match	Match
False positive	Non-match	Match
True negative	Non-match	Non-match
False negative	Match	Non-match

Deterministic record linkage (DRL) is based on classification rules to determine whether pairs of records are links or non-links. Within DRL, exact matching requires all linkage variables to be identical. But more generally, DRL allows for some degree of variation provided that the match is still definitive i.e. based on set rules and with no alternative competing record pair. DRL is most suitable in highly discriminative datasets with unique identifiers and/or high concordance between linkage variables. But, because of its high specificity, it can result in a high rate of missed matches (false negatives).

Probabilistic record linkage (PRL) is based on assigning match weights to variables of interest to represent the likelihood that a record pair is a true match, given the degree of agreement between the variables, and an agreed threshold for classifying the link as true or false. PRL is traditionally supported with clerical review to set the thresholds for classifying matches/non-matches. Sometimes a grey area is set aside for manual decision making to limit the rate of incorrectly assigned matches. Depending how the weights and thresholds are set, the resulting sensitivity of PRL can tend towards a high rate of incorrectly assigned matches (false positives) [1, 2].

However, the distinction between DRL and PRL is not dichotomous, and both have many valid applications providing they are tuned appropriately. Both are able to accommodate uncertainty and partial agreement between record pairs, and there is always a trade-off between false positives and false negatives [3]. Often an optimised strategy is used in practice to combine both DRL and PRL techniques.

In addition to PID-based linkage, there are also methods that do not rely on using identifiable data for record comparison. Privacy-preserving record linkage (PPRL) refers to data linkage using hash-encoded PID, with or without other variables of interest, to create linked record pairs. As a general principle, the PID is subjected to a one-way cryptographic hash-function, converting it to a string of data points such as a combination of 1s and 0s, with positional and frequency variations distinguishing one string of encrypted PID from another, and acting as the basis for record linkage. However, PPRL methods are not as long-established as DRL and PRL and are the subject of on-going research endeavours [4].

The choice of data linkage method will depend on various factors. But also, in setting out to link administrative datasets, it must be remembered that the data were generally not collected with linkage in mind. As such, the presence of a common, unique, cross-dataset, entity identifier is the exception rather than the norm in the UK. For example, healthcare records have an NHS number, employment data have an NI number, and school records have a pupil number. These numbers are not indexed for cross-comparison. Furthermore, there is often no accessible ground truth, or ‘gold standard’ listing against which to resolve identities definitively. As well as this, administrative data have other inherent challenges, including, missing records, missing variables, and entry errors [5]. We will return to some of these issues in the discussion.

2.2 Legislative and regulatory backdrop

Within the UK, the main data protection legislation is the Data Protection Act (2018) [6] enacting the EU General Data Protection Regulation (2016) [7]. There is a basic requirement that the processing of general personal data must be able to rely on a lawful provision, set out in Article 6 of the GDPR, and special category data on the provisions of Article 9. Personal data are defined in Article 4 as *'any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.'* Special category data are those revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sex life or sexual orientation, and the use of genetic data and biometric data to uniquely identifying a natural person.

Data that have been anonymised, such that an individual is no longer identifiable, fall outside data protection law. However, the GDPR includes a definition of pseudonymisation, and this refers to processing data in such a way that *'the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures.'* For these reasons, and the fundamental ethic of respecting individual privacy, the provisions of data protection legislation are highly relevant to data linkage processes.

There are various established data linkage enterprises in the UK that make use of administrative data for research and evaluation. In general, datasets are used in anonymised format, subject to a suite of controls applied to both the data and the data environment. Data linkage is most usually carried out using PID, before the anonymization processes take place., but sometimes data providers require anonymization prior to linkage. For datasets derived from healthcare records, the use of an NHS-based trusted third party between the data provider and the data linkage enterprise can address the issues of using PID for data linkage. We will discuss this in more detail in Section 5.4. The Digital Economy Act (2017) [8] excludes health and social care data but makes provision for data sharing from non-health sources for research. Importantly, the provisions of the Digital Economy Act, specifically chapter 5 (64), can be relied upon for matching and linking such administrative data, in addition to those of the GDPR and DPA. These and other pieces of legislation may apply to the primary collection of identifiable data, but as we are making the assumption that administrative data are extant we consider this out of scope.

As well as general legislation, the provisions of the UK Statistics and Registration Service Act (2007) [9] and the Office for Statistics Regulation (OSR) Code of Practice for Statistics (2018) [10] are of particular relevance for work producing official statistics from administrative data. The Code is founded on three pillars: trustworthiness, quality and value, for confidence in producing data, assuring data and supporting society's information needs. However, the USKA/OSR Quality Assurance of Administrative Data (QAAD) toolkit is used more specifically to assess administrative data use. Data linkage is an invaluable tool to maximise the potential of currently standalone datasets provided that this can be accomplished within the bounds of proper data governance to ensure privacy and confidentiality. The choice of data linkage methods to be used

will vary with many factors, but will include the legislative position as to whether PID can lawfully be processed for this purpose.

3. Data linkage methods

3.1 Data linkage steps

There is a series of generic steps in record-to-record data linkage [11]. Privacy and confidentiality are essential considerations at all stages, because PID are required for at least some stages in all methods, and resulting linked datasets are not immune to disclosure risks.

i) Pre-processing is necessary for assurance that both datasets are in a compatible format. This relies essentially on the use of PID.
ii) Indexing/Blocking are used to reduce the number of record pairs compared, by specifying that potential matches must agree on certain criteria, thereby reducing the complexity of the matching process. This requires knowledge of the data characteristics [12].
iii) Record pair comparisons generate potential matches, based either on comparisons of PID or hashed functions.
iv) Classification involves assigning matches, non-matches and potential matches. This is best performed against a gold standard and may involve clerical review.
v) Evaluation of the quality and completeness of the matched dataset provides an assessment of process efficacy.

3.2 Deterministic record linkage

In many ways, DRL using PID is the most straightforward method of record linkage, particularly exact linkage. Whatever linkage variables are chosen, such as NHS number, first name, surname, date of birth, etc., a true match is assigned when the variables are identical in a pair of records. It does not have to rely on complex mathematical processes, but simply on whether or not two records are identical. If they are not identical, the record pair is a non-match. It is easy to see why exact matching has high specificity (resulting in a low rate of false positives), but also that it has low sensitivity (resulting in a high rate of false negatives). The advantages of exact matching are when datasets to be linked are of reliable quality and both contain highly discriminatory variables such as an NHS number.

Beyond exact matching, DRL methods allow for degrees of variability in the linkage variables and these are embodied in a succession of decision rules. For example, following exact matching, comparisons of unmatched records may allow record pairing where all variables agree except one, or the rules may be set to allow more complex variability. In practice, for most administrative datasets with their many inherent imperfections, multiple decision rules will be required and the process becomes non-trivial. Decisions need to be made on the hierarchy to assign to the rules, so they can be applied sequentially, beginning with the most specific and moving towards more flexibility. This is sometimes a subjective process, as well as depending on the linkage variables available and their relative discriminatory power. The resulting record matches are assigned a match rank to express the likelihood that they represent a true match. At the start, specificity will be high and sensitivity low, moving to *vice versa* as rules allow greater flexibility. The balance between the accompanying trade-off

between false negatives and false positives will depend on the requirements for the linked dataset produced.

3.3 Probabilistic record linkage

PRL, like DRL, relies on the use of PID for the data linkage process. The idea that PRL is probabilistic and DRL is not, is somewhat of a misconception since both include likelihood and uncertainty [3]. In PRL, variables are assigned a weighting based on their discriminatory power and statistical theory is used to determine if a set of linkage variables constitute a true match. Basically, two sets of probabilities are produced: *m* represents the probability that a variable agrees on two data sources, given the pair are a true match; *u* represents the probability that the variable agrees on both data sources given the pair are NOT a true match. So, the *m*-probability is a measure of data quality and the *u*-probability measures the distinguishing power of that variable. These are used to assign a match weight to the pair to create a score representing their likelihood of being a true match [13]. The decision of whether to accept a given score as a true match is based on setting thresholds (and often clerical review) which in turn depends on the planned applications of the product dataset, as for DRL.

In practice, there is no reason to use DRL or PRL in isolation. Algorithms using combinations of DRL and PRL can be used by means of successive matching runs. Successive match runs can be very effective in large scale linkage exercises where DRL is efficient at linking the majority of straightforward cases. This allows the greater computational demands of PRL to be reserved for linking more challenging residual cases. In an ideal situation, whatever linkage regime is used, it would be evaluated against a gold standard dataset of verified identities, and dataset purpose based on whether the optimum for the use case is to maximise specificity or sensitivity. Measures to evaluate linkage methods are discussed in Section 4.

3.4 Privacy-preserving record linkage

The key point about privacy-preserving record linkage (PPRL) is that it doesn't rely on PID for the data linkage process beyond the pre-processing stage. Instead, it uses cryptographically hashed bit strings of the linkage variables, but it does, of course, require PID to produce the strings [14]. There are many PPRL methods, but few have been operationally evaluated for use with real-world administrative data [15]. A thorough examination of PPRL is beyond the scope of this article but the general principles are illustrated via the Bloom filters method [16].

A Bloom filter is a probabilistic data structure that can be used in similarity comparisons for record linkage. They are bit vectors consisting of a series of 1s and 0s, with all positions initially set to zero until the hash function is applied to a given variable. An example is shown below:

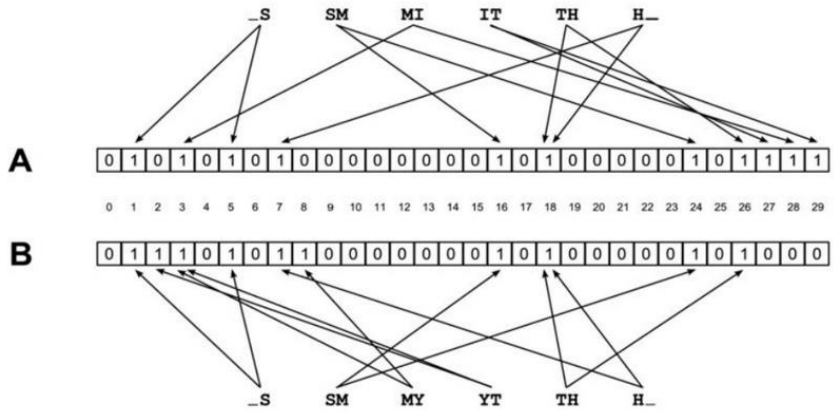


Figure 1: An example of Bloom filters from the names Smith and Smyth for string comparison[1].

The variables of interest, in this case the names Smith and Smyth, are split up into *q*-grams (in this case bigrams: _S, SM, etc.) and the hash function determines the positions of the 1s. It can be seen that where the bigrams are identical between Smith and Smyth, 1s are generated in the same positions, and where they differ (e.g. between MI and MY) they result in differing placements. Hash functions are one-way i.e. once they are applied, they cannot be used in reverse to return to the PID used to create them.

The likelihood of a match is based on the similarity between the strings, with one method of assessing this being the Dice coefficient. The use of such coefficients is not exclusive to PPRL, and can be used to improve linkage in PRL by incorporating partial agreement on linkage variables:

$$D(a,b) = \frac{2h}{(|a| + |b|)}$$

Where, h is the number of shared bigrams, and |a|, |b| is the number of bigrams in the strings a and b. Other variables of interest can be used, including numerical values such as date of birth, and variables can be concatenated to create cryptographic long-term keys (CLKs). The resulting similarity value is based directly on the extent to which the positions of the 1s match exactly. But the assessment is more sophisticated with the addition of more flexible measures of comparison and expectation, which are beyond the scope of this article. Generally, there's a trade off with different types of encryption. Those that retain intelligence to enable comparison of string similarity (as with Bloom filters) are often criticised as being vulnerable to cryptographic attacks. Those that do not retain intelligence for string comparison (e.g. Secure Hashing Algorithms) offer more protection, but less utility to maximise linkage quality.

Despite the fact that PPRL does not rely on PID once it comes to the record linkage process, the process and the resulting dataset are not immune to privacy and confidentiality risks. PID are still used to create the hashed strings and, even if all (or

[1] After Schnell, 2009 [16].

the majority of) possible quasi-identifiers are encrypted, linked datasets are still subject to disclosure risks [17, 18]. It may not be practical or desirable to go to the length that all attributes be encrypted, and controls beyond those applied directly to the data need to come into play to maintain data utility.

3.5 Computer science approaches

In recent years there have been major developments in what can be grouped into computer science approaches, made possible as a result of more accessible and larger compute capacity. These are based on data mining or machine learning techniques, and can be supervised or unsupervised. Supervised learning requires training data to verify true matches, whereas unsupervised learning does not. Both may use clustering or graph-based approaches to classify record matches.

These differ from 'traditional' methods in that they don't rely on classifying pairs of records individually but on clusters of similar records. But there are some major hurdles to overcome. For example, good training data with known true matches and non-matches are rarely available in practice and may have to be generated. This can be costly, risky for data quality and burdensome in terms of work effort. Active learning processes, or semi-supervised learning, aim to overcome this problem by selecting samples of record pairs to be classified manually. The method is then iteratively trained and improved based on the manually labelled data. Unsupervised models operate a step further, without training data or labelled data [19].

Graph-based data linkage and management approaches are gaining momentum in some quarters but are still subject to much research. These methods are based on relationships between groups of records, rather than individual pairs. Basically, rather than store the data in relational databases, which might not adequately capture the underlying nature of the data, they are stored in a graph database, having a network of nodes (records) connected by edges (comparisons). This structure is purported to better capture the relationships between data elements. As such, it is not a linkage approach *per se* but is a management approach to enable novel linkage approaches.

The typical end to end workflow for the graph approach is as follows:

- Data cleaning
- Adding data to system and assigning identity
- Calculating deterministic links and adding equivalence links to database
- Performing self-linkage and adding de-duplication layer to database
- Performing cross-linkage and adding cross-linkage layer to database
- Identifying thresholds
- Identifying clusters containing non-reviewed edges within thresholds
- Reviewing/resolving these clusters
- Extracting resulting clusters
- Assigning cluster identity and generating change reports
- Returning information to client

However, a graph-based approach may also result in there being no explicit cluster identity because the cluster created depends on the thresholds applied. For this reason, some see the use of graph databases as a departure from traditional thinking in terms of a master list or definitive record identity, to one where the notion of 'truth' is somewhat 'relaxed'. None the less, graph-based approaches are proposed to have greater flexibility and advantages over relational approaches [20]. Privacy and

confidentiality issues need to be properly taken into account as for any data management and linkage approach.

3.6 Successive data linkage

So far, we have considered linkage of two datasets by record-to-record comparison[2]. But in practice, there will often be a need to link more than two datasets to produce data for research. An ideal way of ensuring consistent successive record linkage is by creating and retaining a unique linking key for each individual (or other entity) represented in the dataset. However, this might not always be possible. While there are other ways to approach this issue, they may incur difficulties in assigning dataset precedence and may lead to greater numbers of records remaining unmatched (false negatives). For example, when dataset A is linked to dataset B producing dataset C, data imperfections and linkage trade-offs, will mean potential matches are left behind. When attempting to link a successive dataset to dataset C, depending where datasets A and B reside, it might not be possible to go back and conduct the new linkage on the full sets of records, but only on dataset C. This will result in the loss of potential matches for further successive linkage, as was the case in linking A and B, and so on.

Depending on the method used in successive linkage, there can be considerable issues with transitive closure, especially if not all PID variables agree exactly for all records relating to a single entity. Transitive closure relates to the concept of reachability between data nodes. For example, for records, a1, a2 and a3, independent match decisions can lead to contradictions as follows: a1 and a2 are designated a match, a1 and a3 are designated a match, but a2 and a3 are designated a non-match [22]. As a result of issues such as this, linkage optimisation in succession is still the focus of on-going research. The problem can be avoided if a linkage key, unique to each individual represented in the dataset, can be retained.

Data linkage enterprises might not be authorised to hold identifiable data in order to create a linking key. In such cases they may use a trusted third party (TTP) to carry out the matching process and to remove the PID. Alternatively, this may be done by the original data controller. This process (ideally) verifies the identities in a dataset of interest against a gold standard reference dataset (where this exists). It then replaces the PID with a linkage key, unique to each individual represented in the dataset. The anonymous key can be used to link any number of datasets rendered devoid of their PID. This principle will be illustrated using a working example in Section 5.4.

As noted above, any datasets prepared by record linkage might still risk privacy and confidentiality disclosures, even if they are ostensibly anonymous. This does not rely on reversing the process of PID removal; it depends on piecing information together from ‘quasi-identifiers’. There are various terms for attacks on such datasets, with common ones being ‘jigsaw’, ‘attribute re-identification’ and ‘implicit linkage’. It may be attempted from the linked dataset alone or in combination with external pieces of information in the attacker’s possession or awareness. There is a whole realm, sometimes referred to as ‘re-identification science’ [23], dedicated to developing and testing linkage attacks with a view to minimising residual risk. This is particularly challenging if the means available to highly motivated intruders are taken into account, such that some view the creation of a totally anonymous, useful dataset as impossible.

[2] As well as record-to-record approaches, there are also clustering and graph-based methods to assess potential links and to manage data [19-21].

Crucially, controls on the data environment are highly relevant, not solely those applied to the data.

Some organisations choose to use synthetic datasets, since they may be acceptable for certain purposes requiring broad trends by retaining the distributions of key variables. Synthetic data may be created *de novo* or derived from extant data. In either case, data linkage can still be applied to synthetic data provided that the datasets contain one or (preferably) more variables in common, with sufficient reliability and discriminatory power to be used for linkage. Again, a consistent unique key would be ideal.

4. Evaluating data linkage efficacy

The usual measures to assess linkage efficacy are precision and recall. Precision is similar to positive predictive value and is defined as the number of true matching pairs (True Positives (TP)) divided by the total number of pairs designated as matching (TP plus False Positives (FP)):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity, is calculated by dividing TP by the number of all true matching pairs (TP plus False Negatives (FN)):

$$Recall = \frac{TP}{TP + FN}$$

The estimation of precision and recall requires a ground truth, or gold standard, for comparison. Alternatively, it can be established through clerical review of stratified samples of matches & pairs classified as non-matches. They are augmented by other measures such as the harmonic mean between precision and recall, which is known as the *F*-measure [24, 25]. The *F*-measure can be useful in combination with other measures, but caution has been recommended since the *F*-measure can also be expressed as a weighted arithmetic mean of precision and recall, with weights which depend on the linkage method being used. This reformulation reveals that the *F*-measure has a major conceptual weakness since it depends on the relative importance given to precision and recall, and means that different linkage methods are being evaluated using different measures. In effect, the measure being used to evaluate match performance depends on the mechanism being evaluated [26].

For DRL and PRL which rely the use of PID for data linkage, we take for granted here that identifiable data must only be processed in accordance with data protection legislation and regulations, and with data provider permissions for appropriate data handling to ensure privacy and confidentiality. This must be the case whether data processing is carried out at source, by a TTP, or by a data linkage enterprise. Whatever technique is used, PID must be surrounded by a suitable regime of controls and safeguards to prevent data breaches and misuse, as for working with personal data for any purpose.

At first glance, it could be thought that the use of PPRL techniques which de-identify data at source, avoid the issues of privacy and confidentiality. However, this is not the

case, as alluded to above. PPRL techniques are designed to mitigate the risks associated with the use of PID in the record linkage stages. But even with PPRL and federated linkage, once linked datasets are produced which have quasi-identifiers and a breadth of granular data variables, it can be relatively straightforward to re-identify at least some of the data subjects. When thought of in this way, the value afforded by PPRL is limited, since the extra protection it provides initially does not carry through to linked data usage.
Consequently, PPRL techniques are evaluated via a combination of scalability, linkage quality and privacy [17, 24]. Scalability relates to process efficiency and compute demand, since this can be particularly high when using PPRL. Linkage quality is the measure of linkage precision (similar to positive predictive value) and recall (sensitivity). Privacy in this context is a measure of security risk of disclosure of PID during the process. Overall performance refinements are made as a trade-off between these three factors depending on use requirements and risk appetite.

Using our example of Bloom filters for PPRL, the main types of disclosure attack that can be made are:

- Dictionary attack: If an attacker knows the hash function, they can encrypt a large set of quasi-identifiers until matching masked values are found.
- Frequency attack: If a relevant set of population quasi-identifiers is known, frequency distributions can then be analysed to determine likely corresponding *q*-gram mappings (even without knowing the hash function).
- Cryptanalysis attack: This can occur where patterns in bit distributions allow an adversary to learn the characteristics of hash functions that are used (e.g. for common names) to map record values into a Bloom filter.
- Composition attack: With auxiliary information (background knowledge) about the individual datasets and/or certain records in the datasets, a composition attack can be successful by using the information in combination to learn sensitive values of certain records.
- Collusion: Where a TTP is used for data linkage, there is scope for collusion between one data provider and the TTP to uncover information on the other database owner's data.

It must be noted, however, that these attacks would require effort and motivated action. They would not occur through poor data governance practice where data are accidentally released or naively misused, as could be the case with PID-based data linkage unless a robust suite of controls is in place. Even so, there have been successful attacks and further developments (and work continues) to harden PPRL methods to enhance their security. But it should always be borne in mind that disclosure risk is unlikely to ever be totally eradicated.

5. Linkage methods in practice

Among the existing data linkage enterprises, there is a variety of data linkage and operational models. We set out illustrations from Australia, Canada and the UK to show how data linkage works in practice. We have not attempted to be exhaustive, but have chosen a variety of models as illustrations[3].

[3] We have, therefore, not included all data linkage systems in the UK, Canada and Australia.

5.1 Australian Federal and State-level Linkage

Australia has a long history of data linkage enterprise, beginning in the 1970s. The Population Health Research Network (PHRN) and its Centre for Data Linkage (CDL) were established in 2009 to provide researchers with access to an array of state and federal health and non-health administrative datasets [27]. Linkage of federal-level ('national') datasets is carried out by the CDL, whereas state-level linkage is conducted by data linkage units within the specific jurisdictions.

There is no centralised data repository at the CDL and data providers only release data on a project-by-project basis. The CDL uses demographic data to create and hold an index of linkage keys but does not hold content data (clinical details, events, etc.). CDL uses PRL without a gold standard reference dataset to create the index of keys. Similarly, state-level linkage units such as the Centre for Health Record Linkage (CHeReL), which operates in the Australian Capital Territory and New South Wales, maintain state-level indices of links. CHeReL uses a combination of DRL and PRL methods building on well-established methods in several Australian states [28]. Data providers send approved de-identified datasets to researchers, and the CDL provides them with project-specific keys to enable linkage and simultaneously disenable collusion and linkage to other datasets that may be held as part of another project. In contrast to the CDL, state-level units may also hold de-identified versions of complete datasets and enable access to them for research[4]. Where this occurs, there is a strict separation between the linkage and research functions.

State-level units also use differing data linkage approaches: for example, the Western Australia Data Linkage Branch (WADLB) is moving to a cluster-based data linkage system [29, 30]. This new system is referred to as DLS3 (Data Linkage System 3) and has been designed in-house to replace the legacy linkage system which used proprietary file-based linkage software (FLS). DLS3 has been designed to overcome the limitations in the FLS that became evident as the scope of data linkage in the WADLB grew. For example, the FLS did not retain information about matching decisions, it was limited to a set order, it required onerous clerical review, and it could not concurrently consider more than two records for potential matching. DLS3 is being implemented using a phased replacement of the legacy system to maintain compatibility in the interim [30].

5.2 Population Data BC, Canada

Population Data BC (PopData) operates data linkage and holds a repository with a breadth of administrative data in British Columbia [31, 32]. PopData is authorised to hold identifiable datasets and to function as a TTP in carrying out data matching and linkage in order to provide access to de-identified data within a secure research environment (SRE). It does not have a research programme of its own, but enables research by other parties subject to approvals. PopData operates within a strictly-controlled, high-security environment with zones to enact separation principles and to control access. Data linkage is performed using a combination of DRL and PRL techniques. Linkage keys are transformed into project-level identifiers to preclude linkage to other datasets.

[4] Some Australian data enterprises operate a secure research environment +/- a model of external data release to researchers.

Researchers access their approved linked datasets within the SRE. This takes place via a fire-walled virtual private network (VPN) with two-factor authentication of data users using a Yubi-Key. Researchers are prevented from downloading identifiable or row-level data. Instead, information for intended release is managed via a transfer programme to enable PopData to keep an audit trail.

5.3 Secure Anonymised Information Linkage Databank, Wales

The Secure Anonymised Information Linkage Databank (SAIL) was established in 2007 as a repository of linkable de-identified datasets to be made available for research in anonymised form [33]. SAIL holds many health and non-health administrative datasets such as: GP, hospital (in-patient and out-patient), screening services, ONS births and deaths, education, housing, and fire & rescue services datasets. SAIL does not handle PID, but makes use of a TTP that carries out the matching process and the creation of a unique, consistent identifier for each person represented in the data. Linkage includes both DRL and PRL techniques in sequence, beginning with exact matching. A strict separation principle is operated where a data provider divides their data into two components: demographic and payload/content. The demographic data are sent to the TTP and the content data to SAIL. Recombination of de-identified datasets is made at SAIL and the unique key allows linkage across datasets.

SAIL enables data access within a SRE subject to approvals and a suite of technical, physical and procedural controls, including two-factor authentication of data users using a Yubi-Key, and a fire-walled virtual private network (VPN). Researchers cannot remove or alter the underlying data and requests to export analysis products are scrutinised and managed by a data guardian.

5.4 Unpacking data linkage at SAIL

The data linkage process in place for SAIL (for both health and non-health administrative data) begins with a fully identifiable dataset at the data provider setting. Having divided the data into demographic and content, the provider sends the demographic data to the TTP and the content data to SAIL. This strictly applied separation principle is key to ensuring privacy and confidentiality in the data linkage process, since it is based on PID.

The TTP in this case is the NHS Wales Informatics Service (NWIS): a national organisation providing statistics and IT solutions to the NHS in Wales. The TTP uses the NHS number (where present), plus first name, surname, date of birth, gender and postcode, as the linkage variables. Since the datasets can be large, blocking is used to reduce matching complexity. Beginning with exact matching and progressing to other DRL and PRL techniques, it matches the demographic data against an administrative register referred to as the Welsh Demographic Service (WDS.) The WDS acts as the gold standard since it is a maintained database of everyone in Wales who has engaged with the NHS, and therefore approximates a demographic database.

If an exact match is not found, the algorithm progresses to techniques such as Soundex and lexicon matching. Soundexing relies on converting a variable (such as surname) to a short sequence e.g. J520, S530, based on the first letter and the consonants present in the name. In this way it allows variability and matching based on ‘sound’ similarities, such Smith, Smyth and Smythe which have the same (S530)

Soundex code. Lexicon matching allows for variations in names via the use of diminutives, such as Robert to Bobby or Bob. The rules and thresholds have been optimised for use with health and non-health administrative data arising from Wales. However, it is worth noting that Soundexing and Lexicon matching can be less reliable and do not perform as well on non-Anglo Saxon names. They would need to be optimised for the population of study.

Each match is assigned a probability, and these are based on likelihood ratios calculated using a Bayesian approach of prior and posterior odds.

The posterior odds are calculated as: Posterior odds = prior odds x likelihood ratio

The likelihood ratios are calculated as follows:

Firstly, where the demographic variables match (e.g. on surname), the agreement weight is given by:

$$p(\text{match}|\text{records relate to the same person}) = m$$
$$p(\text{match}|\text{records relate to a different person}) = u$$

$$\text{Agreement weight} = m/u$$

And where the demographic variables do not match, the disagreement weight is given by:

$$p(\text{non-match}|\text{records relate to the same person}) = 1-m$$
$$p(\text{non-match}|\text{records relate to a different person}) = 1-u$$

$$\text{Disagreement weight} = (1-m)/(1-u)$$

Generally, the likelihood ratio is calculated for each record pair in logarithms, more specifically log (base 2), as the sum of the logged agreement & disagreement weights.

Assigning match probabilities enables decisions on the acceptable cut-off to be made depending on the use purpose, making the balance between high specificity or high sensitivity. Once matching is complete, the TTP replaces the PID with an identifier based on an encryption of the NHS number. This is referred to as the Anonymous Linking Field (ALF) and it is unique to each person. This is not limited to datasets arising from healthcare and can also be applied to those without NHS numbers. NWIS holds a record of NHS numbers in the WDS and by using the matching variables (first name, surname, date of birth, gender and postcode) can, therefore, assign an NHS number to records where it is absent [34].

The ALFs and minimal demographic details (week of birth, gender and Lower Layer Super Output Area of residence) are sent to SAIL for recombination with the content data. This uses a simple record number maintained in both components when the data were divided by the provider. SAIL re-encrypts the ALF to ensure that no one at SAIL or the TTP can reverse the process leading to identities. The double encrypted ALF is used to link other datasets at SAIL processed in this way. Data are linked when

required for research studies and the ALF can be further encrypted at project level as an additional safeguard.

All the examples illustrated in this section use a persisted unique identifier for each person (or other entity) represented in a dataset. They all use, and have honed, their data linkage methods using a combination of DRL and PRL. In addition, the Australian CDL has carried out significant work on the development of PPRL methods based on Bloom filters [35, 36, 37]. Work is ongoing to operationalise them at the Australian CDL, in PopData and for use with SAIL.

6. Discussion

This article illustrates the principles of record linkage, and emphasises the importance of privacy and confidentiality at all stages of the process. It is important to note that some measures of disclosure risk are specific to data linkage, but the process needs to sit within an overarching data governance framework to protect the data, as for any work using person-based data. The data linkage process is not special in this respect.

In accordance with statistical disclosure control (SDC) theory [17], if an individual can be re-identified, or if sufficient information can be derived to attribute personal information to an individual, there is a privacy risk. Risks may be classified as record-level or attribute-level disclosure, and they are usually assessed by conducting attack tests and evaluations[5]. But as shown in the five safes framework, safety is a measure, not a hard-and-fast quantifiable state [41]. As we have strongly emphasised, there must be controls around the data as well as within the data, since there are limits to the measures that can be applied to data if their utility is to be retained. Many countries engage in large-scale linkage for population census assessment [42], and even when using privacy-preserving methods, risks have been uncovered leading to the recommendation for protection of the data, through a suite of physical, security and procedural controls upon the data environment [43].

As well as the remaining privacy risks, PPRL techniques have been shown to limit the utility and quality of any resulting linked datasets. This is a significant issue and one which must inevitably be factored into choices of data linkage approach. A three-year study to inform the work of NHS Digital resulted in PPRL not being recommended for use above PID-based methods. This was largely because of the potential of currently available PPRL methods to undermine linkage accuracy and hence data utility. It was proposed that a more rational solution would be to avoid anonymisation at source, and to improve linkage accuracy by having regional and national trusted linkage environments with expertise in using sophisticated data linkage methods to improve and evaluate linkage accuracy [44].

Another limiting factor that needs to be considered in PPRL is the burden that data pre-processing can place on the data provider: both in terms of effort and expertise. This can be onerous and challenging, but in some cases PPRL might be the only way providers will share their data. In such cases, a Catch-22 may apply, where the use of PPRL circumvents unwillingness to share data, but the burden precludes the process from actually taking place in a timely way, if at all. This is an issue that needs to be carefully navigated if it is to be pursued whilst endeavouring to respect privacy and

[5] We refer the reader to work of Elliott [38], El-Emam [39,40] and the chapter on SDC for further information.

confidentiality. This will necessarily include the nature of the data provider, datasets and proposed data uses.

As we have shown, applying more and more controls to the data will not fully address the problem of residual privacy risks. There is a wealth of research that shows that putatively anonymised datasets can be used to re-identify individuals [23], and it is unlikely that the privacy risk in any meaningful dataset can be zero. This is not a trivial problem, nor one that can be solved by 'balancing' privacy and utility, since these factors are not linearly related, but are more of a scatter, since they vary with use case and questions to be addressed by means of the data.

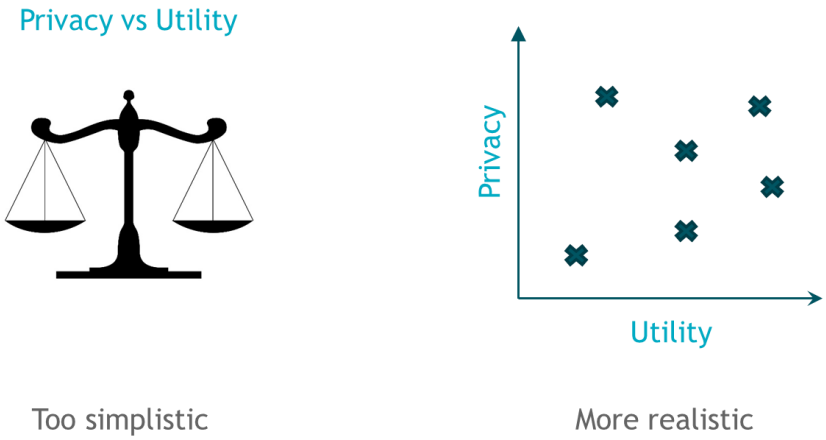


Figure 2: Relating privacy and utility

There comes a point where additional control measures do not add significant safeguards, but do reduce data value. This has been termed 'privacy-protectionism' [45]. It arises in the laudable effort to minimise risks, but what is needed is a recognition of the place of data controls as part of a data governance framework with privacy-by-design. This is true of the data linkage process as well as when using the resulting linked data for research.

6.1 Privacy-by-design case study

The data linkage process in operation for SAIL was described in Section 5.4. Here we elaborate to show how this fits into a privacy-by-design data governance framework, with a particular focus on privacy and confidentiality [46]. We have chosen SAIL because it is a long-established UK example using multiple population-level datasets from health and non-health sources. From this, we will draw out a set of guiding principles to aid decision makers in selecting a data linkage approach.

Many of the datasets that come to SAIL originate in operational IT systems, often in clinical settings. The status, type and level of inter-operability inherent in such systems varies widely, as it does for other administrative data sources. In considering how to approach data linkage, we needed to consider data provider readiness in terms of their compute capacity, the burden that would be imposed on them, and their due diligence processes to share data in line with data protection legislation and regulations.

As we've noted, the quality of administrative data is inconsistent, with inaccuracies and missing values. As such, they are generally not of sufficient standard to use without

verification. For this reason, and to avoid the complications that would be incurred if we needed to handle identifiable data at SAIL, we established a data linkage model using a TTP for identity verification, data matching and anonymisation. Secure data transfer protocols are made use of at all stages, in moving the demographic data to the TTP, the content data to SAIL, and the ALFs and minimal demographics to SAIL.

The data matching method is based on PID and identifiable linkage variables. As such, privacy and confidentiality are assured by a combination of physical, technical and procedural controls as per the TTP data governance framework. By its essence, this is not a statistically quantifiable measure; it relies on industry standards and ultimately integrity in data processing. In the case of NWIS acting as a TTP, demographic data arising from the NHS is processed as a part of its core business in accordance with the EU GDPR and UK DPA. For administrative data from other sources, the provisions of the UK DEA allow data processing for matching and anonymisation by a TTP.

SAIL is underpinned by an infrastructure referred to as the UK Secure Research Platform (UK SeRP). This forms the basis of the SRE via which data are made accessible for *bona fide* research purposes. UK SeRP is an ISO27001 approved independent technology and analysis platform. It can be configured via customisable specifications to allows multiple, complex datasets to be managed, analysed and shared in line with the data owners' Information Governance framework, subject to legislative and regulatory requirements. Data users access UK SeRP remotely via an internet portal subject to permissions and approvals. The SRE for SAIL is referred to as the SAIL Gateway. UK SeRP can be made available for use by other organisations and can be augmented by the implementation of a National Research Data Appliance (NRDA) at data provider organisations. The NRDA is a set of modular data concentrator technologies, that includes automated matching, anonymisation, linkage, data management, metadata capture and data quality assessment based on developments at Curtin University, Australia. These technologies enable organisations who do not have their own data linkage and management infrastructures to share, link and use data in a secure environment[6].

SAIL does not rely solely on technical measures applied to the data or even within the SRE as a whole, but augments these with physical and procedural controls. We will not describe these in detail, other than to say that they include access-controlled zoned areas, and a set of agreements, policies and operating procedures, plus data user training and accreditation. All these measures are necessary simply because although SAIL does not handle identifiable data, there are still inherent risks to consider. In order to mitigate these, we have to go beyond measures applied to the data if utility for research is to be retained. The totality of privacy and confidentiality risk in a SRE such as this cannot be quantified as a coefficient. Instead the risk status of the system is measured by industry best practice and certification: SAIL itself is ISO 27001 certified. Taking into account the environment in which the data are managed, we operate in a position where totality of the data is functionally anonymised [38]. Where there is any doubt that such data could be considered de-identified but not fully anonymised, we rely on the provisions of the GDPR for the work of data management and preparation for research being in the public interest.

Once the data arrive at SAIL, the ALF is subjected to a second level of encryption on a wholescale basis. This ensures that no one at SAIL or at the TTP can reverse the

[6] Further information on UK SeRP and the NRDA is available from the authors.

anonymization algorithm to return to individual identities. Further controls are applied at project level. These are determined based on the use case for the data, and the types of data, being requested. SAIL established an independent Information Governance Review Panel (IGRP) to advise on the suitability of data use proposals and their likely disclosure risks, over and above assessment by skilled SAIL analysts. The IGRP carries out a subjective assessment, based on experience. Members include data guardians of major datasets (such as hospital episodes), ethics professionals and members of the general public from a Consumer Panel to guide on social acceptability in data use[7]. Where risks are perceived, feedback is given to the researcher and, with guidance from a SAIL analyst, the proposal is modified to mitigate risks without adversely affecting data utility. This may involve applying record and attribute level control measures to the data. SAIL uses masking, suppression and aggregation, but does not use data perturbative methods.

Following analysis within the SRE, the researcher needs to make a request to a SAIL data guardian in order for results to be released externally. This involves a process of manual scrutiny by the data guardian to ensure the products of analysis are appropriate for export and do not introduce risks to privacy and confidentiality. Row-level data are not allowed to leave the SRE, unless participant consent and all relevant permissions and regulatory approvals are in place. Instead, we release statistical coefficients, tables with no cell counts below a limit, charts, graphs and other summary measures.

SAIL periodically carries out in-house internal audits, including testing security protocols and ensuring controls are of the highest standard across the system. This is in addition to being externally verified via ISO 27001. Since individuals with access to restricted data may choose to go beyond their contract and attempt to piece together information leading to uncovering identities, SAIL tracks queries written by researchers, can suspend data access and initiate penalties for data misuse. Security protocols can only go so far, and there is a point at which there has to be an element of trust. This is not a concept unique to working with administrative data in a SRE, but is common wherever researchers have access to meaningful data about people. But the need to rely on trust should always be minimised as far as possible. At the end of a research study, data prepared for the project are retained by SAIL, and they can, if necessary, be archived when the data provider and governance approval conditions do not permit re-use.

Altogether, and from start to finish along the pathway from data provider to project completion, the suite of technical, physical and procedural controls in place at the SAIL databank comprise a privacy-by-design model to ensure privacy and confidentiality have paramount importance.

6.2 Selecting a data linkage approach

It is evident that the choice of data linkage approach cannot be made in isolation. It will depend on a number of inter-related factors which must be considered in combination, since the involvement of other parties and the nature of the data environment(s) are key to feasibility and, crucially, to privacy and confidentiality in data linkage [47].

Data provenance is an important consideration since it will dictate the type of linkage method that can be used. An essential first question is whether the PID are allowed to

[7] SAIL has a Consumer Panel to provide a public perspective across the scope of our work.

move. If the answer appears to be no, then we need to find out the reasons why, and determine if this can be remedied. Sometimes this is a matter of due diligence within an organisation concerning internal policies, or it might be that there is genuinely no lawful provision. Some government data can move under the provisions of the DEA, other datasets are subject to the creation of (temporary) legal gateways for particular purposes, or further limitations.

It will always be worth seeking to allow the PID to move since this opens up the opportunity to use a TTP for identity verification, matching and the creation of a consistent linkage key via PID-based methods. If it is not possible for the PID to move, it may be possible for the TTP functions to be carried out in-house, depending on feasibility and providing there can be agreement on an algorithm to create a shared key for onward linkage. This, however, is prone to risk and would require additional security measures, since at face value a common key would enable an organisation to uncover the identities in another organisation's dataset. Where PID cannot move, the only option might be to use PPRL methods to anonymise at source.

IT systems in administrative organisations are frequently under pressure, not compatible intra- or inter-organisationally, and often not at the cutting edge in compute capacity or performance. This situation, along with inconsistent and sometimes low data quality, will influence the choice of data linkage approach. The demand on the data provider can be minimal or significant. Considering the operational burdens incumbent on administrative services in general, as well as the lack of experience in data linkage, it would be wise to minimise their active role in the process wherever possible. This, and the likely insufficiencies in IT systems, would point preferentially towards transferring PID to a TTP for data linkage, rather than doing this in-house. Or it may be possible to implement a PPRL hashing function at the data provider, providing the pre-processing step can be accomplished and as long as the compute demand was limited. Even so, the next steps (blocking, classification, evaluation) would need to rely on a TTP in a three-party protocol.

The nature of the datasets and the main intended use cases of linked data will influence the selection of linkage method. Data quality in terms of accuracy and completeness will need to be factored into the decision, taking into account the degrees of tolerance in data use. To some extent this is a balancing act on whether it is most important to have high specificity (and incur relatively more false negatives) or high sensitivity (and incur relatively more false positives). But it is only really an issue where data quality is poor, not where data quality is high or can be pre-processed via identity verification, effectually increasing the quality of the linkage variables. Again, this comes back to whether there is a gold standard dataset or if ground truth is accessible. In practice this issue is addressed by algorithms carrying out linkage steps in sequence, and providing measures of matched pair likelihood, albeit via single or combinations of DRL and PRL[8].

Intended data use cases may be well-defined or more variable, and the nature of the decisions to be made from the outputs is an important deciding factor on whether to lean towards specificity or sensitivity. For the majority of administrative data uses it is usually advantageous to prioritise the former for greater accuracy, but with reference to distributions across the dataset to avoid selection bias.

[8] PPRL methods are not excluded here since DRL and PRL refer to linkage being deterministic or probabilistic, not specifically the use of PID.

The available infrastructure for data linkage, management and access, along with the involvement of other parties will play a major role in determining which approaches can be used. In terms of other parties, we refer to the availability of a suitable TTP and other data providers being able to engage in data linkage and sharing. This will depend on legislation and accreditation as well as on due diligence processes to ensure privacy and confidentiality are respected. The use of a TTP is a valuable (sometimes invaluable) component in many data linkage methods. But there are alternatives. Some PID-based or PPRL mechanisms can be conducted without a TTP and are referred to as two-party protocols. The NRDA which operates as part of the UK SeRP includes identity verification and matching in the automated modular system by incorporating the gold standard dataset. But in effect, it is functioning as the TTP: a reliable go-between but without the involvement of a third-party organisation. As we have observed, some TTP functions can be carried out by data providers in-house, but privacy and confidentiality need to be considered carefully to avoid inappropriate disclosure to outside agencies.

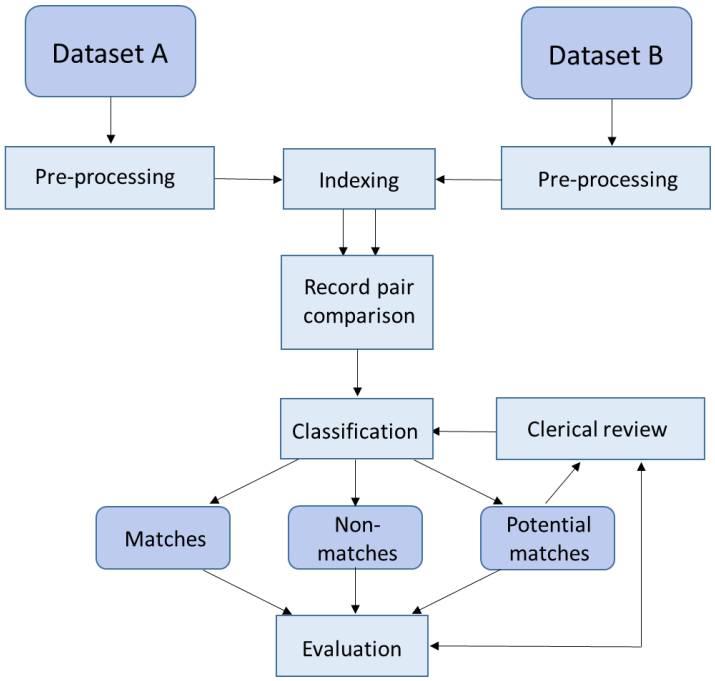
This leads us to the data environment itself and whether linked data are to be released externally to researchers or accessed within a managed SRE. We illustrated these options briefly via case studies (Section 5). In the former example, there is a requirement for a co-ordinating department or organisation to retain an index of linkage keys to enable linkage of datasets provided externally to researchers. In the latter, data are usually held in a central repository[9] and accessed for analysis in a virtual environment.

The privacy and confidentiality considerations hinge on whether it is acceptable for linked data to be exported or if they are required to remain within a specified setting. Where data are exported, control measures must be applied to mitigate re-identification risk and to stipulate researcher behaviour by means of procedural controls and approvals. If data are held in a repository, then all necessary physical, technical and procedural measures must be in place to protect the data. Neither of these are trivial exercises and it is necessary to evaluate their implementation feasibility including factors such as compute capacity, technical expertise, engagement with other parties and costs/resource implications.

Ultimately, there are decisions to be made on the many elements directly and indirectly involved in the process of data linkage whatever overall model is chosen. This will encapsulate what can be done, with which datasets, by whom and for which purposes. At all stages this must be carried out to ensure privacy and confidentiality, whilst simultaneously maximising data utility. This necessarily includes controls around the data as well as applied to the data. It is not something that can be fully captured in statistical coefficients, but relies on both quantitative and qualitative measures.

Having considered these wider factors, we can now put the generic steps in the data linkage process (Section 3.1) into context, so that privacy and confidentiality are built into selecting an approach[10].

[9] It is also possible to access distributed data held in separate repositories via a SRE.
[10] Diagram modified after Christen [11]



Generic stages: (i) pre-processing; (ii) indexing/blocking; (iii) record pair comparison; (iv) classification +/- clerical review; and (v) evaluation +/- clerical review.

Authorisations and permissions for the use of data in identifiable form at any stage.

Status of security protocols and data governance frameworks of all engaged parties, including data providers, TTPs, data linkage enterprises and linked dataset destinations.

Physical, technical and procedural disclosure controls applied to the data and around the data.

Quantitative (statistics-based) and qualitative (accreditation-based) measures of privacy and confidentiality.

Figure 3: Generic steps in data linkage

7. Recommendations and options

Having drawn out these observations, we propose a high-level set of recommended questions to be considered in making decisions on options for data linkage, so that privacy and confidentiality are built in.

'LINKAGE'
Guideline questions for data linkage

L: Legislative position – Which are the key legislative instruments and lawful provisions for data processing, and what are the due diligence processes to be followed?

I: Information systems – What is the status and readiness of data provider IT systems to supply data, and what additional demands are required to engage in data linkage?

N: Nature of datasets – How do the datasets measure up in terms of data quality, completeness and update frequency?

K: Knowledge-base – What level of expertise is held by data custodians and what would be the demand upon them to engage in data linkage?

A: Aims and purposes – Allowing for flexibility, what are the main anticipated purposes for the linked data?

G: Ground truth – Is there a gold standard reference dataset, or is possible to access ground truth via clerical review, for identity verification and matching?

E: Environment – Which data management and access models are permissible and feasible considering the range of relevant factors?

The answers to these questions need to be taken into account when deciding on a data linkage approach. These feasibility and practicality issues are to be navigated in combination via an iterative process to assess their relative contributions to the overall model. The desired result should be an effective data linkage approach that sits within a robust data governance framework to ensure privacy and confidentiality.

We expand on each of these questions to enable data linkage options to be considered.

L: Legislative position –
The first question concerns the pertaining legislative position and whether the PID component of the dataset can legally move from its source to be processed for data linkage. It is understood that for some government department datasets, PID are not permitted to move, and that in other cases, there is a position of risk aversion such that organisational policy makes data movement difficult. If there is no legal gateway, then it might be that the only feasible option is PPRL. But for all other cases, we would recommend that every effort is made to support data providers in their due diligence processes so that the PID can be moved to an appropriate agent for de-identification, matching and DRL/PRL.

I: Information systems –
The capacity of data provider information systems must be considered since the legislative position and due diligence will to a large extent determine the demand to be placed upon them. If only PPRL is allowed, then data provider systems will need to carry out pre-processing to prepare the data for hashing, and then apply the hash functions. This may create quite a demand on administrative information systems not designed with such processing in mind. The requirements upon data provider systems in engaging in PID-based record linkage are generally less onerous.

N: Nature of datasets –
Data quality, completeness and update frequency also need to influence the choice of linkage approach. Administrative data are subject to many inconsistencies, and these will affect data linkage efficacy unless the data can be pre-processed effectively, and preferably matched against a reliable reference dataset or subjected to clerical review. Again, this is far more straightforward to enact with PID-based than with PPRL methods, if sub-optimal results are to be avoided.

K: Knowledge-base –
Engaging in data linkage requires some action by data providers, whichever method is used. In the PID-based case studies (in Section 5), we showed that this can vary depending on the operating model in place. In most of the Australian models, data are called from providers by a data linkage unit by means of an index of links. PopData are authorised to hold the full identifiable datasets in their repository. SAIL operates a straightforward separation principle where data providers divide their datasets into two components, with the demographic data going to a TTP and the content data going directly to SAIL. Each of these methods is functioning and enabling data linkage to take place without over-burdening data providers. PPRL generally requires additional expertise and effort from data providers, unless personnel from the data receiving organisation come in and run the processes (assuming available compute capacity

exists). But this, of course, raises privacy and confidentiality issues, since a third party would need a level of access to the PID in order to do this.

A: Aims and purposes –
The proposed uses of the data will play a part in the choice of data linkage method. This is partly to do with the trade-off between specificity and sensitivity, in terms of the relative importance of having a smaller set of highly-accurate linked data, or a larger set of data with possibly greater variability in match quality. This will depend on the types of decisions to be made from the linked data and the implications of inaccuracy vs missingness. In practical terms, data are unlikely to be destined for one narrow purpose, and so it is essential that data linkage is optimised to meet requirements. Again, this is less complex using PID-based methods than PPRL. Data quality is crucial as an ethical issue if reliable findings are to be produced for decision-making.

G: Ground truth –
Whether there is a reliable reference dataset or the availability to carry out clerical review for record matching will play an important part in the choice of data linkage method. These options are important to ensure data quality and maximise the rates of high quality linkage. There are privacy and confidentiality issues to take into account in terms of who is allowed to compare records to evaluate linkage steps. In some cases, such as the method used by NWIS (Section 5.4), the process is automated and not viewed by individuals, even though NWIS staff are authorised to view the PID as required. However, clerical review against ground truth is sometimes done manually using a sub-set of identifiable records, which means that appropriate authorisations must be in place for personnel carrying out the process.

E: Environment –
Linked data are managed and accessed under different operating models and data governance regimes. Which of those are permissible and feasible will depend on jurisdictional legislative and regulatory frameworks, but may also depend on organisational policies. Some data providers may or may not agree to linked data (incorporating their datasets) being released externally. Alternatively, this may be preferable, provided that all relevant approvals are in place in the absence of a suitable repository. As described in the case studies, Australian data linkage centres commonly follow the former model, whereas those of PopData BC and SAIL follow the latter. In each of these cases, data are prepared for researchers with limitations on successive linkage by curtailing the data and the use of project-specific linkage keys, with a requirement on researchers to behave with integrity.

The principles, recommendations and options we draw from this article are highly relevant to the UK Government Statistical Service and resonate with those identified in the recent Office for Statistics Regulation report, namely: value, quality and trustworthiness. Many important societal questions cannot be answered without using linked data. A greater willingness to share and link data is essential to enable official statistics to add value by addressing society's pressing questions. Data quality is paramount to avoid misleading information and ensure reliable conclusions to be drawn. Data custodians must demonstrate trustworthiness in safeguarding public data throughout data sharing and linkage processes [48].

In summary, taking into account the current state-of-play in the various record linkage approaches, our recommendation is to endeavour to use PID-based linkage wherever

possible, with an optimised combination of DRL and PRL, and to surround the data with privacy-by-design. This should be coupled with a clear awareness of the information needed at each step of the linkage pathway to improve the transparency, reproducibility, and accuracy of linkage processes, and ultimately the validity of data analyses and the interpretation of results [49]. It may well be the case that PPRL approaches will develop extensively in the near future to the point where they are as reliable as PID-based methods. But this does not depend only on advances in the approaches themselves, but on the accompanying factors required for their effective implementation, including the implications for data providers.

We recommend effort being placed into working with data providers to draw their attention to the problems of pushing for anonymisation at source where there is no absence of a legal gateway for moving data. This could be framed in terms of the risks to data quality, the fact that PPRL does not eradicate disclosure risks, and the benefits of tried and tested PID-based linkage methods operating within a privacy-by-design framework.

8. Conclusion

This article has focused on privacy and confidentiality in data linkage. It set out the basic principles of linking data and the legislative context for data protection within the UK. From this, it outlined the main data linkage methods, with their practicalities and evaluation methods, and used case studies as illustrations of methods in practice. There is a range of inter-related factors that need to be taken into account in selecting a data linkage approach and we propose a set of high-level questions and options to aid this process. There are challenges to overcome to link and use administrative data lawfully, safely and in line with social-acceptability. It is simply not enough to rely on lawful provision, since legality is not the same as social licence, and due public engagement is essential [50]. Consequently, it's easy to find reasons for not using data, but the *status quo* is a dangerous position.

As well as the paramount importance of using data safely and effectively for public benefit, an international case study has demonstrated the serious harms to individuals and society when data are not used, i.e. the non-use of data. Taken globally, this phenomenon results in the loss of hundreds of thousands of lives and $billions in societal financial burdens [51]. We recommend a clear recognition of the limitations and risks in relying solely on control measures applied to data. There is a real need to surround data in a robust, proportionate, privacy-by-design data governance model, to maintain and maximise data utility and simultaneously protect privacy and confidentiality in data linkage operations.

9. Annotated references

1) Christen, P. (2012). Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Part I(1) Introduction. Springer-Verlag, Berlin.

Short history of data matching. Aims and challenges in data matching: privacy and confidentiality; lack of unique entity identifier; data quality; computation complexity; lack of training data/true match status. Data integration and link analysis. Example application areas: national census; health; national security.

2) Harron, K., Goldstein, H. and Dibben, C. (2016). Introduction, In: Harron K, Goldstein H and Dibben C, editors. Methodological Developments in Data Linkage, Wiley, UK

State of play in data linkage. Data preparation and data linkage methods. Defining deterministic and probabilistic linkage. Linkage error and its implications. The future of data linkage.

3) Doidge, J.C. and Harron, K. (2018). Demystifying probabilistic linkage: Common myths and misconceptions. IJPDS, 3:1, DOI https://doi.org/10.23889/ijpds.v3i1.410

'Many of the distinctions made between probabilistic and deterministic linkage are misleading. While these two approaches to record linkage operate in different ways and can produce different outputs, the distinctions between them are more a result of how they are implemented than because of any intrinsic differences.' 'We aim to explain how the outputs of either approach can be tailored to suit the intended application.'

4) Boyd, J.H, Randall, S.M. and Ferrante, A.M. (2015). Application of privacy-preserving techniques in operational record linkage centres. In: Gkoulalas-Divanis A, Loukides G, editors. Medical Data Privacy Handbook. Springer International Publishing Switzerland.

A review of current practice, processes and developments for maintaining security and privacy as applied in existing record linkage centres. Record linkage infrastructures. Models for role separation and data flows. Privacy challenges and data governance. Evaluation and requirements for an effective privacy-preserving record linkage protocol.

5) Harron, K., Dibben, C., Boyd, J. *et al.* (2017). Challenges in administrative data linkage for research, Big Data & Society July–December 2017: 1–12. doi: 10.1177/2053951717745678

Providing 'an overview of challenges in linking administrative data for research. We aim to increase understanding of the implications of (i) the data linkage environment and privacy preservation; (ii) the linkage process itself (including data preparation, and deterministic and probabilistic linkage methods) and (iii) linkage quality and potential bias in linked data.'

6) UK Government (2018) Data Protection Act http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

7) EU (2016) General Data Protection Regulation https://gdpr-info.eu/

8) UK Government (2017) Digital Economy Act http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted

9) UK Government (2007) Statistics and Registration Service Act https://www.legislation.gov.uk/ukpga/2007/18/contents

10) Office for Statistics Regulation (2018) Code of practice for Statistics https://www.statisticsauthority.gov.uk/wp-content/uploads/2018/02/Code-of-Practice-for-Statistics.pdf

11) Christen, P. (2012). Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Part I(2) The data matching process. Springer-Verlag, Berlin.

Overview of the data matching process in 5 main steps: pre-processing, indexing/blocking, record-pair comparison, classification and evaluation.

12) Christen, P. (2012). Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Part II (4) Indexing. Springer-Verlag, Berlin.

The aim of indexing in data matching is to reduce the number of record pairs that are compared in detail as much as possible, by removing that are unlikely to correspond to true matches. At the same time, all record pairs that possibly represent true matches must be kept for detailed comparison. Without indexing, the matching of two datasets *m* and *n* would result in *m* x *n* record pairs for comparison, which would not be feasible for large datasets. Indexing is the process to filter records into blocks based on selected criteria so that similar records are moved closely together.

13) Winkler, W.E. (2016). Probabilistic linkage, In: Harron K, Goldstein H and Dibben C, editors. Methodological Developments in Data Linkage, Wiley, UK

Introducing and defining PRL. An overview of methods: Felligi-Sunter model; learning parameters; Expectation-Maximisation algorithm; linkage +/- training data; alternative machine learning models; string comparators; blocking; indexing; parsing; linkage error.

14) Bonomi, L., Fan, L. and Xiong, L. (2015). Overview of privacy preserving mechanisms for record linkage, In: Gkoulalas-Divanis A, Loukides G, editors. Medical Data Privacy Handbook. Springer International Publishing Switzerland.

A broad review of recent works in PPRL from a privacy-centric perspective, summarising a comprehensive framework of the PPRL process (including data transformation, blocking and matching) and describing the privacy assurance techniques for each step. The review also categorises existing PPRL works with a taxonomy that includes privacy, scalability and linkage quality.

15) Vatsalan, D,, Christen, P. and Verykios, V.S. (2013). A taxonomy of privacy-preserving record linkage techniques. Information systems, 38:946-969

An overview and characterisation of privacy-preserving record linkage techniques.

16) Schnell, R. (2016). Privacy-preserving record linkage, In: Harron K, Goldstein H and Dibben C, editors. Methodological Developments in Data Linkage, Wiley, UK

Linking with and without personal identifiers. A description of the state of play and the most widely used PPRL techniques, with particular focus on Bloom filters. The use of trusted third parties. Challenges. Types of attack and privacy considerations.

17) Vatsalan, D., Christen, P. et al. (2014). An evaluation framework for privacy-preserving record linkage. Journal of privacy and confidentiality, 6(1):35-75.

An evaluation framework for privacy-preserving record linkage (PPRL) solutions that enables assessment and comparison of different solutions in terms of the three main properties of PPRL, which are scalability, linkage quality, and privacy.

18) Christen, P. (2012). Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Part III (8) Privacy aspects of data matching. Springer-Verlag, Berlin.

Privacy and confidentiality challenges. Data matching scenarios. Privacy-preserving techniques. Practical considerations for research.

19) Christen, P. (2018). Linking administrative databases: Recent developments and research challenges. ADR conference, Belfast, June 2018 http://users.cecs.anu.edu.au/~christen/publications/christen2018adrn.pdf

An introduction to data linkage. Challenges of linking administrative databases. Techniques for scalable data linkage. Automating the linkage process. Privacy aspects in data linkage. Evaluating linkage quality and completeness. Dealing with uncertainty in linked datasets. Towards an end-to-end linkage framework. Graph-based group linkage.

20) Farrow, J.M. (2016). Using graph databases to manage linked data. In: Harron K, Goldstein H and Dibben C, editors. Methodological Developments in Data Linkage, Wiley, UK

Focuses mainly on data management in graph databases, with some information on the relationship between data linkage and how data are contained.

21) Gollapalli, M. (2015). Literature Review of Attribute Level And Structure Level Data Linkage Techniques, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No. 5. http://aircconline.com/ijdkp/V5N5/5515ijdkp01.pdf

'We identify problems as four major research issues in Data Linkage: associated costs in pairwise matching, record matching overheads, semantic flow of information restrictions, and single order classification limitations. The purpose for this review is to establish a basic understanding of Data Linkage, and to discuss the background in the Data Linkage research domain. Particularly, we focus on the literature related to the recent advancements in Approximate Matching algorithms at Attribute Level and Structure Level. Their efficiency, functionality and limitations are critically analysed and open-ended problems have been exposed. '

22) Christen, P. (2012). Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Part II (6) Classification. Springer-Verlag, Berlin.

Classification of candidate record pairs using a variety of techniques.

23) Ohm, P. (2009). Broken promises of privacy: responding to the surprising failure of anonymization http://epic.org/privacy/reidentification/ohm_article.pdf

‘Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques for protecting the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated they can often “re-identify” or “de-anonymize” individuals hidden in anonymized data with astonishing ease. By understanding this research, we will realize we have made a mistake, laboured beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. We must respond to the surprising failure of anonymization, and this Article provides the tools to do so.’

24) Ferrante, A. and Boyd, J. (2012). A transparent and transportable methodology for evaluating data linkage software. JBI, 45(1):165-172

‘Although evaluations of DL software exist; most have been restricted to the comparison of two or three packages. Evaluations of a large number of packages are rare because of the time and resource burden placed on the evaluators and the need for a suitable “gold standard” evaluation dataset. In this paper we present an evaluation methodology that overcomes a number of these difficulties. Our approach involves the generation and use of representative synthetic data; the execution of a series of linkages using a pre-defined linkage strategy; and the use of standard linkage quality metrics to assess performance. The methodology is both transparent and transportable, producing genuinely comparable results. The methodology was used by the Centre for Data Linkage (CDL) at Curtin University in an evaluation of ten DL software packages. It is also being used to evaluate larger linkage systems (not just packages). The methodology provides a unique opportunity to benchmark the quality of linkages in different operational environments.’

25) Christen, P. (2012). Data Matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Part II (7) Evaluation of matching quality and complexity. Springer-Verlag, Berlin.

Methods for measuring matching complexity and quality.

26) Hand, D. and Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. Journal of Statistics and Computing, 28(3): 539-547. DOI: 10.1007/s11222-017-9746-6

‘Due to the generally high class imbalance in record linkage problems, standard accuracy or misclassification rate are not meaningful for assessing the quality of a set of linked records. Instead, precision and recall, as commonly used in information retrieval and machine learning, are used. These are often combined into the popular F-measure, which is the harmonic mean of precision and recall. We show that the F-measure can also be expressed as a weighted sum of precision and recall, with weights which depend on the linkage method being used. This reformulation reveals that the F-measure has a major conceptual weakness: the relative importance assigned to precision and recall should be an aspect of the problem and the researcher or user, but not of the particular linkage method being used. We suggest alternative measures which do not suffer from this fundamental flaw.’

27) Population Health Research Network http://www.phrn.org.au/

28) Centre for Health Record Linkage (CHeReL) http://www.cherel.org.au/

29) Western Australia Data Linkage Branch https://www.datalinkage-wa.org.au/

30) Eitelhuber TW (2018) Western Australia unveils its new data linkage system, IJPDS [in press]

31) Pop Data BC https://www.popdata.bc.ca/

32) Hertzman, C.P., Meagher, N. and McGrail, K.M. (2013). Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. J Am Med Inform Assoc. American Medical Informatics Association, 1;20(1):25–8, https://pdfs.semanticscholar.org/ddaf/02838caaa73d95b9dc813b59a601a3669be9.pdf

'Population Data BC (PopData) is an innovative leader in facilitating access to linked data for population health research. Researchers from academic institutions across Canada work with PopData to submit data access requests for projects involving linked administrative data, with or without their own researcher-collected data. PopData and its predecessor — the British Columbia Linked Health Database — have facilitated over 350 research projects analyzing a broad spectrum of population health issues. PopData embeds privacy in every aspect of its operations. This case study focuses on how implementing the Privacy by Design model protects privacy while supporting access to individual-level data for research in the public interest. It explores challenges presented by legislation, stewardship, and public perception and demonstrates how PopData achieves both operational efficiencies and due diligence.'

33) Ford, D.V., Jones, K.H. et al. (2009). The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Services Research 2009,* 9:157.

A set of objectives was identified to address the challenges and establish the Secure Anonymised Information Linkage (SAIL) system in accordance with Information Governance. These were to: 1) ensure data transportation is secure; 2) operate a reliable record matching technique to enable accurate record linkage across datasets; 3) anonymise and encrypt the data to prevent re-identification of individuals; 4) apply measures to address disclosure risk in data views created for researchers; 5) ensure data access is controlled and authorised; 6) establish methods for scrutinising proposals for data utilisation and approving output; and 7) gain external verification of compliance with Information Governance.

34) Lyons, R.A., Jones, K.H. et al (2009). The SAIL databank: linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making* 2009, 9:3.

The aim of this work was to develop and implement an accurate matching process to enable the assignment of a unique Anonymous Linking Field (ALF) to person-based records to make the databank ready for record-linkage research studies. An SQL-based matching algorithm was developed for this purpose. Firstly, the suitability of using a valid NHS number as the basis of a unique identifier was assessed. Secondly, the algorithm was applied in turn to match primary care, secondary care and social services datasets to the NHS Administrative Register (WDS), to assess the efficacy of this process, and the optimum matching technique.

35) Randall, S.M., Ferrante, A.M., Boyd, J.H., Bauer, J.K. and Semmens, J.B. (2013). Privacy-preserving record linkage on large real world datasets, Journal of Biomedical Informatics 50 (2014) 205–212

‘Record linkage typically involves the use of dedicated linkage units who are supplied with personally identifying information to determine individuals from within and across datasets. The personally identifying information supplied to linkage units is separated from clinical information prior to release by data custodians. While this substantially reduces the risk of disclosure of sensitive information, some residual risks still exist and remain a concern for some custodians. In this paper we trial a method of record linkage which reduces privacy risk still further on large real world administrative data. The method uses encrypted personal identifying information (bloom filters) in a probability-based linkage framework. The privacy preserving linkage method was tested on ten years of New South Wales (NSW) and Western Australian (WA) hospital admissions data, comprising in total over 26 million records. No difference in linkage quality was found when the results were compared to traditional probabilistic methods using full unencrypted personal identifiers. This presents as a possible means of reducing privacy risks related to record linkage in population level research studies. It is hoped that through adaptations of this method or similar privacy preserving methods, risks related to information disclosure can be reduced so that the benefits of linked research taking place can be fully realised.’

36) Boyd, J., Ferrante, A., Brown, A., Randall, S. and Semmens, J. (2017). Implementing privacy-preserving record linkage: welcome to the real world, IJPDS, Issue 1, Vol 1:134, Proceedings of the IPDLN Conference (August 2016). DOI https://doi.org/10.23889/ijpds.v1i1.153

‘PPRL techniques that operate on encrypted data have the potential for large-scale record linkage, performing both accurately and efficiently under experimental conditions. Our research has advanced the current state of PPRL with a framework for secure record linkage that can be implemented to improve and expand linkage service delivery while protecting an individual’s privacy. However, more research is required to supplement this technique with additional elements to ensure the end-to-end method is practical and can be incorporated into real-world models.’

37) Irvine, K., Smith, M., De Vos, R., Brown, A., Ferrante, A., Boyd, J. and Thackway, S. (2018). Real world performance of privacy preserving record linkage, IJPDS, Vol 3 No 4: 399. Proceedings of the IPDLN Conference (September 2018). DOI: https://doi.org/10.23889/ijpds.v3i4.990

‘Compared to the gold standard probabilistic linkage using full personal identifiers, the PPRL techniques produced quality metrics of precision, recall and F measure in excess of 0.90. When configured to leverage pre-existing links between emergency department, hospital and mortality data, quality metrics around 0.98-0.99 were achieved. Lower rates of linkage quality were associated with missing demographic information and some residual variation in linkage quality across practices was observed. PPRL using Bloom filters is a promising technique for achieving high quality linkage across primary and secondary care in Australia. Further evaluation will assess scalability and quality in Australia but international collaborations are encouraged to more rapidly develop the evidence base and tactical approaches to support real world implementations.’

38) Elliott, M., Mackey, E., O'Hara, K. and Tudor, C. (2016). The Anonymisation Decision-Making Framework, UKAN Publications, Manchester. http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

'We have learnt that we have to deploy effective anonymisation techniques and assess re-identification risk in context, recognising that there is a wide spectrum of personal identifiability and that different forms of identifier pose different privacy risks. This authoritative and accessible decision-making framework will help the information professional to anonymise personal data effectively. The framework forms an excellent companion piece to the ICO's code of practice. It is easy to say that anonymisation is impossible and that re-identification can always take place. It is just as easy to be complacent about the privacy risk posed by the availability of anonymised data. It is more difficult to evaluate risk realistically and in the round and to strike a publicly acceptable balance between access to information and personal privacy. The guidance in this framework will help information professionals to do that.' Functional anonymization asserts that a holistic, contextual approach should be used to determine anonymisation, taking into account the data environment, not just the status of the dataset. This includes the presence of other data, the agents accessing the data, the data governance model, and the infrastructure in place.

39) El-Emam, K., Jonker, E., Arbuckle, L. and Malin, B. (2011). A Systematic Review of Re-Identification Attacks on Health Data PLoS ONE 6(12): e28071. doi:10.1371/ journal.pone.0028071 https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0028071&type=printable

'Some recent articles in the medical, legal, and computer science literature have argued that de-identification methods do not provide sufficient protection because they are easy to reverse. Should this be the case, it would have significant and important implications on how health information is disclosed, including: (a) potentially limiting its availability for secondary purposes such as research, and (b) resulting in more identifiable health information being disclosed. Our objectives in this systematic review were to: (a) characterize known re-identification attacks on health data and contrast that to re-identification attacks on other kinds of data, (b) compute the overall proportion of records that have been correctly re-identified in these attacks, and (c) assess whether these demonstrate weaknesses in current de-identification methods.'

40) El-Emam, k. (2008). De-Identification Reduce Privacy Risks When Sharing Personally Identifiable Information. De-Identification Whitepaper 2 Privacy Analytics Inc. © Copyright 2008-2009. http://www.ehealthinformation.ca/wp-content/uploads/2014/08/2009-De-identification-PA-whitepaper1.pdf

'Today we live in a world where our personal information is being continuously captured in a multitude of electronic databases. Details about our health, financial status and buying habits are stored in databases managed by public and private sector organizations. These databases contain information about millions of people, and can provide valuable research, epidemiologic and business insight. For example, examining a drug store chain's prescriptions can indicate where a flu outbreak is occurring. To extract or maximize the value contained in these databases, data custodians must often provide outside organizations access to their data. In order to protect the privacy of the individuals whose data is being disclosed, a data custodian will "de-identify" information before releasing it to a third-party. De-identification ensures that data cannot be traced to the person about whom it pertains. What might

seem like a simple matter of masking a person's identifiers (name, address), the problem of de-identification has proven more difficult and is an active area of scientific research.'

41) Desai, T., Ritchie, F. and Welpton, R. (2016). Five Safes: designing data access for research, Economics Working Paper Series 1601, University of the West of England, Bristol

The Five Safes model is a popular framework for designing, describing and evaluating access systems for data, used by data providers, data users, and regulators. The model integrates analysis of opportunities, constraints, costs and benefits of different approaches, taking account of the level of data anonymisation, the likely users, the scope for training, the environment through which data are accessed, and the statistical outputs derived from data use.

42) Abbott, O., Jones, P. and Ralphs. (2016). Large-scale linkage for total populations in official statistics In: Harron K, Goldstein H and Dibben C, editors. Methodological Developments in Data Linkage, Wiley, UK

Current practices in record linkage for population censuses. Case studies of countries that operate a population register. Case study of the England and Wales 2011 census and then the beyond 2011 programme. Challenges, linkage quality, next steps.

43) Culnane, C., Rubinstein, B.I.P. and Teague, V. (2017). Vulnerabilities in the use of similarity tables in combination with pseudonymisation to preserve data privacy in the UK Office for National Statistics' Privacy-Preserving Record Linkage https://arxiv.org/pdf/1712.00871.pdf

'In the course of a survey of privacy-preserving record linkage, we reviewed the approach taken by the UK Office for National Statistics (ONS) as described in their series of reports "Beyond 2011". Our review identifies a number of matters of concern. Some of the issues discovered are sufficiently severe to present a risk to privacy. The issues discovered are as follows, in order of severity, from least to most severe: 1. Incorrect cryptographic assumptions have been made, in combination with incorrect statements regarding the required entropy for HMAC (hash-based message authentication) keys. The consequence is an overstatement of the security of the solution; 2. The provision of similarity tables with HMAC'd names exposes the approach to frequency attacks; and, 3. Plaintext similarity scores provide an index to HMAC'd names, permitting plaintext recovery of names.'

44) Goldstein, H. and Harron, K. (2018). 'Pseudonymisation at source' undermines accuracy of record linkage. Journal of Public Health, Volume 40, Issue 2, 1 June 2018, Pages 219–220, https://doi.org/10.1093/pubmed/fdy083

'Pseudonymisation is one element of a range of measures that can be used to protect the privacy of individuals. 'Pseudonymisation at source' is a technique used by data providers to avoid identification of individuals before data are linked for secondary uses such as service evaluation or research. The technique involves replacement of direct identifiers, known as 'personal data' or 'confidential patient information', such as NHS number, date of birth and postcode, with a pseudonym, which does not reveal a person's real world identity. Use of the same pseudonymisation key for multiple data sources before data are shared enables data sources to be linked together without using 'personal data' and therefore avoids the need for patient consent or other legal provision under the Data Protection Act or the General Data Protection Regulation. As

we discuss below, however, this limits the utility and quality of any resulting linked datasets.'

45) Allen, J., Holman, C.D.J., Meslin, E.M. and Stanley, F. (2013). Privacy protectionism and health information: any redress for harms to health? J. Law Med. 21(2):473-85.

'Health information collected by governments can be a valuable resource for researchers seeking to improve diagnostics, treatments and public health outcomes. Responsible use requires close attention to privacy concerns and to the ethical acceptability of using personal health information without explicit consent. Less well appreciated are the legal and ethical issues that are implicated when privacy protection is extended to the point where the potential benefits to the public from research are lost. Balancing these issues is a delicate matter for data custodians. This article examines the legal, ethical and structural context in which data custodians make decisions about the release of data for research. It considers the impact of those decisions on individuals. While there is strong protection against risks to privacy and multiple avenues of redress, there is no redress where harms result from a failure to release data for research.'

46) Jones, K.H., Ford, D.V. *Et al.* (2014). A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation, Journal of Biomedical Informatics: special issue on medical data privacy, doi: 10.1016/j.jbi.2014.01.003. https://ac.els-cdn.com/S1532046414000045/1-s2.0-S1532046414000045-main.pdf?_tid=bfc64cea-1ff1-4390-bf7e-06b90cb95c15&acdnat=1532797834_55a9f98c2c7feab9839055d3bac55696

'With the current expansion of data linkage research, the challenge is to find the balance between preserving the privacy of person-level data whilst making these data accessible for use to their full potential. We describe a privacy-protecting safe haven and secure remote access system, referred to as the Secure Anonymised Information Linkage (SAIL) Gateway. The Gateway provides data users with a familiar Windows interface and their usual toolsets to access approved anonymously-linked datasets for research and evaluation. We outline the principles and operating model of the Gateway, the features provided to users within the secure environment, and how we are approaching the challenges o fmaking data safely accessible to increasing numbers of research users. The Gateway represents a powerful analytical environment and has been designed to be scalable and adaptable to meet the needs of the rapidly growing data linkage community.'

47) Dibben, C. and Elliott, M. t al. (2016). The data linkage environment In: Harron K, Goldstein H and Dibben C, editors. Methodological Developments in Data Linkage, Wiley, UK

'Working with data that is strictly non-personal (i.e. absolutely anonymous) guarantees protection but is rarely achievable in a data linkage context, and indeed may even be impossible with any useful data.' 'What is required is the construction of a data linkage environment in which the process of re-identification is mad so unlikely that the data can be judged as functionally anonymous.' This chapter describes the key features of the data linkage environment, with examples for illustration.

48) Office for Statistics Regulation (2018) Joining up Data for Better Statistics https://www.statisticsauthority.gov.uk/wp-content/uploads/2018/09/Data-Linkage-Joining-Up-Data.pdf

'In 2017, the Office for Statistics Regulation launched an investigation of the UK statistics system's ability to provide greater insight to users via data linkage. We are the independent regulators of the UK's official statistics system and our interest in this area is underpinned by the Code of Practice for Statistics' three pillars of trustworthiness. Statistics add value when they answer society's questions. Many questions cannot be answered without sharing and linking data. As a result, a greater willingness and ability to share and link data is an essential prerequisite for improved official statistics. Without a focus on the quality of the data – their source, how they have been collected and processed, any biases and incompleteness in the data – the results could be misleading. Custodians of public data must demonstrate their trustworthiness by safeguarding data robustly during and after the sharing and linkage process, and by being open to public scrutiny. Organisational trustworthiness is at the core of OSR's work and is a key component of the first pillar in the Code of Practice'

49) Gilbert, R., Lafferty, R., Hagger-Johnson, G. *et al.* (2017). GUILD: GUidance for Information about Linking Data sets, Journal of Public Health, pp.1–8 doi:10.1093/pubmed/fdx037

'GUILD aims to improve the quality of data processing, linkage, analyses and research reports by raising awareness about detailed information that could be shared at each step of the linkage pathway. The guidance also aims to highlight the responsibilities of data providers, linkers and analysts, not just report writers, to make this information available.' 'GUILD highlights the choices and decisions made during data processing that affect linkage error and hence the results of analyses. Sharing information along the data linkage pathway could improve the transparency and reproducibility of research, promote the use of improved methods to address linkage error, and improve the interpretation of studies based on linked data.'

50) Laurie, G., Ainsworth, J., Cunningham, J., Dobbs, C., Jones, K.H., Kalra ,D., Lea, N. and Sethi, N. (2015). On moving targets and magic bullets: Can the UK lead the way with responsible data linkage for health research? International Journal of Medical Informatics 08/2015; 84(11)

Discussion centres around lessons learned from previous successful health research initiatives, identifying those governance mechanisms which are essential to achieving good governance. This article suggests that a crucial element in any step-increase of research capability will be the adoption of adaptive governance models. These must recognise a range of approaches to delivering safe and effective data linkage, while remaining responsive to public and research user expectations and needs as these shift and change with time and experience.

51) Jones, K.H., Laurie, G., Stevens, L.A., Dobbs, C., Ford, D.V. and Lea, N. (2017). The other side of the coin: harm due to the non-use of health-related data. IJMI, 97:43-51

An international case study approach to explore why data non-use is difficult to ascertain, the sources and types of health-related data non-use, its implications for citizens and society and some of the reasons it occurs. It does this by focussing on issues with clinical care records, research data and governance frameworks and associated examples of non-use. There is ample indirect evidence that health data non-use is implicated in the deaths of many thousands of people and potentially

£billions in financial burdens to societies. The need to keep benefits and limitations in perspective, to move steadily towards socially responsible reuse of data becoming the norm to save lives and resources.