Beyond 2011: Matching Anonymous Data
July 2013

Background

The Office for National Statistics is currently taking a fresh look at options for the production of population and small area socio-demographic statistics for England and Wales. The Beyond 2011 Programme has been established to carry out research on the options and to recommend the best way forward to meet future user needs.

Improvements in technology and administrative data sources offer opportunities to either modernise the existing census process, or to develop an alternative by re-using existing data already held within government. Since methods for taking the traditional census are already relatively well understood most of the research is focussing on how surveys can be supplemented by better re-use of 'administrative' data already collected from the public.

The final recommendation, which will be made in 2014, will balance user needs, cost, benefit, statistical quality, and the public acceptability of all of the options. The results will have implications for all population-based statistics in England and Wales and, potentially, for the statistical system as a whole.

About this paper

This paper provides a summary of research undertaken to ascertain the level of quality loss that is incurred by anonymising data prior to record linkage.
The paper is the first publication of our proposed methods for matching anonymous data in Beyond 2011. Research into the feasibility of matching anonymous data is continuing throughout the Programme, and further developments will be published in the future.

This document is one of a series of papers to be published over coming months. These will report our progress on researching and assessing the options, discuss our policies and methods and summarise what we find out about individual data sources.

For more information

Paper references (Paper M1, O1, R2 etc) used throughout refer to other papers published by Beyond 2011 all of which are available on the Beyond 2011 pages of the ONS website.

Search Beyond 2011 @ www.ons.gov.uk or contact : beyond2011@ons.gov.uk

BEYOND 2011
M9
METHODS
& POLICIES

Office for National Statistics

Table of Contents

Office for National Statistics

1 Executive Summary

The administrative data options being evaluated in Beyond 2011 include large scale record linkage between administrative sources and surveys. We are aiming to optimise the quality of record linkage under these options as efficient and accurate matching is likely to improve the estimation process in the production of population estimates.

It is recognised that our planned approach of matching multiple administrative sources might elevate the associated risks relating to the privacy of data about people and households. We have therefore taken a decision to anonymise the data to ensure that high levels of anonymity and privacy are maintained. More information on this process can be found in 'Beyond 2011: Safeguarding Data for Research: Our Policy' (Paper M10).

It is important to note that there is a distinction between "anonymisation" and "pseudonymisation". The approach that we are using is more strictly described as "pseudonymisation". However, for the remainder of the paper the word "anonymisation" will be used.

This paper provides a summary of research undertaken to ascertain the level of quality loss that is incurred by anonymising data prior to record linkage. Having developed a method to match data under these conditions, a series of comparison exercises have been undertaken to measure the additional matching error that is incurred.

Quality loss has been measured by comparing the links made by the Beyond 2011 matching approach with those made by an optimised match routine which includes conventional score based and clerical methods. Referred to as a 'gold standard' matched dataset, this controlled comparison provides the basis for measuring the level of error that is incurred by anonymising data prior to record linkage. Two comparison exercises are reported in section 4 of this report, the first relates to the linkage of records from the NHS Patient Register (PR), (see Beyond 2011: Administrative Data Sources Report: NHS Patient Register (S1)),to the 2011 Census, the second relates to the linkage of records from the student Higher Education Statistics Agency Student Record dataset (HESA) to the PR. Headline estimates are summarised below.

- For PR to census matching the error rate relating to false positives[1] is estimated to have increased by 0.2 % to 0.6 %.

- For PR to census matching the error rate relating to false negatives[2] is estimated to have increased by 0.9 % to 3.3 %, depending on the local authority (LA) being matched.

- For the HESA to PR match the error rate relating to false positives was estimated at 1.0 % and the error rate for false negatives was estimated at 2.5 %.

Overall, the quality of matching reported in these comparison exercises is very encouraging. Ongoing research continues within the Programme to improve the accuracy of matching, particularly in reducing the number of false negatives and to explore whether an adjustment can be made during the estimation process (which is outlined in 'Beyond 2011: Producing Population Statistics Using Administrative Data: In Theory' (Paper M8), published simultaneously with this

[1] Links between records that have been made in error by our matching approach
[2] True match pairs that have been left unlinked by our matching approach

report). We would welcome any comments relating to the methods presented in this paper and any suggestions that might improve the accuracy or efficiency of the methods proposed.

2 Introduction

2.1 Background

Population and socio-demographic statistics are currently based on a 10-yearly census of the population. There is a clear, ongoing need for high quality statistics and, whilst the 2011 Census was successful, the traditional census is becoming increasingly costly and difficult to conduct. There is also a demand for more frequent statistics from some users.

Improvements in technology and administrative data sources offer opportunities either to modernise the census process, or to develop an alternative by re-using existing data already held within government. Whilst this would require investment in the short term, it should deliver real-term savings in the longer term compared to the 2011 Census, which cost £480m over 10 years.

In May 2010 the UK Statistics Authority asked the Office for National Statistics (ONS) to investigate the possible alternatives for England and Wales. Whilst the UK Statistics Authority is an independent, non-Ministerial department, the final decision will be taken by Parliament because of funding and legislative requirements. A recommendation will be made in 2014.

During the first phase of the programme, running from 2011/12 to 2014/15 the programme will assess user requirements for small area population and socio-demographic statistics and the best way to meet these needs. The outcome of the first phase will be a full business case underpinning the recommendation, setting out the costs and benefits of the options considered.

A full public consultation allowing stakeholders to contribute views on the key issues and options is planned for September to November this year.

2.2 This paper

We are currently looking at options for the production of population and small area socio-demographic statistics for England and Wales. Three of the six options that we are exploring involve the linkage of multiple administrative sources at the individual record-level. This approach would elevate the associated risks relating to the privacy of data about people and households if it were not mitigated, and we have therefore taken a decision to anonymise the data to ensure that high levels of anonymity and privacy are maintained. More information on our policy to safeguard our data can be found in ‘Beyond 2011: Safeguarding Data for Research: Our Policy’ (Paper M10).

This paper outlines the methodological challenges of developing a matching process that gives suitably low levels of matching error to inform the discussion as to whether data can be anonymised in this way for any future implementation phase.

The body of this report is structured over four sections. Section 3 looks in more detail at the challenges associated with matching anonymised data, and the methods that have been developed by the Beyond 2011 team during feasibility research. Section 4 presents the results from two quality assurance exercises where links made by Beyond 2011 matching algorithms have been compared with those obtained by conventional matching methodologies. Section 5 outlines some of the future research that the team will focus on to improve the quality of matching throughout the research programme. Section 6 concludes by considering the impact of matching error in estimating the population and the feasibility of using an anonymisation approach in Beyond 2011.

This paper is the first publication of our proposed methods for matching anonymous data in Beyond 2011. Research into the feasibility of matching anonymous data is continuing throughout the Programme, and further developments will be published in the future.

3 Beyond 2011 matching methodology

Data matching has a vital role in the administrative data options being researched in Beyond 2011. One of the major challenges for the administrative data options is to exclude individuals that are listed on the administrative data but who are no longer resident in the population. Data matching between the administrative sources is used to identify individuals that appear on multiple administrative sources, and therefore have a higher likelihood of being in the *usually resident* population[3]. The matching process is crucial in determining rules to construct a 'Statistical Population Dataset' (SPD) that can be used as an auxiliary for estimating the population. More information about the construction of SPDs and the rules that have been formulated from administrative data matching is reported in 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice' (paper M7).

Additionally, data matching is likely to play a significant role in coverage adjustment when estimating the population. In order to make accurate adjustments for under-coverage and over-coverage on the administrative data it is necessary to match the SPD to the Population Coverage Survey (PCS) to a very high standard. Further information about the PCS design is reported in 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory' (paper M8).

Matching error is of particular concern when matching the SPD to the PCS, as it has the potential to inflate or deflate the population estimates depending on the number of false positive or false negative matches. For the 2011 Census, census records were matched to the 2011 Census Coverage Survey (CCS) using a combination of exact matching, score based matching, clerical matching and clerical searching. A summary of these methods is provided below.

- Exact matching – automatically linking pairs of records that are identical on all matching fields (for example name, sex, date of birth and postcode).

- Score based matching – scoring pairs of records for their overall level of agreement and automatically linking those that score above a specified threshold.

- Clerical matching – the manual review of pairs of records that are classified as potential matches based on their overall agreement scores. These records have scored lower than the auto-match threshold and a trained matcher is required to make a decision based on the evidence available as to whether or not the two records should be linked.

- Clerical searching – individually taking the 'residuals' (unmatched records) on one of the datasets and querying the database of the second dataset for a matching record. Where potential match pairs are identified, a clerical decision is made by the matcher as to whether or not to link the two records.

In the 2011 Census to CCS matching, the role of clerical matching and clerical searching was very important for the optimisation of matching accuracy. By using these methods, it ensured that the

[3] We are currently adopting the UN definition of *usually resident* – that is the place at which the person has lived continuously for at least the last 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months (United Nations, 2008).

number of false positives and false negatives were kept to a minimum, thereby enabling accurate coverage adjustments to be made in the estimation process.

In Beyond 2011, the anonymisation process 'transforms' the data in a manner that prevents the use of clerical matching and clerical searching, as well as restricting the viability of conventional score based matching. The focus of our research to date has therefore concentrated on the development of new methods that perform well in an anonymous research environment, as well as measuring the quality of matching using anonymised data when compared to optimised methods that include clerical and score based matching.

3.1 The anonymisation process, Beyond 2011

In order to mitigate the privacy related risks associated with the administrative data options, all datasets with person identifiable information is held in a Statistical Research Environment (SRE) with strong security safeguards. Each dataset is loaded separately onto the reception server where it undergoes a series of data pre-processing steps including geo-referencing, variable standardisation, match-key creation (outlined in section 4.4) and the construction of similarity tables (outlined in section 4.5). Only one dataset is ever held on the reception server at any point in time, and it is after the pre-processing stage that data is anonymised. Once the data has been transformed through the anonymisation process, it is moved into the SRE core environment for matching with other datasets. The SRE reception server is then sanitised in advance of receiving the next dataset for pre-processing.

We have used a cryptographic hash function[4] to anonymise person identifying information including names, dates of birth and addresses. The hash function, which converts a field into a condensed representation of fixed value, is a one-way process that is irreversible – once the hashing algorithm is applied it is not possible to get back to the original information without significant effort and the use of tools that are not available in the research environment.

It should be noted that the Information Commissioner's Office (ICO's) Code of Practice http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation describes the difference between "anonymisation" and "pseudonymisation". The hashing technique that we are using is more strictly described as "pseudonymisation". However, for the remainder of the paper the word "anonymisation" will be used. Further information about the policy for safeguarding data during the research phase of the Programme is reported separately in 'Beyond 2011: Safeguarding Data: Policy for Research' (paper M10).

3.2 Implications of matching anonymised data

We need to optimise the quality of record linkage under the administrative data options as high quality matching will be essential in the production of accurate population estimates. The nature of the hashing process means that only in cases where two records are identical - i.e. where names, dates of birth and addresses are recorded in precisely the same format, will an automatic match be possible on the hashed values. In cases where there are spelling errors or inconsistencies between two records relating to the same individual (for example the names John and Jon), the hash values will not be identifiable as being similar. Figure 1 provides some examples of hashed values to demonstrate the differences arising from slight inconsistencies between two strings.

[4] SHA-256 hash function used in combination with a secret cryptographic key

Figure 1 – Examples of name and date strings transformed into hash values

String/Value to Hash	Hashed Value
John	8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
Jon	86 1A 42 1C 1A 05 E0 E8 FA 24 A1 53 41 59 69 1F
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
John	8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61
Smyth	CB 36 9F C9 0A 3B A0 2E E9 9C A0 5E E0 69 84 FB
Jonathan	F4 5C C5 B7 A6 59 23 79 B8 5B 81 81 AA AD 38 50
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
Jonny	ED ED 5C 0E 56 00 83 84 AA 03 8F E7 02 AA AB E3
27/01/1965	4F 6E B0 E4 55 84 BC 0A 8B A3 89 B5 16 F4 49 9A
26/01/1965	2C 5A 2C 3D 80 D1 48 35 70 24 6A D8 E5 2C 94 17
27/01/1965	4F 6E B0 E4 55 84 BC 0A 8B A3 89 B5 16 F4 49 9A
27/02/1965	EC 67 CC 6D C7 23 40 84 09 E7 B5 7C DE 79 6B D4
27/01/1966	90 BF F8 D3 C5 DD 3F DB 3C 6D DC 39 AD EE E6 46
1965	10 B2 57 F8 08 7E 72 F1 2A E6 96 E4 A1 E4 26 DE
1966	5C C9 4A 4C CA E8 48 75 B5 52 68 E0 B0 C5 F3 CA
1965	10 B2 57 F8 08 7E 72 F1 2A E6 96 E4 A1 E4 26 DE
1966	5C C9 4A 4C CA E8 48 75 B5 52 68 E0 B0 C5 F3 CA

Once the information has been hashed, it is not possible to compare likenesses with other similar words. Examples of circumstances where this might occur include the following:

- use of abbreviations or nicknames;
- data capture errors (resulting from discrepancies in spelling, handwriting and typing);
- recall errors, for example inaccurate reporting of postcodes.

Figure 2 illustrates the types of matches that can be made when using conventional score based approaches or clerical resolution. In this example, the first pair of candidate records differ on surname only. Typically the decision to match or not would be resolved by a string comparison algorithm that assigns a similarity score for agreement between the two surnames. The decision to accept these two records as a match would be automated in this case because the match score exceeds a certain threshold (for example 95 %). The second pair of records, which have a lower match score due to a number of inconsistencies, would fail automatic matching and would be passed for clerical investigation.

Figure 2 – Examples of score based auto-matching and clerical matching

Source 1				Source 2			
Forename	**Surname**	**DoB**	**Postcode**	**Forename**	**Surname**	**DoB**	**Postcode**
Sarah	Johnston	22/10/1982	PO15 1HS	Sarah	Johnson	22/10/1982	PO15 1HS
Michael	Smyth	13/02/1970	PO15 1HR	Mike	Smith	13/02/1970	PO11 4FF

Agreement Scores				
Forename	**Surname**	**DoB**	**Postcode**	**Match Score**
1	0.87	1	1	0.97
0.32	0.91	1	0.65	0.72

In an anonymous ('hashed') environment, the name 'John' may be assigned a hashed value of 'XY140127'[5]. This assignment will be consistent across all data sources for the name 'John', allowing exact matching to another field with same value. However, the assignment of a hashed value to a similar variant 'Jon' would result in a different hashed value (for example 'MY7793812'). There is no longer any basis for comparing similarity once the transformation has taken place.

Consequently some of the well-established methods for finding true matches are redundant once data has been hashed in this way. Matching techniques that use string comparison algorithms to identify records that are similar are ineffectual in this context, as are resolutions that are made clerically by directly comparing candidate pairs of records side by side.

The challenge has been to develop alternative approaches that can automate the matching of anonymised records that do not exactly match. This is also a major goal in the wider field of record linkage and various methodologies have been reported that attempt to tackle the issue, including third-party methods (Lyons R A *et al*., 2009), Bloom filter encryption (Schnell *et al*., 2010; Schnell *et al*., 2009), and n-gram comparators of encrypted values (Churches & Christen, 2004). However these methods, at least in the short term, are not appropriate for our existing data storage environment. The cost and logistic complexity of setting up a third party arrangement is not practical during feasibility research, and recent innovations, such as bloom filter encryption have not been fully explored from an accreditation perspective.

Hence for the research phase, the focus has been on identifying matches that are similar but not in exact agreement when using a secure hashing algorithm. We have explored a number of alternative data pre-processing techniques that attempt to resolve these inconsistencies between candidate match pairs. These are outlined in further detail in the next section of this report.

3.3 Summary of Beyond 2011 methods

The matching methods used in Beyond 2011 can be categorised under the following:

- deterministic, or 'rule based' matching using match-keys;
- score based matching using logistic regression.

Crucial to the success of these methods are the steps undertaken during the data pre-processing stage when data is imported into the SRE. An outline of the match-key process is outlined in detail

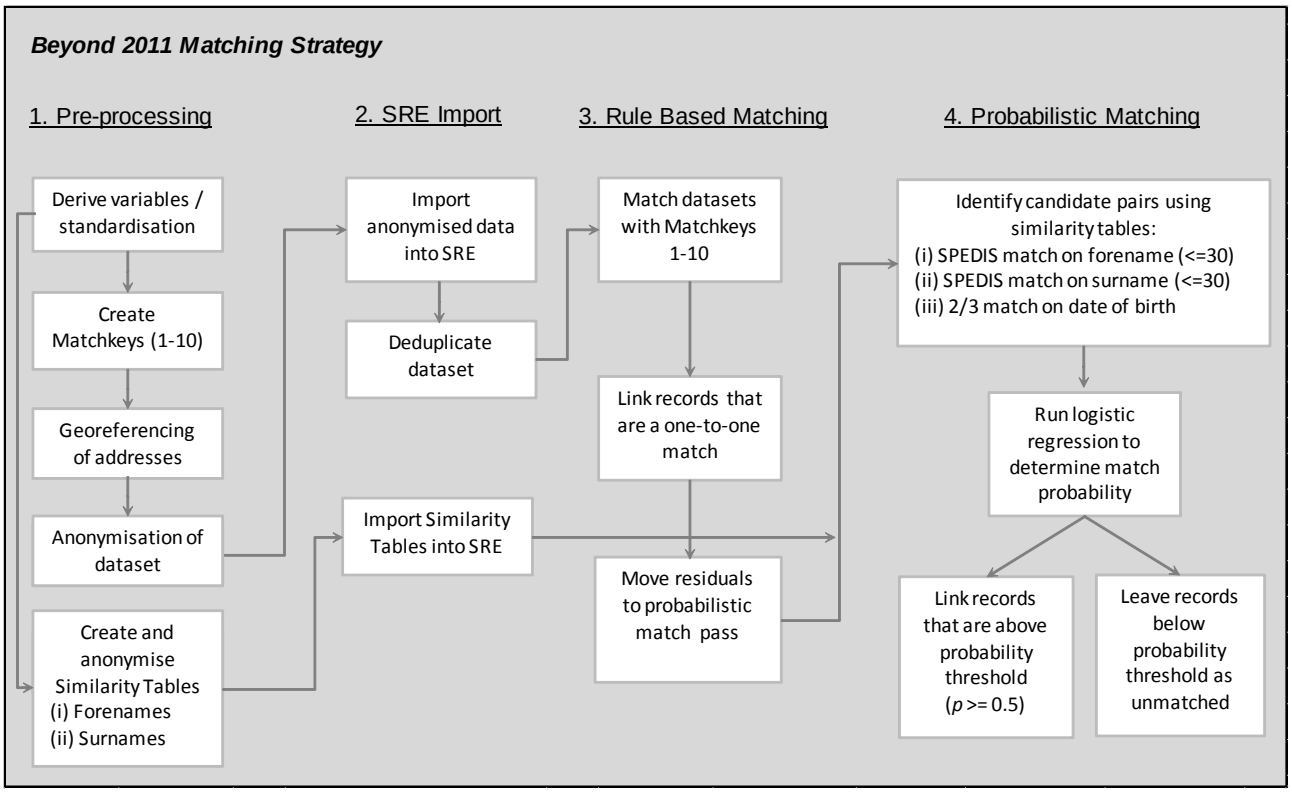
[5] Examples of hash value are illustrative only

in sub-section 3.4. In order to initiate a score based method, similarity tables for names and dates of birth are also constructed during pre-processing and this is explained in detail in sub-section 3.5.1. The modelling of decision making through logistic regression is then outlined in section 3.5.2.

Other pre-processing methods, including the use of name clustering and Soundex algorithms that have been tested during the research phase are summarised in appendices A to C at the end of this report. These methods have been rejected from use in our matching, either because they have been superseded by a preferred method, or because they have a limited capacity to improve the algorithm.

Figure 3 is a flow diagram of our matching strategy.

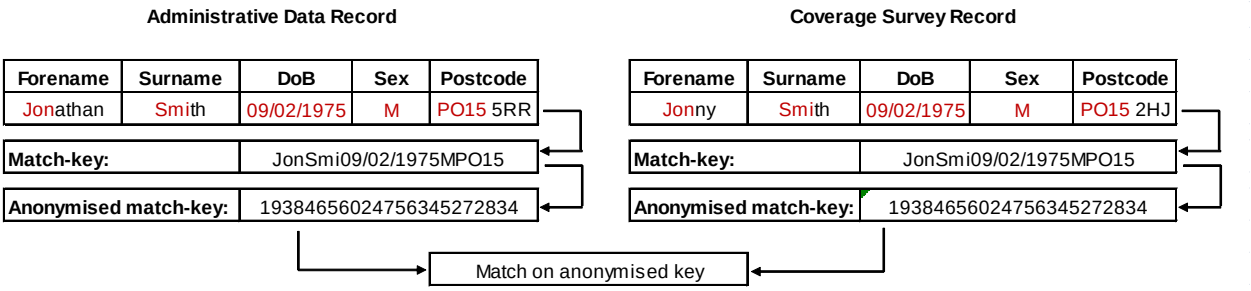
Figure 3 – Beyond 2011 matching strategy which takes place in the SRE



3.4 Match-keys

Match-keys are created by putting together pieces of information to create unique keys that can be hashed and used for automated matching, with the intention of elimination some of the discrepancies that might otherwise prevent an automated match. For example, a match-key might be constructed from the first three characters of an individual's forename and surname, combined with their date of birth, sex and postcode district. Figure 4 shows how such a key is created, with the highlighted text compressed to form a single string. The resulting string is then 'hashed' and can be used to link records between datasets in the anonymous data research environment.

Figure 4 – Example of match-key creation for linking records



Our approach is to construct a series of these match-keys, each of which is designed to resolve a particular type of inconsistency that often occurs between records belonging to the same individual. These match-keys are presented as matching fields in a hierarchical, or stepwise, linkage process, each forming a separate 'match pass' essentially forming the deterministic phase of the overall matching strategy (Karmel *et al*., 2010; Gomatam *et al*., 2002). Such hierarchical deterministic approaches are prevalent in linkage studies across epidemiology (e.g. Li *et al*., 2006; Pacheco *et al*., 2008) and the match-key approach has been seen to perform well in an Australian community health care study (Karmel *et al*., 2010).

A major requirement when using this approach is to ensure that the resulting match-key retains a high level of uniqueness for the majority of records to be matched. A good example of a match-key produced in this way is a concatenation of forename initial, surname initial, sex, date of birth and postcode district[6]. Having undertaken frequency analysis of the PR, it is estimated that 99.55 % of people in the UK have a unique match-key when data is concatenated in this way. Furthermore, the potential for disagreement between two matching records is significantly reduced when matching on this information - provided the first letter of the forename and surname have been documented correctly and the individual has accurately reported the characters of their postcode district, the potential for inconsistency between the two sources has been significantly reduced.

Inconsistency between matching variables can occur in a number of different forms. A single match-key alone cannot resolve all of the inconsistencies that occur between data sources. Frequency analysis of the PR has been undertaken for a range of variable concatenations resulting in a series of match-keys which are being used in our matching approach, all of which are designed to resolve particular inconsistencies between match pairs. Figure 5 presents the structure of each of these match-keys and the uniqueness of those keys when they are created for all records held on the PR. It also summarises the type of inconsistency that each match-key is designed to resolve. These match-keys capture, on average, 95 % of the available matches.

[6] Postcode District is the inward part of the postcode e.g. SW19

Figure 5: Uniqueness of match-keys derived from the PR and the inconsistencies they resolve between true match pairs

Match-key	% Unique records on PR	Inconsistencies resolved by match-key
(1) Forename, Surname, DoB, Sex, Postcode	99.99%	None - exact agreement
(2) Forename initial , Surname initial, DoB, Sex, Postcode District	99.55%	Name / postcode discrepancies
(3) Forename bi[▽]-gram, Surname bi-gram, DoB, Sex, Postcode Area	99.44%	Name discrepancies / movers in area
(4) Forename initial, DoB, Sex, Postcode	99.84%	Surname discrepancy
(5) Surname initial, DoB, Sex, Postcode	99.44%	Forename discrepancy
(6) Forename, Surname, Age, Sex, Postcode Area	99.46%	Dob discrepancy / movers in area
(7) Forename, Surname, Sex, Postcode	99.19%	DoB missing / incorrect
(8) Forename, Surname, DoB, Sex	98.87%	Movers out of area
(9) Forename, Surname, DoB, Postcode	99.52%	Sex missing / incorrect
(10) Surname, Forename, DoB, Sex, Postcode (matched on Key 1)[*]	99.99%	Forename / surname transpositions
(11) Middle name, Surname, DoB, Sex, Postcode	99.90%[◊]	Forename / middle name transpositions

In an attempt to reduce the risk of false positive matches, records are only linked on a match-key if it is unique on both datasets (i.e. one-to-one match). If multiple records match on a particular match-key then the link is not made and candidates are passed on as a residual to the next match pass. The hierarchical nature of the whole matching process has implications. Matches that are made at an early stage of the process are linked and removed, with only the residuals being passed to the next stage. This means that once two records are linked there is no process to review that match status and 'unlink' the records on the basis that the true match pair is identified at a later stage in the process. However, future methodological refinements of the algorithm may allow for false positives to be unlinked in favour of a more probable match at a later stage in the process.

3.5 Score based matching

There are two challenges with the implementation of score based methods in our anonymous research environment. The first is in obtaining string comparison scores between anonymised names and dates of birth. Since the anonymisation process does not allow the comparison of similar information such as when an abbreviation is used in one source and the full name in another, we have developed the use of 'similarity tables' that are created during the pre-processing stage to serve as lookups for names that are similar. The second issue is how to use 'similarity scores' to determine whether the two records are in fact a match.

[▽] In this context the bi-gram comprises the first two characters of the name

[*] Key 10 has a transposition between forename and surname. A match is searched for on key 1 of the other dataset to match individuals that may have used their forename and surname interchangeably.

[◊] Key 11 is not produced if middle name is not provided. This is an estimate of uniqueness for the key from records that include middle name. Matching key 11 on key 1 from the other dataset can also be used to match individuals that may have used their forename and surname interchangeably.

3.5.1 Similarity tables

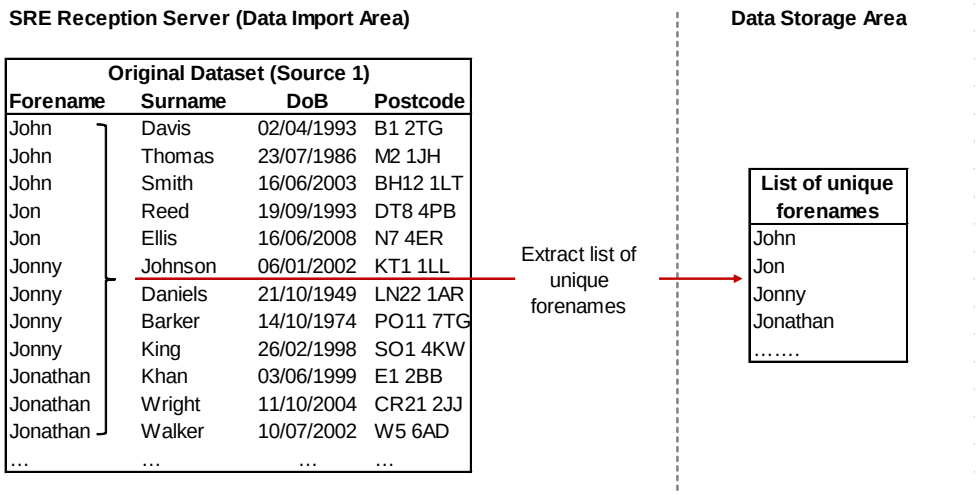
The construction of 'similarity tables' is founded on the principle that the matching of single variables one at a time (prior to hashing) is non-disclosive on the grounds that individual persons are not identifiable from a forename (or surname) alone. For example, the isolated matching of a list of forenames contained on two datasets cannot be traced back to an individual record. A similar technique, utilising a public reference table of names, has been proposed by Pang & Hansen (2006), although their method has been shown to perform poorly against other approximate string matching methods (Bachteler *et* al., 2010).

However, we had a clear need to develop a method that provided opportunities to perform more sophisticated score based matching between candidate records. The development of this method required selecting a suitable string comparison algorithm, which for the purposes of initial research was the SAS[7] proprietary SPEDIS function, which operates as an edit distance function similar to Levenstein Distance (Yancey, 2005). This method was chosen over standard Levenstein Distance and the string comparison metric Jaro-Winkler (Yancey, 2005; Cohen *et al*., 2003) due to its ease of use during research.

The series of diagrams below demonstrate how the method we developed, centred around the pre-hashing creation of SAS SPEDIS edit distance tables, can be constructed from single variable matching, and how candidate match pairs can be identified by hashing the tables and using them as a lookup reference in the anonymous research environment.

a) Dataset 1 import: A complete list of all forenames is extracted and deduplicated from dataset 1 to create a list of forenames.

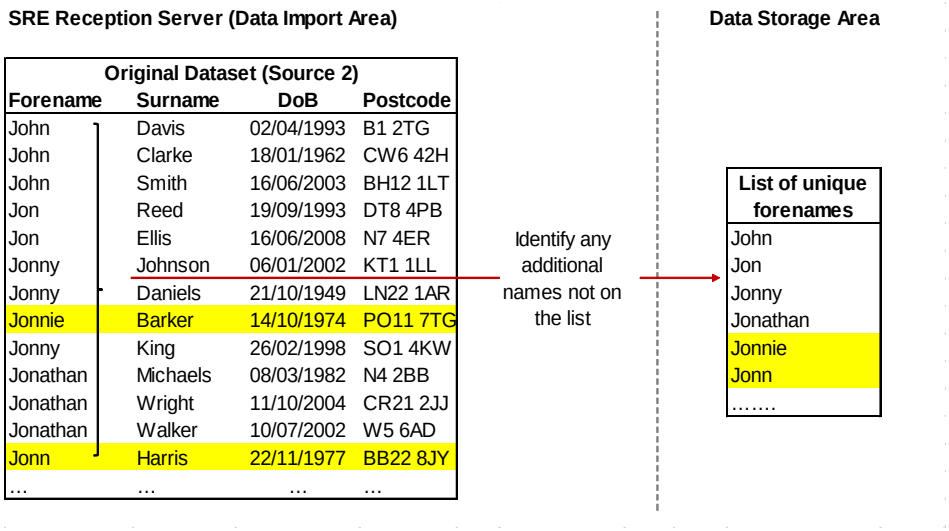
Figure 6 – Forename extraction from dataset 1



b) Dataset 2 import: Any forenames on the second dataset that are not already on the list are added to create a complete list of forenames between the two datasets.

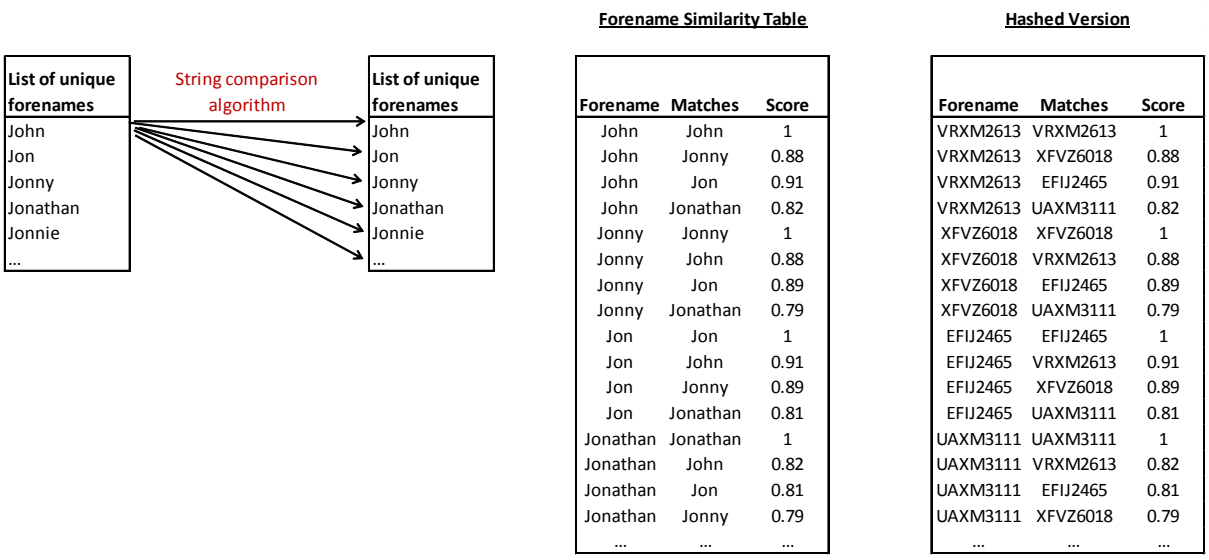
[7] Statistical Analysis System (SAS), SAS Institute Inc.

Figure 7 – Forename extraction from dataset 2



c) **Create similarity table:** A copy of the list of forenames is created and a string comparison algorithm is run between the names contained in the two tables. In the example below we have used the SAS SPEDIS function to score levels of similarity between all possible pairings of names. Those names that match within a specified threshold are retained in the table with an agreement score calculated from the algorithm. This forms the basis of a similarity table, which is hashed for use in the anonymous research environment.

Figure 8 – String comparison and hashing



d) **Identify candidate match pairs:** Hashed similarity tables are also produced for surnames and dates of birth. Using them as lookup tables for exact matching in the anonymous research environment enables the identification of candidate match pairs between the two data sources.

Figure 9 – Undertake 3-way join using similarity tables

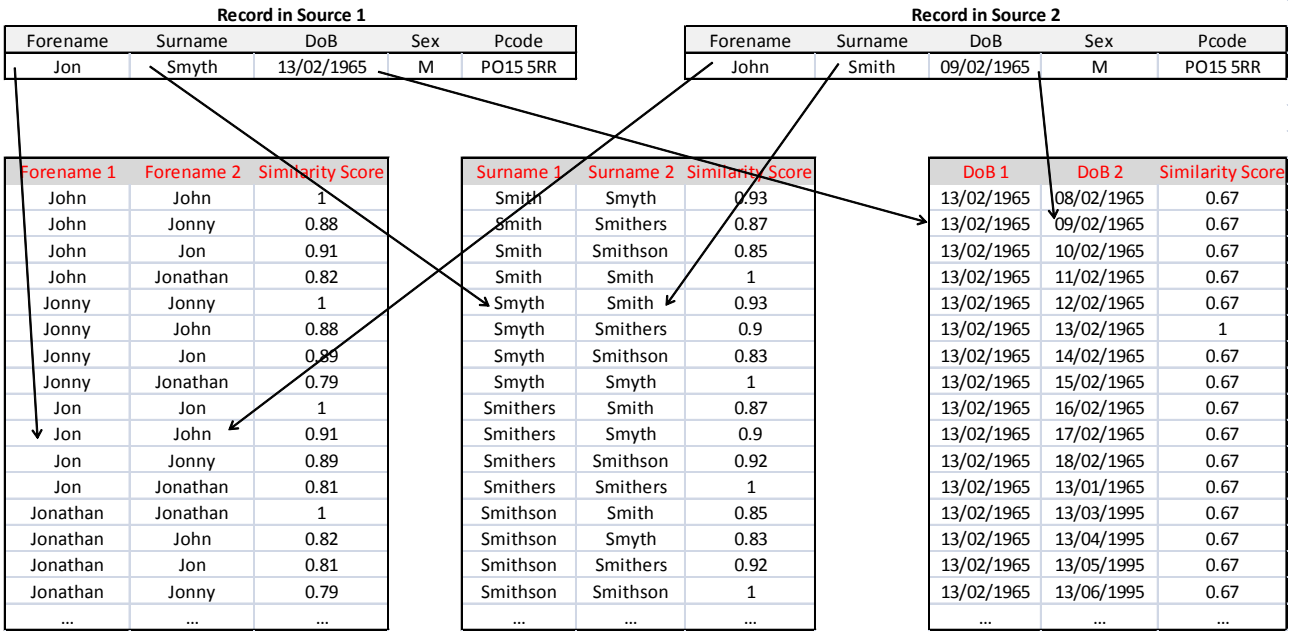
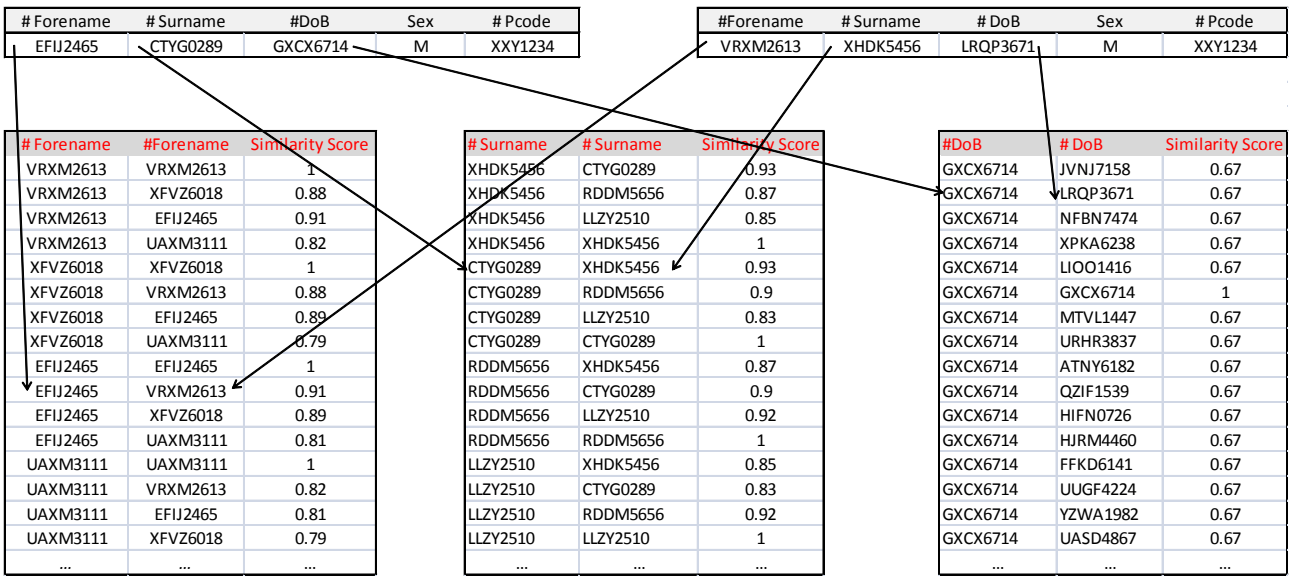


Figure 9 above demonstrates the process with 'un-hashed' data for illustrative purposes. The names and dates of birth of records on source 1 will always be located in the left column of the similarity tables since they contain the full list of names and dates of birth that exist across the two datasets. The adjacent columns hold all of the names and dates of birth that are similar in accordance with the specified threshold, and it is in this column that matches from source 2 are sought. Where a match is found on all three tables, the two records are brought together as potential match candidates. Figure 10 demonstrates how the process would appear in the anonymous research environment using 'hashed' data.

Figure 10 – Undertake 3-way join using similarity tables ('hashed' version)



When this information is combined, the following set of scores can be derived as demonstrated in figure 11.

Figure 11 – Agreement scores to be used in score based matching

Candidate pair ID	Forename	Surname	DoB
Candidate pair_1	0.91	0.93	0.67

The advantage of this approach is that candidate pairs that have multiple inconsistencies can still be identified as potential matches based on the level of agreement between forename, surname and date of birth. Having agreement scores for these variables, as well as a measure of geographic distance and commonality of names provides opportunities for score based matching. The measures of agreement between the matching variables, commonality of names and distances between locations can all be stored in the anonymous research environment as an array of modelling data as shown in figure 12.

Figure 12 – Example of modelling data for match candidates derived from similarity tables

Candidate pair ID	Forename agreement score	Surname agreement score	Forename % Frequency	Surname % frequency	Postcode agreement score	Sex agreement	DoB agreement score	Distance between LSOA[8] centroids
Candidate pair_1	91%	93%	0.025%	0.014%	4	1	67%	2.1
Candidate pair_2	78%	93%	0.025%	0.001%	2	1	33%	6.1
Candidate pair_3	69%	100%	0.025%	0.110%	3	1	33%	31.2

The construction of similarity tables is of particular importance for identifying matches on names that have been misspelt as a result of the data collection or capture process. The similarity table for forenames will also be supported by a name thesaurus that identifies matches on nicknames and established abbreviations[9].

This section considers how decision making regarding the match statuses of candidate pairs can be modelled with supervised training methods.

3.5.2 Matching with logistic regression

Where two records have been identified as a potential match from the similarity tables, a method is required to automate the decision whether or not to declare the two records as a match. We have tested the use of regression models to develop a prediction function to estimate the predicted probability of a pair of records being a match. The use of prediction functions based on statistical models can be found in many areas of research, including finance and biomedical sciences. However, there is limited literature available on the use of prediction functions based on statistical modelling of a match event.

A more conventional approach in score based matching is to use the agreement scores and their associated weights to assign an overall composite score or probability that the two records are a match. Based on incremental sampling of the score distribution and clerically checking the match statuses of candidate pairs, an upper and lower threshold is identified where all matches above or

[8] LSOA – Lower Layer Super Output Area – total population between 1,000 and 3,000 people, average 1,600 people. For an introduction to the different types of geography see http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/index.html

[9] See appendix C – it is intended that Beyond 2011 matching will incorporate nicknames into the similarity tables

below those scores can be automatically passed as matches and non-matches respectively. This is the basis of the formal mathematical model of probabilistic record linkage, known as the Fellegi-Sunter method (Winkler 2006; Fellegi & Sunter, 1969). This incremental approach to sampling is problematic given the anonymous research environment in which we are working. Furthermore, the remaining candidates that have scores lying between the two thresholds would ideally be sent for clerical resolution. Even if we were using identifiable data, this would entail a clerical matching exercise on a very large scale and would be extremely resource intensive. We have therefore had to refine automated matching by introducing a single score threshold for identifying whether pairs match or not.

Crucial to identifying a single score threshold is the way that an overall agreement score is calculated. The agreement scores between the matching variables need to be appropriately weighted to generate an overall score that can be determined as a match or non-match depending on whether it is above or below the single score threshold. As an initial exploration, we have tested the use of logistic regression to do this. Logistic regression is one of various supervised techniques that can effectively be 'trained' to determine whether a particular candidate is a match or not (Köpcke & Rahm, 2008). The predictor variables for the model consisted of the following:

Forename similarity score
Surname similarity score
Date of birth agreement code (4=full, 3=M/Y, 2=D/Y, 1=D/M, 0=none)
Sex agreement code (1=Yes, 2=No)
Postcode agreement code (4=full, 3=postcode sector, 2=postcode district, 1=postcode area, 0=none)
Forename weight (% frequency)
Surname weight (% frequency)
Name count (forename and surname combined)
Distance between LSOA centroids

We are currently working through the strategy to have access to a sample of unhashed candidate records in a way that does not compromise our policy to protect data by appropriate privacy and security safeguards. This sample would then be used to make a clerical decision (match or non-match) which is used as the Y variable for the logistic regression. The fitted model can then serve as a training dataset for the remaining candidate pairs. Logistic regression has been tested initially because it was seen to be one of the ways of developing a binary classification model which is based on a single score threshold. Firstly, the model coefficients calculated in the regression procedure serve as weights for each match variable to generate an overall predicted probability of a match. Secondly the regression equation can be used as a prediction function to automatically classify the remaining candidates as a match or not, depending on whether they are above or below the single threshold score ($p = 0.5$).

This technique has been applied to a number of test datasets with consistent results achieved. On training datasets the model is able to predict the match status in concordance with the clerical decision in approximately 97 % of cases. When the model equations are then run on an independent sample of pairs, which are then clerically matched for comparison, the agreement level is usually higher than 95 %. Some initial sensitivity analysis has been undertaken to identify whether the optimum threshold for classifying matches or non-matches is located somewhere other than at 0.5 in the score distribution. The early indications are that the optimum threshold can very depending on the sample that has been selected as training data. For a small number of samples drawn from PR to census matching, the optimum threshold score was always located between 0.45 and 0.55, resulting in a small increase in the percentage of correct decisions being

made. Further analysis based on the plotting of ROC curves[10] will further inform this research in the future.

3.6 Alternative methods for modelling decision making

We have not yet finalised the method for determining score based matches in Beyond 2011. Whilst logistic regression models derived from training data appear to be viable based on the evidence of research to date, there are a number of classification techniques that could be used in its place. It is intended that research will continue throughout the programme to identify the optimum method for classifying pairs into matching and non-matches. Alternative methods could include decision trees, Support Vector Machines, Bayesian classifiers or even the Fellegi-Sunter method if the problems noted above can be overcome.

4 Results of quality assurance

One of the essential requirements of matching research in Beyond 2011 is to establish the degree of quality loss that is associated with hashing the data. The challenges of running string comparison algorithms and the unavailability of clerical resolution makes it inevitable that there will be an increase in the rate of false positives and false negatives. Both of these error types could have a direct impact on the quality of population estimates, depending on the estimation method. For example, if we use a dual-system estimator, a 1 % positive bias would lead to a 1 % bias in the population estimates. For more information about estimation methods, please refer to 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory' (Paper M8).

4.1 Matching student records to the PR

Early in the Beyond 2011 Programme, we undertook a comparison exercise to match 10,000 student records from the HESA dataset to the PR. This exercise was undertaken as an initial piece of research to ascertain whether it was worth pursuing a linkage option given the challenges of linking anonymised data. Whilst there were some limitations with the design of this comparison study, the results were encouraging and suggested that the matching of two administrative data sources was viable in an anonymous research environment.

The relevance of this study was to ascertain whether administrative datasets could feasibly be matched to one another to obtain a SPD. A more detailed discussion of this concept can be found in the paper 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice' (Paper M7), and its use in estimating populations in the paper 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory' (Paper M8). It should be noted that while the matching quality here needs to be good, otherwise the creation of SPDs would not be feasible, it does not have to be outstandingly high, as a certain amount of under and over-coverage in the SPD can be adjusted for using the PCS.

Our matching approach was considerably under-developed at the time of this comparison. Only a small number of match-keys had been created at this point, and the matching algorithm outlined in section 3.4 had not yet been fully developed. This resulted in higher rates of error than could now be achieved.

[10] A receiver operating characteristic (ROC), or ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied.

Due to the data processing demands it was necessary to subset the PR to a manageable size, and this was achieved by randomly selecting 20 dates of birth for the study. The implication of doing this was to impose a caveat that all matches made in this study would have to agree exactly on date of birth.

In order to estimate for the number of false positives and false negatives it is useful to have a 'gold standard' matched dataset for comparison. This comprises a set of perfectly matched records between two datasets, where all of the true links have been identified and all of the true non-links are left unmatched. Whilst it is difficult to obtain a gold standard without a common identifier between the two sources, the limited size of the study and the pre-arranged restriction to agreement on date of birth made this a feasible proposition. The matching was first carried out using rule based and score based matching, followed by the clerical resolution of all the pairs in a widely drawn range of scores deemed not to be high enough to be certain matches and not to be low enough to be certain non-matches. In this way it was possible to create a matched dataset which it is reasonable to label as 'gold standard'.

4.1.1 Comparison study design

For the gold standard match, an exact match was run on three passes. The cumulative match rate after each pass is reported in the table below:

Figure 13 – Exact match passes for gold standard

Match pass	Exact match variables	Match rate (cumulative)
Pass (1)	Forename / surname / DoB / term-time postcode / sex	46.2%
Pass (2)	Forename / surname / DoB / domicile postcode / sex	60.8%
Pass (3)	Forename / surname / DoB / sex	83.5%

A score based method was then used to identify a further 160 automated matches increasing the match rate to 85.1 %. A residual set of 516 candidate pairs was then investigated by a clerical team to produce a final match rate of 89.9 %.

An estimate for the expected proportion of HESA students that are registered with a General Practitioner (GP) is not known. It is assumed that a significant number of foreign students may not have registered with a GP, and there may also be circumstances where UK students are not registered. It follows that a match rate of well below 100 % is to be expected. With the caveat that all matches have exact agreement on date of birth, it is assumed that the gold standard match has identified all of the possible matches between the two datasets.

One of the advantages of the student dataset for matching purposes is that both domiciliary and term-time address information is collected by the universities. This is very useful for the resolution of lags on the PR, where students are more likely to move further distances and not register with a GP in their student towns.

For the Beyond 2011 method, the first three match passes were identical to those used in the gold standard comparison. An additional two passes were run using match-keys based on initials, date of birth, sex and postcode district:

Figure 14 – Exact match passes for Beyond 2011 methods

Match pass	Exact match variables	Match rate (cumulative)
Pass (1)	Forename / surname / DoB / term-time postcode / sex	46.2%
Pass (2)	Forename / surname / DoB / domicile postcode / sex	60.8%
Pass (3)	Forename / surname / DoB / sex	83.5%
Pass (4)	Forename initial / surname initial / DoB / term-time postcode district /sex	86.0%
Pass (5)	Forename initial / surname initial / DoB / domicile postcode district / sex	87.6%

An algorithm that identified candidate pairs through the use of similarity tables was then run on the remaining 1,240 residuals. An additional 50 matches were identified using this approach producing a final match rate of 88.1 %.

4.1.2 Findings

Our method matched 88.1 % of the 10,000 HESA records, compared with a match rate achieved by the gold standard of 89.9 %. However, this does not mean that all the matches made by the Beyond 2011 methods were correct.

To evaluate exactly how well our method was performing it was necessary to make a direct comparison between the matches made by the gold standard and by the Beyond 2011 method. Since the gold standard match is assumed to be completely correct, those matches made by our method that also appear in the matches made by the gold standard are assumed to be true. These are labelled *true positive matches*. Those matches made by Beyond 2011 that do not appear in the gold standard match are assumed to be false positive matches, i.e. matches made that are not true. The matches made by the gold standard match that are not made by the simulation are false negative matches, because they are assumed to be true matches that Beyond 2011 failed to make.

It is then possible to calculate the *precision* and *recall* of the algorithm, which can be summarised as the algorithm's freedom from false positives and false negatives respectively:

Figure 15 – Calculating precision and recall

$$\text{Precision} = \frac{\text{True positive matches}}{\text{True positive matches} + \text{False positive matches}}$$

$$\text{Recall} = \frac{\text{True positive matches}}{\text{True positive matches} + \text{False negative matches}}$$

Precision for the Beyond 2011 methods was calculated at 99.0 % (8708 true matches out of a total of 8793 that were made) and recall at 97.5 % (8708 true matches out of a total of 8932 that were possible).

It should be noted however that the conditions under which the comparison was made may have generated some bias. Firstly, it is certainly not the case that for all true match pairs, date of birth agrees. Due to recording and transcription errors, inconsistencies may occur between datasets.

Secondly the clerical exercise undertaken as part of the gold standard was conducted by the Beyond 2011 team, and was therefore not an independent comparison.

4.2 Comparison with census quality assurance matching

A key part of the quality assurance process for the 2011 Census was the assessment of the age-sex census population estimates compared with administrative data. The Census QA matching team undertook a large scale matching exercise for 36 local authorities, using a combination of exact, probabilistic and clerical methods, including the clerical examination of difficult cases to determine whether or not they are matches. A considerable focus of the work was based on a sample of local authorities selected for matching the PR to the 2011 Census, initially concentrating on matching PR records within the same LA as the census. This work has been used as the basis for a second study. The comparison was based on taking records from the PR in the (approximately) 1 % of postcodes that were sampled in the 2011 CCS, and attempting to match them to the wider LA on the census.

The motivation behind this second study was to examine how our matching approach performs in the task of matching an administrative data set to a survey. After administrative data sets have been combined in some way to become an SPD, the SPD can itself be viewed as an administrative data set since all its records are derived from administrative data sources. As discussed in the paper 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory' (Paper M8), this has then to be matched to a survey, the PCS. In the study reported here, the PR was a proxy for an SPD, and the census, limited to the postcodes that formed part of the CCS sample, is a proxy for the PCS.

The matching of the SPD to the PCS is likely to be a vital part of the our population estimation process, and as pointed out will require high matching accuracy if it is assumed that individual level matching is required for use in the estimation process, such as that required for a Dual System Estimator (DSE).

The methods used by the Census QA team included exact matching, probabilistic matching (Fellegi-Sunter), clerical resolution and clerical searching. As explained in sections 4.1and 4.4.2, the last three of these do not appear to be feasible when the data is anonymised prior to record linkage, so the comparison provides a good way of measuring the level of quality loss incurred when data are anonymised and our method is used.

4.2.1 Comparison study design

The following local authorities were selected as the basis for comparison; Birmingham, Newham, Lambeth, Southwark, Westminster, Powys, Aylesbury Vale and Mid Devon. The latter three local authorities were included to gauge the accuracy of the matching algorithm in areas of low population churn. The others were selected on the basis of being city areas with high population churn. In such areas, the likelihood of definitional differences with administrative sources is greater as a result of the increased impact of lags in updating details. We ran our matching approach on these eight local authorities.

The linked records were compared with those originally made by the Census QA team. These cannot quite be described as a gold standard comparison for Beyond 2011 matching. However, we had invaluable ongoing support from the Census QA clerical matching team, with every discrepancy between the matches made being carefully evaluated using all the census data sources available to them. Where the original Census QA match decision was reversed, this decision was recorded and became a part of a new count referred to here as the *updated Census*

QA matches. Unlike the study described in section 5.1, the final clerical decision of the matches was made by the Census QA team, which is therefore a genuinely independent comparison.

Within the updated Census QA matches there are likely to be very few false positives. Approximately 100 were found over the eight local authorities when the original Census QA matches were compared with our matches and these were examined. It is assumed that most of the false Census QA matches were discovered by this process and that therefore very few remain among the matches that are common to both sets.

It should be noted that even after updating, Census QA matches are likely to have missed a very small number of true matches. Further work has been carried out for the local authorities of Westminster and Mid Devon to estimate this number. This involved using a much more relaxed blocking strategy[11] to identify any other matches that were not already included in the updated Census QA matches. As expected, these were found to be in very small number, with an additional six matches found in Mid Devon and 18 in Westminster.

In view of the above, the number of the updated Census QA matches is best viewed as a lower bound for, but reasonably close to, the true gold standard.

4.2.2 Findings

Figure 7 shows the match rates (as a percentage of PR records) for the updated Census QA matches and the Beyond 2011 matches. It also shows the rate of false positive error (as a percentage of all the matches we made) and false negative error (as a percentage of all updated Census QA matches made). A table reporting the number of records matched, on which these rates are based, is provided at appendix D at the end of this report. A percentage breakdown of matches made at match pass by the Beyond 2011 algorithm is also provided at appendix D.

Figure 16 – Table of match results - Census QA comparison

LA	Census QA match rate	B2011 match rate	B2011 false positives	B2011 false negatives
Birmingham	82.0%	80.2%	0.4%	2.5%
Westminster	65.1%	63.6%	0.4%	2.7%
Lambeth	64.0%	62.9%	0.5%	2.2%
Newham	68.3%	66.5%	0.5%	3.0%
Southwark	66.3%	64.3%	0.4%	3.3%
Powys	94.3%	92.9%	0.2%	1.7%
Aylesbury Vale	89.9%	89.1%	0.2%	1.1%
Mid Devon	88.6%	88.3%	0.6%	0.9%

The comparison exercise shows that our matching methods have the capacity to match administrative data to a survey despite the challenges imposed by hashing the data. Indeed, considering the simplifications imposed on the data in the study to match student records to the PR, it has performed rather better than for that study.

[11] Blocking attempts to restrict comparisons to just those records for which one or more particularly discriminating identifiers agree.

The closeness of the updated Census QA matches to a gold standard makes the columns for false positives and false negatives indicative of the degree of quality loss that is associated with hashing the data. These levels of error are considerably higher than that currently achieved with census matching, which are estimated to be less than 0.1 % for false positive and 0.25 % for false negatives[12]. However it should be noted that the 2011 Census was designed to be matched with the CCS, and that targeting such accuracy when matching administrative sources to a coverage survey would be unrealistic.

Matching error is notably higher in areas where there is high population churn, namely the London Boroughs and Birmingham. Our method misses approximately two to 3 % of the potential matches that were available in these areas. At this stage however there are still many opportunities to refine the matching strategy to try and reduce these false negatives further. The figures reported above are an initial comparison between Beyond 2011 and Census QA matching approaches. Developing additional techniques to identify more of the true links available between the two datasets is the focus of ongoing work in matching research and is discussed in more detail in sections 5 and 6 of this report. It is inevitable that there will be a certain number of match pairs that cannot be identified as a direct result of the requirement to hash the data.

Further work is being done to reduce the size of the matching error thorough improvements to the matching methods and improving data capture, and to explore whether an adjustment for matching error can be made during the estimation process. We are continuing to build upon existing methods by investigating the links that have been made in error as well as those that have been missed.

5 Next steps

There are a number of avenues for further research that the Beyond 2011 team will be investigating. In particular, the focus will be on attempting to reduce the number of false negatives to a level that is compatible with our estimation approach, which is still under development. In Census QA matching, it is apparent that previous addresses collected on the census forms were a very useful piece of information to draw upon when making clerical decisions. To date we have done a small test to look at the impact of using previous addresses to improve matching rates, and this has shown some small improvements on matching error (approximately 0.6 % reduction on false negatives), and this research will continue.

More recently we have started some research around the concept of associative matching. Whilst a method has not yet been fully developed, the initial indications are that a number of additional matches might be made possible by drawing upon the strength of another individual who has already been matched within the same household, for example a spouse, partner, child or parent. This is particularly useful for people that have recently moved addresses, allowing for a relaxation on matching criteria at the individual level.

6 Conclusion

The comparison exercises that we have been undertaken so far have provided us with a robust framework to develop and quality assure data matching in Beyond 2011. Having access to a comparison from Census QA matching has been invaluable and will continue to provide an

[12] To be reported in 'An Assessment of the Quality of the Matching between the 2011 Census and the Census Coverage Survey' (ONS, 2013 – currently awaiting publication).

ongoing basis for improving the algorithm. We will continue to develop the algorithm and attempt to reduce the number of false positives and false negatives that are generated.

Based on the evidence currently available, the proposed approach to data anonymisation appears viable for Beyond 2011 matching. Whilst it is crucial to us to understand how the reported increase in matching error will impact on the accuracy of population estimates, it should also be noted that the efficiency of the methods that have been developed is very much a practical requirement of any future implementation. The scale of record linkage proposed in Beyond 2011 is considerably higher than the matching exercises that were undertaken for the 2011 Census. In practice, it would be very difficult and resource intensive to run similar clerical exercises between all of the administrative sources that are used to develop the SPD.

We are developing the estimation framework for the linkage model in parallel with matching research, and the tolerance for error in the matching process has not yet been fully understood. For SPD creation, which requires the matching of administrative sources, the tolerance for matching error is considerably higher than for PCS matching. Failure to match administrative records when constructing the SPD will result in under-coverage and a similar situation occurs with non-respondents to the census. For census estimation, non-response is adjusted for by matching census records to the CCS and using the results of the matching to apply a DSE. In principle, the same method can be used with administrative data that has under-coverage, however this will require a very high standard of matching between the SPD and PCS, where tolerance for error is much lower. For the 2011 Census, the error rate for false positives was estimated to be less than 0.1 % and less than 0.25 % for false negatives.

It is recognised however that census and CCS matching is very different from matching administrative sources to a coverage survey. Firstly, census and CCS data is collected over a 6-8 week period, thereby increasing the likelihood that respondents will be recorded correctly at the same place on both sources. This cannot be controlled for with administrative data, where lags in people's interaction will result in them being recorded in different places. Secondly, the data collected from census and CCS field exercises were designed to be matched, whereas administrative data is collected for operational purposes without any design consideration for record linkage. Considering these factors it can be assumed that the error rate for false positives and false negatives will increase when matching the SPD to PCS. Whilst our expectation is to lower the error rates nearer to those achieved with census and CCS matching, a realistic target for false positive and false negative errors in the Beyond 2011 context has yet to be developed. This target will need to be compatible with the estimation approach that we take, which is also still under development and which will explore whether an adjustment for matching error can be made during the estimation process. Further information about the estimation approaches we are considering is reported in 'Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory' (Paper M8).

Lowering the number of false negatives is our priority for future development and more recent research suggest that the false negative rate can be brought down further by using previous addresses and associative matching. Having trialled some of these methods as a continuation of the Census QA comparison exercise, we have already observed that they can substantially reduce the number of false negatives missed by the algorithm, with matching error for some of the trial areas already getting below 1 %.

It is also important to look beyond the matching algorithm as the only basis for improving the accuracy of linkages. Improving the quality of survey data by re-designing forms and the mode of collection has the potential to greatly improve matching accuracy in Beyond 2011. Current plans for the PCS include a greater emphasis on internet collection and the possible use of handheld

devices by field teams that have in-build validation checks for possible misspellings and inconsistencies, and to clean address entries.

Understanding bias will be a major focus of research throughout the remainder of the programme. The error rates reported in this paper are likely to be more prominent for particular sub-groups of the population, and this could have an adverse affect on small area estimates or age / sex disaggregations at the LA level. There is also a need to evaluate our capacity to accurately match records where the location of the individual is recorded in different local authorities. The Census QA comparison reported in this paper was limited to matching records within the same LA and we are currently in the process planning a comparison of records matched nationally.

We will need to plan carefully how we will measure the quality of our matching in the future. Currently the matching exercises undertaken by the Census QA team serve as an excellent basis for measuring the precision and recall of the algorithm. Whilst methods can be implemented in the anonymous research environment to measure precision and recall, they are unlikely to be as robust as the measures obtained in the Census QA comparison. The Census QA matching team used a process of clerical searching in an attempt to identify all of the possible matches that were available. This was important for making links between true match pairs that were too low scoring to be brought together as candidate match pairs for clerical resolution. As a result we have confidence that our comparison work with the Census QA team gives us a reliable estimate for the number of false negatives.

We are currently working through the strategy to have access to a sample of ‘unhashed’ records in a way that does not compromise our policy to protect data by appropriate privacy and security safeguards, and this sample may be used to train the models in the score based matching, and could be used to quality assure our methods. However obtaining measures of false negatives may still be problematic under the proposed approach. For individual records that are still left unmatched at the end of the algorithm, there will be no way of relaxing the blocking criteria to search the other dataset for all of the possible true matches. In theory, this problem can be overcome by using blocking techniques that ensure that all candidate match pairs are identified by the similarity tables and are therefore entered into the logistic regression. In this way, a measure of false negatives can be calculated by sampling from the algorithm decisions and observing the number of rejected matches that were made in error. However, the similarity tables by necessity need to have a cut off point for names and dates of birth that are considered similar. It is inevitable that some true matches will be not identified for the logistic regression, however it is expected that we can keep this number to a minimum.

Overall significant progress has been made to demonstrate that there are a variety of pre-processing techniques that can help to overcome the challenges of linking anonymised data. The results look very encouraging for the matching of records between administrative data sources as well as for the matching of survey data to administrative records. One future piece of research is to undertake another comparison study to ascertain whether matching two surveys will achieve similar results.

Research into the implications of matching anonymised data will continue, and new methods may emerge from research being undertaken outside of ONS and from the literature available. We will continue to subject research findings to methodological review and to communicate the work of the team to wider public interest groups.

Appendix A: Soundex algorithm

The most commonly used phonetic coding schemes are Soundex and NYSIIS (New York State Identification and Intelligent System). Initial research has focussed on the Soundex coding scheme. Soundex operates by converting a name string into a 4 character alpha-numeric code that consists of a letter followed by three numerical digits. The letter is the first letter of the name, and the digits encode the remaining consonants. Similar sounding consonants share the same digit so, for example, the consonants B, F, P, and V are each encoded as the number 1.

A demonstration of Soundex in operation is displayed in the table below. The surname examples that were illustrated in the previous section have been coded using Soundex with differing results:

Name	***Soundex version***
Johnson	J525
Johnston	J523
Smyth	S530
Smith	S530

When comparing Johnson and Johnston, the Soundex versions have taken on slightly different values, whereas the code for Smith and Smyth is indicative of a match between the names. Only in the latter case would agreement be indicated between the surnames once data have been hashed and transferred into the research environment.

The main advantage of using Soundex in Beyond 2011 is that it targets the standardisation of name variants that are phonetically similar. This is of particular use for matching data that has been collected by doorstep interviewers, as is the case with a survey like the 2011 CCS[13]. The strength of Soundex coding is most apparent where names have been misheard or unverified by the person collecting information. The resulting transformation will standardise names where they are phonetically similar. One of the limitations is that Soundex has principally been designed to identify similarities between Anglo-Saxon names. A large volume of matches between sources covering the England and Wales population will need to be made on names from a variety of ethnic origins and phonetic coding will be of limited use in these circumstances.

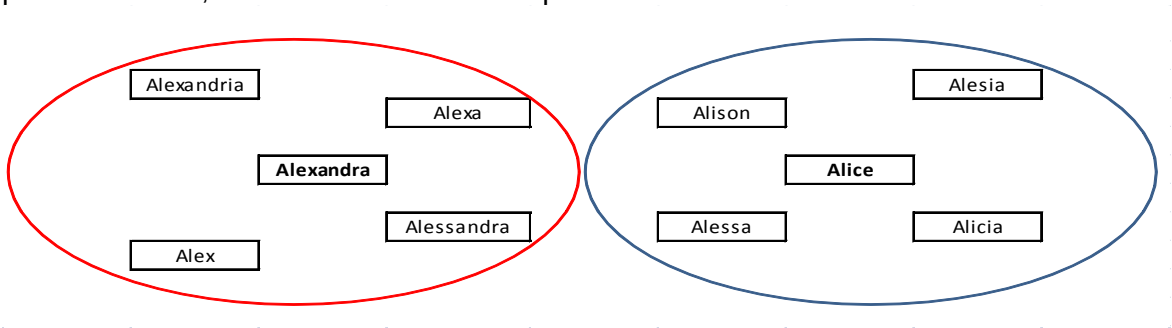
Soundex has been tested in Beyond 2011 matching, but was found to identify only a small number of additional true matches after exact matching. It may be used later in the matching algorithm to further reduce the number of unmatched residuals.

[13] The CCS was a post enumeration survey carried out after the 2011 Census in selected parts of the country to help assess the coverage of the population counted by the census. In considering administrative data based options, a similar coverage survey is likely to be needed, although the mode of collection may differ.

Appendix B: Name clustering

The method of name clustering involves the creation of central repositories for all forenames and surnames (separately) and undertaking edit distance functions between the names to identify those that are similar in string composition. A clustering algorithm is then used to form groups of names that represent similar name entities, which can then be assigned a cluster ID. For example, edit distance functions that have been undertaken on the following set of names: Alexander, Alexandria, Alexia…etc would be identified as belonging to the same group and would be assigned a cluster ID such as 'A10000'.

This approach has considerable benefits where there are misspellings (or recording errors) in the name variables. Data dictionaries will only identify matches between accepted and established name variants, whereas this approach can group names that contain erroneous spellings or data capture errors. The drawbacks with this method are processing time; the clustering algorithm itself requires a lot of processing power and iterations to complete the process of grouping names. In addition there are also likely to be a number of cases where the formation of clusters divorces the potential match, as illustrated in the example below:



In this case the true match was between the names Alessa and Alessandra. Due to the way the cluster has formed with Alexandra and Alice at the centre, the algorithm has separated both names into different clusters. The resulting cluster IDs that are assigned to each record will therefore not exactly match in the research environment. There may be ways around this issue by introducing proximity between clusters, for example assigning cluster IDs of A10000 and A10001, however this will inflate the risk of false positives.

This approach has been tested in Beyond 2011 with some success. However the subsequent development of similarity tables has superseded this approach on the basis that the scores derived facilitate the use of more sophisticated score based methods.

Appendix C – Name thesaurus

A name thesaurus (also known as data dictionary), in the context of name matching represent a form of lexical mapping whereby abbreviations, nicknames and other derivatives of names can be collectively grouped together or assigned to a 'root' name. We have considered the possibility of developing our own data dictionary for names by undertaking frequency analysis of all matched records on the census that have been resolved clerically and where the forenames or surnames do not exactly agree. Whilst this option may still be considered in future research, it is recognised that a significant number of matches within ethnic groups may not have been passed by clerical teams where they do not have sufficient experience of naming conventions used by different cultures. For this reason a specialist name thesaurus is being procured for the purpose of lexical mapping, the testing of which is commencing in the second half of 2013. Consideration will need to be given as to how the data dictionary will function in practice, the examples below illustrate that some name abbreviations will be derivatives of multiple root names:

Dan	Danny	**Daniel**	Danielle	Daniella	etc…	
Mo	**Mohammed**	Mohemmed	Mohummed….etc.	**...Morgan**	**Maureen**	**Morris**

In the above example the name Mo could in fact be an abbreviation of one of four root names. A singular one-to-one mapping process in these circumstances could inflate the number of false links where abbreviations have been assigned to the incorrect root name. Conversely an approach that would enable multiple lookups against the data dictionary may require very intensive processing, in the example of Mohammed, our research indicates that there are 79 different variants of the name – all of which will need to be run separately. Although a name thesaurus' can provide a list of potential matches for nicknames and spelling variants, there is limited capacity for resolving typing errors in the data. It is inevitable that the operational processes that underpin data collection and data capture will result in erroneous entries that cannot be resolved with a data dictionary. For this reason it is necessary to support the name thesaurus with an additional set of potential candidates that have been derived from the data sources themselves. The process of constructing similarity tables is outlined in section 3.5.1.

Appendix D: Table of match results – Census QA comparison

Number of records matched by LA: Beyond 2011 and Census QA

LA	PR Count	Census QA matches	Beyond 2011 matches	B2011 True Positives
Birmingham	21,313	17,482	17,101	17,038
Westminster	9,626	6,268	6,121	6,098
Lambeth	10,532	6,740	6,624	6,589
Newham	13,461	9,193	8,956	8,914
Southwark	9,993	6,627	6,429	6,405
Powys	1,648	1,554	1,531	1,528
Aylesbury Vale	2,732	2,455	2,435	2,429
Mid Devon	613	543	541	538

Percentage breakdown of Beyond 2011 matches made at each pass by LA

Match Pass	Birmingham	Lambeth	Southwark	Newham	Westminster	Mid Devon	Aylesbury Vale	Powys
Forename, Surname, DoB, Sex, Postcode	60.2	73.2	54.5	55.7	52.3	67.5	53.8	69.0
Forename initial , Surname initial, DoB, Sex, Pcode District	19.8	13.0	21.8	22.7	24.9	17.8	25.0	16.7
Forename bi-gram, Surname bi-gram, DoB, Sex, Pcode Area	4.2	0.7	2.7	3.6	3.3	2.4	1.6	1.2
Forename initial, DoB, Sex, Postcode	5.3	3.7	7.0	6.1	6.3	4.9	6.5	4.7
Surname initial, DoB, Sex, Postcode	3.5	3.5	4.4	4.6	4.9	2.3	4.1	3.4
Forename, Surname, Sex, Postcode	3.1	3.0	3.4	2.6	3.0	2.5	2.7	2.5
Forename, Surname, DoB, Postcode	0.3	0.4	0.2	0.3	0.2	0.2	0.4	0.3
Forename, Surname, DoB, Sex	0.1	0.6	1.2	0.1	0.2	0.5	1.0	0.4
Middle name, Surname, DoB, Sex, Postcode	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Surname, Forename, DoB, Sex, Postcode	0.2	0.0	0.1	0.1	0.4	0.0	0.1	0.0
Logistic regression	3.1	1.9	4.9	4.1	4.4	1.8	4.8	1.8
Total	100	100	100	100	100	100	100	100

Office for National Statistics

Glossary

Abbreviation	Meaning
GP	General Practitioner
HESA	Higher Education Statistics Agency Student Record
LA	Local Authority
L2	Lifetime Labour Market dataset
PCS	Population Coverage Survey
PR	Patient Register
NHS	National Health Service
NISRA	Northern Ireland Statistics and Research Agency
NRS	National Records of Scotland
SPDs	Statistical Population Datasets

References

Bachteler T, Schnell R, & Reiher J (2010) '*An empirical comparison of approaches to approximate string matching in private record linkage*', in *Proceedings of Statistics Canada Symposium*. Ottawa, Canada: Statistics Canada.

Beyond 2011 'Producing Population Statistics Using Administrative Data' *Methods & Policies Report (M6*), ONS (2013). Being published simultaneously with this report. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/methods-and-policies/index.html

Beyond 2011 'Producing Population Statistics Using Administrative Data: In Practice' *Methods & Policies Report (M7)*, ONS (2013). Being published simultaneously with this report. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/methods-and-policies/index.html

Beyond 2011 'Producing Population Statistics Using Administrative Data: In Theory' *Methods & Policies Report (M8*), ONS (2013). Being published simultaneously with this report. Available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/methods-and-policies/index.html

Churches T & Christen (2004) 'Some methods for blindfolded record linkage', *BMC Medical Informatics and Decision Making*, 4, pp 9

Fellegi I P & Sunter A B (1969) 'A theory for record linkage', *Journal of the American Statistical Association* 64, pp 1183-1210

Gomatam S, Carter R, Ariet M & Mitchell G (2002) 'An empirical comparison of record linkage procedures', *Statistics in Medicine* 21 pp 1485-1496

Karmel R, Anderson P, Gibson D, Peut A, Duckett S & Wells Y (2011) 'Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study', *BMC Health Services* 10, pp 41 doi:10.1186/1472-6963-10-41

Köpcke H, & Rahm E (2008) 'Training selection for tuning entity matching', *QDB/MUD*, 3, pp 12

Li B, Quan H, Fong A & Lu M (2006) 'Assessing record linkage between health care and Vital statistics databases using deterministic methods', *BMC Health Services Research* 6, pp 48 doi:10.1186/1472-6963-6-48

Lyons R, Jones K, John G, Brooks C, Verplancke J P, Ford D, Brown G & Leake K (2009) 'The SAIL databank: linking multiple health and social care datasets', *BMC Medical Informatics and Decision Makin*g 9, pp 3

Pacheco A G, Saraceni V, Tuboi S H, Moulton L H, Chaisson R E, Cavalcante S C, Durovni B, Faulhaber J C, Golub J E, King B, Schechter M & Harrison L H (2008) 'Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil', *American Journal of Epidemiology* 168, pp 1326-1332 doi:10.1093/aje/kwn249

Pang C & Hansen D (2006) '*Improved record linkage for encrypted identifying data*', in *Proceedings of the 14th Annual Health Informatics Conference* pp 164-168

Schnell R, Bachteler T, & Reiher J (2009) 'Privacy-preserving record linkage using Bloom filters', *BMC Medical Informatics and Decision Making*, 9, pp 41

Schnell R, Bachteler T, & Reiher J (2010) '*Private record linkage with bloom filters*', in *Proc. of Statistics Canada Symposium*, pp 304-309

United Nations (2008). Principles and Recommendations for Population and Housing Censuses. Statistical Papers, series M, No.67/Rev.2. New York

Winkler W E (2006) '*Overview of record linkage and current research directions*', in *Bureau of the Census*

Yancey W E (2005) '*Evaluating string comparator performance for record linkage*', in *Statistical Research Division Research Report* http://www.census.gov/srd/papers/pdf/rrs2005-05.pdf