Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 101-124

Methodology and Applications of Building a National File of Health and Mortality data

Leicester Gill
University of Oxford, Department of Public Health and Primary care
Unit of Health-Care Epidemiology, Oxford

Abstract: National collections of historical administrative and other health data can number hundreds of millions of records, with new data being added at the rate of tens of millions of records each year. Although improvements in computing and storage technology have to some extent kept pace with this accelerating growth in the datasets, there has been little development over the past few decades in the way in which probabilistic record linkage is undertaken, particularly in respect of the match acceptance thresholds and the clerical review processes, which are required to make decisions about matches which are doubtful.

This paper describes the major features of the Oxford Record Linkage Study (ORLS), and the developments in probabilistic matching methods and the use of intelligent and data mining methodologies to select potential links between pairs of records.

The ORLS linked file was developed using a collection of linkable abstracts that comprise a health region in the United Kingdom. The ORLS file contains 12 million records for 6 million people and spans 39 years. This dataset is used for the preparation of person linked health services statistics, and for epidemiological and health services research. The policy of the ORLS is to comprehensively link all the records rather than prepare links on an ad-hoc basis.

The ORLS have been developing improved techniques for deterministic and probabilistic linkage and developing algorithms for reducing the amount of clerical review, which is time consuming, expensive, and of variable quality. The methodology has been extended and refined for matching and linking other large UK government datasets, in particular the National Health Service Central Register (60+ million records), a number of disease and local authority registers, and more recently, for the development of a UK National File of Linked Hospital Episode Statistics and Mortality data. This file spans 4 years and currently holds 52 million records and will increase by 14 million records per annum.

Since the implementation of the Data Protection Act (1998) in the UK, all names and address have been stripped from the health files. Matching and linkage is undertaken using the national NHS number and other partial identifiers. The matching methodology described in this paper is for linking such datasets using various combinations of the partial identifiers.

Zusammenfassung: Nationale Datenbestände von historischen administrativen und anderen Gesundheitsdaten können in die Hunderte von

Millionen Sätzen gehen; die Zahl der jährlich hinzukommenden Sätze können in der Größenordnung von 10 Millionen Sätzen liegen. Obgleich Verbesserungen in den Rechner und Speicher Technologien mit diesem Wachstum einigermaßen Schritt halten konnten, hat es während der letzten Dekaden nur geringe Fortschritte in der probabilistischen Record Linkage Technik gegeben, besonders was das Festlegen von Schranken für die Match-Entscheidung und die manuellen Entscheidungsprozesse anbelangt, die bei zweifelhaften Matches anzuwenden sind.
Dieser Aufsatz beschreibt wesentliche Elemente der Oxford Record Linkage Studie (ORLS), die Entwicklungen bei den probabilistischen Record Linkage Techniken und den Gebrauch von intelligenten und Data Mining Verfahren zum Auswählen potentieller Matches.
Die ORLS Datei wurde aus einer Sammlung von Sätzen aus einem Bezirk der Gesundheitsverwaltung des UK entwickelt. Die ORLS Datei enthält 12 Million Sätze von 6 Million Personen aus 39 Jahren. Dieser Datensatz wird für die Herstellung von personenbezogenen Statistiken zur Krankenversorgung und für epidemiologische und das Gesundheitswesen betreffende Forschungen verwendet. Die Idee der ORLS ist es, einen umfassenden Datenbestand bereitzustellen, um nicht speziellere Datenbestände nach dem Bedarf spezieller Fragestellungen erstellen zu müssen.
Die ORLS hat verbesserte Techniken für deterministisches und probabilistisches Record Linkage entwickelt; auch wurde Algorithmen entwickelt, um die manuellen Entscheidungsprozesse zu verringern, die zeitraubend, kostspielig, und von unterschiedlicher Qualität sind. Die Methoden wurde erweitert und adaptiert für das Zusammenführen anderer großer Verwaltungsdatenbestände des UK, insbesondere das Nationale Health Service Central Register (mehr als 60 Million Sätze), eine Anzahl von Kranken-Register der lokalen Behörden und, vor kurzem, für die Entwicklung einer UK-weiten nationalen Datenbestandes mit Hospitalisierungs- und Sterbedaten. Diese Datei deckt vier Jahre ab und enthält zurzeit 52 Million Sätze; sie wird um 14 Million Aufzeichnungen pro Jahr wachsen.
Seit der Implementierung des Data Protection Act (1998) im UK wurden alle Namen und Adressen aus den Gesundheitsakten entfernt. Das Matchen erfolgt mit Hilfe der nationalen NHS Zahl und anderen Indikatoren von persönlichen Merkmalen. Die Matching Methode, die in diesem Artikel beschrieben wird, ist für das Zusammenführen solcher Datenbestände geeignet und arbeitet mit verschiedenen Kombinationen von partiell identifizierenden Merkmale.

Keywords: Exact, Probabilistic, Statistical matching, medical record linkage, data linkage, data fusion

1 Introduction

Administrative records and data linkage or data fusion techniques are becoming increasingly important in academic research and projects commissioned by central and local government. Record linkage is defined as 'the bringing together of information from two different records that are believed to belong to the same person, family or entity' (Gill, 2001). These records may come from a single dataset or multiple files. Wholly new and very large datasets are being created with direct relevance to health, performance and policy issues. There is, however, nothing new or complicated about data linkage. The value of record linkage has long been recognized in the field of medical studies. Gill (2001) reminds us that the Oxford Record Linkage project, originally used to link patient records automatically across hospitals in the Oxford area, dates from the early 1960s (Acheson, Truelove and Witts). The Office of National Statistics (ONS) Longitudinal Study, based on a 1% sample of census records since 1971 is linked into the national registers of births, deaths and new cases of cancer.

2 Current usage of linked administrative records

Although automatic or semi-automatic data linkage is not a new methodology, in the past few years central records have increasingly become computerised and systematised, and therefore indexing and linking schemes have been built into the systems for service and management purposes. Thus, many government departments have developed a unique numbering scheme at individual level that potentially allows individuals to be tracked across services and linked over time if, for example, there are repeat contacts at a later date. These are primarily for treatment, management and research purposes (i.e. to avoid duplication, prevent over-counting etc). Sometimes these records include full identifiers, which can be used as the search mechanism. While the inteniont is to track individual cases, the result has often been to build up impressive and comprehensive sets of data.

3 Medical Record Linkage

The greatest use of record matching and linkage has been in health studies, where the most frequent application is in searching large files of morbidity and mortality records. Other uses include the preparation of disease registers, the provision of additional health information for an individual, and purging files of duplicate entries (Gill, 2001).

The term record linkage, first used by HL Dunn (Dunn,1946, Gill and Baldwin, 1987), expresses the concept of collating health-care records into a cumulative personal file, starting with birth and ending with death. Dunn also emphasised the use of linked files to establish the accuracy or otherwise of the recorded data. Newcombe (Newcombe et al., 1959; and Newcombe, 1967, 1987 and 1988) undertook pioneering work on medical record linkage in Canada in the 1950s and thereafter, Acheson (1967) established the first record linkage system in England in 1962.

When the requirement is to link records at different times from different sources, in principle it would be possible to link such records using a unique personal identification number. In practice, a unique number was not generally available on records in the UK of interest in medicine and therefore other methods such as the use of surnames, forenames, dates of birth have been necessary.

The fundamental requirement for correct matching is that there should be a means of uniquely identifying the person on each and every record to be linked. The matching may be based on exact methodologies where there is a unique high quality identifier, which is universally available, fixed, easily recorded and verifiable. Few, if any, identifiers meet all these specifications, however ciphers and personal numbers can be generated that meet most of the criteria, the UK National Health Service (NHS) number and the national Insurance Number (NINO) are one such examples. In circumstances where unique numbers or ciphers are not universally used, obvious candidates for matching variables are person's names, date of birth, sex and other supplementary variables such as address postcode and sex, and the use of probabilistic or statistic matching methods are required.

Each person has anatomical and physiological characteristics that are unique to them, such as fingerprints, eye colour, retinal patterns, voice graph and DNA samples. Although some of these methods in medical practice are usually ethically unacceptable or therapeutically undesirable, over the past few years much work has been undertaken on the recording of DNA characteristics, fingerprints, voice prints, facial features, facial heat patterns, retinal patterns and gait. Some of these characteristics are being further developed in various countries including the UK, for visa application, national identification and passports.

Identifying numbers are often made up, in part, from stable features of a person's identification set, for example, sex, date of birth and place of residence, and so can be constructed in full or in part even if the number is lost or forgotten. In the United Kingdom, the NHS number is an arbitrarily allocated 10 digit integer, almost impossible to commit to memory and which cannot be reconstructed from the person's identifiers.

Matching and linkage in established datasets usually involves comparing each new record on the data file with existing records on the master file. The files are ordered or blocked in particular ways to increase the efficiency of searching. In a fashion similar to that used in the telephone directory, the matching software must be able to cope with variations in spelling and recording. Algorithms that emulate the 'see-also' method are used in the matching process, and the match is determined by the amount of agreement or disagreement between the identifiers on the incoming record with those on the master file record.

The reliability and efficiency of matching is very dependent on the way in which the initial grouping or 'file blocking' step is undertaken. It is important to generate blocks of the right size. If the blocks are too large, many unproductive matches may be made, and if the blocks are too small some matches may be missed since the corresponding records are in adjacent blocks.

3.1 Methods for matching records and the experience of the ORLS

There are three primary stages in linking records together. The first stage requires the data to be cleaned and formatted, and the potential match-pairs to be brought together for comparison by sorting the file into various orders. The second stage involves comparing the potential match-pairs to decide whether they belong to the same individual/entity. This would generally use a deterministic or a probabilistic matching method, although other matching algorithms have been developed in recent years. The third stage involves the collation of the person linked records into a person-linked file either by sorting or creating an index in a database system.

Dealing with typographical error is a crucial step in record matching. Where there are discrepancies between the two variables being compared, it is normal to use an algorithm that measures the differences in terms of the number of deletions, insertions, transpositions and substitutions to convert one string into the other. The traditional methods are based on modifications of the Knuth-Morris-Pratt (KMP) algorithm, such as the Winkler-Jaro and the ORLS algorithm (Gill, 2001). Newer developments include the use of the Bigrams, although in its orginal form the algorithm cannot compensate for character transpositions. (Porter and Winkler, 1999).

3.1.1 The exact method (also known as deterministic or, all-or-none matching)

To undertake exact matching, it is necessary for the records on both files to have a unique, universally available, high quality identifier. Examples of such identifiers in the UK are the National Health Service Number (NHSNUM) or the National Insurance Numbers (NINOs). In this method, the selected identifier is compared across records and a link is made where the identifiers agree exactly. At its most basic, the output of the match is straightforward: either the records agree or they do not. A less strict version of exact matching simply requires an almost-exact match, where the number of variables that agree are used to decide whether or not two records should be linked. Where a unique identifier does not exist, but there is a group of identifiers which may be combined to form a composite identifier, the exact method can still be considered. The problem arises where a similar combination of identifiers can be generated by two different individuals, for example, in the use of date of birth, sex and postcode, young, same sex twins would generate identical identifiers. It is recommended that other identifiers are also used for confirmation of the match when using the exact method.

3.1.2 The probabilistic method

The probabilistic method is used when data and master files contain errors and omissions or where there is no unique identifier. This type of matching is based on the assumption that no single match between variables common to the data or master file cannot identify an individual or entity with complete reliability. Instead, the probability that two records belong to the same individual - rather than matching by chance or due

to coding error - is estimating by calculating probability weights that indicate how powerful a particular variable is in determining whether records are from the same person. The probability of a match increases when a greater number of identifying variables from the two datasets are alike. Variables that might agree by chance in unlinked record pairs are features that do not divide the population into many subclasses, for example, sex or marital status. Date of birth, however is far more suitable for use on probabilistic matching since it will divide the file into 75*365=27375 classes. With the probabilistic method there is a three-part classification of linkages: true match, definite non-match and possible or query match.

Other matching methodologies includes keyword matching that works by pattern matching the characters in the data record to those on the master file records. Rule based or Boolean search is a more sophisticated method that allows the user to combine keywords with Boolean operators such as AND, OR and NOT. These methodologies have been developed and refined for use in the search engines currently in use on the Internet (Gill, 2001)

3.1.3 Determination of match threshold

In exact matching the decision is clear-cut, either the records match or they do not. In probabilistic or statistical matching a number of approaches have been developed for the determination of the match threshold. The methods are based on the summation of the individual match weights or using a combination of Boolean rules (Gill, 2001). Over the past fifteen years, ORLS (Gill et al, 1993,1997) have developed an approach in which a two dimensional orthogonal array is prepared, in which the algebraic sum of the names weights form the 'X' axis and the algebraic sum of the non-names weights form the 'Y' axis. The results of previous, clerically scrutinised matches are stored in the cells. These empirical probabilities are further smoothed across both axes using regression and cubic-spline methods (Hays, 1974). Over 400,00 record pairs were rigorously checked by very experienced clerical staff and three counts were stored in each cell in the matrix, referenced by summed names weight and summed non-names weight. The counts are: the total number of matches for that particular cell, the number of good matches, and the number of non-matches. When two records are compared in a subsequent matching run, the probability weights are calculated and summed, the array is interrogated, and the decision made whether the two records match, or do not match, or should be sent for clerical scrutiny..

3.1.4 Results of matching

The matches made using a computer are prone to two types of errors: false negatives (the procedure does not identify the matching record, although it exists) and false positives (an erroneous match is produced). In the trade-off between the two types of errors most researchers prefer to allow more false negatives than false positives. To measure the extent of mismatch, one possibility is to extract a random sample and check the accuracy of linkage by comparing it with the results of a manual review. A second option employs assumptions regarding the dependency of mismatch on the amount of

potential matches before the linkage decision is taken, and the estimation of error on the basis of the actual frequency of ambiguous matches (Blakely and Salmond, 2002).

4 Types of data linkage

Many types of record linked files have been prepared over the past 50 years, and include the following datasets:

- Administrative data consisting of names and addresses
- National health and mortality records
- Development of cancer registries and other disease registers
- Census and other large file development
- Cross matching of national registration files, for example the NHS Central Registry

In this paper only the hospital episodes and mortality linkage will be described.

4.1 Administrative data to administrative data at individual level

Linking administrative data at individual level takes two principle forms. First, the methodology could involve linking extracts of one type but from different time points to one another at an individual level to create a longitudinal database. An example of this would be using NHS Number (encrypted or perturbed in a standardised way) to link together extracts of health-care over a number of years to track geographical movements or movements in and out of the system, or to track an individual through many healthcare systems. Secondly, the scope could be extended by linking data extracts either cross-sectionally or longitudinally.

Longitudinal analysis would require a matching variable, typically a NHSNUM or NINO, although in principle, within government or for those working for government, it would be possible to match using the type of matching techniques described in detail by Gill (2001) where names and addresses or other identifiers may be used.

4.2 The Oxford Record Linkage Study file

Perhaps one of the most well known examples of record linkage in the health and vital registration is, The Oxford Record Linkage Study (ORLS). This study consists of computerised, anonymised abstracts of records of hospital inpatient care, births and deaths in the four counties of the former Oxford health region between 1963 and 1999 (a population of about 2.5 million). Initially it was hoped to link these data together using the NHS number, but after discovering that only a minority of records contained this information it was decided to use a probabilistic method, adapting the methodology

developed by Howard Newcombe at Statistics Canada, and including the NHS number among the identifiers.

Although resolving the problem of a lack of unique identifier, such as the use of a person's identifiers, means that preserving individual confidentiality is a more difficult task, since the names and addresses are included in the matching algorithm and deleted only after the linkage was accomplished. The ORLS data file is split into two files, the first file contains only those variables that will be used for matching, and the second file only those variables that will be used for analysis. The matching file does not contain any clinical data, and the statistics file any of the matching data. Some data, for example, date of birth, sex and postcode are required for analysis. These variables are aggregated up to age or local authority district level to minimise any attempt to identify an individual from the ensuing analyses.

The records on the ORLS files are encrypted, only the anonymised record header is left un-encrypted to support sorting and file manipulation. During the match run, each record is decrypted when loaded into the memory of the computer, and after matching is completed the record is permanently deleted from memory. All the output files are encrypted. The print files are decrypted at a later stage, printed under clerical supervision, and stored in locked cabinets. After clerical intervention has been completed, the paper files are shredded and pulped by security approved recycling agencies. Since October 2001, all the names and addresses have been removed from all the files and backup systems and permanently destroyed.

The ORLS person linked file was used by Goldacre et al. (2000) for the study of co-occurrence of several diseases. Goldacre highlights the utility of the ORLS for medical research, advocating the importance of detailed health care statistics, accessible for the research community. For other applications of the ORLS, see: Goldacre et al. (2002); Roberts and Goldacre (2002); Goldacre et al. (2003).

4.3 Scottish health file

Similarly the Scottish Record Linkage System, which has been built up over the last ten years or so, makes it possible to track event histories of NHS patients in hospitals in Scotland. The system is being used for the analysis of episodes, stays and patients, analysis of readmission rates, and analysis of mortality and the modelling of outcomes. Again, probabilistic matching is used because a unique patient identification number has not been in general use. Given the success of the Scottish Record Linkage project in following the event histories of patients over ten years, Walsh et al. (2001) highlight the importance of electronic patient records and identification numbers.

4.4 Studies on matching the national HES dataset with the ONS mortality dataset (work of the NCHOD project)

The National Centre for Health Outcome Development (NCHOD) is a unit within the University of Oxford, Unit of Health-Care Epidemiology (UHCE) and has been commissioned by the UK Department of Health (DH) to study the long-term trends in hospital admission rates for individual specialties and clinical conditions; age, period

and cohort effects in the incidence of peptic ulcer and of fractured proximal femur; postoperative mortality (deaths in the first 30 days after surgery) and other adverse outcomes of care, suicide risk after discharge from psychiatric care, associations between different clinical conditions, (for example colon cancer risk after cholecystectomy, breast cancer risk after abortion), the use of hospital care by the elderly. To complete this work, NCHOD was further commissioned to undertake a pilot study for national record linkage of hospital statistics and mortality data covering the period 1998-2002.

The pilot study was designed to match and link records from the national hospital inpatient episode statistics dataset (HES) against themselves and with records from the mortality dataset provided by the Office of National Statistics. The aim of this study was to gain insights into the matching rates that may be generated using a number of different combinations of matching variables from each of the two datasets.

The hospital episode datasets (HES) were supplied by the DH and covered the period 1st April 1998 to 31st March 2002, and contained 48,940,133 records. The mortality dataset was supplied by the Office for National Statistics (ONS) and consisted of extracts of death registration data for the period 1st April 1998 to 31st January 2003, and contained 2,614,576 records.

The mortality records were reformatted so that the corresponding fields on the ONS mortality record and on the HES record were stored in the same character position in the record. Two unique and arbitrary check-digited numbers were appended to each record on the file. The first, a serial or accession number, uniquely identifies the record, this number is never ever changed, and is used as the major key in all the record identification and merging processes. The second number functions as a person number. Initially each record is given a different person number, and after the matching and linking has been completed, all records that belong to the same person will have the same person number copied across them. A series of tags were also added to each record to track the versions of the coding systems used in the record.

The unmatched HES-ONS file was split into two files. The first file, referred to as the master matching file (MMF) contains only those variables that are used for matching, and does not contain any clinical or other statistical data. This file is encrypted and stored on a different computing system to the statistical file (see below). Some variables, for example, date of birth, sex and postcode are required for analysis and have been aggregated up to age or local authority district level to minimise any attempt to identify an individual from the ensuing analyses.

The second file, referred to as the statistical analysis file, contains the variables that are not used for matching but are used for statistical analysis. The two files of HES records are linked together using the unique accession number. The method of splitting the identifying information from the clinical information into separate files is good practice, and meets the requirements of the United Kingdom Data Protection Act (1998).

4.4.1 Data quality and editing

A series of computer runs were carried out to check the frequency of the matching variables, and to identify the erroneous values and outliers. Records that had such errors were tagged to prevent the erroneous variables being used for matching. The variables

contain a variety of errors and omissions especially where they are recorded as 'not known' or 'had not been collected'. These default values were identified, corrected where possible and the records tagged.

NHS Number. The UK NHS number is 9 digits long together with a check digit. This internal check digit is validated using the Modulus 11 algorithm that will detect the transposition of two adjacent digits (a common error where humans are involved) and 91% of random errors where two or more digits are invalid.

Date of Birth. The date is entered and stored in the European format, DDMMCCYY. Two default dates have been used where the date of birth had been recorded as 'not known', and on HES records this field had been set to: 15/10/1852 or 01/01/1901. The 1852 date can be filtered out since only dates later than 1875 are accepted for matching, however the 01011901 date does cause problems since it overlaps with records having the genuine date of birth of 01/01/1901. Other problems have been identified where the DD and MM have been transposed, as have the DD and YY

Sequence of dates within a hospital episode of care. Problems were found in the sequence of dates within an episode. All of the dates in an NHS Trust hospital episode should be in a logical date sequence, starting with the start of episode or date of birth and ending with end of episode or date of discharge/transfer/death. Any overlaps of dates in a multiple-episode spell can only be rectified after the episodes have been correctly matched and linked together.

Sex (gender). In a small number of records (n=75,569, 0.2%), the sex codes have been set to not known. The value of the sex code can be borrowed across all records in a matched person set when matching has been completed.

Postcode. The postcode is supplied as a proxy for usual address (which is not supplied), and consists of an 8 character field. The first part of the code (inward code) is left-justified in the field, and the second part of the code (outward code) is right-justified in the field,

for example OX37LF is stored as OX3∇∇7LF (Note: ∇ = space characters)

The first two letters of the inward code were carefully checked with the codes issued by the UK Royal Mail, since only about 120 pairs out of the possible 256 letter pairs are used. Any postcodes with initial letter pairs not on the approved list were set to not known within the matching software.

The four years of HES (48 months) contain both finished and un-finished consultant episodes covering the years 1998 to 2002 and five years (58 months) of mortality records covering the years 1998 to 2003. The percentage of complete and valid NHS numbers on the hospital file increased from 74.1% in 1998 to 84.1% in 2001, with the four year average being 81.9%. The date of birth was 99.7% complete although the two types of invalid date amounted to 0.26% of the file. The postcode was available on 96.4% of the file of which 3.15% were for postcodes either not known or outside the UK. The percentage of NHS numbers on the mortality file is very high at 99.8%. The death certificates were coded using the computing systems and manual cleaning methods at the National Health Service Central Register (NHSCR) in Southport. The date of birth, sex and valid postcode are 99.7% complete.

4.4.2 Matching the national file using the NHS Number

The first series of runs involved matching together all record pairs that have the same NHS number. The file was split into two parts, those with valid NHS numbers (n=42,241,828, 81.93%), and those without NHS numbers (n=9,312,881, 18.07%). The second part was not matched at this stage. The matching file was sorted into: NHS number, date of birth, sex and postcode order. All records having the same NHS number were collected into a block and matched against each other using a recursive method, that is, matching first with all the remaining records in the block, second with the remaining records, and so on. The exact method of record matching was used, although it was augmented by a set of Boolean logical rules, this permitted the minor variations in date of birth and postcode to be taken into account. Each block was matched eight times using various combinations of the matching variables (Runs 1-8), and then further matched using a random sequence of Runs 1-8. It was found that there were slight differences in the overall match rate, depending on the order in which the runs were carried out and the above procedure reduced these to very small values. The results of all the matches in each block were 'OR'ed' (logically added together) and checks made on the overall match for consistency.

Table 1: Matching the 4 Year National File using NHS number as the main key

Matching variables:	NHS	DOB	SEX	PCD	PRO/LOPATID	
Run 1:	Y	Y	Y	Y	N	(4 from 5)
Run 2:	Y	Y	Y	N	Y	(4 from 5)
Run 3:	Y	Y	Y	N	N	(3 from 4)
Run 4:	Y	N	Y	Y	N	(3 from 4)
Run 5:	Y	Y	N	Y	N	(3 form 4)
Run 6:	Y	N	Y	N	Y	(3 from 4)
Run 7:	Y	Y	N	N	Y	(3 from 4)
Run 8:	Y	N	N	Y	Y	(3 from 4)

We found that the NHS number had been occasionally issued by the hospital Trust to two different people, it is therefore essential to rigorously check that two records having the same NHS number do in fact belong to the same person. Similar checks were carried out on the other variables in each record. The logical rules used for checking the match are shown in Table 2. The two records were considered to be correctly matched together when the set of logical of rules were found to be TRUE. The following rules are presented in decreasing order of reliability. The rules cannot be used where some of the identifying variables used for the verification may be absent or only partially present on either or both of the records in the pair.

Table 2: Logical rules to be applied when matching records together

NHS Number.	*	Exact match only on the 9 most significant characters. Check for default NHS numbers (e.g. 222222222).
	*	Check whether the same NHS number has been recorded for both the Mother and her baby.
Sex	*	Both records should have the same valid SEX code.
Date of Birth.	*	The year of birth should be between 1875 and current year. Exercise caution in using 01011900 or 01011901 since both of these dates have been used as defaults for not known. Caution should also be exercised in using dates of type CCYY0101.
	*	Some dates have been entered in USA format e.g. MMDDCCYY.
	*	Some dates have the DD and MM swapped while others have the DD and YY swapped. These are checked at run time and all possible transpositions are tested. In the comparison of two dates, the following discrepancies between the two dates have been found: Days: 1,2,3,4,5,10 days discrepant Months: 1,2,10,11,12 months discrepant Years: 1,2,3,10 years discrepant
	*	Record-pairs in which the two years of birth are more than 15 years apart will not be matched together. This will prevent Mother/Daughter and Father/Son linkages.
	*	Partial matching can be undertaken using any two parts of the three components of the Date of Birth.
POSTCODE.	*	Postcode must be complete and within the valid range.
	*	Check for default postcodes like OVERSEAS etc.
EPISODE	*	Accept date of discharge after date of death by up to 3 days
DATES	*	Wrong episode date recorded

The match rates for the NHS number match are shown in Table 3. The first block of counts shows 33,376,479 (31,761,841+1,614,838) records matched to another exactly with the same NHS number, date of birth and sex. The second block shows that 381,483 records matched to another using the NHS number and two or more parts of the date of birth, postcode or the PROCODE/LOPATID composite variable.

Examples of two or more people being allocated the same NHS number were found, indeed one NHS number was shared by 171 different people (independently verified using date of birth, sex, postcode and LOPATID). There were 798 pairs in which both the mother and the child had been allocated the same NHS number, and 49 pairs where the father and son were given the same NHS number. This error is easy to detect since the years of birth are more than 15 years apart. A total of 2019 cases were found in matching ONS mortality to HES for elderly patients where there were changes in year of birth, due to bad recording or poor memory.

Where two records have been successfully matched together and all the validity checks have been satisfactorily completed, the person numbers on the two records are compared, and the lowest number person number of the pair is copied across both records. This process is repeated on all the record pairs in turn. This method ensures that all the records in the same NHS number block that have matched together, receive the same person number. Any records in the block that have not matched at this time retain their original person number.

Table 3: Counts of records matched using the NHS number

Total HES records	=	48,940,133
Total ONS records	=	2,614,576
Total HES and ONS records	=	51,554,709
Total records with a valid NHS number (81.9%)	=	42,241,828
Numbers that failed with invalid NHS number (18.1%)	=	9,312,881
Number of record pairs that matched with another exactly where		
Number of HES records	=	31,761,841
Number with same NHS numbers	=	31,761,841
Number with same exact DOB	=	31,634,794
Number of ONS mortality records	=	1,614,838
Number with same NHS numbers	=	1,614,838
Number with same exact DOB		1,598,689
Number of record pairs that matched with another where a combination of the other matching variables was used	=	381,483
Number of HES records	=	322,351
Number with NHS numbers	=	322,351
Number of DOB+other fields	=	322,351
Number of ONS mortality records	=	59,132
Number with NHS numbers	=	59,132
Number of DOB+other fields	=	59,132

4.4.3 Matching using the Date of Birth, Sex, Postcode (DOB/SEX/PCD).

The second series of matching runs involved matching all record pairs that have the same date of birth, sex and postcode (DOB/SEX/PCD). The input file contained all the records that had previously been matched on NHS number together with the file of that did not contain an NHS number. The file was sorted into: date of birth, record type, sex, postcode and NHS number order. All records having the same DOB/SEX/PCD were collected into the same block. The records in the block were matched against each other using a recursive method. The deterministic or exact method of record matching augmented by a set of Boolean logical rules designed to accept the permitted variations in the date of birth and postcode variables.

Each block was matched three times using various combinations of date of birth, sex, postcode and the persons hospital record number (Runs 1-3), and then further matched using a random sequence of Runs 1-3,. The order in which the matches were carried out are shown in Table 4: Records that did not have a valid postcode were excluded from the run, together with those records with dates of birth before 1875 or had the date 19010101. Records with a date of birth 19010101 were processed in a separate run.

Since there are errors and omissions in the composite matching variable DOB/SEX/PCD, it is essential to verify that the two records do match, and it is good practice to use other matching variables: for example, NHS number (if present on both records in the pair), GP practice code, dates of the start and end of the episode and date of death.

Table 4: Order of matching the DOB/SEX/PCD blocks on the 4 year file

Matching variables:	NHS	DOB	SEX	PCD	PRO/LOPATID	
Run 1:	?	Y	Y	Y	Y	(4 from 4)
Run 2:	?	Y	Y	Y	N	(3 from 4)
Run 3:	?	Y	Y	N	Y	(3 from 4)

Since twins below the age of 18 normally live in their parent's house, that is, have the same postcode, and the same date of birth, it is highly likely that false positive matches would be generated, especially where they are both male or both female. Normally, twins are allocated adjacent NHS numbers or hospital numbers by the national or local systems. Examination of either of these variables to determine if the two values are adjacent is a powerful technique for resolution of these erroneous matches.

The match rates for the DOB/SEX/PCD match are shown in Table 5. 35,083,755 records matched to another exactly with the same DOB/SEX/PCD and the matches were cross-checked using the NHS number.

Table 5: Counts of records matched using the Date of birth, sex, postcode

Total HES records	=	48,940,133
Total ONS records	=	2,614,576
Total HES records with a valid date of birth/sex/postcode (97.4%)	=	48,791,986
Total ONS records with a valid date of birth/sex/postcode (99.9%)	=	2,614,574
Number of records that matched with another exactly where:		
the records had a valid DOB/SEX/PCD	=	35,083,755
Number of HES records	=	35,083,755
Number with a valid NHS number	=	30,134,806
Number of records with valid DOB/SEX/PCD	=	35,038,575
Number of ONS mortality records	=	1,478,076
Number with a valid NHS number	=	1,472,164
Number of records with valid DOB/SEX/PCD	=	1,475,119

Examples of two or more people sharing the same DOB/SEX/PCD were found, in some cases the pairs represented twins and as explained above, these cases were easily identified and corrected. In other cases, the errors were due to rounding in the day and month of birth, or parts of the postcode being set to spaces. The first two letters of every postcode were checked against the list of genuine postcodes issued by the Royal Mail, but no attempt was made to correct the file for changes in postcode. For a file covering a longer period of time some method of updating the postcodes must be incorporated into the process.

Where a pair of records have been matched together and all the validity checks have been completed satisfactorily, the person numbers on the two records were compared and the lowest number then copied across both records. This process is repeated on all the record pairs in the block of records in the block. This method ensures that all the members of the DOB/SEX/PCD block that have matched together receive the same person. This process also brings together those record pairs that were matched in the NHS number stage and also matched in the DOB/SEX/PCD stage. Any records in the set that did not match at this time retain their original allocated person number.

4.4.4 Corrections and updates applied to the National 4 year file

The corrections and updates were merged together and applied in the same computer run to ensure that any logical inconsistencies could be trapped and corrected. In this run, the following corrections were applied, and the records updated:

All the very large blocks were printed out and clerically scrutinized. Any reords for people who had been wrongly matched, were separated into individual records and each given a new person number, however all the records retained their original accession number. These records were resubmitted to the matching process.

Identify and possible separate any wrongly matched twins and other multiple births. Where it is possible to identify the successive records for each twin, the person numbers are copied onto the appropriate records. Where this is not possible the records are given a new person number. As before, all the records retain their original accession number.

Identify and separate any persons with two or more ONS mortality records. Checks are carried out on the sequence of dates of event, and the best match between the HES record and the ONS record was chosen clerically. The ONS mortality record that did not match to the HES record is given a new person number and the record is re-submitted to the matching process

4.4.5 Analysis of all the matched pairs

The linked file was analysed by taking all matched pairs of records that have the same person number. Examination of these pairs will provide statistics on those pairs that,

- matched together (good matches)
- did not match together when they should not (non-matches)
- matched to the wrong person when they should not (false positives, Type II errors)
- did not match together when they should match (false negatives, Type I errors).

Since a hospital spell can consist of several episodes of care, only the first episode in each spell i.e., the admission record was used for these analyses since it was assumed that the identification details would have been copied from the first record in the spell through all the other records in the same spell.

In Table 6, is shown the number of person record pairs created. The counts for the HES-HES and the HES-ONS linkages are shown separately. There are 68,150,786 pairs of HES-HES and 4,812,216 HES-ONS. Analysis of the pairs shows that 81.6% of the HES-HES pairs matched with a genuine NHS number on both members of each pair while the corresponding figure for HES-ONS is 85.5%

Table 6 Count of the NHS number in the matched pairs

	HES-HES	HES-ONS
Number of record pairs	68,150,786	4,812,216
Same valid NHS number on both records	55,615,596 (81.6%)	4,114,600 (85.5%)
Different (but valid) NHS on both recordsm	129,375 (0.19%)	8,746 (0.19%)
Record pairs with blank/invalid NHS number	12,405,815 (18.2%)	688,870 (14.3%)

Examining the pairs with a valid NHS number on both members of the pair shows that the number of matches based on NHS number is 98.9% (HES-HES) and 96.7% (HES-ONS) respectively. These totals are shown in Table 7.

Table 7: Count of matched pairs containing NHS number

	HES-HES	HES-ONS
Number of matched pairs	68,150,786	4,812,218
Number of pairs with valid NHS number	55,744,973 (99.76%)	4,123,346 (99.78%)
Different (but valid) NHS on both records	129,375 (0.24%)	8,746 (0.22%)

In Table 8 is shown a count of all matching variables and combinations of matching variables in record pairs that have the same valid NHS number

In the HES-HES linkage, there were 81.6% pairs that matched on NHS number. Of these,

134,479 failed to match on sex (0.18%)
135,360 failed to match on date of birth (0.31%)
3,138,360 failed to match on postcode (5.7%)
3,243,776 failed to match on either sex, OR date of birth OR postcode (6.1%)

Using the NHS number as the matching key and verifying the match using other identifying variables, 19.6% of all hospital to hospital matches would be missed using NHS and date of birth, and 24.9% using NHS number and postcode.

In the HES-ONS linkage, there were 85.5% pairs that matched on NHS number, of these,

5,167 failed to match on sex (0.01%)
79,990 failed to match on date of birth (2.0%)
479,585 failed to match on postcode (12.7%)
486,856 failed to match on either sex OR date of birth OR postcode (11.8%)

Using the NHS number and verification using other identifying variables, 16.5% would be missed using NHS number and date of birth, and 25.8% using NHS number and postcode together. This evidence shows that it is crucial to collect NHS number on all the HES records.

Table 8: Counts of the matching variables in matched pairs that have the the same NHS number

	HES-HES	HES-ONS
Number of record pairs	68,150,786	4,812,218
Same NHS number	55,615,598 (81.6%)	4,114,600 (85.5%)
Same date of birth	67,865,570 (99.6%)	4,718,816 (98.1%)
Same sex	67,962,218 (99.7%)	4,805,373 (99.8%)
Same postcode	64,082,185 (94.0%)	4,274,619 (88.8%)
PROCODE/LOPATID	51,920,119 (76.2%)	0 (0.0%)
DOB/SEX/PCD	63,732,344 (93.5%)	4,196,352 (87.2%)
DOB/SEX/PCD + NHS	52,260,112 (76.7%)	3,568,231 (74.15%)
DOB/SEX/PCD + LPD	49,466,842 (72.6%)	0 (0.0%)

Pairs with the same NHS number	55,615,598	4,114,600
Same sex,	55,481,119 (99.8%)	4,109,433 (99.9%)
Different sex	134,479 (0.18%)	5,167 (0.01%)
Pairs with the same NHS number	55,615,598	4,114,600
Same valid date of birth	55,480,230 (99.7%)	4,034,610 (98.0%)
Different date of birth	135,360 (0.31%)	79,990 (2.0%)
Pairs with the same NHS number	55,615,598	4,114,600
Same valid postcode	52,476,776 (94.3%)	3,635,015 (88.3%)
Different postcode	3,138,822 (5.7%)	479,585 (12.7%)

In Table 9 is shown the count of all matching variables and combinations of matching variables in record pairs that have the same valid DOB/SEX/PCD.

In the HES-HES linkages, 82% (n = 52,260,112) had the same NHS number while 21.5% (n = 11,336,581) of the HES-HES pairs did not have an NHS number, the corresponding figure for the HES-ONS being 14.7%.

Table 9: Counts of the matching variables in matched pairs which have the the same DOB/SEX/PCD

	HES-HES	HES-ONS
Number of matched record pairs	68,150,786	4,812,218
Same valid DOB/SEX/Postcode	63,732,344 (93.5%)	4,196,352 (87.2%)
Of these:		
Have the same valid NHS	52,260,112 (82.0%)	3,568,231 (85.0%)
Have a different valid NHS number	135,651 (0.21%)	10,761 (0.26%)
Have a blank or invalid NHS number	11,336,581 (17.8%)	617,360 (14.7%)
Have the same PROCODE/LOPATID	49,466,842 (77.6%)	0

The percentage of HES-HES record pairs having the same hospital record number is 77.6% and this gives reassurance that this combination of hospital site code and hospital patient number can be used for matching, or for checking the existing matches.

4.4.6 Preparation and analysis of the linked file.

The matched records for a given person were collated together, sorted into date order and any date clashes or other logical errors in the person set were resolved. Where they are missing, some variables can be borrowed across the set, for example sex and date of birth, and checks can also be done to ensure the consistency of the various dates on and between the records.

The matched file was sorted into person number order, and further sorted on record type and on episode dates within each person number set. The sort order ensured that

the last record in every set was the ONS mortality record or the latest discharge/transfer. In this way the time sequencing of the records could be checked.

Where the records overlapped, further checks were performed to establish whether the matching of the episodes within the person set was correct, and where there was evidence of wrong matching, all the records in the person set were tagged. These person sets could then be extracted and if required, could be printed out for clerical scrutiny. Where the records have been wrongly matched together, the individual records are separated, corrected where possible, given new person number from a new range of such numbers and re-matched.

Table 10. Analysis of the linked file

Total number of records on the file	=	51,554,709
Total number of people	=	22,672,262
Number of exact duplicates	=	67,784
Sex clashes within a person set	=	28,321
Date of birth after first record in set	=	2,478
Discharge after death (> 3 days)	=	3,714
Clash between spell dates	=	18,937
Overlap of spell dates	=	68,079
Records after disposal dead	=	970
Duplicate (same Hospital/dates diff LOPATID)	=	24,654
Two deaths for same person	=	90
Wrong HES-ONS matches due to poor data	=	90

Table 10 shows that after all the merging and linking has been carried out the linked HES and ONS mortality file has 51,554,709 records for 12,961,775 people with one record and 9,710,487 people with two or more records. At this stage it is possible to examine all the matches for a given person.

The file contains 67,784 duplicated records and these can be removed or tagged as required. Checks on the duplicates were carried out and it was reassuring to find that each member of the duplicate was allocated the same person number. The 28,321 sex clashes within the file indicates that either the sex has been wrongly recorded or the matching has brought together two different people together. Sampling the file has shown that wrong recording of sex was the major error. There were 2,478 people who had a date of birth after their admission to hospital. The discrepancy between the two dates is usually small, just a few days apart and are normally attributed to errors in date recording. It is thought that the Trust computer system date has been entered into the date of birth field.

The error of discharge from hospital within three days after death is common and usually arises from deaths that occurred over a weekend and the person was administratively discharged from hospital at some later date.

The 18,937 clashes between spell dates and the 68,079 overlap of spell dates means that spells overlapped and need to be carefully separated. There were 24,654 duplicates

where the record had the same hospital episode dates but the LOPATID numbers were different. This error could be attributed to bad matching or that a different set of notes had been used for the patient.

The 90 cases of two deaths for the same person is due to small errors or rounding in the dates of birth or incomplete postcodes. All these records were printed out for clerical scrutiny and it was found that the 90 deaths were wrong matches mainly due to errors and omissions in the data. Only one of each death record was retained and the other records were allocated new person numbers.

All the erroneous records were tagged with the above information and in future analyses, decisions could be made whether to include or exclude these records.

4.4.7 Preparation of the analysis files

The analysis files contain all the administrative and clinical data, but they do not contain any of the matching variables. Variables used for analysis like sex, age and area code are aggregated and stored on the analysis file in such a manner that the record cannot be back-linked and the person identified. A record on the analysis file is referenced to its counterpart on the matching file using the accession number, and the matching file records are encrypted using the PGP algorithm (Pretty Good Protection) with a 1024 bit key.

The analysis files were prepared after some corrections were applied to the records. These corrections consisted of:

- Removing duplicates
- Resolving all the overlaps and date clashes

4.4.8 Estimate of matching rates

The linked file was sampled using random sampling methods. The samples were printed out and clerically scrutinised. In parallel with this procedure, a file of all ONS mortality records was merged with a file of all HES records that had a discharge destination of 79 (discharged dead). This file was used to count all those people who had died in hospital and who had an ONS mortality record and conversely to count all people who had place of death as an NHS hospital on the ONS mortality file and should have a corresponding HES record.

Work is continuing on estimating the false negative matching rate. This work is of international importance since in the preparation of any census type of file, the false negative matches will inflate the file. Where a person has two records that do not match together, the person will have two entries in the census, each under the different set of matching variables, and so will be counted twice.

4.5 Future prospects: opportunities and constraints

A number of the current data linkage projects have been listed above. This is a rapidly developing field and the possibilities for further study are numerous. Until about 1996, the concept of routinely extracting and analysing data from government datasets would have been both technically at the edge of possibility and also firmly ruled out by administrative decision. Exceptionally, a study supported by the DSS and carried out at the Centre for Research in Social Policy (CRSP) at Loughborough, using a paper extract of local DSS case level data, demonstrated the potential of such analysis (Dobson et al., 1996). The climate has altered significantly since 1997 as the advantages and power of such administrative data are realised. It seems that such analysis for research purposes, given appropriate safeguards, can be undertaken in ways that meet the requirements of the Data Protection Act.

4.6 Technical issues

The methodology group of the Government Statistical Service Task Force on record matching and data sharing commissioned an evaluation of current best practice methodology for automatic record linkage. The report by Leicester Gill (Gill, 2001) provides a simple yet comprehensive review of data linkage techniques, and is perhaps the best source of information on the practicalities of record linkage. In light of this we will only highlight the most important technical issues here.

4.7 Legal and ethical concerns

Data linkage raises legal and ethical concerns as well as technical issues which are discussed in more detail below. At one extreme there are examples where extensive data systems, depending on data linkage, have been set up. At the other extreme, there are cases where researchers working for local authorities have been stopped from linking data. Thus, what may be possible for central government departments (or in some cases for one department but not for others) - and possibly by extension to the 'agents' of this department (including researchers working under contract) - may not be possible for researchers acting on their own account, or under charitable or research council support.

The overall impression is that, on occasions, a strict blanket interpretation has ruled out activities that would be acceptable to the Data Protection Registrar (now Information Commissioner). Although the central focus of Gill's (2001) report concerns technical matters, he does address the legal and ethical issues of data linkage, and it is worth repeating his guidance – and some additional advice - in some detail. Before carrying out projects involving data linkage the following aspects should be considered:

- The ethics of linking the datasets, depending on the source and contents.
- The public good resulting from research that employs record linkage, especially when studying health (see Goldacre, 2003) or poverty and welfare policies (see Noble, Smith, et al., 1998; 2003) might exceed the costs of risking individual privacy, although this risk ought to be minimized in any case.

- Confidentiality of data about individuals and businesses (although individual anonymity can be assured in the original datasets, when linked identities may be worked out from the combination of attributes. In the UK and the US, data protection refers to all types of information that make an individual identifiable. For example information on the date and place of residence, sex and race are considered to be identifiable information and are treated according to the privacy legislations (see Kingsbury, 2001).
- Physical safeguarding of confidentiality by providing a secure computer system. This extends to paper listings produced during linkage or analysis.
- Compliance with the Data Protection Act 1998, which applies to individuals, and the Statistics of Trade Act 1947, which applies to businesses. Other datasets may be subject to specific legislation.

The Office of National Statistics Code of Practice sets out the key principles and standards for official statisticians to follow and uphold. The Code is supported by twelve Protocols describing how it should be implemented in practice. These documents are a sound basis for good practice in data matching.

5 Acknowledgements

The Unit of Health-Care Epidemiology is directed by Professor Michael Goldacre and is funded by the NHS Research and Development programme. The research into record matching and linkage methodologies and the preparation of the linked databases has been undertaken by L. Gill, G. Bettley, M.Griffith and S.Flynn.

References

E.D. Acheson. *Medical Record Linkage*, Oxford University Press, London. 1967.

T. Blakely and C. Salmond. Probabilistic record linkage and a method to calculate the positive predictive value', *International Journal of Epidemiology*, no. 31, pp.1246-52. 2002.

H. Dunn. Record Linkage, *American Journal of Public Health*, 36, pp 1412-1416. 1946.

L.E. Gill and J.A. Baldwin. Methods and technology of record linkage: some practical considerations. In: Baldwin.J.A., Acheson.E.D., and Graham W.J., (eds) *Textbook of Medical Record Linkage*. Oxford (Oxford University Press, 1987), pp 39-54. 1993.

L.E. Gill, M.J. Goldacre, H.M. Simmons, G.A. Bettley and M. Griffith. Computerised linkage of Medical Records: methodological guidelines. *Journnal of Epidemiology and Community Health*. 47 (1993) pp 316-319. 1993.

L.E. Gill. OX-LINK, The Oxford Medical Linkage System. In: *Record Linkage Techniques 1997, Proceedings of an International Workshop and Exposition, Arlington*, VA, March 20-21, 1997. Washington Federal Committee on Statistical Methodology, Office of Management and Budget, Washington.DC. 1997.

L.E. Gill. Methods for Automatic Record Matching and Linking and Their Use in National Statistics, *National Statistics Methodology Series*, No. 25, London: Office of National Statistics. 2001.

M.J. Goldacre, M. Griffith, L. Gill and A. Mackintosh. In hospital deaths as a fraction of all deaths within 30 days of hospital admission for surgery: Analysis of routine statistics, *British Journal of Medicine*, no. 334, pp. 1069-70. 2002

M.J. Goldacre, L. Kurina, D. Yeates, V. Seagrott and L. Gill. Use of large medical databases to study associations between diseases, *Quarterly Journal of Medicine*, no. 93, pp. 669-75. 2000.

M.J. Goldacre, S.E. Roberts and D. Yeates. Case fatality rates for meningococcal disease in an English population, *British Journal of Medicine*, no. 326, pp. 193-4. 2003.

T.J. Hayes. *Algorithms for curve and surface fitting. In: Software for numerical mathematics*. London and New York: Academic Press. pp. 219-233. 1974.

N.R. Kingsbury. *Record Linkage and Privacy Issues in Creating New Federal Research and Statistical Information*, United States General Accounting Office, GAO-01-126SP. 2001.

H. Newcombe, J. Kennedy, S. Axford, A. James. Automatic Linkage of Vital Records. *Science* 130 (3381): 954-959. 1959.

H.B. Newcombe. The design of efficiency systems for linking records into individual and family histories. *American Journal of Human Genetics* 19: 335-339. 1967.

H.B. Newcombe. Record linking: the design of efficiency systems for linking records into individual and family histories. In: Baldwin JA, Acheson ED, and Graham WJ (eds), *Textbook of Medical Record Linkage*. Oxford: Oxford University Press. pp. 39-54. 1987

H.B. Newcombe H.B. *Handbook of record linkage methods for health and statistical studies, administration and business*. New York: Oxford University Press. 1988. (out of print)

M. Noble, M. Evans, C. Dibben and G. Smith. Changing Fortunes: Geographic Patterns of Income Deprivation in the late 1990s, *Report for SEU/DETR/DSS*, DLTR, London. 2001.

M. Noble, G. Smith et al. *Lone Mothers Moving In and Out of Benefits*, York: Joseph Rowntree Foundation. 1998.

Office for National Statistics. *Longitudinal Study 1971-1991: History. Organisation and Quality of Data*, TSO (London).1995.

E.H. Porter and W.E. Winkler. Approximate String Comparison and its Effect on an Advanced Record Linkage System. In: *Record Linkage Techniques – 1997*. Washington DC: National Academy Press. pp. 190-199. 1999.

S.E. Roberts and M.J. Goldacre. Time trends and the demography of mortality after fractured neck of femur in an English population, 1968-1998: database study, *British Journal of Medicine*, no. 327, pp. 7418-771. 2002.

D. Walsh, M. Small and J. Boyd. *Electronic health summaries – Building on the foundation of the Scottish Record Linkage System*. 2001.

T. Wilson and P. Rees. Linking 1991 population statistics to the 1998 local government geography of Great Britain, *Population Trends* 97. 1999.

W.E. Winkler. Record *linkage software and methods for merging administrative lists*, http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/papers/winkler.pdf 1999.

Author's address:

Leicester Gill
Unit of Helth-Care Epidemiology
Department of Public Helth and Primary care
University of Oxford
Old Road Campus, Headington
Oxford OX3 7LF
Great Britain

Tel. +44 1865 227017
Fax +44 1865 226993
Elect. Mail: Leicester.Gill@ihs.ox.ac.uk
http://www.uhce.ox.ac.uk/