# ENGLISH NATIONAL RECORD LINKAGE OF HOSPITAL EPISODE STATISTICS AND DEATH REGISTRATION RECORDS

**REPORT TO THE DEPARTMENT OF HEALTH, 2003**

**GILL, LE and GOLDACRE, MJ**

**NATIONAL CENTRE FOR HEALTH OUTCOMES DEVELOPMENT,**
**UNIT OF HEALTH-CARE EPIDEMIOLOGY,**
**UNIVERSITY OF OXFORD**

# Contents

# 1. Introduction and outline of the report

## Record linkage

Record linkage is simply the bringing together of information from two different records that are believed to belong to the same person. The records may come from a single data file or from multiple data files. They may relate to persons or to families. Where the two records agree on all the variables, and are unlikely to have done so by accident, the level of assurance that the link is correct will be high. Conversely, if most of the variables disagree there will be little doubt that the linkage is wrong. For intermediate situations the record matching methodology must predict whether the link is true or false or indeterminate.

The introduction of errors and variation in a dataset is unavoidable during the transcription and keying of such data items as names, addresses and dates. This makes the task of record linkage far more difficult since it is not possible to be absolutely sure whether two matching records have identical values on the selected variables purely by chance or that the two records genuinely correspond to the same person.

The data files may have been generated through administrative procedures or from surveys or censuses. A file may represent an entire defined population or only a sample from a defined population. The files may have the same or different time references. In many record systems, input files may contain duplicates, and by making linkages both within and between files these duplicates may be removed, also, missing data items may be copied from one record to another.

Automatic record matching and linkage using computer methodology involves striking a balance between the efficiency of the process and the quality of the matched file. While the technology is becoming more powerful and less expensive, enhancements of the mechanics of file organisation and record matching will reduce the time spent in file processing as well as providing an improvement in overall performance and match rate.

## Outline of the report

Chapter 2 contains an overview of the record matching and linkage processes.

Chapter 3 describes the techniques employed to pre-process input files ready for matching.

Chapter 4 describes the exact or deterministic methods of record matching.

Chapter 5 describes the probabilistic methods of record matching.

Chapter 6 describes methods for collating the matched records into the matched file.

## 2. Overview of record matching and linking

### Introduction

The basic concepts of record linkage should be familiar to all of us. We apply them whenever we look for a number in a telephone directory, a service in the yellow pages or a product in a catalogue. We start with certain information that could be a name (although we may be uncertain about the spelling), a description, a town' or a street and number. The blocking and sorting techniques used in the compilation of the directory limit the scope of our search. To find a telephone number we examine the telephone directory for the appropriate geographic area and select the section for individuals or for business and professional organisations. We then search for the name in the alphabetical index since the telephone number for individuals is blocked and sorted on surnames followed by initials and address.

Where we cannot find a unique entry that is in full agreement with our matching criteria of surname, initial and street address, we look for entries that are in partial agreement. We place some of these in the *possible match category* and make implicit judgements about how much weight to give to the names in the directory for each of the matching variables. These possible links are then resolved by telephoning the subscribers, starting with the one deemed most promising, until a positive link is established.

### Record Linkage in Oxford

A major resource used for the preparation of this report has been the development of the Oxford Record Linkage Study (ORLS). The ORLS have successfully linked hospital discharge data with mortality data using and adapting the methodologies developed at Statistics Canada by Howard Newcombe (Acheson 1967).

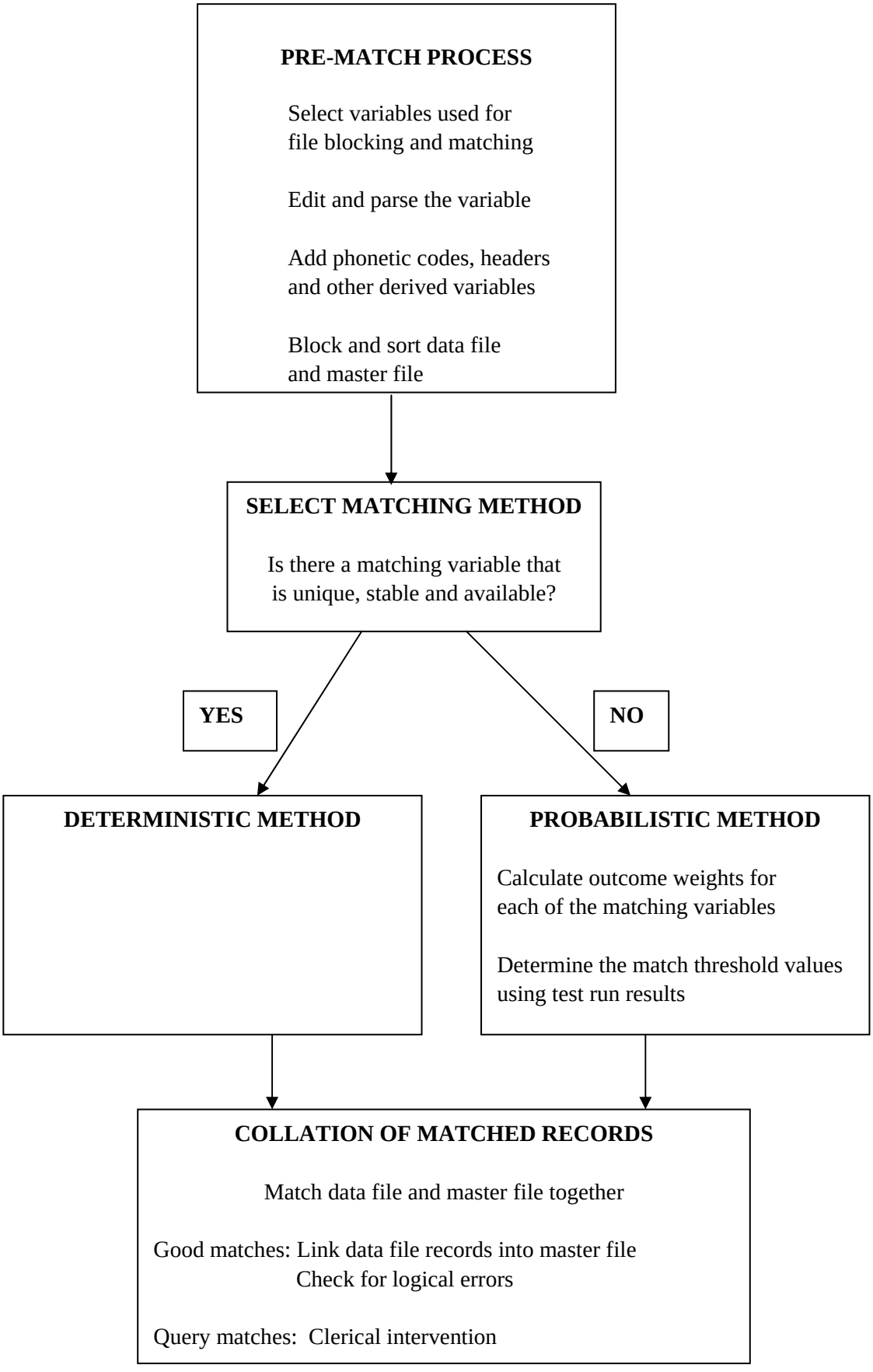
Initially, the ORLS data were limited to brief extracts of each birth, hospital in-patient discharges and deaths for a population of about 350,000. At first it was hoped to link these data together using the National Health Service (NHS) number. In practice, only a minority of the records contained the NHS number and the decision was made to adopt and use the probabilistic method being developed and used in Canada (Newcombe, Kennedy, Axford and James, 1959). Over the period 1963-1999, the study was extended to the whole Oxford health region with a population of 2.5 million.

The Oxford work has achieved the goal of comprehensively matching and linking the hospital discharge and death records. We have drawn heavily on this expertise for building and using these linked file systems.

### Record matching and linkage methods

Record matching and linkage involves three stages, which are shown in Exhibit 1:

- Exhibit 1: Stages for record matching and linkage



Pre-match process bringing potential matches together so that they may be compared.

Comparison of record pairs to decide whether they belong to the same person using:

- exact (deterministic) matching method
- combination of the various partial identifiers to compute scores called weights for each potential match based on probabilities (probabilistic matching).

Collation of the matched records into the matched file.

Record linkage involves bringing together records on the data file with the potential matching records on the master file and assessing whether there is enough evidence to assert a true linkage between the two records. The record on the data file is usually called the query or data record and all the potentially matching records on the master file are called the candidate or master file records.

Each record is composed of variables or identifiers that contain the information on which the match will be evaluated. The identifying variables on the data file must have the equivalent variables on the master file such as surname, address or postcode. To use a computer for the task of matching, a data record with records from the master file, the process must be reduced to an evaluation of each record pair on the basis of some computable function or measure. The differences between the methods largely depends on how this measure is calculated for each pair, and how the measure is evaluated to determine whether the pair should be regarded as a true match or not.

The pre-match process is the most labour intensive of the three stages since over 75% of the effort for record matching and linking two files together is expended in cleaning and parsing the two input files. The variables judged to be the best for file blocking are given additional editing and checking since the success of the matching process is highly dependent upon these items being as accurate as possible thereby permitting potential match pairs to fall into the same block. The next 5% of the work represents the matching and linking processing. The remaining 20% reflects the efforts of clerically checking that the computer matching is correct.

**Pre-match process**

Independent of the matching method used, it is far more efficient to undertake record matching where the input and master files are converted to the same standard format. The major element of the pre-matching process is the editing of the matching variables for errors and omissions and for range checking the numerical variables. This process can be long and tedious but the overall result of the matching process is highly dependent on getting this step as accurate as possible.

The main elements of the pre-matching process are:

- select which variables are to be used for file blocking and record matching

- edit and parse the matching variables and reformat the records if required
- add any numbering systems to identify the record or the person
- add any phonetic codes, headers or other derived variables
- block and sort both the data file and the master file.

The variables used for record matching and blocking the file need to be carefully selected. If no unique identifier(s) are available, a composite identifier may be defined which consists of a combination of partial identifiers. Where the exact matching method is selected for the record matching process, the simple key or the combined key needs to be as complete as possible since only records with absolutely identical keys will be blocked and matched together. In the case of a combined key, there is the possibility of a collision since many records could generate exactly the same key from the component identifiers. For example, where the key is a combination of date of birth/gender/postcode, identical keys could be generated from the records of a pair of twins who are young, live at the same address and have the same sex. Where collisions are expected to occur then additional variables need to be added to the combined set and these could include forename, birth order or some other identifying variable to uniquely identify and separate the people with identical keys.

Matching two files together is more efficient if the format of the data files and the master files are identical and the corresponding variables have the same field length and coding status. It is normal practice to add a header portion to each record that contains the blocking keys, although this is largely determined by the requirements of the sorting package. Each record will need an accession or serial identification number, and a reference number for the person to whom the record belongs. Adding an accession or serial record number to each record will readily and accurately identify the record for further processing. In similar fashion, the addition of a person number can be used for the linkage between the index record and all other records for this person.

Before any attempt is made to match the data and master files the variables in each record must be standardised. Firstly, the variables that have been selected for blocking and matching should be rigorously edited and any names and address fields should be rigorously parsed. The accuracy of the match is dependent upon the quality of these variables and without this procedure many true matches would be lost. Secondly, the other variables in each record that are not used in the record matching process should also be checked at this time for accuracy, completeness and that the range of values are within the predefined limits.

In practice, the header record should contain all the blocking and sorting keys. A header record is essential where the variables in each record are encrypted, since such encrypted records cannot be blocked and sorted in a satisfactory manner. The unmatched records can be re-blocked and re-sorted using a number of different key combinations, and then re-matched, the appropriate header records should contain all the variables necessary to enable the process of re-blocking and re-sorting to be repeated until most of the matches are found.

When all the editing and cleaning stages have been completed both the data and master files need to be sorted into the same sequence for blocking and matching. The blocking and

sorting of the files is dependent upon whether the data file is to be matched to the master file or whether it is to be matched to itself.

Normally, records on the data file are matched against records in the corresponding block on the master file. Where the data file contains two or more records and each matches to the same record on the master file, they can be regarded as matching to each other and belong to the same person. Where there is no corresponding record on the master file, there is no opportunity for the records to match with each other, therefore a different methodology is required to effect the data-data matching.

There are two methods of matching the data file against a master file depending on whether the records on the data file are to be matched against themselves or not. The methodology depends on matching a single file against itself or matching two files, one against the other. In the two file method the data and master files are held as two separate files, blocked and sorted in the same order. Each record on the data file is recursively matched against the records in the corresponding block on the master file. The data records are matched against all the potential records on the master file but may not be matched with each other unless internal cross matching of each of the input files is undertaken.

In the one file method the data file and the master file are merged, blocked and sorted into one combined file and the data file records and master match file records suitably tagged. Every record in a given block is matched with every other record in the same block in a triangular fashion (first with the rest, followed by the second with the rest, and so on). In the same matching run, the data records are matched against all the other records in the same block. The data records can also be matched with each other and, if required, the master file records could also be matched against each other, which would result in data/data, data/master and master/master matches.

**Selection of matching method**

The two main methods for matching are the exact and the probabilistic. The selection of the appropriate matching method depends on a number of factors, the most important of which are the availability, stability and uniqueness of the variables in the dataset. The general guidelines for the choice of the matching method are:

- Where the records on both the data file and the master file have a matching key on every record that is stable, has low error and high discrimination, then use the **exact** method.

- If a unique key does not exist but there are a group of high quality partial identifiers that in combination form a unique variable, the **exact** method can still be used.

- Where the data are noisy and contain random errors but there is an array of partial identifiers usable for blocking and record matching, use the **probabilistic** method.

**Exact matching (deterministic, all-or-none methods)**

The main requirement for exact matching (sometimes called deterministic or all-or-none

matching) depends upon the records on both files containing a variable or characteristic of a person that is ideally:

- universally available
- fixed
- easily recorded
- unique to that individual
- readily verifiable.

Few, if any, variables meet all these requirements, though several come sufficiently close to be usable. The perfect variable would be an integral feature of the individual such as a personal trait or a group of traits that together form a set that is unique to that individual. Alternatively, unique identification numbers may be assigned to individuals at birth or a unique number or cipher allocated to the person or object using a highly reliable and accurate procedure such as national insurance numbers, bank or other national or catalogue numbers.

Systems of numbers or other ciphers can be generated which meet the above criteria within any given setting. In the UK healthcare setting, the roll-out of the new ten digit NHS number has improved the prospects of exact linkage, and it is a requirement that the new NHS number has to be incorporated in all health care records from 1997, by 2002, 86% of HES records contain a viable NHS number. (Secretaries of State, 1989; National Health Service and Department of Health, 1990).

Exact matching generates links that are based on the exact agreement of the selected identifying variables on each record in the pair. In its simplest form the result of the match is clear cut, either the records match or they do not. This tactic simplifies the record linkage methodology, making it more practical for rapid processing on small computing systems. For the comparison of records that contain many variables in which there is a possibility of the variables having a low error rate, another version of the exact matching process can be used in which the exact match criterion is slightly relaxed. The number of variables that agree are used to determine whether the record pair should be linked and whether an almost-exact match is all that is required. This is quite different from probabilistic matching since it uses a very simple matching method and does not require the generation and use of probability weights.

**Probabilistic matching**

Person names are not normally unique except where they are combinations of very rare surnames and forenames. Names when spoken, written or captured in a computer system are subject to considerable variation in recording and spelling. Even if the names stored in the computing system are carefully cleaned and regarded as accurate, the names from the data file will come from the real world and be subject to the usual error and variation.

Non-name data such as dates of birth, address and postcode are all subject to error, truncation

and incompleteness. The use of an exact matching method would fail to bring together records that contain even small amounts of error, omission or over or under qualification in any one part of the identifying set. Complications arise when the identifying set contains spelling errors, data preparation errors, use of synonyms and nicknames, anglicisation of foreign names, initials, truncation and abbreviation, missing words and extra words. Certain names like Smith are 10,000 more frequent than Szabo, so there will be many more surname/forename combinations to be searched for a common name combination than a rare name combination. For these reasons the non-exact or probabilistic methods of record matching have been developed.

Probabilistic matching methods are used for matching together data and master files that contain errors and omissions and for which there is no unique or universally available high quality identifier. If all the selected variables on the data and master file record agree, and are unlikely to have done so by accident, the level of assurance that the records belong to the same person will be high. Conversely, if they all disagree and are very unlikely to have done so in truly linked pairs of records, there will be little doubt that the records in the pair are wrongly matched. For intermediate cases the evidence must be evaluated to decide whether the records possibly match.

Probabilistic record matching is so named because it relies on calculating matching scores based on the probabilities. The method involves measuring the agreements and disagreements between the corresponding variables in the two records. A matching score is computed based on the number of agreements between the variables and the the number of disagreements and is used for determining whether the record pair should be regarded as truly linked or not. This is similar to the process used in exact matching.

Probabilistic record matching carries these calculations a little further. Either from previous experience of record matching in similar areas of application, or based on a preliminary matching exercise carried out on the current data, how likely is it that the variables which matched in the current record pair would have done so by chance even if the records were not correctly linked? This is compared with how likely the agreement would be in correctly linked record pairs.

Clearly in any reliable record matching procedure it is essential to use those agreements between variables which are more typical of correctly linked pairs rather than those which might well have occurred by chance in unrelated records. The features that might agree by chance in unlinked record pairs are those that are quite common, and therefore have a low discrimination, and don't sub-divide the population into many groups, good examples being: gender or marital status. A better variable , with a larger discrimination, and it has a larger range of values, is the date of birth. Birthday alone (ignoring the year) divides the population into 365 groups. Clearly, birthday is more useful as a component of a matching variable than gender or marital status, although it is not sufficient on its own because on average $1/365^{th}$ of the population will have the same birthday, whereas the use of the full date of birth would divide the population into 365*72 = 26,280 blocks.

Probabilistic matching requires some preliminary matching to have been carried out or experience from a previous study to be utilised. For each variable in the matching record pair, the probability that it would agree in truly linked records is calculated. This value is then

compared to the probability that it would agree by chance or agree despite coding error in unlinked records. Scores based on ratios of these probabilities (or relative frequencies) are calculated and called ***frequency ratios*** after Newcombe et al (1959).

These scores are transformed (by taking logarithms) to give what are termed weights and then algebraically summed over all the identifying variables to give an overall score for the record pair. This score will almost always have positive values for weights indicative of fields that agree and negative values for those that disagree, since non-linkage is usually more likely when the variables disagree and linkage more likely when they agree. The score is then compared against some threshold value, determined *a priori*, to determine whether the overall weight for this record pair is high enough to classify it as a true match.

There is one further important refinement to the methodology. Newcombe et al (1959) introduced methods for using the specific values that the matching variables take in the calculations of the weights. To illustrate this using forenames, a pair of records containing the forename Leicester will have a higher probability of belonging to the same person than a pair of records containing the forename John, since the frequency of Leicester is about 1/3000 of the frequency of John. In this example the forename Leicester has more discriminating power than John. So the weight is different for each possible outcome of a specified identifying variable. The term that is used to describe this refinement is ***outcome-specific weights.***

For probabilistic record matching to be cost effective and efficient it is necessary to block together records that are likely to refer to the same person, thereby reducing the time spent searching the file and only making productive comparisons within these blocks.

Clearly, the reliability and efficiency of the matching procedure is highly dependent upon the manner in which the file blocking is carried out. One important consideration is that the number of records in each block is small enough to avoid too many unproductive comparisons and yet large enough to prevent records for the same person spilling over into different blocks and so failing to be compared. The balance between the number and size of the blocks is particularly important when matching large files.

**Collation of the matched records**

There are three ways in which the data records can be matched with the master file.

- One to one matches. This method is used where a data record is to be matched to a record on the master file, and this is the method normally used for person matching. This method of matching seeks to establish unique pairs of records, both referring to the same person.

- Many to one matching**.** This method is normally used for removing duplicates from a file that contains many identical copies of the same record.

- One to many matching**.** In this method a data record is matched against a master file and all the matches for the person are selected. This method would be used, for example, in culling all records for a person who has many health care contacts.

The matched data records are merged and then collated with the master file, or incorporated into the database index. The records should be inserted into the proper temporal sequence and any overlaps or other logical inconsistencies should be checked at this stage. Where there are logical inconsistencies, such as hospital discharge after death or the person matched as being in two types of care at the same time, both the data record and all the master file records for this person should be extracted for clerical intervention, since the sequence could have been generated from bad matches from previous runs, or could contain records for a different person. These errors usually result from poor data quality or where two different people have almost identical identifying sets, for example in the case of same sex twins. Other problems arise where the persons are elderly and have changed their names several times or where the identifying set provided by the patient is different from that supplied by the next of kin on a death record.

In the use of the exact method, the result is clear cut, either the records match or they do not match. Where the keys are absent or partially present it is very unlikely that the records will match. Where a combination of partial identifiers are used for an exact or almost-exact match, clerical intervention will need to be required to improve the match rate and when this cannot be achieved the general rule will be to leave the records unmatched.

Where probabilistic record matching is used, a third category of outcome is generated, the query match. This category is best printed out for clerical intervention. However, lessons should be drawn from this match that can then be used for trimming the match acceptance threshold and then re-running this data set and using the new thresholds for this and future data sets. Any corrections to the computer matching arising from the clerical intervention should also be applied at this time

The matched output will consist of those data records that matched well with the records on the master file, those records that did not match and those records that are possibly matched and will require clerical scrutiny. The 'possibly matched records' are printed out in a specialised format so that the clerical staff can inspect the two data records together with the computer generated result. Where the clerks are sure that the two records (one data record and one master file record) belong to the same person, a correction is applied and the two records are matched together, and where any doubt exists the records are left unmatched.

The result of matching one record against another record will result in one of the following four outcomes:

- **Good match**. The records have matched together and refer to the same person or family. A random sample of these matched record pairs should be checked for accuracy in the matching, especially for those matches that are near the threshold cut-off values.

- **No match**. The records are deemed to belong to different people or families and therefore are not matched together. A random sample of these record pairs should be checked for accuracy of the matching, especially those matches that are near the threshold cut-off values.

- **False negative or Type I error.** This type of error arises when the records that should

have been matched together have not matched since the overall matching criterion falls below the preset threshold and the record pairs are left unmatched. These records could be matched in a clerical exercise, or in a subsequent match run, each record in the pair would match to a third record and in this way be matched together.

- **False positive or Type II error**. This type of error arises when the records for two different people have been wrongly matched together, possibly, since the record pair contains matching variables that have similar or low discrimination values. Suitable logical checks in the linking stage should find these erroneous matches and any erroneous links can be broken. Otherwise, sampling will have to be used to estimate the false positive rate.

A random sample of the linked file is culled and used to estimate the numbers of records that have been wrongly matched. The random sample consists of sets of all records linked under the same person number. The set is then clerically examined to determine whether all the records do in fact belong to the same person or belong to two or more different people. Since the records for twins and other multiple births have identifying variables that have the same values, the above method could be used to correct the file.

Another method is to examine all the records for those people who have large numbers of events on the file and check for inconsistencies such as changes in common items, for example, date of birth, gender or first forename which are usually regarded as stable.

# 3. Pre-match processing

## Introduction

The data file and the master file may have been compiled at different occasions, in different settings and are often in different formats. Before the files can be matched it is important that the two files are converted into the common standard format and any errors that can possibly affect the matching process are edited out. Furthermore, the size of the files may be very large and direct comparison of each of the records on the data file with each of the records on the master file would be prohibitively expensive even with the most modern and powerful computing systems. The two files need to be partitioned into manageable blocks. This partitioning of the data files into mutually exclusive subsets or blocks is with respect to a selected set of key matching variables constructed from the individual matching variables.

The techniques employed to pre-process the input files before matching are:

- selection and definition of matching variables
- identification of errors that occur in matching variables
- editing, parsing and standardisation of the matching variables
- creation of linked files
- blocking and sorting of the data files
- structuring and organising files.

## Selection and definition of matching variables

The choice of the matching variables depends on the type and contents of the data file and the master file. The main requirement is that any variable selected as a matching variable must occur both on the data file and the master file. The precise definitions of the matching variables and the quality of reporting and capture of these are crucial for the success of the matching and linkage process.

If the data and the master files already exist, potential matching variables must be evaluated to determine whether a linkage of acceptable quality is feasible or not. If one or more of the input files have yet to be created, it may be possible to influence file development in ways that will assist record linkage, for example adding the variables needed for linkage to other files and by using operating definitions and formats that are compatible with those files.

To illustrate the choice of matching variables and the issues involved, consider the problem of linking a file of person records based on common identifying information such as NHS number, names, address, postcode, gender and birth date.

The NHS number, subject to rigorous confidentiality constraints, is one of the most important linking variables that can be used for person matching purposes, since it is issued to almost all of the population of England and Wales and issued separately in Scotland and Northern Ireland. National Health Service (NHS) numbers are issued at a person's birth registration or when they have registered as an immigrant with the NHS. The new-style NHS numbers have been issued since 1997 and have been back-loaded into general practitioner, health authority and hospital trust master indexes. It is expected that the NHS number will become a universal identifier throughout the whole of the health care sector. Since it has an embedded check-digit, the NHS number is accurate and stable enough to be used for exact matching. The presence of the internal check digit in the NHS number enables the identification of errors by a simple examination of the number itself. In Scotland the community health index number is also used in the record linkage process.

All babies born in the UK are issued with the new NHS number as part of the birth registration process. If the baby (and its parents) leave the country for a long period of time, on re-entering the UK they will be required to re-register with the NHS and a new set of NHS numbers may be issued. In this way, the NHSCR has generated duplicates in the allocation of the new NHS number from the old numbers and some people may have two or more numbers possibly as a result of changing their marital status or registering with two different GPs or two different primary care trusts. Despite the low frequency of the problem, multiple NHS numbers can largely be ignored unless one is looking at longitudinal information stretching back to the early days of the NHS.

Where the new NHS number is either missing or proves unusable records have to be linked using other matching variables. As a general rule none of the other variables used for matching are unique and all of them are subject to errors and omissions either in reporting, transcription or keying. The identifying variables can be considered in the following six quite separate groups.

- Group 1. This group consists of the proper names of the person that rarely change during the lifetime of a person, with the exception of the present surname where women traditionally adopt their husband's name on marriage or revert to their former surname when the couple get divorced. The readily collected names are birth surname, present surname, first forename or first initial, second forename or second initial and other forenames.

- Group 2. This group consists of the non-name personal characteristics that are fixed or issued at birth and very rarely change during the course of a person's lifetime. This group includes gender, date of birth, birth order, place of birth (address where parents were living when the person was born), NHS number, community health index (CHI) number allocated in Scotland, National Insurance number and ethnicity.

- Group 3. This group consists of socio-demographic variables that may change many times during the course of a person's lifetime such as street address, postcode, general practitioner, marital status, social class, unique numbers allocated by a primary care trust, NHS trust, GP or other specialised systems.

- Group 4. This group consists of variables that may be used in the compilation of

special registers such as clinical specialty, diagnosis, surgical procedure code, cancer site, drug idiosyncrasy or therapy, occupation, date of death and other dates.

- Group 5. This group consists of variables that may be used for family record linkage such as those names described under Group 1 together with mother's birth surname, father's surname, marital status, date of marriage, number of marriages, number of births, birth order, birth weights and dates of delivery.

- Group 6. This group consists of arbitrarily allocated numbers that uniquely identify this record (accession or serial number) or this person (person number), historical markers for the data in the record, and markers to indicate the edition or version of the codes used within the record.

It is normal practice to match together records that contain matching variables selected from one or more of the above groups, normally from groups 1,2,3 and 6. It is also possible to concatenate two or more of the matching variables to construct a compound matching variable, for example date of birth/gender/postcode, and use it in an exact match. This enables the user to exploit the power, speed and flexibility of the exact match without the benefit of having a unique number.

Still other variables could be used for matching such as ethnicity, marital status and telephone number. Ethnicity is a variable that has similar characteristics to gender except not nearly as well reported (unless it is recorded as black or non-black). Telephone numbers while potentially of enormous value are not universally available or stored in the UK health and administrative systems. Some of the population do not have a home telephone number.

**Identification of errors that occur in matching variables**

Errors in the matching variables may creep in during the capture and processing of these variables. Other sources of errors in the matching variables are:

- variations in spelling

- data coding and preparation

- use of phonetic name compression methods

- use of name synonyms and nicknames

- anglicisation of foreign names (usually mid-european)

- use of initials, truncation and abbreviation of names and addresses

- use of compound names, missing words and extra words (since during data entry extra words being entered in a given field can overflow and fill the adjacent field and in this way create extra and erroneous fields).

The errors occurring in commonly used matching variables are described below.

*Present surname*

Name changes due to marriage or divorce are the main difficulty in using the present surname for matching. For some ethnic groups there can be many surnames and the order of their use may vary. Concatenations of the birth surname and the marriage or partnership name into a compound (or hyphenated name) are quite common so that both parts are required for matching purposes. Spelling variations in surnames are quite common due to the effects of transcription of the names through the various systems. In some cultures there is no exact equivalent of a surname and members of a family may have no single name in common. It is common for Sikhs to add Singh to their name sequences while the addition of Kaur often means that the person is an unmarried girl although in other circumstances it can equally apply to married women.

*Other surnames*

People may have many different surnames, which includes their birth surname, and the name that they use may vary from occasion to occasion. In the UK, with the increasing number of single parent families this practice is becoming more common. For example, an unmarried mother may give her birth surname when registering the birth of her child, and her common-use surname, which may be that of her current partner, when registering with a doctor. It is the latter name that is used to generate a birth registration and the NHS number. Consequently, this creates problems when linking birth registrations to hospital records.

*First forename*

There are wide variations in the spelling of forenames due to recording and transcription errors. It is fashionable to modify existing forenames either to modernise them or to copy the names used by television or sports stars such as Rebecca which becomes Rebeka. Other widespread problems include the use of nicknames and contractions. Some are readily identifiable, such as Jim for James, Wm for William or Liz for Elizabeth. Others, can not be readily identified such as Ginger for Paul or Blondie for Jane. Some records may just record the fact that the person is a baby or a twin and, until such time as the birth is registered or the baby is named, the records may contain the name baby or twin.

*Other forenames*

It is quite common for a person to have many forenames and the forenames that they use may vary from time to time. Elderly people in particular may have a large number of forenames, some of which are recorded in hospital computing systems while others are recorded in the death registration processes at the person's death (normally the information is supplied by the next-of-kin). Where the person is young and living at home the differences in the forenames can be used to detect similar sex twins, when the only variables that differ are the forenames. Where the person is much older, it is more difficult to decide whether the two records are for the same person if the only variables that are different are the forenames. It is difficult to decide whether the person is using a number of different forenames at various times or as the circumstances change. Without further investigation the match between such records must be regarded as dubious and further information sought.

*Swapping of names and surnames*

Occasionally the surnames and forenames are swapped around, or there may be spelling variations in the names due to transcription and data preparation errors. If the name is matched against a population index it should be possible to improve the quality of the name. For example, it should be possible to get extra variables like second forename, or obtain better spellings for the names. The names can be corrected or their order changed (switch forenames and surnames) and tests performed to check whether the person has died or migrated.

*Address and postcode*

These are excellent variables for confirming otherwise questionable links. However, disagreements are hard to interpret because of address changes, address variations and differences between mailing addresses (usually all that is available in administrative files) and physical addresses (generally all that is obtained in a household survey). Much research on this variable has been undertaken by Childers and Hogan (1984).

*Gender (sex)*

Gender is generally well reported and, except for transcription and recording errors is a very reliable variable. Some databases do not collect this variable although it can be generated through the recoding of forenames, although this recoding cannot be done with complete accuracy.

*Date of birth*

The day and month of birth are generally well reported even by proxy respondents. Year of birth can be used alone or in combination with day and month of birth as a matching variable. If the variable is collected along with the present age of the person, checks can be performed at the data collection stage to verify that the year of birth and the computed age agree. Problems have been experienced where the DD and the MM have been transposed from the European format (DDMMYY) to the US format (MMDDYY) largely through the use of US computing systems and software. With the increasing use of hospital master index systems this error is now quite small. In some cultures there is no concept of date of birth and in these cases the date is recorded as 0101YEAR in which the person was born. Problems arise where parts of the date of birth are missing. For example, missing day of birth could be substituted by using 01, 15 or 99, and where day and month are missing the possible default values are 0101, 1506, 3006, 0107, 1507 or 9999.

*Key variable truncated*

Where some of the key matching or blocking variables are missing, partially recorded or not known, they can be replaced with a standard default value, examples of which are shown in Exhibit 2. Where a variable has been partially recorded due to some limitation in the transcription process or in the computing storage allocation, methods will have to be devised to compare the corresponding data and master file variables which may have strings of different lengths. A good example of this is where the forenames have been captured at different times and stored using different storage standards. The forenames may have been

captured and stored in a 6 byte field, or in other formats such as 8 bytes or higher.

**Exhibit 2: Examples of variables that are set to unknown values**

| | |
|---|---|
| **Administrative dates**: | set to 0101YY, 010199, 999999 |
| **Date of Birth** | 0101YY, 1506YY, 3006YY, 0107YY, 1507YY, 0101YEAR |
| **Names:** | set to spaces, NK, UNKNOWN, or ZZZZ<br>BABY, MALE, FEMALE, TWIN, TRIPLET, INFANT |
| **Other variables:** | set to 9, 99, 9999, -1<br>NK (Not Known)<br>NA (Not applicable)<br>NC (Not coded)<br>U (Unknown) |

Where the forenames that have been captured, truncated and stored in a fixed length field, say for example six characters, when matched against a record having a longer forename only the first six characters can be used. For example:

| | |
|---|---|
| CHRISTINA | stored as CHRIST (truncated to 6 characters) |
| CHRISTIAN | stored as CHRIST (truncated to 6 characters) |
| CHRISTOPHER | stored as CHRIST (truncated to 6 characters) |

The common stem (CHRIST, the first six characters in this example) can only be compared with the same number of characters in another full forename.

*Variables have low discriminating power*

Where the records contain a combination of the most common forenames and surnames, for example the surname is Smith or Jones and the forenames are either John or Margaret, the name identifiers alone cannot provide a large enough measure of discrimination to separate two John Smiths or the two Margaret Jones. The computed outcome weights will reflect this low discriminatory power and consequently the weight will be allocated a low value. For the match to be accepted other ubiquitous variables will be needed to bolster the total outcome weight so that this total weight may exceed the preset threshold value and the match can then be accepted.

Some of the names recorded in the forename variable can be non-specific and can be applied to a wider group of people. The examples presented in Exhibit 3 show the different kinds of names that have a low discrimination power for the purposes of record matching.

**Exhibit 3: Examples of names or titles that have low discriminating power and will therefore generate low outcome weights**

1. Names that contain certain cultural titles, for example:
   Bibi, Kaur or Singh.

2. Temporary names given to babies or very young children in hospital, for example:
   Baby, Babyone, Babytwo, Boy, Girl, Twin, Twin1, Twin2, Triplet.

3. Names given to patients in order to preserve their identity, for example:
   ZZZZ, anon, anonymous, anonymised patient, unknown patient.

When these name strings are encountered as part of the matching algorithm, they should be identified using the parsing procedures and then allocated a very low outcome weight by the system. If these errors are not trapped, all babies or adults given temporary or non-identifiable names would be matched together

*Embedded titles in the names variables*

The surname and forename strings may contain embedded titles such as those presented in Exhibit 4.

The names string needs to be parsed and the various components identified and separated out. Some of the titles can be used for linkage purposes, for example those associated with marital status, academic title and hyphenated names.

**Exhibit 4: Examples of titles that are usually embedded in or appended to the surname or forename variables.**

| | |
|---|---|
| Marital status: | Mr, Mrs, Ms, Miss, Master, Son, Daughter |
| Peerage: | Lord, Lady, Baron, Viscount |
| Civil honours: | Hon, Sir, Dame, Lady |
| Political titles: | MP, Councillor, Mayor |
| Academic title: | Dr, Professor |
| Academic degree: | MA, MD, PhD |
| Military: | Major, General, Lt.Col |
| Church: | Bishop, Monsignor, Rabbi, Father, Brother, Sister |
| Family order: | Fred Jr, Bill Sr, Hiram 3, Hiram Third, Hiram III |
| Hyphenated names: | Baden-Powell, Twistleton-Fiennes |
| Concatenated names: | Joan Brown Smith |

Source: ORLS, 1999

*Missing or partially given matching variables*

Even if the value of the matching variable may be missing or set to some arbitrary value, it can still be used for matching provided that the outcome weights and the match thresholds are adjusted accordingly. Occasionally, one or more of the matching variables may be partially complete, for example, where a forename is truncated to a fixed number of characters, or just the initial may have been recorded. It some cases the variables may have been set to spaces (or blanks) or to some agreed default value for not known or not collected. Although not perfect, the corresponding matching variables in both records may still be compared and the amount of agreement or disagreement computed. Where there are small deviations in, for example, the date of birth the calculation of the agreement is based on the empirical results compiled from matches using sample or similar datasets rigorously checked by experienced clerks.

**Editing, parsing and standardisation of the matching variables**

The matching variables are also used to partition the files into blocks. Some variables may not be suitable for this purpose since they may have low discrimination or have a high error rate. Rigorous editing and parsing of these matching variables is required to reduce the errors, since there is the possibility that many true matches would be split over different blocks and wrongly designated as non-links.

Record linkage practitioners must pay particular attention to the formatting and standardisation of the matching variables since they can improve the chances of correctly linking two records. In most cases record linkage would be impossible without the initial re-formatting of records. Consider matching files by name where the order of the names are different, for example, one file has surnames first and the other has forenames first. Unless the variables are reordered, the matching algorithm will need to cross match all the names against each other and then select the best fit. Differences in the presence or absence of titles (Mr, Mrs) can also cause difficulties as can the inclusion of junior/senior. Some caution should be exercised since standardisation can result in distortions (truncation of names down to a common and standard length) and loss of information which will increase the likelihood of designating some pairs of records as positive links when, in fact, they do not match.

DeGuire (1988) presented the concepts needed for parsing and standardising addresses but the same techniques could also be applied to the parsing of names. The procedure involves identifying the constituent parts of the matching variables and representing them in a common standard way through the use of look up tables, lexicons or phonetic coding systems. The standardised individual elements can then re-arranged into a common order. Specific examples are described below.

*Standardisation of surnames and forenames*

The basic uses of the standardisation of names are to:

- replace the many spelling variations of commonly occurring names and addresses with the correct or standardised spellings or agreed abbreviations in common use.

- use the key words generated during the standardisation process as hints for the further

development of editing and parsing sub-routines.

The purpose of name standardisation in record matching is to support name-matching software to work more efficiently by presenting names in a consistent fashion and by separating out parts of the name that would have little or no value in matching.

In the standardisation process, variations in the spelling of forenames such as LIZ and BETTY might for consistency be replaced with the original or formal spelling such as ELIZABETH. It is also possible to convert identifying stem words such as FRED although these could equally be associated with ALFRED and FREDERICK. Other procedures used in formatting names include the removal of punctuation or blanks. For example, O'BRIEN becomes OBRIEN, Le MESURIER becomes LEMESURIER and Van DAMM becomes VANDAMM. As previously described, dictionaries and lexicons have been developed that can relate commonly used nicknames and name contractions to formal names (BOB and ROBERT, BETSY and ELIZABETH) and to link the common variations in spelling (SMITH, SMYTH, SMYTHE or HORTON, HAWTON, HOUGHTON).

In some cases, the order of surnames and forenames in a record is not clear, since during the data capture and data preparation phases the order of some names may not have been recorded in a consistent fashion. There may be records in which the names and surnames have been swapped, for example JOHN SMITH recorded as SMITH JOHN. The ORLS have developed lexicons for identifying most common forenames and surnames combinations and use it for identifying and correcting the name swaps. Problems occur where both the forename and the surname can be used equally as a surname or forename, for example JAMES DOUGLAS could be DOUGLAS JAMES or JIM LORD could be LORD JIM.

*Use of look-up/conversion tables, lexicons and dictionaries*

While a conventional dictionary primarily provides definitions of words and phrases it may also be used to provide a list of synonyms. This property may be exploited in the conversion of a given name to another name, or to retrieve a forename from a list of nicknames or contractions, and this technique provides a good method for comparing names in both exact and probabilistic matching techniques.

**Exhibit 5: A sample of a Lexicon for converting names to a formal name**

| | |
|---|---|
| WILLIAM | WILLY, BILLY, WILL, BILL, WM |
| JOHN | JOHNNY, JON |
| ELIZABETH | LIZ, LIZA, BETTY, LIZZY, BETH, BETSY, ELISABETH |
| MARGARET | MAGGIE, MAGGY, MEGGIE, MEG, PEGGY, MADGE, MARGE |

Source: ORLS, 1999

The name conversion process involves two stages. The first stage involves checking the spelling of the name against all the relevant entries in the lexicon. If valid, the name undergoes a second stage where the various synonyms are returned for use in phonetic coding schemes or table look-up methods. Errors that have occurred when the original valid name was transformed into a second valid name using the lexicon, such as ELIZABETH from LIZA or MARGARET from PEGGY, are inherently undetectable in this type of scheme and therefore it is non-reversible.

Lexicons are also be used for the conversion of a name string to a code or a service name to a service code. A sample of such a lexicon is presented in Exhibit 5.

The lexicon which may be used for equivalencing the various spellings of common forenames, and an example of which is shown in Exhibit 6.

*Use of phonetic coding schemes*

The links lost through problems in the method of blocking the files using names in a number of mortality studies ranged from 2% through to 25% (Lalonde, 1992). For this reason the files may be blocked by the phonetic code and the matching is then carried out only between records that fall within the same coded blocks. This method offers a very efficient method of matching small blocks that contain most of the possible variations of the surname.

A common procedure for parsing surnames has been to encode surnames phonetically and to use the encoded values for blocking the file. Although the phonetic code of the names are used to position both the data file and the master file, the original names are retained in full on both the records and are used for the matching process.

Two of the procedures in common use for the phonetic coding of names are the Russell Soundex Code (Soundex) and the New York State Identification and Intelligence System (NYSIIS). The Soundex system, which is the older of the two, creates a four-character alphanumeric code which uses the first letter of the surname for the first character of the code and a further three digits (Lynch and Arends, 1977). ( )

The NYSIIS code is a fixed length alphanumeric code that divides the population of North American and European surnames into groups that vary less in size than those associated with the Soundex codes (Lynch and Arends, 1977). Both coding systems are designed so that surnames of similar sound have the same code and frequently encountered errors of reporting do not cause changes in the code (Howe and Lindsay, 1981).

**Exhibit 6: Some typical variations on the spelling of ABIGAIL**

| | | | | | |
|---|---|---|---|---|---|
| ABAGAYLE | (1) | ABAIGEAL | (3) | ABBEGAIL | (2) |
| ABBEYGAIL | (1) | ABBEYGALE | (1) | ABBIE-GAYLE | (1) |
| ABBIEGALE | (1) | ABBIGAIL | (7) | ABBIGALE | (2) |
| ABBYGAIL | (3) | ABBYGALE | (1) | ABBYGAYLE | (1) |
| ABBYGEL | (1) | ABBYGIEAL | (1) | ABEGAIL | (2) |
| ABEGALE | (1) | ABIAGUIL | (1) | ABIEGAIL | (2) |
| ABIGAEL | (2) | ABIGAIL | (1479) | ABIGAIL-LOUI | (1) |

The number in brackets indicates the frequency of the name on the ORLS master files

Source: ORLS, 1999

The main objective of the use of phonetic coding is to create blocks which:

- place all variations in spelling of a given surname into the same block.
- limit the number of the blocks, and dividing the file of records over the number of available blocks.
- Ideally, create blocks that contain few dissimilar surnames
- require minimal computer processing without the need to employ large look-up or conversion tables.

Most of these systems have two features in common:

- The vowel information is either partially or wholly suppressed because of its instability. Some phonetic algorithms retain the position of the vowel in the names string but convert them into the letter A.
- Certain consonants with similar sounds (or groups of letters with these consonants) are replaced by a standard character or group of characters representing that sound or phoneme. For example the letters M and N are grouped together, as are D and T.

The Soundex and NYSIIS algorithms possess the above characteristics. However, NYSIIS retains information on the sequence of and position of the vowels in the name by changing them all to the letter A, whereas Soundex removes the vowels. All of the algorithms are capable of revealing similarities between names even where the coded forms do not agree precisely. Finally, where shortened forms of the names are needed for compactness, name compression may be desirable because it loses only the vowels and the redundant consonants.

A development undertaken by the Oxford Record Linkage Study (Gill, 1986, 1993, 1997), referred to as the Oxford Name Compression Algorithm (ONCA), uses an anglicised version of the NYSIIS method of compression as the initial or pre-processing stage and the transformed and partially compressed name is then Soundexed in the usual way. This two-stage technique has been used successfully for blocking the files of the ORLS and overcomes most of the unsatisfactory features of pure Soundexing while retaining a convenient four-character fixed-length format.

The file blocks produced using the ONCA procedure vary from quite small and manageable for the less common surnames to very large and uneconomic for the more common surnames. Further sub-division of the phonetic blocks on the file is usually effected using gender, first initial or date of birth either singly or in combination.

Occasionally, phonetic coding is used to reveal similarities between names even where the codes themselves may differ. A further use of such codes to reduce the physical sizes of the names is rarely undertaken, but it is used extensively for the preparation of constant length blocking keys.

*Standardisation of addresses*

Standardisation of addresses operates in a similar fashion to to that used for the standardisation of names. Abbreviations Rd or Cres should be replaced by appropriate expansions to Road or Crescent or to a set of standard abbreviations commonly used by the organisation. For example, when a variation of a rural address, such as Hill Top Farm or Sunny-Side Nursing Home, is encountered the software should use a set of parsing routines different from those associated with house-number and street-name addresses. Where reference files containing town, county and postal codes are available from the post office or from some other source, the town names in the address lists can be reformatted into a standard form that is consistent with the reference list.

Parsing divides the free-form address variable into a common set of components that can be compared, for example street number, street name, town and county. Parsing algorithms often use words that have been standardised. For example, words such as Street or Road would cause parsing algorithms to apply different procedures than words such as High or Oxford. While character-by-character comparison of the standardised but unparsed names would yield no matches, the use of the components in the address might help designate some pairs as links. Commercial software packages such as PAAS software (DeGuire 1988) and are excellent at parsing and standardising addresses. The unresolved cases can be parsed manually as human beings are best at comparing the many types of addresses because they can associate the corresponding components in free-form addresses.

*Standardisation of postcodes*

The UK postcode is issued and supported by the Royal Mail and was designed to route the delivery of the mail. In recent years it has also been used as a proxy measure of geographic location in the absence of other measures. The National Grid Reference System used in the generation of the Ordnance Survey maps is the most accurate geo-spatial identifier, although most people are unlikely to know their grid references. The postcode consists of two parts, the inward code that designates the postal town and the outward identifying the postman's round.

The format of the postcode consists of one or two letters that designate the main postal town, followed by two or three numbers then two further letters. The code can be between five and seven characters long and sometimes a space character is inserted between the inward code and the outward code, for example OX3∇7LF (where ∇ denotes a space character). Errors occur where the code, for example OX12∇9EL, is truncated to fit into a seven character variable. Instead of removing the embedded space character the last character is normally

chopped off resulting in OX12∇9E. All postcodes used for matching purposes should have the embedded space character removed and the whole code left justified and stored in a seven character wide variable as shown below.

| | | | |
|---|---|---|---|
| OX12∇9EL | ➔ | OX129EL | (should be OX129EL) |
| B1 2SP | ➔ | B12SP∇∇ | (Where ∇ denotes a space character) |

**Creation of linked files**

*Creating additional records on the master file when different names are recorded in subsequent records for the same person*

The matching methodology must be capable of matching together all the records that have different surnames and forenames that any person may provide over the period of the file or the lifetime of the person. In most cases the need to change a name will arise from marriage, divorce or deed poll changes. Removing the old name from the system cannot be done since such records will contain historical data. To maximise the ability to match records the most effective method is to add the additional names entries to the index or master file for the person. It might also be advantageous to indicate which record contains the current preferred name or registration. This is analogous to adding extra cards in a card index, one under the present surname and another under the birth surname, and so on.

The fundamental decision in creating the blocks for record matching is whether the blocks based on the phonetic code of the names contain all the possible variants of the names for a given person. Where the phonetic codes are different the records may be spread over many different blocks. In this case the organisation of the file is more difficult and the only solution is to create multiple copies of records (one in each of the phonetic blocks) or multiple pointers to each record.

The multiple records on the master file are generated from all the combinations of present and birth surnames and forenames. To illustrate the generation of such extra records consider the example presented in Exhibit 7. Mrs Hall would have a master file record included in each of the eight ONCA/Initial blocks. A data record containing any combination of the above names would generate an ONCA/initial code similar to any one of the eight, which would then have a high probability of matching using any of the name variations at the matching stage. All eight entries on the master file, as shown in Exhibit 7, and would have the same person number and accession number since they are exact copies of the original record.

*Record headers and other variables that may be added to each record prior to matching*

It is good practice to add a record header to every record on both the master file and the data file. The use of such header records will improve the efficiency of file handling, blocking and sorting. A typical file header should contain those keys that will be used for the various matching processes, and an example is shown in Exhibit 8.

*Use of a record header when record matching using an encrypted file*

The ORLS have developed a record matching system that can match records in which the names and other identifiers have been encrypted. This is done in two stages. In the first stage, a header record is added to every record on the dataset, in which is stored all the blocking information, and example of the ORLS header is shown in Exhibit 8. The data in the record header are computed from the encrypted names and date of birth, and the resulting codes are generated and stored as plain text in the header area, while the variables in the rest of the record are left encrypted. Using this header the records are sorted in the appropriate blocks in both the data file and the master file. When the two records are to be compared, each is read from the input files into the computing system, they are decrypted and stored in the internal buffers of the matching program. Only in this way are the decrypted text identifiers made available to the matching programme.

At the end of the matching run all the buffers are cleared. The problem then arises about the security and storage of the computer listings that are generated for clerical checking. These listings should be stored in a locked environment and shredded and pulped when all the clerical checks have been completed.

**Exhibit 7: Creating extra records on the master file where there are many variations of the surnames and the forename**

Consider the record for a woman who has the following names information:

| | |
|---|---|
| Birth surname: | SMITH |
| Present surname (married surname): | HALL |
| First forename: | LIZ (contraction of Elizabeth) |
| Second forename: | PEGGY (contraction of Margaret) |
| Year of birth: | 1948 (old enough to be married) |

If file blocking is based on the phonetic code and the further divided by the first initial, eight identical records would be generated on the master file. Each record would be indexed under a different combination of the phonetic code (in the ORLS case the ONCA code) and first initial, as follows:

| | |
|---|---|
| blocked under the Present surname | HALL: i.e. (ONCA H400), |
| H400L | for Liz |
| H400E | for Elizabeth (formal version of Liz) |
| H400P | for Peggy |
| H400M | for Margaret (formal version of Peggy) |
| blocked under the Birth surname | SMITH, (i.e. ONCA S530),: |
| S530L | for Liz |
| S530E | for Elizabeth |
| S530P | for Peggy |
| S530M | for Margaret |

*Addition of historical pointers/markers*

In practice the ORLS have found it advantageous to add three further variables to the main part of each record for the allocation of arbitrary identification numbers, and to identify the coding systems used in each record.

The **accession number** is an arbitrary number allocated from a pool of such numbers and is absolutely unique to this record. The number is never ever changed and is used for the absolute identification of the record during the correction and amendment stages. It is also used as the main key where the files are separated into a number of sub-files, for example, name and address file (which contains no administrative or clinical data) and an analysis file (which contains no names or address data). In the ORLS system this number is check digited using the modulus 97.

The **person or system number** is an arbitrary number allocated from a pool of such numbers. This number must be the same on all the records that belong to the same person. The number can be changed or replaced according to preset rules where the data file record matches with the master file record. This number is check digited using the modulus 97.

| | **Exhibit 8:** | **Details of a typical ORLS record header** |
|---|---|---|
| 1. | Bytes 1-8 | The primary name blocking keys are generated using both the ONCA or the NYSIIS compressions of the present and the birth surname. |
| 2. | Bytes 9-16 | The full date of birth (CCYYMMDD), |
| 3. | Bytes 17-24 | Sex and postcode |
| 4 | Bytes 25-28 | First forename or initial |
| Note 1: | | Where the birth surname is present on the record and is not the same as the present surname, as would be the case for a married woman, a further record is generated on the master file under the phonetic code of birth surname and again sub-divided by the initial - (a process termed *expanding* the file) |
| Note 2: | | The secondary keys are normally generated using the sex or the initial letter of the first forename. Where this forename is a nickname or a known contraction of the 'formal' forename, the initial of the 'formal' forename is used. For example, if the recorded forename was Bill, the 'formal' forename would be William, and the initial used for blocking would be W. In practice two records would be generated, the first blocked on the initial B (for Bill) and the second blocked on the initial W (for William). |

Indicators are included that document the various editions of the coding frames used for each record such as the International Classification of Diseases (ICD) , surgical procedure codes (OPCS), or local codes e.g. social class. These indicators ensure that the correct coding edition is always recorded on each record and reliance is not placed on a vague range of dates

or other clues.

**Blocking and sorting data files**

When matching large files, it is not possible to compare all the record pairs since the number of possible match pairs is the product of the number of records on each of the two files. Even small files of say **m** and **n** records each would generate **m*n** pairs to compare. In practice, to reduce the time spent in attempting unproductive matches each of the input files should be partitioned and sorted into blocks prior to the matching stage. The records in a block on the data file would only be compared with those records in the corresponding block on the master file. Special provisions may be needed if multiple matches, each with a different blocking structure, are to be performed.

Consider the matching of a file against itself for the purposes of removing duplicates. If the file contained just 1,000 records the number of comparisons would be (1000*1000) / 2 = 500,000. Since the match between record x and record y is the same as the match between record y and record x only half of the matches are necessary. In this example it is expected that 1000 match pairs would be found and the remaining 499,000 would be unmatched pairs. Where the files are substantially larger, say one million records, the number of comparisons would be 500,000,000,000 and so on. This number of comparisons, even on the fastest machine, would prove to be uneconomic in terms of manpower and machine utilisation. For example, if the computing system could achieve one million matches per minute, the elapsed time to cross match the file of one million records would be 347 days.

If the file consisted of 1000 records and these were divided into ten blocks of 100 records each, the number of comparisons would be 10 * (100 * 100) / 2 = 50,000 or roughly one tenth of the effort required in the example in the previous paragraph. For the file with 1,000,000 records the corresponding record pairs that require matching would be 10,000*(100*100) / 2, = 50,000,000 or roughly one ten thousandth of the effort. Using a computing system that could achieve one million matches per minute the task would be completed in 50 minutes, compared with the 347 days for the unblocked file..

In practice, a file of one million records blocked by date of birth will have about 10,000 blocks of 100 records each and the processing time will be about 60 minutes.

The surname (or phonetic compression of it) is not as well distributed as that of the date of birth file and it is expected that there will be some very large blocks and some very small blocks. Using the name distribution, typically found on the ORLS files, and with a file of about of one million records, and based on 1 million matches per minute, it is expected that:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 100 blocks with 2,000 records | = | 200,000 records | = | 100,000,000 | = | 100 minutes |
| 200 blocks with 1,000 records | = | 200,000 records | = | 100,000,000 | = | 100 minutes |
| 1,000 blocks with 200 records | = | 200,000 records | = | 20,000,000 | = | 20 minutes |
| 8,000 blocks with 50 records | = | 400,000 records | = | 10,000,000 | = | 10 minutes |

Therefore, the elapsed time to complete all the 230,000,000 matches would be about four hours.

The use of blocking methods both support and expedite the matching process by reducing the number of potential match pairs that have to processed in any one phonetic or other type of block. Blocking precludes some of the possible matches since only those that fall within the same block are matched. The number of non-matches (missed matches) will be increased since many of the potential match pairs may have different blocking criteria and would fall outside the range of blocks being compared. To reduce this failure to match, additional match pairs are generated from the given set of records and inserted into the appropriate parts of the file. The selection of the blocking variable(s) is a crucial step in the record linkage process and it is necessary to check that a blocking variable does not have any unusually large values so creating very large blocks since these would greatly increase the processing resource required.

The choice of an optimum set of matching variables is basically a balance between reliability and discriminating power. These are essentially independent characteristics of any identifier when used for matching. Reliability serves to keep the non-matching losses down and discriminating power is needed to keep the amount of computer processing costs down. The definition of best matching implies a balancing of reliability against discriminating power. Usually, some limit will have to be placed on the acceptable number of comparison pairs of records that must be examined by the computer to do a matching job. Unless both files are small this precludes the matching of every data record with every record being searched. At the other extreme, a requirement for precise agreement on a set of identifiers will often result in an unacceptably high number of errors in the form of missed linkages. The optimum lies somewhere in between.

Using a larger number of matching variables generally improves the efficiency of the matching stage. However, there is a point at which the errors and omissions in the larger matching set outweigh the benefits of using the additional variables. It is preferable to use names variables that have more distinguishing power than using non-names or address variables, and their use generates smaller and more efficient blocks.

Since the blocking variable can contain errors and omissions it is usually necessary to re-block the file using other blocking variables. The files can then be sorted on a new combination of the matching variables, and re-matched. The matched outputs can be combined with those from previous runs and the record pairs merged together.

**Structuring and organising files to be record matched**

The two file structures that are normally used in record matching are:

- clustered or blocked flat file

- indexed database when matching will be undertaken between pairs of records in the database and the index modified to reflect that they belong to the same person.

The records in a clustered flat file are so organised that those records having similar content

defined by some specified criteria are located close together and are readily accessible as a group. Because retrieval from a bibliographic or image file is inexact and relies on similarity judgements, this clustered structure is widely used in information retrieval work. It is the type of file structure most commonly encountered in record matching and linkage systems.

The development of a clustered file depends on the combination of the matching variables used in the file blocking technique. Using statistical techniques it is possible to assess, on the basis of record characteristics, the probability that a group of records that relate sufficiently to the same person will be clustered together. A fundamental question in forming the clustered file is whether the clusters or blocks have common members, or are disjoint. In the latter case, organisation of the file is more difficult and may require multiple copies of records to be stored in the file or multiple pointers to each record. This is the approach adopted by ORLS for the preparation of an expanded file to cope with records that contain many name variations and where the records would fall into the adjacent blocks. Maintenance of a file with overlapping clusters or blocks is difficult because of the need to coordinate the multiple copies or reference pointers each time the file is changed or updated.

# 4. Exact or deterministic methods of record matching

## Introduction

Exact matching is the method of choice where a unique or a near unique identifier (single matching variable or a combination of partial matching variables) exists and the quality of data is relatively high. The method relies on the comparison of the identifier on the query file being matched against the identifier on the master file. In its simplest form the method has been implemented in the popular software packages SPSS® and SAS® under the option 'file merging and sorting'.

The key issues are:

- defining a unique identifier
- blocking, sorting and matching
- checking match validity
- resolving uncertainties
- avoiding risks associated with wrong matching.

## Defining a unique identifier

The main requirement for exact matching is the availability of a characteristic of a person that is unique, universally available, fixed, easily recorded and at the same time readily accessible and verifiable. The perfect identifier would be an integral trait of the person, or numbers allocated to the person by means of a highly reliable matching procedure. A unique identifier could be defined by using a series of numbers large enough to encompass all the members of the population.

Many numbering systems have been devised and these fall into three broad groups:

- Serial numbering systems in which a unique number is assigned to each individual from a central allocating point. Serial numbering systems have three main advantages in that they are simple to use, easy to automate and do not depend upon non-unique characteristics of the individual. The new NHS number can be used for this purpose since it is unique, almost universal and checkable and it is issued from one central allocation point.

- Derived numbering systems in which the number is derived using the readily available unique characteristics of the person. The advantage of such numbers is that they can be derived at any place and at any time without reference to a central allocation point. The disadvantage is that they depend on the person's stated characteristics and the risk

that two or more people will share the same characteristics.

- Composite numbering systems that are combinations of the serial and derived numbering systems. They use a central allocation point to obtain parts of the number and the non-unique characteristics to derive the other part(s).

All three types of number allocation schemes are prone to errors in the recording of the numbers, whether by speech, handwriting or keying. To improve the match rate and reduce the errors and missed matches it is desirable to incorporate a checking device in the serial number. One such method is to incorporate check-digits or check-characters into this key matching variable (Wild, 1968; Hamming, 1986; Gallian, 1989; Gill and Baldwin, 1982; Sethi, 1978; Dass, 1984; Brown, 1973; Holmes, 1975).

Where records contain these unique checkable numbers automatic matching can be rapid, reliable and inexpensive. Since the file is sorted on the blocking key and only records with identical identifying sets or keys are matched, this results in very fast matching.

Where unique numbers or ciphers are not available or not collected by the system, obvious candidates for use as matching variables are combinations of names, date of birth, gender and perhaps other variables such as address or postcode. The exact match will then depend on their use in combination. Much thought has to be given to the order of such partial identifiers, since the most reliable identifiers should be used for the more significant end of the combination key and the less reliable identifiers used for the least significant end. In concatenating date of birth, sex and postcode, the most reliable and discriminatory identifier would be the date of birth and the least reliable identifier would be postcode. The order of the concatenated identifier would then be date of birth/gender/postcode, and the system could then cope with partially complete postcodes.

**Blocking, sorting and matching**

Both the data and the master files should be ordered sequentially on the values of the single or compound identifier before attempting matching. The reliability and efficiency of the matching procedure is highly dependent on the manner in which the blocking is carried out. The simplest way to compare the identifier on the two files is to use brute-force (BF) algorithm. It consists merely of trying all possible character positions in the text string. For each such position it verifies whether the characters match at that position. When all the characters and the numbers agree the two records can be regarded as matched together. Conversely, if any characters in the string do not agree the records are deemed to belong to different people.

The outcome of the matching process is clear cut when using unique identifiers . Either the records match or they do not, depending upon whether the identifiers match or not. If there are any differences between the two strings then the match is deemed to fail and the two records are judged to belong to two different people.

Where the key has been built up from a number of partial identifiers some relaxation in the matching criteria could be used to bring together pairs that do not match exactly. The criteria for accepting the match would be in the form of a simple Boolean rule, for example records having the same date of birth same, sex same and just the first part of the post code (in area code). This relaxation would depend upon the number of master file records with which the data record matches and could only be resolved using clerical assistance. It is important not to confuse this with the threshold setting used in probabilistic matching.

**Checking the validity of the match**

Having matched the file on the unique identifier, the comparison of sex, date of birth and forename will usually suffice to check that the records have been correctly matched and do belong to the same person. However, problems will arise in the case of similarly named and same sex twins or for elderly people who may have a number of surnames (through many marriages) or they may have two or more forenames, synonyms or nicknames. Where the identifier is composed of partial variables like date of birth, sex and postcode, further checking will need to be undertaken using other variables in each record that are almost universally available.

**Resolving uncertainties**

The match can only determine whether the data record matches with the master file record or not. When this record is subsequently linked into the master file, and the logical checks fail, the output will need clerical intervention. Where there is any doubt about the match the data record should be regarded as unlinked.

Errors in the data will normally cause matching to fail and it will show up in the form of non-matches or generate ties. In these cases it would be more appropriate to use the exact matching method as the first step, followed by the use of probabilistic matching. It is common practice to combine the two record matching approaches. Exact techniques may be used as a first stage producing files that can be analyzed to generate probabilistic weights for further matching. Roos and Wajda (1991) and Wajda et al (1991) provided some specific examples of exact matching (also see Kendrick and Clarke, 1993).

**Risks associated with wrong matching**

The major risks in using data that are badly matched is that the data record will be linked up with records for a different person. Since the only matching options available in the exact match are True or False, the creation of a wrong match indicates that either the blocking or matching keys are wrong.

Where the matching key is an allocated number there may be two reasons for this error:

- wrong digit has been recorded in the identifying number as a result of a transcription error and hopefully this should be trapped using the check digit algorithm

- person has been issued with a number that had been issued previously to another different person .

Error or omissions in any one component part of the key built up from a number of partial identifiers, will result in collisions and wrong matching. In many cases this bad match can be detected when a logical check is performed across all the records for any given person. Sometimes it is almost impossible to determine whether the record belongs to the same person or not. In these cases the link cannot be made and the data and master file records are regarded as belonging to two different persons.

A number of major matching applications are concerned with improving coverage in surveys and censuses. In these cases a false negative is a catastrophic error because each non-match is added to a list as a new person.

Using keys based on serial numbers issued from some central allocation point and incorporating a check digit, it is estimated that the match rate should be about 95.6% for modulus 11 and 99.95+% for modulus 97 or the use of a double 11 modulus check digit. Some typical results are shown in Exhibit 9.

**Exhibit 9: Typical results obtained from exact matching datasets containing names or other matching variables.**

| | | |
|---|---|---|
| 1. | Primary matching of ORLS dataset<br>Exact name matching (Gill, 1987) | 87-90% |
| 2. | De-duplication of NHSCR<br>Exact name on surname/forename/sex/date of birth (Gill, 1994) | 75-85% |
| 3. | De-duplication of NHSCR<br>Exact matching using the NYSIIS code (Gill, 1997) | 90-93% |
| 4. | Exact match of a dataset against a part of the NHSCR<br>Exact matching using the NYSIIS code (Gill, 1998) | 87-90% |
| 5. | Exact matching of an ORLS (HES) dataset against itself using<br>Date of birth/sex/postcode | 97% |

Source: Various

## 5. Probabilistic methods of record matching

### Introduction

Probabilistic record matching and linkage is a process of statistical assessment and collation of information from two files that consist of records that might belong to the same person. This is of particular value when many of the records do not have universally available and unique identifiers. Two major complications may arise:

- All computerised records are prone to error, and these errors may occur because incorrect information has been obtained from the person or the data has been transcribed or keyed incorrectly. Because of such errors the records for the same person may not agree and therefore would not be considered to be matched at this time. Conversely, two records that do agree may belong to two different people.

- Parts of the record might be missing or are encoded in a different way. The missing data may occur in a random fashion, as happens when some data items are unreadable or lost. The loss could also be systematic, such as the lack of a second forename for a person who has only one name, or only one name has been recorded by the system.

The examples in Exhibit 10 illustrate the types of record pair in which exact and probabilistic matching are appropriate. In the example 1, each of the six identifiers is exactly the same, and the use of the exact matching methodology will report that the two records may belong to the same person. If there are no errors, or at the most very few errors in coding or transcription, then exact matching can be used. In example 2, the records appear to be for the same person, although three out of the six variables are different. This pair would generally fail to be linked using exact matching methodology, and a probabilistic method would probably be more appropriate.

**Exhibit 10: Cases where probabilistic matching is appropriate**

**Example 1:**

| | | | | | |
|---|---|---|---|---|---|
| HALL | STEPHEN | JOHN | Male | 220738 | 14 High Street |
| HALL | STEPHEN | JOHN | Male | 220738 | 14 High Street |

**Example 2:**

| | | | | | |
|---|---|---|---|---|---|
| HALL | STEPHEN | JOHN | Male | 220738 | 14 High Street |
| HALL | STEVEN | | Male | 220838 | 14 High Street |
| **p1** | **p2** | **p3** | **p4** | **p5** | **p6** |

The basic concept of probabilistic matching is best described in Smith (1984)'s description:

- Agreements of various identifying variables will generally argue in favour of a linkage, whereas disagreements will argue that the records relate to different people. Numerical weights can be used to quantify the fact that rare names, rare birthplaces and such carry more discriminating power when they agree than do their common counterparts.

To illustrate the basic theory of probabilistic linkage it is best to start with a simple example. For this exercise only full agreement or disagreement outcomes of specific variables within the identifying sets of variables are recognised and all partial agreements are disregarded. As a further simplification it is assumed that the identifiers selected for the comparison are always present on both the data file and the master file.

Probabilistic linkage uses weights based on frequency ratios, which give the likelihood that the records in the pair under consideration are truly linked, relative to the likelihood that the record pair is unlinked. Each frequency ratio in a sense represents the evidence in favour of a true match.

$$\text{FREQUENCY RATIO (for agreement)} = \frac{\text{relative frequency of agreement (x,y) among LINKED pairs}}{\text{relative frequency of agreement (x,y) among UNLINKED pairs}}$$

$$\text{FREQUENCY RATIO (for disagreement)} = \frac{\text{relative frequency of disagreement (x,y) among LINKED pairs}}{\text{relative frequency of disagreement (x,y) among UNLINKED pairs}}$$

where x indicates the value of the identifier on the query or data record**,** and
y indicates the value of the identifier on the master file record.

These are also termed *global frequency ratios* (as opposed to output-specific frequency ratios which will be discussed below) since the identifiers can take any value. These global frequency ratios give some indication of the value of the particular identifier for providing evidence of a true link between the records. Exhibits 11 and 13 give some illustrative examples. Exhibit 11 shows that agreement on surnames is more persuasive evidence of a true link (965:1) than agreement on forename or year of birth (which themselves are of approximately equal value having frequency ratios of 88:1 and 70:1 respectively). Disagreement on any of these variables is suggestive of a non-link but the evidence against a link is much less strong. This is probably due to errors in the coding of these variables. Exhibit 13 illustrates that agreement on sex alone provides very weak evidence for a link (+1) whereas disagreement on sex is powerful evidence for a non-link (-12). These examples illustrate the usefulness of frequency ratios as indicators of the value of agreements and disagreements on particular variables as evidence of a true link and non-links respectively.

To calculate global frequency ratios you need to know the frequencies of the various outcomes in a file of linked pairs of records and in a corresponding file of unlinked records.

Since matching and linkage will not have been done at this stage, the user will need to employ one of the following methods to obtain the frequencies :

- An initial linkage of pairs of records from a similar file (after first sorting and blocking to improve efficiency) could perhaps be carried out manually by exact matching or by some other means. This linkage does not have to be perfect nor the numbers of records large. About 1000 records would normally be sufficient.

- Where a previous linkage exercise has already been undertaken on other but similar files, the outcome frequencies from this process may be used instead. Care must be taken to ensure that the two files are of the same quality and represent the same population.

- Frequency ratios for some identifiers like date of birth can readily be calculated from their known distribution in the population. For example, month of birth has only 12 values and since birthday is approximately uniformly distributed throughout the year, the probability of matching in the absence of errors is approximately 1/12 in truly unlinked pairs and in linked pairs is 1. The probability of disagreement in truly linked pairs is 0 and in truly unlinked pairs is approximately 11/12. In this way it is usually possible to predict what the expected frequency ratio for unlinked pairs will be. The frequency ratio for agreement from these calculations is 12 and for disagreement is 0. In practice there would usually be some coding or transcription error rates to take account of, so these values would be slightly modified to allow for these errors.

- Another modification to this procedure is to remove the requirement to calculate the denominator frequency ratios from unlinked pairs. Often the relative frequencies in the denominator can, with little loss of accuracy, be calculated using pairs of records drawn at random from the file rather than specifically unlinked pairs. So the relative frequencies can be calculated from a single file, sometimes called a global file. This is an approximation, but for many purposes it is adequate.

- Another source of frequency ratios for name and address identifiers is to use previously published values. Newcombe (1998) gives some typical values that are shown in Exhibit 12 that take into account whether the particular values the identifier takes are common or rare events in the population. The intuition underlying outcome- specific frequency ratios is that surnames such as Zacharias occur less often than Smith. Thus pairs of records both containing the name Zacharias are more likely to indicate a true linkage than pairs of records both containing the name Smith.

Outcome-specific ratios are not difficult to use in principle with the following definition:

$$\text{OUTCOME-SPECIFIC FREQUENCY RATIO (for agreement on JOHN)} = \frac{\text{relative frequency of agreement (x,y) among LINKED pairs (where the identifier has a specified value, say JOHN)}}{\text{relative frequency of agreement (x,y) among UNLINKED pairs (where the identifier has a specified value, say JOHN)}}$$

where x indicates the value of the identifier on the query or data record, and
y indicates the value of the identifier on the master file record.

**Exhibit 11: Global frequency ratios for agreements and disagreements of selected identifiers on matched pairs of records.**

| Identifier | Comparison outcomes | Relative frequencies (%) | | Global frequency ratios |
|---|---|---|---|---|
| | | Links | Non-links | (links/non-links) |
| Surname | Agree | 96.5 | 0.1 | 965/1 |
| | Disagree | 3.5 | 99.9 | 1/29 |
| Forename | Agree | 79.0 | 0.9 | 88/1 |
| | Disagree | 21.0 | 99.1 | 1/5 |
| Year of birth | Agree | 77.3 | 1.1 | 70/1 |
| | Disagree | 22.2 | 98.9 | 1/4 |

Source: Newcombe (1988)

The mathematical basis of such intuitive assessments is really quite simple. The greater the ratio of the linked/unlinked frequencies, the greater will be the mathematical weight attached to any particular kind of agreement. Furthermore, one can simplify the process of the calculation using the proportions of occurrences of the string John, say on both the global and the matched file and without the need to create a file of unlinked pairs, since the difference between relative frequency among the unlinked pairs and that among all pairs (the statistic in Exhibit 12) will generally be very slight.

**Exhibit 12: Frequencies of specific values of identifiers**

| Type of identifier | Value of the identifier on the search record | Relative frequency in the file being searched |
|---|---|---|
| Present surname | SMITH | 1.22% |
| | HALL | 0.27% |
| | ZACHARIUS | 0.0001% |
| First forename | JOHN | 0.14% |
| | MARGARET | 0.06% |
| | LEICESTER | 0.0000076% |
| Month of birth | MARCH | 31/365 = 8.49% |

Source: ORLS, 1997

The specific issues that need to be addressed in the use of probabilistic matching are:

- generating match weights from frequency ratios
- combining weights over all the variables in the identifying set
- generating outcome specific weights
- blocking the file to reduce the number of unproductive matches
- file blocking and re-matching where the variables blocking key variables exhibit an error
- reducing the number of unproductive comparisons by constraints on matching
- matching the record pairs
- setting the matching threshold
- resolving uncertainties
- combining results from many match runs using different blocking keys
- reducing risks associated with wrong matching.

**Generating match weights from frequency ratios**

The various centres that have undertaken record linkage on a large scale have invested substantial development resources in the methods of generating weights. These weights are a measure of the logarithm of the likelihood-ratio that pairs of records, containing arrays of partial identifiers that may be subject to error or variation in recording, or do or do not belong to the same person. Decisions can then be made about the overall level of this ratio to accept or reject the pair for linkage. The use of these weights is an attempt to reduce the probability of:

- not matching the records that should be matched together
- matching together those records which should not be matched, (Type 1 and 2 errors) (Winkler, 1995; Scheuren and Winkler, 1996: Holmes, 1975; Jaro, 1995; Gill, 1997).

So far, frequency ratios have been determined for particular identifying variables to give some indication about their usefulness for matching. If the frequency ratio for agreement is very low then the variable is of little use. If it is very high then agreement on that variable is likely to be strongly indicative of a true match. The figures in Exhibit 12 show that agreement of surname is much more discriminatory than agreement on month of birth. When the agreement on surname occurs with a rare surname it is even more powerful for matching.

**Exhibit 13: Binit weights for some common identifiers**

| Agreements or disagreements | Relative frequency in linked pairs (L) | Relative frequency in unlinkable pairs (U) | Ratio L/U | Binit weight log2 L/U |
|---|---|---|---|---|
| ***Agreements*** | | | | |
| male Sex | 1/2 | 1/4 | 2 | +1 |
| Initial 'J' | 1/16 | 1/256 | 16 | +4 |
| Initial 'Z' | 1/1000 | 1/1000000 | 1000 | +10 |
| ***Disagreements*** | | | | |
| Sex | 1/8000 | 1/2 | 1/4000 | -12 |
| Initial 'J' | 1/40 | 32/40 | 1/32 | -5 |

Source: Newcombe (1988)

In practice, the frequency ratios are normally transformed using logarithms. The reason for taking logarithms is that multiplying frequency ratios (which as shown below are necessary when combining the evidence for linkage of a specific record pair over several variables) corresponds to the simple algebraic addition of the weights. Following the practice used in information theory, logarithms to the base 2 are used and these are called Binit weights, that are simply:

$Wt = \log_2 (L/U)$, where:

L = relative frequency of agreement/ disagreement among linked pairs

U = relative frequency of agreement/ disagreement among unlinked pairs

The Binit weights calculated using $\log_2$ are generated from standard $\log_{10}$ tables in the following way:

$$\log_2 (L/U) = \log_{10} (L/U) / \log_{10} 2 = \log_{10} (L/U) /0.30103$$

Some typical values for these weights are shown in Exhibit 13. Another way of thinking about them is as the power of 2 that equals the frequency ratio. In general, Binit weights for agreements will have positive values and for disagreements the weights will be negative.

When a variable is subject to coding or transcription error, the L relative frequency or probability is effectively one minus the error rate. The more reliable the measurement of the variable the closer to one the estimated probability will be. The U probability is the probability that a variable agrees given that the record pair is for different people selected at random. Since there are many more unlinked pairs than linked pairs this probability is, in most cases, effectively the probability that the values of the variable agree at random.

**Combining weights over all variables in the identifying set**

The question now arises of how to use the frequency ratios (and their equivalents – the single variable Binit weights) in a practical matching exercise. What is required is to obtain the equivalent of a single frequency ratio for all the variables in the identifying set that were used in the matching exercise.

Unsurprisingly, frequency ratios are used to calculate overall match rates**.** The frequency ratios for the agreements and disagreements that actually occur in each field are multiplied together.

Suppose that the frequency ratios in each of the three variables used in a particular matching exercise are fa1, fa2, fa3 and fd1, fd2, fd3 respectively for agreement and disagreement and, for the particular record pair being evaluated, there was agreement on variable 1, disagreement on variable 2 and agreement on variable 3. Then multiply together fa1, fd2, fa3 to obtain an overall frequency ratio or score, fa1*fd2*fa3, for what was observed in the specific record pair. This score can be interpreted as the likelihood of true linkage relative to the likelihood of non-linkage for this record pair based on this evidence. For example, the variables could be surname, sex and year of birth, then fa1=965 (see Exhibit 11), fd2=1/4000 (Exhibit 13), fa3=70 (Exhibit 11), the overall relative frequency would be 965*(1/4000)*70 = 16.9. So from this evidence, despite the fact that there was a disagreement on sex, a linkage is more likely than a non-linkage.

This calculation of the overall frequency ratio is based on the assumption that there is no correlation between the variables that constitute the fields. This is probably true for some variables, for example month of birth is usually independent of forename. However, some forenames and surnames are likely to be correlated in some populations. For example, in the population in the UK forenames and surnames that are characteristic of ethnic groups or particular nationalities are more likely to occur together. Marcel is more likely to go with Duchamp than Smith. However, in most practical matching exercises this complication is ignored and the frequency ratios are multiplied together. One solution to the problem of correlated variables is to use the combined variable as a single variable in the calculation of frequency ratios. Another example occurs with a change of address and hospital code. If a family move their home, one would expect that variables such as house number, street, town, family doctor and hospital code to change all at once. The frequency ratio can again be calculated considering all the variables as a single unit.

Of course, multiplying the frequency ratios corresponds to adding together the Binit weights. The information contained in every matching variable, when over all variables by adding the Binit weights, and the sum indicates which records should be identified as linked and which not. Each variable generally provides a part of the information but some variables, such as NHS number, provide more information than others such as sex. Nevertheless, it is usually the case that, taken together, the weights for all the variables would normally determine whether the two records match or not much better than any individual variable alone.

The total Binit weight now indicates, on a logarithmic scale, the ratio of the likelihood that

two records should be linked versus the likelihood that they should not be linked based on the information from all the variables in the identifying set (Smith, 1984; Howe and Lindsay, 1981). A total weight of zero represents a relative likelihood of 1:1 that the linkage is a correct one, each added weight unit doubling the relative likelihood and each subtracted unit halving it. For example, weights of + 1 and + 2 represent likelihood ratios of 2:1 and 4:1 respectively in favour of a correct match, whereas weights of -1 and -2 represent likelihood ratios of 1:2 and 1:4 and so argue against a correct match. The weights represent potential links in decreasing order of certainty (Howe and Lindsay, 1981; Newcombe et al, 1983).

*An example using matching on month and day of birth*

To clarify this process, consider a simple example in which matching is being done on month of birth in two records that belong to the same person. The variable is available 100% of the time and any differences will be due to coding or transcription errors and let us assume that the error is about 3%. The L probability will then be:

$$L = (1 - \text{error}) = (1 - 0.03) = 0.97.$$

Therefore the fields for month of birth for any given matched record pair will match 97% of the time. The U probability is the frequency of agreements of two records that do not belong to the same person. Since there are 12 months in a year the random agreement U will occur 1/12 or 8.3% of the time. The likelihood ratio of two records agreeing on month of birth will then be 0.97 / 0.083 = 11.64. This means that agreement on the month of birth alone increases the likelihood that the two records belong to the same person by a factor of 11.64.

In the comparison of each record pair, a composite weight may be computed as the sum of the individual Binit ($\log_2$) weights. Where the same variable agrees on each of the pair of records the outcome specific weights are computed as above. For a variable that disagrees in each of the pair of records being matched, the disagreement weight will be computed as:

$$\log_2 [ ( 1 - L) / (1 - U) ]$$

Therefore, the disagreement weight for month of birth will be,

$$\log_2 [ ( 1 - 0.97) / (1 - 0.0833) ]$$

$$= \log_2 [0.03/0.917] = \log_2 [0.0327] = -4.93$$

If the same calculation is carried out for the day of birth (assume an error rate of 5% and an average of 30 days in each month) the agreement weight will be

$$= \log_2 [ ( 1 - 0.05) / 0.03333) ] = \log_2 (28.50) = +4.83.$$

The disagreement weight will be

$$= \log_2 [ 0.05 / (1.0-0.03333) ] = \log_2 (0.0517) = -4.27.$$

**Exhibit 14: Probability ratios based on day-of-birth and month-of-birth**

| Outcome of comparison | Probability ratio | Log of ratio (base 2) |
|---|---|---|
| Month of birth (MOB) agrees | 11.64 | +3.54 |
| Month of birth disagrees | 0.0327 | -4.93 |
| Day of birth (DOB) agrees | 28.50 | +4.83 |
| Day of birth disagrees | 0.0517 | -4.27 |
| MOB and DOB agree | 331.74 | +8.38 |
| MOB agrees / DOB disagrees | 0.602 | -0.733 |
| MOB disagrees / DOB agrees | 0.932 | -0.102 |
| Neither agrees | 0.00169 | -9.21 |

Finally, if it is assumed that the month of birth and day of birth are independently distributed in the population and that reporting errors for matched pairs are independent, the probabilities shown in Exhibit 14 can be calculated. Usually, given the independence assumption, the probability ratio is broken up into a series of ratios, one for each agreement or disagreement, and logarithms are taken (to the base 2 in this example). The larger (more positive) the total the more likely it is that the pair is a match. Conversely, the more negative the sum the greater the likelihood that the two records are not for the same person.

In this example it is only when both day of birth and month of birth agree that the sum of the logarithms is highly positive (+8.38 obtained from 3.54 + 4.83). As one would expect, the strongest evidence in favour of a non-match occurs when both day-of-birth and month-of-birth do not agree. For this outcome the Binit value of the probability is about -9.21.

This example illustrates nicely the fact that outcomes that are frequent in the population do not add very much to one's ability to decide if the pair should be treated as linked. However, if there are disagreements on such variables and the reporting is reasonably accurate, a combination of the variables may have a great deal of power in identifying comparison pairs that represent non-links.

**Generating outcome-specific weights**

There is a further refinement about the weights that corresponds to the distinction made about frequency ratios. There are two types of probabilistic matching that differ according to how the logarithmic weights are generated. A linkage with global weights**,** as described above, generates the outcome weights based on whether a given identifier agrees or disagrees with its counterpart on the other file.

Probabilistic linkage with outcome-specific weights resolves matches more accurately. Although it is often based on the proportion of the links and non-links in the file, it involves

more intricate and sophisticated weight calculation tailored to the specific pair of outcomes under consideration, as discussed earlier when outcome specific frequency ratios were defined. Each outcome among several is assigned a weight.

In the introduction to this chapter, a formula is given for outcome-specific frequency ratios that can be used only when the results of a prior matching exercise are available. At the start of many matching applications there is no prior information on the population of linked and non-linked records nor the proportions of agreement and disagreement, so these weights cannot be calculated. Instead, approximate outcome-specific weights are generated. A good approximation to the outcome-specific weights (in units of $\log_2$, or Binits) for agreements is given by:

$$Wt = \log_2 (1 / p_x), \text{ where } p_x \text{ is the proportion of the given outcome in the population.}$$

In particular, note that the approximate outcome-specific weights are based just on the proportions of the particular outcome in the population. This procedure assumes that the variables are measured without error (the numerator is one) and that the global frequency (ignoring linkage status) is a good approximation to the probability of the outcome assuming no link.

In some situations, the tables of proportions can be created on-the-fly using the files actually being matched (Winkler, 1989c) while in others the tables of proportions are created a priori using large reference files. The advantage of on-the-fly tables is that they can use different relative proportions in different geographic regions, for example matching records with Asian surnames in Yorkshire and Leicestershire or Arabic names in London. The disadvantage of on-the-fly tables is that they must be based on files that cover a large percentage of the target population. If the data files contain samples from a population then the outcome specific weights should reflect the proportions of the appropriate populations.

*Characteristics of variables for record matching purposes*

Exhibit 15 shows the use and availability of variables in terms of discrimination, liability to change over a person's lifetime, liability to error and their availability in routinely collected data.

*Calculation of outcome-specific weights for surnames.*

The ORLS found that the outcome-specific weights calculated from the frequency of the first letter in the surname (26 different values) was too crude for matching together files that contained over one million records. The weights for SMITH, SNAITH, SNEATH, SMOOTHEY, SAMUDA and SZABO (which all fall into the same ONCA block S530) would all be set to some low value calculated from the frequency of the initial letter S in the population and would be based on the frequency of SMITH and ignoring the frequency of the much rarer SZABO.

Using the frequencies of all of the one million or so different surnames on the master match file is too cumbersome, too time consuming to keep up-to-date and operationally difficult to

store and access during a match run. The list would also contain most of the one-off surnames generated by poor transcription and bad spelling. A compromise solution has been devised by calculating the Binit weights based on the frequency of the ONCA block (roughly 8000 values), with a cut-off value of 1 in a 100,000 in order to prevent the very rare or one-off names from carrying very high weights (see Exhibit 16). Although this approach does not get round the problem of the very different names that can be found in the same ONCA block, it does provide a higher level of discrimination and can to some extent accommodate the low frequency or erroneous names.

**Exhibit 15: Relative characteristics of identifying variables used for person matching in health-related datasets.**

| Identifying variable | Characteristics | | | | Notes: |
|---|---|---|---|---|---|
| | Dp | St | Err | Cp | |
| Present surname | Av | Av | Hi | Hi | Always available |
| Birth surname | Av | Hi | Hi | Lo | Limited availability |
| Forenames | Av | Hi | Hi | Hi | Use at least two forenames |
| Date of birth | Av | Hi | Hi | Hi | Should be checked against age |
| Place of birth | Av | Hi | Av | Lo | Parent's residence at birth |
| Gender (sex) | Lo | Hi | Lo | Hi | Only variable that is unambiguous |
| Marital status | Av | Lo | Av | Lo | Useful for family linking |
| Usual address | Av | Lo | Av | Hi | Full postal address |
| Postcode | Av | Lo | Av | Hi | Full postcode |
| NHS number | Hi | Hi | Hi | Av | Must not be recorded from memory |
| GP | Av | Lo | Av | Hi | Useful for checking address |
| Date of marriage | Av | Hi | Av | Lo | For family record linkage |
| Mother birth surname | Av | Hi | Av | Lo | For family record linkage |
| Hospital unit number | Av | Hi | Av | Av | Valuable if has check character |
| Hospital site code | Av | Hi | Av | Av | Used with hospital unit number |

**Notes:**

| | | |
|---|---|---|
| Dp | = | Discriminating power |
| St | = | Stability, i.e. liability to change |
| Err | = | Liability to error |
| Cp | = | Availability, results of capturing this variable |
| Hi | = | High |
| Av | = | Average |
| Lo | = | Low |

Source: (Baldwin, 1972; Gill & Baldwin, 1987)

**Exhibit 16: Calculation of outcome-specific weight for a surname**

| | | |
|---|---|---|
| Frequency of **TAYLOR** | | 326,669 / 57,963,992 = 0.0056357 |
| Outcome-specific weight | = | $\log_2$ (1 / 0056357) = $\log_2$ (177.439) = 7.47 |
| Frequency of **SZABO** | = | 329 / 57,963,992 = 0.000005675 |
| Outcome-specific weight | = | $\log_2$ (176,182.347) = 17.43 |
| Truncated outcome-specific weight | | = 17.0 |

Source: NHSCR (1997)

A string comparison algorithm is normally used to compare the names on the two variables, one from the data file and the other from the master file, and to compute a weight modification factor (N) for weights computed from the frequency of the surnames (see section on string comparison methods). This algorithm computes the amount of agreement and disagreement in the two name strings and uses:

- length of the shortest of the two names being compared
- difference in length of the two names
- number of characters agreeing
- number of characters disagreeing
- number of character transpositions

The product of the agreement weight and the outcome-specific weight (Exhibit 16) would generate a modified weight that reflects the outcome-specific weight for the two names multiplied by the amount of agreement between the two names. If the two name strings are absolutely identical, the weight will be computed as +2N (N is the Binit weight) but would be reduced to -2N where the amount of disagreement is quite large or the names are dissimilar.

In cases where the birth surname and present surname are swapped with each other, or other surnames have been are recorded, expansion of the file would enable the location of, and access to, the blocks that contain the records stored under the different versions of the surnames. For a given record, the outcome-specific weight for the PSN/PSN (PSN=Present surname) and the BSN/BSN (BSN=Birth surname) pairings are first calculated, then the PSN/BSN and the BSN/PSN pairings are also calculated. The higher of the two values is used in the subsequent calculations for the derivation of the outcome-specific weight.

Where the marital status is recorded as single, or the sex is male, or the sex is female with an age less than 16 years, it is normal practice in the UK for the present surname to be the same as the birth surname. For this reason, only the weight for the present surname will be used for the determination of a match. It is assumed that where the birth surname is the same as the present surname no new information is being recorded.

*Calculation of outcome-specific weights for forenames*

The weights derived for the forenames are usually based on the frequency of the initial letter of the forename in the population. However, the weight may be generated in a similar fashion to that described above for surnames and examples are shown in Exhibit 17. The distributions of male and female forenames are different. There is usually one set of weights for male forenames and a different set for female forenames. Since the forenames can be recorded in any order, for example first forename (FN1) and second forename (FN2) in the first record and FN2 and FN1 in the second record, the weights for all the forename combinations are calculated and the highest values used for the match.

Where there are wide variations in the spelling of the forenames, the ORLS are evaluating the use of the Daitch-Motokoff version of Soundex for phonetically compressing and weighting the forenames similar to that used for the surnames.

*Calculation of outcome-specific weights for non-names variables*

The outcome-specific weights for date of birth (see Exhibit 18), sex, place of birth and NHS number are calculated using the frequency of the variable in the population. To allow for expected errors the weight for the year of birth comparison can been extended. For example, only a small deduction should be made where two date of births differ by exactly one or ten years. The weight is substantially reduced where two date of births differ by, say, seven years.

The calculated weight for the street address is based on the first eight characters of the full street address, where these characters signify a house number (31, High Street), house name (High Trees) or a public house name (The Red Lion). In parsing the address, terms like Flat or Apartment can be ignored and other parts of the address then used for the comparison. The calculation of weights for house numbers in the UK, are based on short streets with low house numbers, and are calculated using the proportion of house names in the population. The postcode is treated and weighted as a single variable although the inward and outward parts of the code could be weighted and used separately.

**Exhibit 17: Calculation of outcome specific weight for a forename**

| | |
|---|---|
| Frequency of SARAH | 24,951 / 28,989,996 = 0.0008606762 |
| Outcome specific weight | = $\log_2$ (1 / 00086067) = $\log_2$ (1161.877) |
| | = 10.18 |
| Frequency of LEICESTER | 44 / 28,989,996 = 0.0000015178 |
| Outcome specific weight | = $\log_2$ (1 / 0.0000015178) = 19.33 |
| Truncated weight | = 17 |

**Exhibit 18: Calculation of outcome specific weight for a date of birth**

Frequency of DAY of birth = 1/31

Outcome weight = $\log_2 (1 / (1/31)) = \log_2 (31)$ = 4.95 (usually rounded to 5)

Frequency of MONTH of birth = 1/12

Outcome weight = $\log_2 (1 / (1/12)) = \log_2 (12)$ = 3.58 (usually rounded to 4)

Frequency of YEAR of birth = 1/70

Outcome weight = $\log_2 (1 / (1/70)) = \log_2 (70)$ = 6.12 (usually rounded to 6)

Outcome weight for the full exact date of birth = 5+4+6 =15

Source: ORLS

*Limiting the values of the outcome-specific weights for very rare or unusual names*

The outcome-specific weights for the non-names variables are generated from the distribution of the variable in the population and, with the exception of house name, are evenly distributed. However, this is not the case for names information where the weights can be quite low for common names like Smith or John or very high for Zacharius or Leicester. In the calculation of outcome-specific weights for rare or unusual names the value of the weight needs to be restricted to prevent records being matched together irrespective of any other information that is present in the matched pair. To prevent situations with very high weights from exceeding the threshold value, the following approach has been adopted by the ORLS:

- impose a ceiling on the value of the outcome weight for rare names to an outcome weight of say 17, that is a probability of about 1 in 100,000

- use dual match thresholds, that is to consider the combined outcome weight for names and combined weight for non-names quite separately, in the form of a two dimensional orthogonal decision table.

*Calculation of outcome-specific weights where two variables do not agree*

The string matching algorithms are used to detect and quantify the differences between the two text strings. The outcome-specific weight will be positive for two similar name strings and negative for two dissimilar strings, for example:

| | | | |
|---|---|---|---|
| G600 | GRAY / GRAY | outcome-specific weight | = +19.0 (17.0) |
| G610 | GREAVES / GRAVES | outcome-specific weight | = +15.5 |
| G620 | GRACE / GEORGE | outcome-specific weight | = -2.0 |
| S530 | SMITH / SZABO | outcome-specific weight | = -12.0 |

The weights are calculated using the differences between the corresponding variables on the data and master files. Some weights can be modified to take into account the more common types of error. For example, the difference in the year of birth on the two records is usually one or ten years but very rarely three years discrepant. Examples of the match between two successive dates of birth for two records being matched on the ORLS file are shown below:

16041957 / 16041957 outcome specific weight = +15 (exactly the same date)

29031996 / 09031995 outcome specific weight = +3 (20 days/1 year discrepant)

11121907 / 16041908 outcome specific weight = -5 (5 days/8 months/1 year)

15102000 / 13041967 outcome specific weight = -15 (all different)

The weight can become negative where there is extreme disagreement between the variable on the data record and the corresponding variable on the master file. For example, in matching street address, postcode and general practitioner the weight cannot become negative, although it can be set to zero because the person may have changed their home address or their family doctor since they were last entered into the system. The latter are changes in family circumstances and not errors in the data and so a negative weight is not fully justified.

Exhibit 19 shows the range of outcome-specific weights that are typically used in the ORLS system based on 38 years of experience. The range of outcome specific weights is shown for each of the identifying variables.

For example, the match on exactly the same surnames would generate outcome weights in the range 2*5=10 up to 2*17=34 depending on the proportion of the surname in the population, i.e. Smith would be 2*5 = 10, and Zacharias would be 2*17=34. Likewise the forenames would be in the range 2*6 = 12 up to 2*170 = 34.

In similar fashion, comparison of two dates of birth would produce a weight of +15 if the two dates were exactly the same and in the range +15 to -15 if different. Where one date of birth is set to some arbitrary value (for example 01011852 or 99999999) the outcome weight would be set to -15.

*String comparison methods use for the calculation of outcome specific weights in names*

The use of algorithms for text searching in strings is a basic part of indexing and is also used for comparing strings for pattern matching. The comparison of two character or text strings is a classical problem to which a wealth of solutions exists and the rest of this section will cover the main algorithms used in record matching.

**Exhibit 19: The outcome specific weights used by ORLS for matching.**

| Identifying variable | | Outcome weight (see note #1) | | |
|---|---|---|---|---|
| | | Exact match | Partial match | No match |
| Surnames: | Birth surname | +2S | +2S to -2S | -2S |
| | Present surname (note #2) | +2S | +2S to -2S | -2S |
| | Mother's birth surname | +2S | +2S to -2S | -2S |
| *(where: common surname S = 5, rare surname S = 17)* | | | | |
| Forenames (note #3) | | +2F | +2F to -2F | -2F |
| *(where: common forename F = 6, rare forename F = 170)* | | | | |
| NHS number | | +10 | NP | -10 |
| Place of birth (code) | | +4 | +2 | -4 |
| Street address (note #5) | | +5, +7 | NP | 0 |
| Post code | | +4 | NP | 0 |
| GP (code) | | +4 | +2 | 0 |
| Sex (note #6) | | +1 | NP | -10 |
| Date of birth | | +15 | +13 -> -15 | -15 |
| Hospital/ unit number | | +7 | NP | -9 |

Source: Gill, 1997

Notes:

1. Where an variable has been recorded as not known, the variable has been left blank, or filled with an error flag, the match weight will be set to 0, except for special values described in the following notes below.
2. Where the surname is not known or has been entered as blank, the record cannot be matched in the usual way, but it is added to the file to enable true counts of all the events to be made.
3. Forename entries, such as boy, girl, baby, infant, twin, or not known, are weighted as -10.
4. Where the weight is shown as NP (not permissible) , this partially known value cannot be weighted in the normal fashion and is treated as a NO MATCH.
5. No fixed abode is scored 0.
6. Where sex is not known, blank, or in error, it is scored -10. (All records input to the match are checked against forename/sex indexes and the sex is set to M or F where it is missing or in error).

Dealing with typographical error is crucially important in record matching and linkage. If the strings are only compared on an exact character-by-character basis many potential matches may be lost. The Post Enumeration Survey (Winkler and Thibaudeau, 1991; Jaro, 1989) reported that among the many true matches, almost 20% of surnames and 25% of forenames disagreed. The ORLS found that only 75-85% of the matched records exactly matched on surnames alone. More than 30% of the matches would have been missed by the computer algorithms where the matching being done on a character by character basis resulting either in substantially increased clerical checking or a poorly matched file.

Where there are discrepancies between the two texts, various measured have been devised. One such measure uses the Levenstein distance, that is a distance measure between two text strings determined by the maximum number of insertions, deletions and substitutions required to transform one string into another string. With minimal modifications the algorithm can be adapted to searching whole words matching with *k* errors.

The algorithm first computes the number of insertions, deletions, transpositions, and the length of the string. The algorithm then uses this information (the length of the shortest of the two names being compared, the difference in length of the two names, the number of letters agreeing and the number of letters disagreeing) to calculate an agreement factor with which to generate the outcome weight. During the comparison adjacent pairs of characters are switched around to test for transpositions that might occur in either of the strings.

String comparison methods are used in the calculation of the outcome-specific weights between two name strings. The final outcome weight in the comparison of two names is based on the theoretical weight calculated from the frequency of the surnames in the population and modified using a method based on the Knuth-Morris-Pratt (KMP) algorithm (Gill, 1997; Stephen, 1994; Gonnet and Baeza-Yates, 1991, 1999; Crochemore and Rytter, 1994 ).

*Brute force string comparison methods*

The brute-force (BF) algorithm is the simplest one to use. It consists merely of trying all possible pattern positions in the text string. It verifies for each such position whether the pattern matches at that position (Baeza Yates, 1999).

Using the first string as the pattern the other string is matched against it. If it matches, fine. If not then the pattern is moved along by one character and another attempt made. The brute force technique involves an unnecessarily large amount of work. If one string contains **m** characters and the other string **n** characters, then in the worst-case **m*n** comparisons are necessary.

The brute force technique requires so much work because it repeats all of the possible comparisons. There is no memory of the comparisons that have been made. If there were some method of recording which comparisons had been made, the algorithm could be altered to omit these comparisons on subsequent occasions.

The following algorithms attempt to do this. Whenever a character in the pattern string is compared with a character in the text string, one of two things can happen, either they match or they do not. If they match, the examination continues with the next character. However, if they do not match the character that was last encountered in the text string is stored. The solution is then to use this stored information to reduce the number of succeeding comparisons.

*The Knuth-Morris-Pratt (KMP) algorithm*

The Knuth-Morris-Pratt known as the KMP algorithm (Knuth, Morris and Pratt, 1977; Baeza Yates, 1999) was the first to use the stored information about previous comparisons, although it is not much faster than the Brute Force algorithm. This algorithm slides a window over the text. It does not try all the window positions as does the Brute Force method. Instead it reuses information stored from the previous checks. The KMP algorithm is built around the

concept of storing the initial sequences of the pattern string. A modification of the KMP algorithm is used for the comparison of the strings in the ORLS system (Gill, 1997) and in the Winkler and Jaro systems (Jaro, 1989; Winkler, 1990b).

*Boyer-Moore (BM) algorithms*

Boyer-Moore algorithms although similar in concept to the KMP are based on the fact that the check inside the window can proceed backwards as well as forwards.

*The ORLS string comparison method*

The comparison algorithm used for the ORLS matching is presented in shown in Exhibit 20. The method is based on counting the number of characters in the two name strings that agree, the number that disagree, the number of character pairs that are transposed, the length of the longest string and the length of the shortest string. The final outcome-specific weight is computed from the initial weight determined from the frequency of the name in the population multiplied by the modification factor.

The algorithm has been refined to give extra weight to the characters that agree at the beginning of the string. Where the two name strings are absolutely identical the weight approaches +2N, (where N is the theoretical initial outcome specific weight based on the value of the variable) but falls down to a lower value of -2N where the amount of disagreement is quite large.

*Winkler-Jaro algorithm*

Jaro (1989) introduced methods for dealing with typographical error such as Smith versus Smoth. Jaro's procedure consists of two steps. First, a string comparator returns a value based on counting insertions, deletions, transpositions and string length. Second, the value is used to adjust a total outcome weight downward toward the total disagreement weight. Jaro's string comparator has been extended by making agreement in the first more important than the last characters of the string (Winkler, 1990b). The original Jaro comparator and the Winkler-enhanced comparator yield a more refined scale for describing the effects of typographical error than do standard computer science methods such as the Damerau-Levenstein metric (Winkler, 1985a, 1990b).

**Exhibit 20: ORLS string comparison method**

**Final weight = initial weight**
**+ { initial weight / #lenmin * (#agree - 3#disagree**
**- 1#diff - 1#transposition)}**

Where:

| | |
|---|---|
| Final weight | is the computed weight taking into account all the differences between the two strings |
| Initial weight | is the theoretical weight based on the frequency of the identifier in the population |
| lenmin | length of the shortest string |
| agree | number of characters that agree |
| disagree | number of characters that disagree |
| transpositions | number of two-character transpositions |

**Exhibit 21: Winkler and Jaro string comparator**

**Jaro(s1,s2) = 1/3(#common/length1 + #common/length2 +**
**0.5*#transpositions/#common)**

The basic Jaro algorithm computes the string lengths, finds the number of common characters in the two strings, and finds the number of transpositions, and is presented in Exhibit 21. The Jaro algorithm has been enhanced by McLaughlin (1993), Winkler (1993) and Lynch and Winkler (1994).

*Use of bigrams for string comparisons*

Another common method of comparing two strings is by comparing the bigrams that two strings have in common, and this algorithm is presented in Exhibit 22. The original algorithm cannot compensate for character transpositions in the strings, a modification of the algorithm has been developed by ORLS to switch the pairs of characters around in each bigram.

The bigram algorithm returns a value between 0 (very different strings) and 1 (very similar strings). Bigrams are known to be very effective in dealing with minor typographical errors and Porter and Winkler have shown empirically that they work well (Porter and Winkler, 1999; Frakes and Baeza-Yates, 1992).

**Exhibit 22: Comparison of two strings using bigrams**

LEICESTER ➔ LE EI IC CE ES ST TE ER
LESTER ➔ LE ES ST TE ER
Number of bigrams in common = 5
Average number of bigrams = (8 + 5)/2 = 6.5
comparison weight = number of bigrams in common / average number of bigrams
= 5 / 6.5 = 0.77

**Blocking the file to reduce the number of unproductive matches**

As previously described, all possible combinations of the record pairs on both files need to be tested against each other. Since the number of comparisons using this method is very large, in practice one or more of the matching variables are used to block the files The practice of blocking provides a practical method of limiting the number of pairs that have to be examined. If both the data and master files are blocked into mutually exclusive blocks and only matches within the blocks are undertaken, the whole process becomes manageable, economic and capable of being completed in a reasonable time. This results in a file which is divided into smaller blocks, and in which the candidate record pairs can be compared in a more efficient and less time-consuming way. One important consideration is that the number of records in each block should be small enough to avoid many unproductive comparisons and yet large enough to prevent records for the same person spilling over into different blocks and so failing to be compared. The best blocking variables are those with the highest number of values, the highest reliability, stability and the lowest error rates. Variables with a high probability of error or change should be avoided, for example, street address. In mathematical terms the variables with the highest weights make the best blocking variables.

**File blocking and matching where the blocking key variables exhibit errors**

All name search and record matching systems exhibit a conflict between performance and reliability. The accuracy and performance of the match depends on the choice and order of the blocking keys. Any key that is unreliable, even though it improves the performance of matching, cannot be used. File blocking should always be organised using the most reliable and accurate variable that is in the data set, provided that the variable has enough discriminating power to divide the file into small but manageable blocks. For example, gender is a well recorded and very reliable variable but it only has enough discriminating power to divide the file into two parts. The year of birth, although having a high discrimination, would be quite useless if it contains a high proportion of errors. ORLS use the phonetic code of the surname which divides the file into roughly 8000 blocks.

Many of the file blocking keys selected from a record will exhibit a small amount of error or omission, for example the date of birth may be a few days different or one month different or one year different. In using geographical identifiers the postcode may have the terminal character missing or the street address could have a spelling error. The portion of the address that is the most reliable and the most discriminatory is the dwelling or business number and

the street address. The full street address has the most discrimination but it also has the most error and lowest stability since people move around the country. The postcode, derived from the street address, will also suffer from this lack of stability.

One way of managing the error in the date of birth is to use the derived age, age group or a year of birth group. The benefit of using the year of birth rather than age is that the year of birth is fixed for all time whereas the age will change with each successive year.

It is often necessary to match two files together using dates for the blocking keys which contain some small errors, for example date of delivery and date of birth (up to three days discrepant) or date of discharge from care and date of death (up to three days discrepant). In these cases the files need to be blocked in overlapping blocks since the date on the first record could be one side of the true value and the date on the second record on the other side of the true value. In this example the records would fall into adjacent blocks. The solution will be to store an exact copy of the records in the appropriate block together with a similar copy of each record in the adjacent blocks. This blocking method would enlarge the size of the file but the probability of matching the two files together will be substantially increased.

Yet another way of blocking the file would be to generate the Julian Date (number of elapsed days from some arbitrary start date) from the date on the record and to generate an index record for this date, and extra records with Julian dates up to 3 days on either side of the index date. This technique has been used quite successfully for matching hospital records for young babies with their other hospital and registration records.

**Reducing the number of unproductive comparisons by matching constraints**

To reduce the number of unproductive comparisons, a data record should only be matched with a record in the corresponding master file block provided that pre-defined rules are satisfied. In the ORLS system records are only matched where the year of birth on both records is within 16 years of each other. This constraint is applied in order to reduce the number of unproductive matches and to restrict matching to those persons born within the same generation. In this way father/son, mother/daughter matches can be minimised. Further constraints could be built into the matching software to limit unproductive matches, for example, matching only within the same sex, logically checking that the dates on the two records are in a particular sequence/range, or that the diagnoses on the two records are in a specified range, such as that required for the preparation of a disease registry file.

**Matching the record pairs**

Within each block, every possible comparison pair from the data file and from the master file must be examined. In the examination of each pair the normal procedure is to observe the extent of the agreement or disagreement separately for each of the matching variables. For any particular variable the comparison outcomes can be coded or classified in a variety of ways.

Suppose, for example, that one matching variable is sex and that every record in both files has one of two codes for sex, male 1 and female 2. The outcome of a match could be restricted to agree or disagree on sex.

However, it may be desirable to extend this and use three categories:

- agree on sex, sex is male
- agree on sex, sex is female
- disagree on sex.

For some matching variables, such as surname, more complex matching rules may have to be devised to cover the presence of common names and the rare names. Although the file is normally blocked on the phonetic compression code (for example, all the records with names that fall in Block S530,) the string comparison algorithm uses the full surname stored elsewhere in the body of the record. In other systems the comparison of the two name strings may only use a limited but fixed number of characters such as the first four characters of the surname (Winkler, 1985b; Howe and Lindsay, 1981).

Other matching systems have been devised that take into account the likelihood of phonetic errors, transpositions of characters and random insertion, replacement and deletion of characters (Jaro, 1985, 1995; Winkler, 1993; Gill, 1997). One or other of the matching variables may either be missing or set to some default value in one or both members of the comparison pair. The string comparator needs to cope with these variations.

Generally, if resources permit all the variables that are judged suitable for matching should be used in the matching comparisons. Where this is not possible, the extra variables can still be used in the manual verification step where the outcome of matching a record pair may be indeterminate. Clerical intervention needs to be carefully limited and closely controlled since it is extremely costly and time consuming. While individual clerical decisions can sometimes be better than those made by computer matching, humans usually lack consistency of judgement and can be distracted by extraneous information resulting in random error.

In most applications of the Fellegi-Sunter model it is assumed that the agreement (or disagreement) on one matching variable is independent of that on any other, conditional only on whether or not the records that are brought together refer to the same person. To make this assumption plausible, special care needs to be taken in computing outcome-specific weights for variables like sex and first name that are inherently related. Fellegi (1985) and Kelley (1986) have undertaken simulation studies to investigate the robustness of the U.S. Census Bureau's linkage system to violations of the independence assumption. For a particular population and a given set of linkage variables, they found that violations of the assumptions may have significant effects on the levels of matching errors. The elements of the combination of geographic variables like address, postcode and general practitioner are also inherently related.

**Setting the matching threshold**

To recapitulate, the heart of the matching system is the procedure that assigns each comparison pair to one of two or more linkage categories, based on the weights computed and assigned in the previous stage. In the Fellegi-Sunter model the outcome weight is a function of the (logarithm to the base 2) likelihood ratio. In a probability-based system the decision

for each pair depends on the sum of the outcome weights for each matching variable. In the Fellegi-Sunter model the match weight is a function of the likelihood or probability ratio:

$$\text{Likelihood ratio} = \text{Probability ratio} = \frac{\text{Probability (result of comparison | given match)}}{\text{Probability (result of comparison | given non-match)}}$$

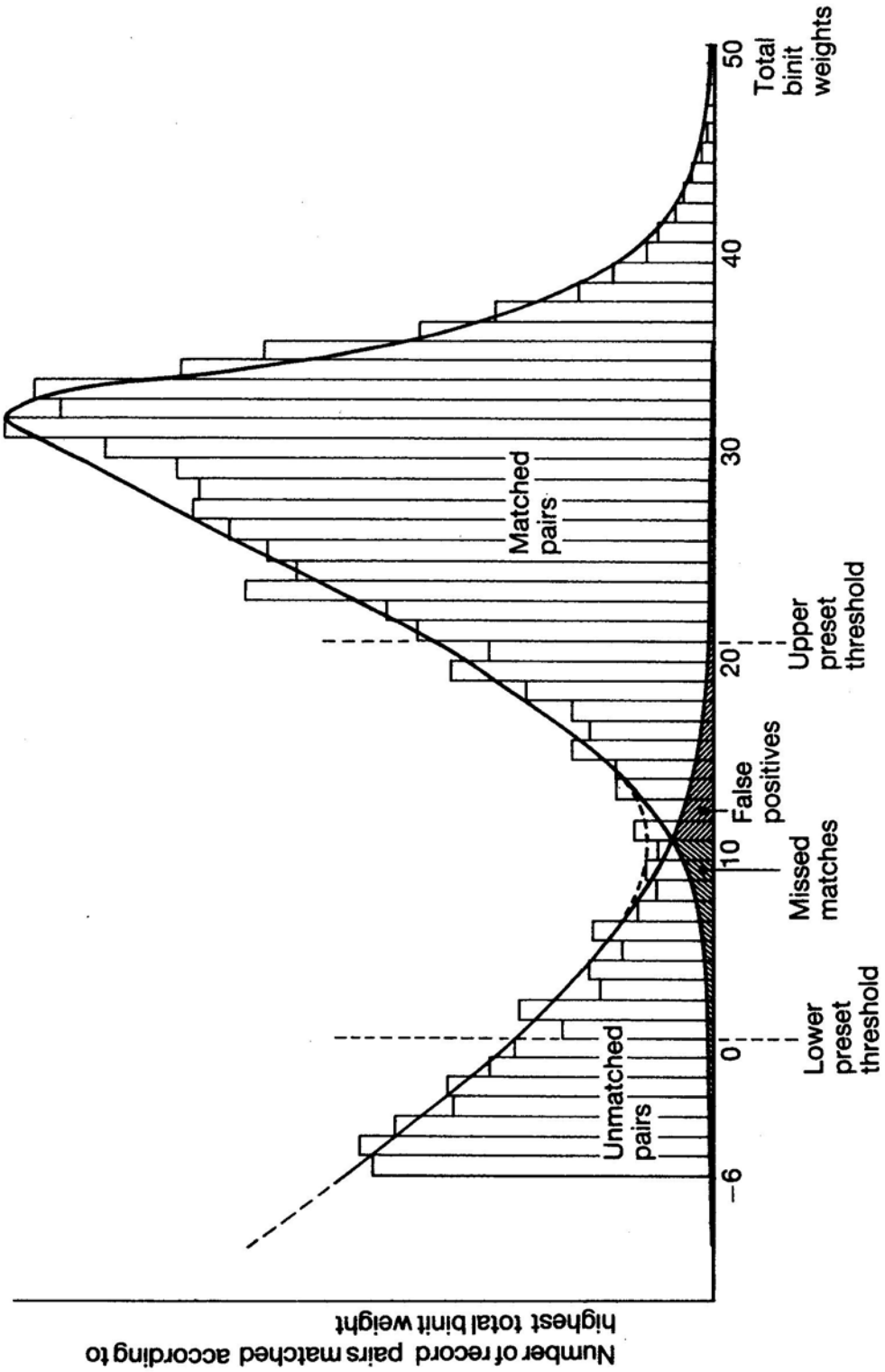
The numerator represents the probability that the comparison of two records for the same person would produce the observed result. The denominator represents the probability that comparison of records for two different persons, selected at random, would produce the observed result. In general, the larger the ratio the greater is the confidence that the two records belong to the same person.

The distributions of the Binit weights (logarithms of the overall likelihood ratios) for linked and the unlinked pairs in a particular study are shown in Exhibit 23. The peak on the right of the graph represents the linked pairs and the peak on the left the unlinked pairs. The left peak is thousands of times larger than the right peak and for simplicity it has been truncated. The two distributions overlap between the points designated as the lower preset threshold and the upper preset threshold.

In the Fellegi-Sunter model the comparison pairs are ordered according to the values of their weights. Two cut-off points are established. The higher of these separates the positive links from the possible links and the lower one separates the possible links from the positive non-links. These are shown as the upper and lower preset thresholds respectively in Exhibit 23. The establishment of appropriate cut-off values is a critical part of any record matching and linkage process. When a new record matching exercise is undertaken all the matches on a small sample of the file are examined clerically. The threshold values are then determined. In subsequent runs these threshold values are used for the determination of matches and non-matches. This process is recursive and the threshold values are so refined. At each stage clerks rigorously check the matches around the threshold values before the new and amended thresholds are used for the next run.

The objective of using the Fellegi-Sunter model is to place upper limits on the proportions of matched and unmatched pairs for which incorrect decisions are made. . In Exhibit 23 the shaded area from 0-10, denoted as missed matches, is a Type I error $\gamma$ and the shaded area from 10 upwards, denoted as false positives, is a Type II error $\mu$. In choosing the target values $\mu$ (the wrongly matched or Type II error) and $\gamma$ (the unmatched or Type I error) for these proportions, one should be aware that the number of unmatched pairs is very much larger (thousands of times) than the number of matched pairs. Therefore, it is usually desirable to make $\mu$ considerably smaller than $\gamma$; otherwise the false matches will tend to swamp the false non-matches Of course, the relative costs, in terms of clerical effor associated with resolving the two kinds of errors may also influence the choice of $\mu$ and $\gamma$.

**Exhibit 23: Frequency distribution of Binit weights for pairs of records.**




Source: Gill, 1997

As noted previously, some applications of model-based record-linkage systems require certain assumptions such as independence of errors in the matching variables. Nevertheless, there is reason to believe that the Fellegi-Sunter procedure is in general fairly robust to departures from independence (but see Fellegi, 1985; Kelley, 1986). Moderate errors in the estimation of weights can lead to different linkage decisions only for comparison outcomes whose weights are close to one of the cut-off points.

Furthermore, there is no theoretical obstacle to extending the underlying models to take into account known dependencies between the linking variables (Kirkendall, 1985). In trying to set up tables of these dependencies there would be significant computational problems. Nevertheless, the approach is entirely workable, especially since the development of advanced linkage software that includes the work of Jaro and his collaborators (Jaro, 1985; Winkler, 1993; Winkler 1989a; 1993b; 2000; Larsen, 1996; Armstrong and Mayda, 1993; Armstrong 2000). Other work to be noted includes GRLS (Hill, 1990), ORLS (Gill et al, 1993, 1997) and Scotland (Kendrick, 1993).

*Threshold setting and Type I and Type II errors*

As previously described, the procedure for deciding whether two records belong to the same person is based on the total outcome-specific weight, derived by algebraically summing the individual weights, each of which is calculated from the comparisons of each pair of identifying variables on the data file and the corresponding variable on the master file. This algebraic sum represents a measure of the (logarithm of the) likelihood that the two records are linked relative to the (logarithm to the base 2) likelihood that they are not linked. By comparing the total weight against a set of values that have been determined empirically, it is possible to decide whether the two records being compared actually refer to the same person.

The first type of error that occurs in matching is false negative, missed matches or Type I errors. These are the more common errors and are due to a failure to match records which refer to the same person, possibly due to erroneous or missing variables on one or both records. The records that should have been assigned to the one person are instead assigned two or more persons and the records are consequently not matched together. Matches may also be missed if the two records fall into different blocks. This may happen where a surname is misspelled, or the phonetic compression algorithm assigns the records into two different blocks.

The second type of error, false positive, wrongly matched records, or Type II errors, is less common but potentially more serious in assigning the same person number to the records for two or more different persons. This type of error arises when the two records belonging to two different people have identifying sets that are almost identical. This situation arises where the two people have very common surnames and forenames, or are similar sex twins. Problems also arise where a person has a number of forenames and they choose to use different forenames on different occasions. Unless other information is available with which to refute the match the records are best left unmatched.

The frequency of both types of error is a useful measure of the reliability of the record matching procedure. These are of course related to the target values $\mu$ and $\gamma$ of the Fellegi-Sunter theory.

In preparing earlier versions of the ORLS linked files a range of outcome weights was chosen and used to select records for clerical scrutiny. This range was confined by the upper and lower pre-set thresholds (see Exhibit 23). The false positive and false negatives are very sensitive to the threshold cut-off weight. If it is too low it gives a very low false positive rate and a high false negative rate. If too high it gives a high and unacceptable false positive rate with a low false negative rate. The values selected for the threshold cut-off are, of course, arbitrary, but must be chosen with care having considered the following objectives:

- minimisation of false positives at the risk of increased missed matches
- minimisation of missed matches at the risk of increased false positive
- minimisation of the sum of false positives and missed matches.

*Threshold problems when combining name and non-name weights*

The simple approach of algebraically summing the outcome weights ignores the fact that the weight calculated for names is based on the degree of commonness of the name and is passed on from other members of the family. Whereas, the weight for the non-names variables are based on distributions of those variables in the population, all values of which are equally probable. An unusual set of rare names information would generate high weights which would completely swamp any weights calculated for the non-names variables in the algebraic total. Conversely, a common name would be swamped by a perfect and identical set of non-names identifiers. This would make it difficult for the computer algorithm to differentiate between similarly named members of the population without resort to clerical assistance.

In the determination of the match threshold, a number of approaches have been developed, the earliest being the two stage primary and secondary match used in building the early ORLS files, through a graphical approach developed in Canada for the date of birth, to the smoothed two dimensional array approach developed by the ORLS and used for all its more recent matching and linking (Gill et al, 1987, 1993, 1997). Other advances in the methodology include the use of the EM Algorithm (Belin and Rubin, 1995; Larsen, 1996; Larsen and Rubin, 2000; Winkler, 1988, 1995; McLachlan and Krishnan, 1997) for the parameter estimation to determine the match thresholds.

*Orthogonal mapping techniques in the ORLS system*

Over the past ten years ORLS (Gill et al, 1993, 1997) have developed an approach in which a two dimensional orthogonal array is prepared, and in each cell of the array is stored the algebraic sum of the names weights forming the abscissa (X axis) and the algebraic sum of non-names weights forming the ordinate (Y axis). In the development of the method, computer runs on sample data were undertaken and the pairs of records rigorously checked by very experienced clerks to determine whether the pairs did or did not match.

The results of all these matches are stored in the cells of the orthogonal array designated by the co-ordinates (summed names weight, summed non-names weight). The empirical probabilities entered into the array were further interpolated and smoothed across the axes using linear regression methods and other curve fitting approaches including the use of cubic splines (Kelly, 1967; Hays, 1974; Borse, 1997).

Very experienced clerks rigorously checked over 400,000 matches and three counts were stored in each cell of the orthogonal array, designated by (sum of names weight, sum of non-names) weight, namely:

- total number of matches for that cell
- number of good matches
- number of non-matches.

A sample portion of this matrix is presented in Exhibit 24. One of the benefits of using two different axes for the matching threshold as described above is that a pair of records that contain a rare set of names or a perfect set of non-names information cannot be matched together unless there is good agreement on both axes. When a pair of records is being matched together, the matching software accesses the array and extracts the probability weight from the cell designated by the co-ordinates, as described above. The array of probabilities can be amended after each successive run, although minor adjustments or tinkering is discouraged. Precise scores and probabilities may vary, at least a little, according to the population and types of record pairs studied. A number of arrays have been prepared by the ORLS for the different types of event pairs being matched, including hospital to hospital records, hospital to death records, birth to hospital records, hospital and health authority records and cancer registry and hospital records.

A graphical representation of the array is shown in Exhibit 25, where each cell contains the empirical decision about the likelihood of a match between the record pair. The good matches are designated as Y, the non-matches as N and the doubtful matches that require clerical intervention as Q. This graph is the positive quadrant of four quadrants where both the names and non-names weights are positive and greater that zero. In the microcomputer implementation of the software this graph is held as a text file and can be edited using word-processing software.

Record pairs with weights that fall in the upper right part of the matrix are shown in Exhibit 25 as Y are considered to be 'good' matches and a 1% random sample of record pairs that fall near the Q-Y boundary are printed out for clerical scrutiny to measure the quality of the match. Record pairs with weights that fall between the upper and lower thresholds and shown in the figure as Q, are considered to be 'query' matches and all the record pairs are printed out for clerical scrutiny and the results keyed back into the computing system. Record pairs with weights falling below the lower threshold and shown on the map as N are considered to belong to two different people and a 1% random sample of record pairs that fall adjacent to the N-Q boundary is printed out for clerical scrutiny.

**Exhibit 24: Sample portion of the threshold acceptance array showing the number of matches and non-matches by outcome weight for names (X axis) and non-names variables (Y axis)**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WT=16 | Percentage | 37 | 41 | 45 | 58 | 83 | 77 | 91 | 98 | 99 | 99 | | |
| | Matches | 198 | 177 | 231 | 255 | 319 | 277 | 413 | 298 | | | | |
| | Nonmatches | 537 | 255 | 298 | 145 | 65 | 83 | 41 | 4 | | | | |
| WT=15 | Percentage | 41 | 38 | 42 | 56 | 61 | 75 | 87 | 98 | 99 | 99 | | |
| | Matches | 190 | 223 | 211 | 316 | 410 | 329 | 218 | 523 | 322 | | | |
| | Nonmatches | 273 | 364 | 293 | 245 | 265 | 109 | 33 | 11 | 4 | | | |
| WT=14 | Percentage | 18 | 25 | 21 | 19 | 31 | 56 | 77 | 93 | 89 | 97 | 99 | |
| | Matches | 113 | 87 | 90 | 110 | 190 | 198 | 660 | 422 | 161 | 377 | | |
| | Nonmatches | 514 | 261 | 330 | 460 | 412 | 162 | 197 | 34 | 19 | 11 | | |
| WT=10 | Percentage | 4 | 7 | 8 | 8 | 14 | 11 | 22 | 26 | | | | |
| | Matches | 17 | 35 | 28 | 34 | 50 | 50 | 69 | 75 | | | | |
| | Nonmatches | 341 | 404 | 284 | 382 | 277 | 295 | 235 | 203 | | | | |
| WT=9 | Percentage | 2 | 4 | 4 | 4 | 8 | 12 | 13 | 15 | | | | |
| | Matches | 18 | 42 | 28 | 47 | 64 | 90 | 90 | 87 | | | | |
| | Nonmatches | 737 | 966 | 637 | 952 | 706 | 644 | 588 | 474 | | | | |
| WT=8 | Percentage | 2 | 7 | 7 | 9 | 12 | 16 | 20 | 22 | | | | |
| | Matches | | 95 | 70 | 118 | 113 | 140 | 147 | 170 | | | | |
| | Nonmatches | | 1,234 | 812 | 1,106 | 785 | 728 | 583 | 588 | | | | |
| WT=7 | Percentage | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 4 | | | | |
| | Matches | 5 | 45 | 43 | 55 | 58 | 57 | 68 | 93 | | | | |
| | Nonmatches | 2,721 | 3,919 | 2,733 | 3,576 | 2,458 | 2,542 | 1,952 | 1,848 | | | | |

Source: Gill, 1997

At the end of each computer run the results of the clerical scrutiny are pooled with all the existing matching results and new matrices can then be prepared. The strategy is to reduce the Q zone to the minimum consistent with the constraints of minimum false positives (Type II errors) and false negatives (Type I errors). This will reduce the number of matches that require clerical intervention, invariably the most costly and rate determining stage.

Further matrices have also been prepared that record the number of match variables used when matching a record pair, for example the number of surnames present, the number of forenames or initials and the numbers of other matching variables. Since the number of matrices can become quite large, intelligent systems and neural net techniques are being developed by ORLS for the interpretation of the N dimensional matrices and the determination of the match threshold (Ripley, 1997; Kasabov, 1996; Bishop, 1995: Baldi and Brunak, 1999).

Special procedures are required for the correct matching of similarly named same sex twins. Where the match weights fall within the clerical scrutiny area and the records are printed out the clerks should identify the two records involved, mark the records in some agreed fashion and amend the match where necessary.

**Exhibit 25: A sample array used for matching hospital records with hospital records.**

```
          30  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
          29  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
          28  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
          27  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
          26  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
          25  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
  ↑       24  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
          23  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
  N       22  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
  O       21  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
  N       20  NNQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYYYY*
  -       19  NNQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYY*
  N       18  NNQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYYY*
  A       17  NNQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYY*
  M       16  NNQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYY*
  E       15  NNQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYY*
  S       14  NNQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYYY*
          13  NNQQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYY*
  W       12  NNNQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYYYY*
  E       11  NNNNNQQQQQQQQQQQQQYYYYYYYYYYYYYYYYYYYY*
  I       10  NNNNNNNNQQQQQQQQQQQQYYYYYYYYYYYYYYYYYY*
  G        9  NNNNNNNNNNNQQQQQQQQQQYYYYYYYYYYYYYYYYY*
  H        7  NNNNNNNNNNNNNQQQQQQQQQYYYYYYYYYYYYYYYY*
  T        6  NNNNNNNNNNNNNNQQQQQQQQQYYYYYYYYYYYYYYY*
           5  NNNNNNNNNNNNNNNQQQQQQQQQYYYYYYYYYYYYYY*
           4  NNNNNNNNNNNNNNNQQQQQQQQQQQQQQQQQQQQQQQ*
           3  NNNNNNNNNNNNNNNQQQQQQQQQQQQQQQQQQQQQQQ*
           2  NNNNNNNNNNNNNNNQQQQQQQQQQQQQQQQQQQQQQQ*
           1  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN*
              1        10        20        30        40
                          Names Weight  ➔
```

Where N = no match
Q = possible match (for clerical checking)
Y = definite match

Special procedures have been developed for elderly people who are recorded in the information system under a given set of forenames but, on a subsequent hospital admission or death, a different set of forenames are reported either by the person or by the next of kin.

*Use of orthogonal matrices to cope with different numbers of matching variables*

To some extent the use of orthogonal matrices based on a particular record type can take into account the vagaries of that dataset. However some of the variables may not be entered into the matching system. For example, some records may contain birth surname and for other types of record or period of time this variable may not have been collected. The summed outcome weight would contain the birth surname contribution for some matches and not for others and so some of the records may fail to be matched together.

Amended orthogonal arrays can be prepared for the number of variables that are entered into the matching process. Birth surname and second forename are variables that are most likely to be missed by some data capture systems but are the most valuable for accurate matching

**Resolving uncertainties**

After the preliminary linkage decisions have been completed there are usually some uncertainties that need to be resolved. These consist of pairs that have been classified as possible links or of multiple links, in other words, groups of linked pairs that have one or more records in common.

In the Fellegi-Sunter procedure, possible links are the pairs that fall between the upper and lower cut-offs. In the ORLS model, these records are assigned to the Q matrix weight. If resources permit these pairs may be reclassified as positive links or non-links either by collecting more data or by clerical review of the record content for these pairs.

If statistical estimates are to be made and the resources needed to seek further information are not available, the potential links may be treated as non-links and a survey-type non-response adjustment may be made (Scheuren, 1980). It is also possible to consider keeping some of the potential links and then conducting the analysis with an adjustment being made for mismatch (Scheuren and Oh, 1975).

Multiple links can occur in the Fellegi-Sunter formulation because the linkage decision is made independently for each pair. As a result a record from either file may be included in more than one pair whose weight exceeds the cut-off for a positive link. In some applications these many-to-one links might be appropriate but usually a further step has to be taken to select the ‘best’ one using a linear assignment algorithm.

*Clerical procedures*

The clerical review is usually the best basis for the final matching decisions, particularly where further information is sought or is available to help in this process. Some automated systems provide preliminary indications of the record pairs judged to be the likely candidates.

The National Death Index (NDI) operating system leaves it to users to resolve indeterminate cases. For each user record they list as possible links all death records that qualify under one or more of 12 sets of matching criteria (e.g. agreement on SSN and first name, agreement on SSN and last name and agreement on month and day of birth and first and last names). Users with small files usually resolve multiple links by clerical review. For large studies some users have developed their own computer algorithms for this purpose (Patterson and Bilgrad, 1985).

Users must also be prepared to determine final match status when only one possible link has been identified for a name submitted to the NDI.

Jaro (1985) suggests a computerised transportation algorithm to solve multiple linkage problems. His approach is most effective when all the linking information has already been computerised and when there are contentious problems in the linkages, that is, *n* records on one file are matching *m* records on another. The procedure is analogous to the constrained matching approach used in statistical matching. It picks a single best set of matches rather than picking the best match for each record in one of the input files. Armstrong (2000) describes an alternate one-to-one matching procedure used at Statistics Canada. It is essentially a greedy heuristic. Winkler (1994) gives details of a generalised assignment procedure that works better than the Burkard and Derigs assignment procedure.

ORLS undertook a clerical review of all the matches that fall in the review area. The doubtful matches are printed out and the clerks compare the variables that were used for the match. In some cases the computer decision can be reversed, especially where the clerk can resolve the partial information. In cases where there is some doubt records are always left unmatched.

Matching is only the first stage in the record linkage process. The second stage is linking the matched record with the other records for this person or entity. Rules can be prepared using Boolean constructs, such that the gender on both records must be the same, or the date of birth must be within one year. If the rules are satisfied the matched record can be linked in with the existing records. Where there is some conflict the whole set of records will need to be printed out for clerical scrutiny. This is especially the case where there is some error in the dates since the record may be inserted in wrong temporal sequence.

Even when many-to-one links are not appropriate, in theory it may be desirable to use the additional information they provide. Especially if conditions do not permit a clear determination of which of the links represent true matches.

Suppose, for example, that a record in File *A* is initially classified as a positive link with each of three records in File *B.* Three linked records could be established, each associating the File *A* record with one of the positive links from File *B.* The outcome weight associated with the File *A* record would be divided among the three linked records. One-third of it might be allocated to each linked record or it might be preferable to allocate it in proportion to the weights used in making the initial linkage decisions.

Difficulties with indeterminate cases can often be traced back to design flaws in the data linkage system. For example, not enough linking information may have been obtained on one or both files to assure uniqueness. The degree of redundancy in the identifiers may have been insufficient to compensate completely for the reporting errors.

In most record-linkage studies the matching is performed in a conservative manner with regard to the links that should be accepted. Sometimes this may mean additional expense to obtain more information or the risk of seriously biasing results by leaving out a large number of the potential links. In any event further research is needed on applying more complex analytical techniques that take explicit account of the false match rate, possibly by examining the errors in variable where the false match rate is estimated (Scheuren and Oh,1975). This

would allow a correction factor to be derived. Attempts to find methods of estimating the false match rate are being undertaken (Belin and Rubin, 1995; Winkler, 1988; McLachlan and Krishnan, 1997).

**Combining results from many match runs using different blocking keys**

File blocking enables all records having the same value in the blocking variable to be compared. One consequence of this strategy is that records not having this particular value in the blocking variable will automatically be classified as a non-match. In fact, if the blocking variable was age, and the age on one of the records was in error, then it would be considered as unmatched. To get around this problem multiple passes are used which are based on different blocking variables.

The blocking strategies for each pass should be as independent as is possible. The data file and the master file should be re-blocked and sorted on a new combination of the matching variables, for example present surname order, birth surname order, date of birth order, forename order or postcode order and so on. The results from each of these quite independent matches can then be combined and duplicate entries removed. The output from these matches will consist of a file of records, each of which holds two person numbers, one from the data file record and one from the master file record. This file can then be used to amend or update the linking of the data file to the master file or to update the index.

When the data record matches with a record on the master file a link is created between the two records, each of which would normally have a different person number. In some systems the two person numbers will be recorded in an index. In others the number from one record will be copied over that of the second, both records would then share the same person number.

Where there is a one-to-one correspondence between the person numbers on the two records the match can be readily accepted. Where there is a one-to-many or a many-to-one arrangement either the best matches (the one with the highest outcome weights) can be accepted or all the matches should be printed out for clerical scrutiny.

Records that cannot be matched together under a given Soundex block (for example Horton (H635) and Hawton (H350)) can be re-blocked and sorted using other identifiers such as date of birth, postcode or forename and matched in the normal way and the results combined.

**Reducing risks associated with wrong matching**

The major risks in using data that are badly matched is that either the data records will not be matched or will be matched to records for a different person. In most cases this bad match can be detected when a logical check is performed across all the records for any given person. Sometimes it is almost impossible to determine whether the record belongs to the person or not. In these cases the link cannot be made and the data record is then regarded as belonging to a new person. The risks can be reduced as follows:

- Select the matching variables and the order in which they will be used in the matching process. The general rule is that the variables that are universally available should be

used for blocking, bearing in mind that the variables should also be fixed, accurately recorded and have a high discriminating power.

- Check that the phonetic coding algorithm can cope with the different surnames in the population, since many spelling variations for common surnames may fall into separate blocks. The variations for any selected surname can then be stored in an agreed block using a lexicon, for example HAWTON(H350), HORTON(H635) and HOUGHTON(H235) , all the variations could point to one variant, say, (H350).

- Use a lexicon for the conversion of forenames and other matching variables to a standardised format.

- Calculate the outcome-specific weights for the population being matched since there are wide variations in the frequency of common surnames and forenames across a geographical area.

- Carry out many test runs with different values of the matching threshold criteria using a random sample of the data records. This will involve clerical checking of the matches made. However, the time spent in this activity will be more than offset by the better matching and the lower future clerical requirement.

During the course of a number of studies on hospital in-patient records and mortality it was found that the number of links lost through problems in the method of blocking the files using names identifiers ranged from 2% to 30%. These estimates were quantified by re-matching the file on a number of different keys and also examining the file for certain events that are expected to happen together such as a matched death certificate for a person who has died in hospital. Less likely are the false positive matches arising from the correspondence of the identifiers from two different people.

Typical results obtained from probabilistic matching of datasets containing names or other matching variables are shown in Exhibit 26.

**Exhibit 26: Typical results obtained from probabilistic matching of datasets containing names or other matching variables.**

| | | |
|---|---|---|
| 1. | Matching the ORLS dataset (all record types)<br>Using ONCA/year of birth (Gill, 1987) | 93-98+% |
| 2. | Matching the ORLS dataset (hospital records vs hospital records)<br>Using ONCA/year of birth (Gill, 1987) | 96-98+% |
| 3. | Matching the ORLS dataset (hospital records vs death records)<br>Using ONCA/year of birth (Gill, 1987) | 93-98% |
| 4. | De-duplication of NHSCR<br>Probability match using surname/forename/sex/date of birth (Gill, 1994) | 70-85% |
| 5. | De-duplication of NHSCR<br>Probability matching using the ONCA code (Gill, 1997) | 85-95% |
| 6. | Probability matching of an ORLS (HES) dataset against itself using the ONCA code | 97-98% |

Sources: Various

## 6. Collation of matched records into the matched file

**Introduction**

Once matched, the data records are merged with the master file in the following manner:

- Merge the matched data file with the existing master file
- Sort the combined file on the person number and other variables within each record. The variables normally used are start and end of episode dates. Check for overlapping dates or other logical inconsistencies within the set for each person.
- prepare the error outputs for the clerical staff to inspect and correct.

**Building the linked files**

It is a common practice to match the data file against the master many times, each time using a different blocking variable. The output from each of these matching runs is a text file that contains details about each pair of matched records. The output files are combined and sorted so that all the records for the same person are collated into the time sequence order, and the entries are then used to update the index, or amend the data file.

One method of applying the match to the data file is to copy the person number from the records on the master file to the data record, and then sort the combined file on the person number. Where there are a number of different person numbers in the person-set, because successive matching runs have been undertaken, the lowest person number is normally copied through all the records in the set. A second method for updating the data file is to use the accession number as the key to an index for all the records that refer to the same person. Using this index, the records can be back loaded into the database and all records for the same person can be grouped and analysed together.

In the ORLS system, the data file is matched against the master file using a series of independent runs, each of which uses the following blocking orders either singly or in combination:

- present surname
- phonetic compression of the present surname
- birth surname
- first forename or initial
- date of birth
- NHS number and other numbers such as hospital code and unit number

- year of birth/sex/postcode
- date of delivery and date of birth.

The format and content of these output files from a typical matching run are shown in Exhibit 27. Each record contains a reference for the data record and the master file record, and the results of the computerised match. The number of records written to the output file for any one person can be very large and is approximately the number of records that matched on the data file multiplied by the number that matched records on the master file. Combinatorial and heuristic algebraic methods have been developed to reduce all similar records in a set down to a small number, ideally one for each match pair (Hu, 1982; Cameron,1994; Lothaire, 1997; Reeves, 1995)

The allocation and use of the person number is crucial to the whole process of person matching. The details of the procedure are as follows.

Suppose, a record on the data file is matched with three other records, as shown in Exhibit 28. If the person has entered the same name(s), sex, date of birth or address on each occasion, all the records in the set 1-4 will have the same values of the matching variables. In this example, four person records are shown each of which has a different person number, (designated as 1451,1796,1845, 8735). These matches may have been the results of four independent blocking and matching exercises, however it can be seen that all four records do in fact belong to the same person. After the matching, merging and linking stages have been completed, the lowest person number is then copied across all four records, which in this case will be 1451. In similar fashion, use of the accession number can be used to indicate that records with accession numbers 1649, 2136 and 3798 all link to the record with accession number 1237. It is better to use the person number than the accession number since collating all the records for a given person is just a simple extraction and sorting process. This is a typical example, where the person has only one identifying set, therefore a single set of links will be set up for this person.

**Exhibit 27: Part outputs from the matching runs in the ORLS system**

| | |
|---|---|
| Details of data record | Person number<br>Accession number<br>Record type |
| Details of master file record | Person number<br>Accession number<br>Record type |
| Details about the match run | Output print stream (good match or query match)<br>Sum of the names weights<br>Sum of the non-names weights<br>Number of variables used in this match<br>Cross-reference to the clerical print-out<br>Clerical matching decision (either Y or N) |

Suppose that the person has changed their name during the span of the file, for example the person gets married, divorced or changes their name, there may be two sets of records on the data file and on the master file, the first set with records containg the original names information, and the second set with records after the name changes.

The matched records for the person will be grouped into two quite separate sets, those recorded under their first name, as shown in Exhibit 29 as records 1-4, and those recorded under the second name, shown as records 5-9). The person number allocated to the set 1-4, will be the lowest of the person numbers for this group, and the person number allocated to the set 5-9 will be the lowest of the person numbers for the second group. After linking, records 1-4 will be linked under person number 1451 and records 5-8 will be linked under person number 5451, and the two sets would be stored separately. This is a typical example where the person has two different identifying sets and so two sets of links will be set up for this person.

Matching on the combination of date of birth and first forename may indicate that set 1 (records 1-4) and set 2 (records 5-9) refer to the same person, however, if there is any doubt, then the two sets 1 and 2 should not be linked together at this time. In further matching runs, records with the present surname will match to set 1 and records with the birth surname will match to set 2, and two links will have been set up.

When a *bridging* record (set 3 in Exhibit 30) is subsequently matched to the file, it will set up independent links to the records from set 1 and set 3, and to the records from set 2 and set 3, and in this way set 1 may be linked to set 2.

**Exhibit 28: Sequence of records for a person with a single linked set**

Result of the record linking using the example below:

Records 1-4 Will be linked together (lowest person number = 1451)

Example of the person records within the linked set:

| | Birth Surname | Present Surname | First Forename | Second Forename | Sex | Date of Birth | Address | Accession Number | Old Person-number | Final |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Smith | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 1237 | 1451 | 1451 |
| 2 | Smith | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 1649 | 1796 | 1451 |
| 3 | Smith | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 2136 | 1845 | 1451 |
| 4 | Smith | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 3798 | 8735 | 1451 |

**Exhibit 29: Sequence of records for a person who has two linked sets**

Results of the record linking using the example below:

Records 1-4 will be linked together (lowest person number = 1451)
Records 5-9 will be linked together (lowest person number = 5451)

Example of person records from two linked sets:

| | Birth Surname | Present Surname | First Forename | Second Forename | Sex | Date of Birth | Address | Accession Number | Old Person-number | Final |
|---|---|---|---|---|---|---|---|---|---|---|
| Set 1: | | | | | | | | | | |
| 1 | | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 1237 | 1451 | 1451 |
| 2 | | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 1649 | 1796 | 1451 |
| 3 | | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 2136 | 1845 | 1451 |
| 4 | | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 3798 | 8735 | 1451 |
| Set 2: | | | | | | | | | | |
| 5 | Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 1256 | 5451 | 5451 |
| 6 | Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 1692 | 6796 | 5451 |
| 7 | Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 2165 | 7845 | 5451 |
| 8 | Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 3998 | 8741 | 5451 |

To illustrate this, using the example in Exhibit 30, records 1-4 (set 1) contain a present surname and so will be linked together under person number 1451. Records 7-10 (set 2) contain a birth surname and will be linked together under person number 5451. Records 5-6 (set 3, the *bridging records*) contain a birth surname and a present surname and so will link to each other and to either set 1-4 or 7-10.

Records 1-4 will match with records 5-6 (lowest person no = 1451)

Records 7-10 will match with records 5-6 (lowest person no = 3735)

Therefore records 1-4 will be matched with records 7-10 (lowest person no = 1451)

In the examples shown in Exhibit 30, recursive matching of all the record combinations will result in 10 * 9 / 2 = 45 runs, the result of each comparison will be a pair of person numbers. These pairs can be reduced recursively using combinatoric and heuristic techniques to the lowest value of the person number, which in this example will be 1451 (Hu, 1982; Cameron, 1994; Lothaire,1997; Reeves, 1995).

Where the person has had many changes of name or marital status, the number of different types of links will increase. Over the 38 year span of the ORLS system links up to five deep have been found. Unless bridging records are recorded in the file systems, for example, birth surname/present surname, present surname1/present surname2 combinations, the person would have their records divided over five parts. In matching exercises on files containing names from some ethnic communities links up to ten deep have been found, since it is difficult to determine the naming practices used by the different cultures.

**Exhibit 30: Sequence of records for a person who might have three or more different linked sets**

Result of the record linking using the bridging record:
First pass

| | |
|---|---|
| Records 1-4 will be linked together | (lowest person number = 1451) |
| Records 5-6 will be linked together | (lowest person number = 3735) |
| Records 7-10 will be linked together | (lowest person number = 5451) |

Then:

| | |
|---|---|
| Records 1-4 will link to 5-6 | (lowest person number = 1451) |
| Records 7-10 will link to 5-6 | (lowest person number = 3735) |

Finally, using the *bridging* record

| | |
|---|---|
| Records 1-6 will link to 5-10 | (lowest person number = 1451) |

Examples of person records from three sets of linked records:

| Birth Surname | Present Surname | First Forename | Second Forename | Sex | Date of Birth | Address | Accession Number | Old Person-number | Final |
|---|---|---|---|---|---|---|---|---|---|
| **Set 1:** | | | | | | | | | |
| 1 | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 1237 | 1451 | 1451 |
| 2 | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 1649 | 1796 | 1451 |
| 3 | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 2136 | 1845 | 1451 |
| 4 | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 3798 | 8735 | 1451 |
| Set 3 *(bridging records)*: | | | | | | | | | |
| 5 Smith | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 9623 | 5735 | 1451 |
| 6 Smith | Hall | Margaret | Elizabeth | F | 220749 | 14,High | 8542 | 3735 | 1451 |
| Set 2: | | | | | | | | | |
| 7 Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 1256 | 5451 | 1451 |
| 8 Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 1692 | 6796 | 1451 |
| 9 Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 2165 | 7845 | 1451 |
| 10 Smith | | Margaret | Elizabeth | F | 220749 | 14,High | 3998 | 8741 | 1451 |

**Sorting and logically checking the records**

The collated linked files are sorted on person number, type of event and date of the event. The sequence of records for any person should start with record at the beginning of the file, or their birth, and end with the last record on the file, or their death. All the intervening records should form a logical time ordered sequence. Various checks can then be performed to ensure that the dates for the person are logically correct such as the date of birth is before the date of death and that the various types of life events are in a logical temporal sequence.

Errors in the system due to administrative procedures need to be treated with caution since on some occasions a person who died in hospital on a Saturday will no be administratively discharged dead until the Monday, so there will be –2 days in the sequence of admission and discharge dates. Other factors that need to be examined include a person being in two types

of in-patient health care at the same time or being admitted to a second hospital before being discharged from the first hospital, although this can happen if the patient is resident in a psychiatry hospital and attends a general hospital concurrently.

**Outputs from the matching process**

Output files are normally extracted from the linked master file, and these can be used for statistical analysis, or for a new master file to matching new data. The statistical extract file will contain some of the original variables that were collected by the various capture systems, supplemented with a selection of derived or grouped variables that will assist in the analysis. To meet the requirements of the data protection legislation and the requirements of the GMC, some of the variables will have to be aggregated so that the identity of an individual is not revealed. The following derived variables are normally calculated from the input data items:

- age of the patient at a specified point in time, other than end of episode
- area classification derived from postcode
- deprivation score derived from postcode
- data borrowed across of all the records for a given person where there are gaps or omissions in the records
- smoothed data for some variables, for example date of birth, where there is some error, or noise in the original variables.
- aggregations of raw variables where this would help in the running of subsequent analyses, when using SPSS® or SAS®.

## Glossary and abbreviations

### Glossary

| | |
|---|---|
| ASCII | Standard codes to represent characters in one byte, defined by the American National Standards Institute. |
| Ad hoc request | Request for information that has not been previously specified. |
| Ad hoc retrieval | Standard retrieval task in which the user specifies information need through a query that initiates a search (executed by the information system) for records which are likely to be relevant to the user. |
| Algorithm | Description of an ordered and logical series of steps to be followed to solve a problem. This is usually in the form of a computer program. |
| Bias (systematic) | Any trend in the collection, analysis and interpretation of data that can lead to conclusions that are systematically different from the truth. |
| Binit Weight | Outcome specific weight computed for the agreement between two variables and represented in the form of logarithms to the base 2. |
| Blocking | Use of sequencing information to divide files into blocks. This method reduces the search for a match to those records in the same block. |
| Boolean logical operator | Group of search terms (eg ‘and’, ‘or’) available on the most common and widely used search engines, which help in refining the search strategy and retrieving information from a database. |
| Boolean model | Classic model of record retrieval based on classic set theory. |
| Browsing | Interactive task in which the user is more interested in exploring the document collection than in retrieving documents which satisfy a specific information need. |
| Byte | Computer representation of a character (q.v.). |
| Character | Unit of alphanumeric data. One of a set of symbols used to denote an alphabetic letter (a-z, A-Z), a number (0-9), or a special character (+=@*, etc). |
| Check characters | Letter or number incorporated into a persons NHS or other identity number, to provide a check on the validity of the number. |
| CHI number | Community Health Index number issued in Scotland. |

| | |
|---|---|
| Clustering | Grouping of records that satisfy a set of common properties. The aim is to assemble together records that are related among themselves. |
| Coding | Substitution of text symbols by numeric codes with the aim of encrypting or compressing text. |
| Compression of text | Study of techniques for representing text in fewer bytes or bits. |
| Concatenation | Concatenation of two strings a and b, ab is the string obtained by appending b to a. |
| Content-based query | Query exploiting data content. |
| Conversion | Change from one form to another, as in converting from analogue to digital. |
| Cumulative file | File that file that contains records for a cohort of people collected at different time points. |
| Data | Raw facts that have been collected, organised and stored in a computer file. |
| Data file | This data file will be matched against the master file to establish links between the records. |
| Data mining | Extraction of new data, relations, or partial information from any type of data file. |
| Data retrieval | Retrieval of variables (tuples, objects, Web pages, documents) whose contents satisfy the conditions specified in a user query. |
| Deterministic match | See Exact match. |
| Document | Unit of retrieval. It might be a paragraph, a section, a chapter, a Web page, an article, or a whole book. |
| Edit distance | Between two strings the minimum number of insertions, deletions and replacements of characters necessary to make two strings equal. |
| Exact match | Mechanism by which only the objects exactly satisfying some well specified criteria are matched together. |
| Expert system | Computer system that uses knowledge and inference procedures to solve problems that would require significant human expertise for their solution. |
| Field | In this context, a field is a piece of identifying information, for example, a name, date of birth etc. The variable is sometimes called a variable or an item. |
| File | Collection of similar records. |
| Flat file | File containing only fixed length records sorted into some arbitrary |

| | |
|---|---|
| | order. |
| Frequency ratio | Frequency of a given comparison outcome amongst correctly linked pairs, divided by the corresponding frequency among the unlinked pairs selected from a random sample. |
| Full text | Logical view of the documents in which all the words which compose the text of the document are used as indexing terms. |
| Fuzzy | Method of reasoning that resembles human reasoning since it permits approximate values and inferences and incomplete and ambiguous data. |
| Fuzzy data | Data that are incomplete or ambiguous. |
| Fuzzy model | Set theoretic model of document retrieval based on fuzzy theory. |
| Hashing | Method of accessing a file in which the address or key is calculated from the key data variable. |
| Heuristic methods | Proceeding to get a solution by trial and error methods. |
| Huffman coding | Algorithm for coding text in which the most frequent symbols are represented by the shortest codes. |
| IF/THEN/ELSE | Programming construct in which one of two possible outcomes is taken, depending on whether the logical condition is TRUE or FALSE. |
| Index | Mechanism by which a program orders the records in a file. |
| Index point | Initial position of a text element that can be searched for, for example a word. |
| Index term | Pre-selected term that can be used to refer to the content of a document. Usually, noun or noun group. In the Web, however, some search engines use all the words in a document as index terms. |
| Information retrieval | (IR) part of computer science that studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language. |
| Intranet | Internet type network built inside an organisation, which may or may not be connected to the Internet itself. |
| Item | In this context, an item is a piece of identifying information, for example, a name, date of birth etc |
| Keyword | See Index term. |
| Levenstein distance | Distance measure between two strings, given by the maximum number of symbol insertions, deletions and substitutions required to |

| | |
|---|---|
| | transform one string into another string. |
| Lexicographical order | Order in which the words are listed in a dictionary or telephone directory. |
| Lexicon | Look-up table for conversion of one string into another string. |
| Linkage | In the broadest sense, record linkage is the bringing together of information from two or more records that are believed to belong to the same person or entity. |
| Log2 | Logarithms to the base 2 that are calculated using logarithms to the base 10 in the following fashion: $\log_2 Wt = \frac{\log_{10} Wt}{\log_{10} 2} = \frac{\log_{10} Wt}{0.30103}$ |
| Longitudinal | Study that looks at the flow or sequence of events occurring over a given period of time to particular groups or cohorts of people. (See cumulative file). |
| Master File | File that is used as the primary source of data for a given job, it is relatively permanent, even though its contents may change. |
| Metasearch | Search technique common on the World Wide Web where a single point of entry is provided to multiple search engines. A meta search system sends a user's query to the back-end search engines, combines the results, and returns a single, unified hit-list to the user. |
| Metadata | Attributes of data or a record, usually descriptive as author or content, often broken up into categories or facets, and typically maintained in a catalogue. |
| Modulus | Number used as a divisor in check digit systems. |
| Natural language | Ordinary spoken language that humans use in everyday conversation. |
| Nearest-neighbour | Query that requests the spatial object closest to the specified object. |
| Near-neighbour | Words that are in close proximity to each other, for example BETSY and BETTY. |
| NHS Number | Number issued to everyone who is registered with a General Practitioner. The Number is issued for England and Wales by the NHS Central Register (NHSCR) in Southport, for Scotland by NHSCR, Edinburgh, and for Northern Ireland by NHSCR, Belfast. |
| Null Hypothesis | Hypothesis put forward when carrying out statistical significance tests, which states that: a) there are no differences between groups being compared, or b) there is no association or relationship between variables, in the studied population. |

| | |
|---|---|
| ORLS | Oxford Record Linkage System is a collection of person linked hospital discharge abstracts and vital records spanning 38 years. |
| Parsing | Resolve a sentence or string into its component parts and describe the various parts grammatically. |
| Phoneme | Any of the units of sound in a specified language that distinguishes one word from another (see phonetic). |
| Phonetic | Deals with the main sound units of speech and provides a direct correspondence between symbols and sound (see phoneme). |
| Population | Population for which the results of a given investigation are to be analysed. The distinction needs to be drawn between the wider population and the sample, this being a subset of the former. |
| Probabilistic model | Classic model of record retrieval based on a probabilistic interpretation of record relevance (to a given user query). |
| Query or data record | Record that is to be matched against the master file. |
| Query language | Computer language to retrieve specific information from a data base. |
| Record | Unit of information that contains a group of related data variables. |
| Rule-based system | Form of expert system in which human knowledge is captured as a series of IF/THEN/ELSE rules concerning objects or events. |
| Semantics | Meaning of words and phrases. |
| Sort key | Key used as the basis for reorganising the sequence of variables in a file or dataset. |
| Stemming | Technique for reducing words to their grammatical roots. |
| Syntax | Rules that are applied to matching words in order to form sentences. |
| Text file | Computer file that contains words and characters. Such files are commonly created in word processing activities. |
| Text structure | Information present in a text apart from its content, which relates its different portions in a semantically meaningful way. |
| Thesaurus | Data structure composed of (1) a pre-compiled list of important words in a given domain of knowledge and (2) for each word in this list, a list of related (synonym) words. |
| Type I error | In the context of record matching, a Type I error is where records that refer to the same person have failed to match together. |
| Type II error | In the context of record matching, a Type II error is where records |

| | |
|---|---|
| | have been matched together that in fact belong to two or more different people. |
| Variable | In this context, a variable is a piece of identifying information, for example, surname, date of birth etc. The variable is sometimes called a field or an item (q.v.). |
| Wild-card character | Character that will match any character or sequence of characters in a name or string. Typical wildcard characters are "*" and "?". |

## Abbreviations

| | |
|---|---|
| ANSI | American National Standards Institute |
| ASCII | American Standard Code for Information Interchange |
| CHI | Community Health Index |
| DETR | Department of Environment, Transport and the Regions |
| DfEE | Department for Education and Employment |
| DH | Department of Health |
| DHA | District Health Authority |
| EPR | Electronic Patient Record |
| FORTRAN | FORmula TRANslator, a high level computing language |
| GP | General Practitioner (Family Doctor) |
| GSS | Government Statistical Service |
| HES | Hospital Episode Statistics |
| KMP | Knuth, Morris, Pratt algorithm |
| LAN | Local Area Network |
| NHS | National Health Service |
| NHS number | National Health Service number issued by the NHSCR (qv) |
| NHSCR | National Health Service Central Register |
| NSTS | National Strategic Tracing Service |
| NYSIIS | New York State Identification and Intelligence System code |

| | |
|---|---|
| ONCA | Oxford Name Compression Algorithm |
| ONS | Office for National Statistics |
| ORLS | Oxford Record Linkage Study |
| RAM | Random Access Memory |
| SES | Socio-Economic Status |
| SQL | Structured Query Language |
| WWW | World Wide Web |

## References and bibliography

*For completeness other important references not quoted in the text are also included.*


Abrahams M. (1998) World Wide Web, Beyond the Basics. New Jersey, USA: Prentice Hall.

Acheson ED. (1967) *Medical record linkage.* London: Oxford University Press.

Acheson ED. (1968) Introduction. In: Acheson ED (ed) *Record linkage in medicine. Proceedings of the international symposium, Oxford, U.K.; July 1967*. Edinburgh: E&S Livingstone.

Acheson ED. (1987) Introduction. In: Baldwin JA, Acheson ED, and Graham WJ (eds), *A textbook of medical record linkage.* Oxford: Oxford University Press.

Alvey W, Aziz F. (1979) *Mortality Reporting in SSA Linked Data:* Preliminary Results. Social Security Bulletin 42 (11): 15-19.

Anderson TW. (1958) *An Introduction to Multivariate Statistical Analysis*, p.39. New York: Wiley.

Arellano MG, Petersen GR, Petitti DB, Smith RE. (1984) The California Automated Mortality Linkage System (CAMLIS). *American Journal of Public Health* 74: 1324-1330.

Arellano M. (1985) An Implementation of a Two-Population Fellegi-Sunter Probabilistic Linkage Model. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service. pp. 255-258.

Armstrong BK, Kriker A. (1999) Record Linkage - a vision renewed. *Australian and New Zealand Journal of Public Health* 23: 451-452.

Armstrong JA. (1992) Error Rate Estimation for Record Linkage: Some Recent Developments. In: *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis.* Carleton University.

Armstrong, JB, Mayda JE. (1993) Estimation of Record Linkage Models Using Dependent Data. *Survey Methodology* 19: 137-147.

Armstrong JB, Saleh M. (2000) Weight Estimation for Large Scale Record Linkage Applications. *American Statistical Association, Proceedings of the Section on Survey Research Methods.* (to appear).

"Ask Glenda". (1997) Soundex history and methods. World Wide Web, http://roxy.sfo.com/~genealogysf/glenda.html

Atallah MJ, Jacquet P, Szpankowski W. (1992) Pattern matches with Mismatches: A probabilistic analysis and a Randomised Algorithm. *Lecture Notes in Computer Science* 644: 27-40.

Atallah MJ. (1999) *Algorithms and the Theory of Computation Handbook.* Boca Raton, Florida, USA: CRC Press.

Aziz F, Kilss B, Scheuren F. (1978) 1973 Current Population Survey - Administrative Record Exact Match File Codebook, Part I, Code Counts and Variable Definitions. *Studies from Interagency Data Linkages*. U.S. Social Security Administration, No.8.

Baase S, Van Gelder A. (2000) *Computer Algorithms*. New York: Addison Wesley.

Baeza-Yates RA. (1989) Improved string searching. *Software Practice and Experience* 19: 257-271.

Baeza-Yates RA, Perleberg CH. (1992) Fast and approximate string matching. *Lecture Notes in Computer Science* 644: 182-189.

Baeza-Yates RA, Navarro G. (1996) A faster algorithm for approximate string matching. *Lecture Notes in Computer Science* 1075: 1-23.

Baeza Yates RA. (1999*) Modern Information retrieval*. New York: Addison Wesley.

Baldwin JA, Gill LE. (1982) The district number: a comparative test of some record matching methods. *Community Medicine* 4: 265-275.

Baldwin JA, Acheson ED, Graham WJ. (1987) Eds. *Textbook of medical record linkage.* New York: Oxford University Press.

Baylis J. (1998) *Error correcting codes, A mathematical introduction*. London: Chapman and Hall.

Beckley DF. (1967) An optimum system with 'modulus 11'. *Comput. Bull*. 11: 213-215.

Beebe GW. (1985) Why are epidemiologists interested in matching algorithms? In: *Record linkage techniques - 1985: Proceedings of a workshop on exact methodologies*. Publication 1299 (2-86), 139-44. Washington DC: Internal Revenue Service, Statistics of Income Division.

Belin TR, Rubin DB. (1995) A Method for Calibrating False-Match Rates in Record Linkage. *Jnl American Statistical Association* 90: 694-707.

Bishop CM. (1995) Three layer networks. In; *Neural Networks for Pattern Recognition.* pp. 128-129. United Kingdom: Oxford University Press.

Blakely T, Woodward A, Salmond C. (2000) Anonymous linkage of New Zealand mortality and Census data. *Aust NZ Jnl Public Health* 24 (1): 92-95.

Borse GJ. (1997) *Numerical methods with MATLAB®.* Boston, MA: PWS Publishing Company, ITP. pp. 363-367, 370.

Brenner H. (1994) Application of capture-recapture methods for disease monitoring: potential effects of imperfect record linkage. *Methods Inf Med* 33(5): 502-506.

Brenner H, Schmidtmann I, Stegmaler C. (1997) Effects of record linkage errors on registry-based follow-up studies. *Stat Med* 16 (23): 2633-43.

Brenner H, Schmidtmann I. (1998) Effects of record linkage errors on disease registration. *Methods Inf Med* 37 (1): 69-74.

Brown DAH. (1973) Biquinary decimal error correction codes with one, two and three check digits. *The Computer Journal* 17 (3): 201-204.

Brown DAH. (1974) Some error correcting codes for certain transposition and transcription errors in decimal integers. *Comput. J.* 17: 9-12.

Camargo KR, Coeli CM. (2000) Reclink: An application for database linkage implementing the probabilistic record linkage method. *Cad Saude Publica* 16 (2): 439-447.

Cameron PJ. (1994a) Graphs, Trees and Forests. In: *Combinatorics*. United Kingdom: Cambridge University Press. pp. 159-186

Cameron PJ. (1994b) Topics, Techniques, Algorithms. In: *Combinatorics*. United Kingdom: Cambridge University Press.

Carpenter M, Fair ME. (1989) Development of a data collection package for long term medical follow-up in Canada. In: Carpenter M and Fair ME (eds), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference,* Ottawa, August 30-31, 1989 (223-238). Ottawa: Statistics Canada.

Carpenter H, Fair ME. (1990) (eds) *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference,* Ottawa, August 30-31, 1989. Ottawa: Statistics Canada.

Chad R. (1993) A Comparison of three different Computer matches. *Special Census/Administrative Record match Working Group in Conjunction with the Year 2000*. Researcher Development Staff, U.S. Bureau of the Census, Washington DC.

Chandrasekaran C, Deming W. (1949) On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association* 44: 101-115.

Childers D, Hogan H. (1984) Matching IRS Records to Census Records: Some Problems and Results. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pp. 301-306.

Clarke EA. (1986) The Ontario Cancer Registry: From administration to research using GRLS. In: Howe GR and Spasoff RA (eds), *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, May 21-23, 1986. Toronto: University of Toronto Press.

Coombs JW, Singh MP. (1988) (eds) *Statistical uses of administrative data: An international symposium, November 23-25, 1987*. Ottawa: Statistics Canada.

Cobleigh C, Alvey W. (1975) Validating the Social Security Number. *Studies from Interagency Data Linkages*, U.S. Social Security Administration, No.4, pp. 89-123.

Coulter R. (1985) An Application of a Theory for Record Linkage. *Record Linkage Techniques – 1985*. U.S. Internal Revenue Service. pp. 89-96.

Copas JR, Hilton FJ. (1990) Record Linkage: Statistical models for matching computer records. *Jnl Royal Statistical Association, Series A* 153: 287-320.

Crochemore M, Rytter W. (1994) Text Algorithms. Oxford: Oxford University Press.

Crow S, Praus B, Thuras P. (1999) Mortality from eating disorders - a 5 to 10 year record linkage study. *Int Jnl Eat Disord* 26 (1): 97-101.

D'Andrea Du Bois NS. (1969) A solution to the problem of linking multivariate documents. *J Am Statist Ass* 64: 163-174.

Dale D. (1990) Linkage as a part of a production system: The Ontario Cancer Registry. In: Carpenter M, and Fair ME (eds), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference*, Ottawa, August 30-31, 1989. Ottawa: Statistics Canada.

Dass BK, Krishnamurthy MA. (1984) Linear unequal error locating codes (UEL) codes. *IEE Proceedings* 131: 52-56.

Day NE, Miller AB. (1988) (eds), *Screening for breast cancer*. Toronto: Hans Huber Publishers.

Devis T, Rooney C. (1997) The time taken to register a death. *Population Trends* 88: 48-55.

Dolby JL. (1970) An algorithm for variable length proper-name compression. *Journal of Library Automation* 3(4): 257.

Dunn H. (1946) Record linkage. *American Journal of Public Health* 36: 1412-1416.

El-Mabrouk N, Crochemore M. (1996) Boyer-Moore Strategy to Efficient String matching. *Lecture Notes in Computer Science* 1075: 24-38.

Evans JM, MacDonald TM. (1999) Record linkage for pharmacovigilance in Scotland. *Br Jnl Clin Pharmacol* 47 (1): 105-10.

Fair ME, Lalonde P. (1991) Application of exact ODDS for partial agreement of names in record linkage. *Computers and Biomedical Research* 24: 58-71.

Fair ME. (1993) Recent Advances in Matching and Record Linkage from a Study of Canadian Farm Operators and their Farming Practices. In: *1993 ICES Proceedings of the International Conference of Establishment Surveys*, American Statistical Association, 1429 Duke Street, Alexandria, Virginia 22314-3402. pp. 600-605.

Fair ME. (1995) An overview of Record Linkage in Canada. In: *Proceedings of the Social Statistics Section of the American Statistical Association 1995*. pp. 25-33.

Fair M, Cyr M, Allen AC, Wen SW, Guyon G, MacDonald RC. (2000) An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. The Fetal and Infant Health Study Group. *Chronic Dis Canada* 21(1): 8-13.

Farr W. (1861) Report on army medical statistics. *Parliamentary paper 366.*

Fellegi IP, Sunter AB. (1969) A theory of record linkage. *Journal of the American Statistical Association* 64: 1183-1210.

Fellegi I. (1985) Tutorial on the Fellegi-Sunter Model for Record Linkage. *Record Linkage Techniques – 1985*. U.S. Internal Revenue Service. pp. 127-138.

Frakes WB, Baeza-Yates R. (1992) (eds) *Information Retrieval: data structures and algorithms*. Upper Saddle River, NJ: Prentice Hall PTR.

Gallian JA. (1989) Check digit methods. *Int J Appl Eng Educ* 5: 503-505.

Gill LE, Baldwin JA. (1987) Methods and technology of record linkage: some practical considerations. In: Baldwin JA, Acheson ED, and Graham WJ (eds), *Textbook of Medical Record Linkage*. Oxford: Oxford University Press. pp. 39-54.

Gill LE, Goldacre MJ, Simmons HM, Bettley GA, Griffith M. (1993) Computerised linkage of medical records: methodological guidelines. *Journal of Epidemiology and Community Health* 47: 316-319.

Gill LE. (1997) 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.

Gill LE, Nichols P, Burd J. *The National Health Central Register. Farmer and Cross*

*revisited.* (paper in preparation).

Gill LE, Graveney MJ. (1996) Medical record linkage and the NHS number. *First report to the IMG NHS Strategic Tracing Service (AR) Programme Board, December 1996.*

Gill LE, Graveney MJ. (1997a) Optimising record linkage for the Initial Tracing Service. *Second report to the IMG NHS Strategic Tracing Service (AR) Programme Board, March 1997.*

Gill LE, Graveney MJ. (1997b) Record linkage optimisation for the strategic tracing service. *Third report to the IMG NHS Strategic Tracing Service (AR) Programme Board, August 1997.*

Gill LE, Graveney MJ. (1998) Record linkage and false positives, a thorough analysis. *Fourth report to the IMG NHS Strategic Tracing Service (AR) Programme Board, March 1998.*

Godber G. (1968) Record linkage. In Acheson, E.D. (ed), *Record linkage in medicine. Proceedings of the international symposium, Oxford, U.K.; July 1967.* Edinburgh: E&S Livingstone. pp.1-4

Goldacre MJ. (1986a) The Oxford Record Linkage Study: Current position and future prospects. In: Howe GR and Spasoff RA (eds), *Proceedings of the workshop on Computerised Record Linkage in Health Research.* Toronto: University of Toronto Press. pp. 97-103.

Goldacre MJ. (1986b) Confidentiality and access to data: the situation in the United Kingdom. In: Howe GR and Spasoff RA (eds), *Proceedings of the workshop on Computerised Record Linkage in Health Research.* Toronto: University of Toronto Press. pp. 106-129.

Goldacre MJ. (1987) Implications of record linkage for health services management. In: Baldwin JA, Acheson ED, and Graham WJ (eds), *Textbook of Medical Record Linkage*. Oxford: Oxford University Press. pp. 305-317.

Goldacre MJ, Simmons H, Henderson J, Gill LE. (1988) Trends in episode based and person based rates of admission to hospital in the Oxford record linkage study area. *BMJ* 296: 583-5.

Goldacre MJ, Seagroatt V. (1990) Case fatality rates as measures of outcome: studies using medical record linkage. *DH Yearbook of Research and Development 1990*. London: Department of Health. pp. 82-4.

Goldacre MJ, Henderson J, Graveney M. (1991) Readmission rates. *BMJ* 302: 414 (letter).

Goldacre MJ, Gill LE. (1995) Interpreting hospital death rates. *BMJ* 310: 536.

Goldacre MJ, Kurina L, Yeates D, Seagroatt V, Gill LE. (2000) Use of large medical databases to study associations between diseases. *Quarterly Journal of Medicine* 93: 669-

675.

Gonnet GH, Baeza-Yates R. (1991) Boyer-Moore text searching. In: *Handbook of Algorithms and Data Structures, 2nd ed*. United States: Addison-Wesley Publishing Co Inc. pp. 256-259.

Greenfield RH. (1977) An experiment to Measure the Performance of Phonetic Key Compression Retrieval Schemes. *Meth Infor Med* 16 (4): 230-233.

Gusfield D. (1997) *Algorithms on Strings, Trees and Sequences*. Cambridge, England: Cambridge University Press.

Hamming RW. (1986) *Coding and Information theory, 2nd ed.* Englewood Cliffs, N.J.: Prentice Hall.

Hassard TH. (1986) Writing the book of life: medical record linkage. In: RJ Brook, GC Arnold, TH Hassard and RM Pringle (eds), *Fascination of statistics.* New York: Dekker. pp.25-46.

Hattersley L, Creeser R. (1995) *Longitudinal Study (1971-1991). History, organisation and quality of the data.* OPCS Series LS No. 7. Office of Population Censuses and Surveys. London: HMSO.

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

Heasman MA, Clarke JA. (1979) Medical record linkage in Scotland. *Health Bulletin (Edinburgh)* 37: 97-103.

Henderson J, Goldacre MJ, Graveney MJ, Simmons H. (1989) Use of medical record linkage to study readmission rates. *BMJ* 299: 709-713.

Hill T. (1990) GRLS-Future directions in record linkage. In: Carpenter M and Fair ME (eds), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31,1989.* Ottawa: Statistics Canada. pp. 161-178.

Hill T. (1993) *GRLS V2 - Generalised Record Linkage Concepts*. Ottawa: Statistics Canada.

Holman D'Arcy CJ, Bass J, Rouse IL, Hobbs MST. (1999) Population-based linkage of health records in Western Australia: development of a health services research linked database. *Australian and New Zealand Journal of Public Health* 23: 453-459.

Holmes WN. (1975) Identification number design. *The Computer Journal* 14: 102-107.

Howe G, Lindsay J. (1981) A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies. *Computers and Biomedical Research* 14: 327-340.

Howe GR, Spasoff RA. (1986) (eds). *Proceedings of the Workshop on Computerized Record Linkage in Health Research.* Toronto: University of Toronto Press.

Howe GR. (1989) The use of computerized record linkage in health studies in Canada: The past, present, and future. In: Carpenter M and Fair ME (eds), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31,1989.* Ottawa: Statistics Canada. pp.3-18.

Hu TC. (1982) Heuristic algorithms. In: *Combinatorial algorithms*. United States: Addison-Wesley Publishing Co. Inc.

Jabine T. (1985) Properties of the Social Security Number Relevant to Its Use in Record Linkages. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service. pp. 213-225.

Jaro M. (1972) Unimatch - A Computer System for Generalized Record Linkage Under Conditions of Uncertainty. *AFIPS, Conference Proceedings.*

Jaro M. (1985) Current Record Linkage Research. *Record Linkage Techniques – 1985*. U.S. Internal Revenue Service. pp. 317-320.

Jaro MA. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 89: 414-420.

Jaro MA. (1995) Probabilistic Linkage of large Public Health data Files, *Statistics in Medicine* 14: 491-498.

Jordan-Simpson DA, Fair ME, Poliquin C. (1990) Canadian Farm Operators Study: Methodology. *Statistics Canada Health Reports* 2(2): 141-155.

Kasabov NK. (1996) Kohonen self-organising topological maps. In: *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering.* Cambridge, Massachusetts, United States: MIT Press. pp. 293-298

Kasprzyk D. (1983) Social Security Number Reporting, the Use of Administative Records and the Multiple Frame Design in the Income Survey Development Program. *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program.* Social Science Research Council. pp. 123-144.

Kelly LG. (1967) *Handbook of Numerical Methods and Applications*. USA: Addison Wesley Publishing Company.

Kelley R. (1984) Blocking Considerations for Record Linkage Under Conditions of Uncertainty. *American Statistical Association, Proceedings of the Social Statistics Section*. pp. 602-605.

Kelley R. (1985) Advances in Record Linkage Methodology: A Method for Determining the

Best Blocking Strategy. *Record Linkage Techniques - 1985.* U.S. Internal Revenue Service. pp. 199-203.

Kelley R. (1986) *Robustness of the Census Bureau's Record Linkage System*. Presented at the August 1986 meeting of the American Statistical Association.

Kendrick SW, Clarke JA. (1993) The Scottish Record Linkage System. *Health Bulletin (Edinburgh)* 51; 72-79.

Kendrick SW, McIlroy R. (1996) One pass linkage: the rapid creation of patient-based data. *Proceedings of Healthcare Computing 96. Current perspectives in Healthcare Computing 1996.* Weybridge, Surrey, UK: British Journal of Healthcare Computing Books. pp. 589-598.

Kendrick SW, Douglas MM, Gardner D, Hucker D. (1997) Best-link matching of Scottish health data sets. *Methods Inf Med* 37(1): 64-8.

Kilss B, Tyler B. (1974) Searching for Missing Social Security Numbers. *American Statistical Association, Proceedings of the Social Statistics Section*. pp. 137-144.

Kilss B, Scheuren F. (1978) The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin* 41 (10): 14-22.

Kilss B, Alvey W. (1985) Record linkage techniques - 1985. *Proceedings of the Workshop on Exact Matching Methodologies, Arlington, VA, May 9-10, 1985*. Washington: Internal Revenue Service.

Kirkendall N. (1985) Weights in Computer Matching: Applications and an Information Theoretic Point of View. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service. pp. 189-198.

Korfhage RR. (1997) *Information Storage and Retrieval.* New York: John Wiley and Sons Inc.

Kraft DH. (1985) Advances in Information Retrieval: Where is that /#*&@$ Record? *Advances in Computers* 24: 277-318.

Knuth DE. (1973) Non-numerical algorithms. In: *The Art of Computer programming*. Vol. 2. USA: Addison-Wesley Publishing Co. Inc.

Knuth DE. (1973) Sorting and Searching. In: *The Art of Computer programming*. Vol. 3. USA: Addison-Wesley Publishing Co. Inc.

Kullback S. (1968) *Information Theory and Statistics.* New York: Dover.

Lahiri P, Larsen MD (2000) Regression Analysis with Linked Data. *American Statistical Association, Proceedings of the Section on Survey Research Methods* (in press).
Larsen MD. (1996) *Bayesian Approaches to Finite Mixture Models*. Ph.D. Thesis, Harvard University.

Larsen MD. (1999) Multiple Imputation Analysis of Records Linked Using Mixture Models. *Statistical Society of Canada, Proceedings of the Survey Methods Section*. pp. 65-71.

Larsen MD, Rubin DB. (2000) Iterative Automatic Record Linkage using Mixture Models. *Statistics Department Technical Report*. University of Chicago.

Lothaire M. (1997) Words and Trees. In: *Combinatorics on words*. United Kingdom: Cambridge University Press. pp. 213-227.

Lynch BT, Arends WL. (1977) *Selection of surname encoding procedure for the Statistical Reporting Service record linkage system*. Washington DC: United States Department of Agriculture.

Maron ME, Kuhns JL. (1960) On relevance, probabilistic indexing and information retrieval. *Journal of the ACM* 3: 216-244.

McLachlan GJ, Krishnan T. (1997) *The EM Algorithm and Extensions*. New York: John Wiley and Sons Inc.

Marks E, Seltzer W, Krotki K. (1974) *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: The Population Council.

Marks E. (1985) Discussion (of paper by W. Winkler). *Record Linkage Techniques – 1985*. U.S. Internal Revenue Service. pp. 205-206.

Meadow CT. (1992) *Text Information Retrieval Systems*. New York: Academic Press Inc.

Miller AB. (1988) The Canadian National Breast Screening Study. In: Day NE and Miller AB (eds), *Screening for breast cancer.* Toronto: Hans Huber Publishers. pp.51-58.

Nathan G. (1967) Outcome probabilities for a record matching process with complete invariant information. *J Am Statist Ass* 62: 454-469.

National Center for Health Statistics. (1990) *National Death Index users manual.* DHSS Pub. No. (PHS) 90-1148. Hyatts-ville, MD: U.S. Department of Health and Human Services.

National Health Service and Department of Health. (1990) *Working for Patients: Framework for implementing systems: The Next Steps.* London: HMSO.

Neutel CI, Johansen HL, Walop W. (1991) New data from old epidemiology and record linkage. *Prog Food Nutr Sci* 15(3): 85-116.

Newcombe HB. (1957) Detection of genetic trends in public health. In: *Effect of radiation on human heredity*. Geneva: World Health Organization.
Newcombe H, Kennedy J, Axford S, James A. (1959) Automatic Linkage of Vital Records. *Science* 130 (3381): 954-959.

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

Newcombe H.B. (1968) Early stages of linked records. In: Acheson ED (ed), *Record linkage in medicine. Proceedings of the international symposium, Oxford, U.K.; July 1967.* Edinburgh: E&S Livingstone. pp. 7-34.

Newcombe HB, Smith ME, Howe, GR, *et al*. (1983) Relia bility of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology al1d Medicine* 13(3): 157-169.

Newcombe HB, Abbatt JD. (1983) *Probabilistic record linkage in epidemiology: Principles employed.* Ottawa: Eldorado Resources Ltd.

Newcombe HB. (1987) 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.

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

Newcombe HB, Fair ME, Lalonde P. (1989) Discriminating powers of partial agreements of names for linking personal records. Part I: The logical basis and Part ll: The empirical test. *Methods of Information in Medicine* 28: 86-91, 92-96.

Nienhuis H, Goldacre MJ, Seagroatt V, Gill LE, Vessey MP. (1992) Incidence of disease after vasectomy: a record linkage, retrospective cohort study. *BMJ* 304: 743-746.

Nigam K, McCallum AK, Thrun S, Mitchell T. (2000) Text classification from Labelled and unlabelled Documents using EM. *Machine Learning* 39: 103-134.

Nightingale F. (1997) *Letters from the Crimea 1854-1856*. Manchester: Mandolin Press.

Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group. (1989) Estimation of the incidence of acute myocardial infarction using record linkage: A feasibility study in Nova Scotia and Saskatchewan. *Canadian Journal of Public Health* 80: 412-417.

Office for National Statistics. (2001) *Mortality Statistics: injury and poisoning 1999.* Series DH4 No. 24. London: The Stationery Office.

Office for National Statistics. (2000) *Mortality Statistics: general 1998.* Series DH1 No. 31. London: The Stationery Office.

Oh H, Scheuren F. (1980) Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data. *American Statistical Association, Proceedings of the Section on Survey Research Methods*. pp. 416-420.

Oh H, Scheuren F. (1983) Weighting Adjustments for Unit Non-response. *Incomplete Data in Sample Surveys, Vol. 2.* Panel on Incomplete Data, U.S. National Academy of Sciences. pp. 143-184.

Patterson J, Bilgrad R. (1985) The National Death Index Experience: 1982-1985. *Record Linkage Techniques – 1985*. U.S. Internal Revenue Service. pp. 245-254.

Pidd M. (1996) Heuristic Approaches. In: *Tools for Thinking, Modelling in Management Science*. England: John Wiley and Sons. pp. 281-310.

Pidd M. (1992) *Computer Simulation in Management Science*. England: John Wiley and Sons.

Porter E H, Winkler WE. (1999) 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.

Post Office. (1962) The function and mathematics of check digits. *Report 170 of the Central Organisation and Methods Branch of the Clerical and Buildings Department.* London.

Raymond NT, Langley JD, Goyder E, Botha JL, Burden AC, Hearnshaw JR. (1995) Insulin treated diabetes: causes of death determined from record linkage of population based registers in Leicestershire. *J Epidemiol Community Health* 49(6): 570-574.

Reeves CR. (1995) Modern Heuristic Techniques for Combinatorial Problems. *Advanced topics in Computer Science.* Europe: McGraw-Hill Book Company.

Rodgers WL. (1984) An evaluation of Statistical Matching. *Journal of Business and Economic Statistics* 2: 91-102.

Rogot E, Schwartz S, O'Conor K, Olsen C. (1983) The use of probabilistic methods in matching census samples to the national death index. *American Statistical Association, Proceedings of the Section on Survey Research Methods*. pp. 319-324.

Roos LL, Nicol JP, Johnson C. Roos NP. (1979) Using administrative data banks for research and evaluation: A case study. *Evaluation Quarterly* 3: 236-255.

Roos LL, Roos NP, Cageorge SM, Nicol JP. (1982) How good are the data? Reliability of one health care data bank. *Medical Care* 20: 266-276.

Roos LL, Nicol JP, Wajda A. (1985) Improving the quality of data banks through linkage. *Chronic Diseases in Canada* 5: 81-82.

Roos LL, Sharp SM, Wajda A. (1989) Assessing data quality: A computerized approach.

*Social Science and Medicine* 28: 175-182.

Roos LL, Roos NP, Fisher ES, Bubolz TA. (1990) Strengths and weaknesses of health insurance data systems for assessing outcomes. In: Gelijns AC (ed), *Medical innovation at the crossroads. Vol I. Modern methods of clinical investigation.* Washington: National Academy Press. pp. 47-67.

Roos LL, Sharp SM, Cohen MM. (1991) Comparing clinical information with claims data: Some similarities and differences. *Journal of Clinical Epidemiology* 44: 881-888.

Roos LL, Wajda A. (1991) Record linkage strategies: Part I. Estimating information and evaluating approaches. *Methods of information in Medicine* 30(2): 17-123.

Roos NP, Montgomery P, Roos LL. (1987) Health care utilization in the years prior to death. *Milbank Quarterly* 65: 231-254.

Roos NP, Shapiro E, Tate R.B. (1989) Does a small minority of elderly account for a majority of health care expenditures? A sixteen year perspective. *Millbank Quarterly* 67: 347-369.

Roos NP, Wennberg JEM, Malenka DJ, *et al.* (1989) Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *New England Journal of Medicine* 320: 1120-1124.

Roos NP. (1989) Using administrative data to study outcomes: Developing control groups and adjusting for case severity. *Social Science and Medicine* 28: 109-113.

Roos NP, Havens BJ. (1991) Predictors of successful aging: A twelve year study of Manitoba elderly. *American Journal of Public Health* 81: 63-68.

Rosenfield and Morville. Information Architecture for the World Wide Web.

Scheuren F, Herriot R. (1975) The Role of the Social Security Number in Matching Administrative and Survey Records - General Introduction and Background. *Studies from Interagency Data Linkages, No. 4.* U.S. Social Security Administration. pp. 1-7.

Scheuren F, Oh H. (1975) Fiddling Around with Non-matches and Mismatches. *American Statistical Association, Proceedings of the Social Statistics Section.* pp. 627-633.

Scheuren F. (1980) Methods of Estimation for the 1973 Exact Match Study. *Studies from Interagency Data Linkages, No. 10.* U.S. Social Security Administration. pp. 1-123.

Scheuren F. (1983) Design and Estimation for Large Federal Surveys Using Administrative Records. *American Statistical Association, Proceedings of the Section on Survey Research Methods.* pp. 377-381.

Scheuren F, Winkler WE. (1991) Regression analysis of data files that are computer matched. *Proceedings of 1991 Census Annual Research Conference.* Washington: Bureau of

the Census. pp. 669-687.

Scheuren F, Winkler WE. (1993) Regression analysis of data files that are computer Matched. *Survey Methodology* 19: 39-58.

Scheuren F, Winkler WE. (1996) Recursive merging and analysis of administrative lists and data. *American Statistical Association, Proceedings of Survey Research Methodology*.

Scheuren F, Winkler WE. (1997) Regression analysis of data files that are computer Matched II. *Survey Methodology* 23: 157-165.

Secretaries of State for Health, Wales, Northern Ireland and Scotland. (1989) *Working for Patients.* CM 555. London: HMSO.

Sethi AS, Rajarman V, Kenjale PS. (1978) An error-correcting coding scheme for alphanumeric data. *Information Processing Letters* 7 (2): 72-77.

Shannon HS, Jamieson E, Walsh C, *et al*. (1989) Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base. *Canadian Journal of Public Health* 80: 54-57.

Smith ME, Newcombe HB. (1980) Automated follow-up facilities in Canada for monitoring delayed health effects. *American Journal of Public Health* 70: 1261-1268.

Smith ME, Silins J. (1981) Generalized Iterative Record Linkage System. *American Statistical Association, Proceedings of the Social Statistics Section.* pp. 128-137.

Smith ME. (1984) Record linkage: Present status and methodology. *Journal of Clinical Computing* 13: 52-69.

Smith ME. (1985) Record-Keeping and Data Preparation Practices to Facilitate Record Linkages. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service. pp. 321-326.

Social Statistics Research Unit. (1998) The ONS Longitudinal Study. *List of publications arising from Research: April 1998.* Longitudinal Study Support Programme. London: City University.

Stephen GA. (1994) Knuth-Morris-Pratt Algorithm. In: *String searching algorithms*. Singapore: World Scientific Publishing Co. Pte. Ltd. pp. 6-25.

Steinberg J, Pritzker L. (1967) Some experiences with and reflections on data linkage in the United States. *Bulletin of the International Statistical Institute 42(2)*: 786-808.

Taft RL. (1970) *Name search techniques*. Research Report. New York: New York State Identification and Intelligence System.

Tepping B. (1968) A Model for Optimum Linkage of Records. *Journal of the American Statistical Association* 63: 1321-1332.

The Caldicott Committee. (1997) *Report on the Review of Patient-Identifiable Information. December 1997*. London: UK Department of Health.

*The NHS Plan – A Plan for Investment, A Plan for Reform*. (2000) CM 4818-1. London: HMSO.

Thibault N. (1989) L'effectif de la population du Quebec en 1986. Une comparaison entre le recensement et le fichier de l'assurance-maladie. *Cahiers Quebecois de Demographie* 18: 323-341.

Thibault N. (1993) The Discrimination Power of Dependency Structures in Record Linkage. *Survey Methodology* 19: 31-38.

U.S. Bureau of the Census. (1973) The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1979 Census. Evaluation and Research Program, Series PC (E), No.7. *Records, Computers, and the Rights of Citizens. Report of the Secretary's Advisory Committee on Automated Personal Data Systems.* United States: U.S Department of Health, Education and Welfare.

U.S. Internal Revenue Service. (1985) *Record Linkage Techniques - 1985. Proceedings of the Workshop on Exact Matching Methodologies, May 9-10, 1985*. Washington DC: Internal Revenue Service.

U.S. Office of Federal Statistical Policy and Standards. (1980) *Report on Exact and statistical Matching Techniques.* Statistical Policy Working Paper 5.

Van Rijsbergen CJ, Harper DJ, Porter MF. (1981) The Selection of Good Search Terms. *Information Processing and Management* 17: 77 -91

Vitter JS, Wen-Chin C. (1987) The Probability Model. In: *Design and Analysis of Coalesced Hashing*. United Kingdom: Oxford University Press. pp. 22-31.

Wajda A, Roos LL. (1987) Simplifying record linkage: Software and strategy. *Computers in Biology and Medicine* 17: 239-248.

Wajda A, Roos LL, Layefsky M, Singleton JA. (1991) Record linkage strategies: Part n. Portable software and exact matching. *Methods of Information in Medicine* 30(3): 210-214.

Wild WG. (1968) The theory of modulus N check digit systems. *The Computer Bulletin* 12: 308-311.

Winkler WE. (1984) *Exact Matching Using Elementary Techniques. Technical Report*. Washington DC: U.S. Energy Information Administration.

Winkler WE. (1985a) Preprocessing of Lists and String Comparison. In: Alvey W and Kilss B (eds), *Record Linkage Techniques – 1985.* U.S. Internal Revenue Service. Publication 1299 (2-86), 181-187.

Winkler WE. (1985b) Exact Matching Lists of Businesses: Blocking. Subfield Identification. Information Theory. In: Alvey W and Kilss B (eds), *Record Linkage Techniques - 1985.* U.S. Internal Revenue Service. Publication 1299 (2), 227-241.

Winkler WE. (1986) *Record Linkage of Business Lists. Technical Report*. Washington DC: U.S. Energy Information Administration.

Winkler WE. (1987) *An Application of the Fellegi-Sunter Model of Record Linkage to Business Lists. Technical Report*. Washington DC: U.S. Energy Information Administration.

Winkler WE. (1988) Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association, Proceedings of the Survey Research Methods Section.* pp. 667-671.

Winkler WE. (1989a) Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Annual Research Conference.* Washington DC: U.S. Bureau of the Census. pp. 145-155.

Winkler WE. (1989b) Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage. *Survey Methodology* 15: 101-117.

Winkler WE. (1989c) Frequency-Based Matching in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association, Proceedings of the Survey Research Methods Section.* pp. 778-783.

Winkler WE. (1990a) *Documentation of Record-Linkage Software*. Unpublished report. Washington DC: U.S. Bureau of the Census.

Winkler WE. (1990b) String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association, Proceedings of the Survey Research Methods Section*. pp. 354-359.

Winkler WE. (1991) Analysis of data from computer linked files. *Proceedings Interface 1991*. pp. 411-414. Fairfax station, VA: Interface Foundation of North America.

Winkler WE. (1991) Error Model for Analysis of Computer Linked Files. *American Statistical Association, Proceedings of the Survey Research Methods Section*. pp. 472-477.

Winkler WE, Scheuren F. (1991) *How Matching Error Affects Regression Analysis: Exploratory and Confirmatory Results*. Technical Report. Washington DC: U.S. Bureau of the Census.

Winkler WE. (1992) Comparative Analysis of Record Linkage Decision Rules. *American Statistical Association, Proceedings of the Survey Research Methods Section*. pp. 829-834.
Winkler WE. (1993a) *Business Name Parsing and Standardization Software*. Unpublished report. Washington DC: U.S. Bureau of the Census.

Winkler WE. (1993b) Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. *American Statistical Association, Proceedings of the Survey Research Methods*

*Section.* pp. 274-279.

Winkler WE. (1994) Advanced Methods for Record Linkage. *American Statistical Association, Proceedings of the Section on Survey Research Methods*. pp. 467-472 (longer version report 94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler WE. (1995) Matching and Record Linkage. In: Cox, Binder, Chinnappa, Christianson, Culledge and Kott (eds), *Business Survey methods*. John Wiley and Sons, Inc. pp. 355-384

Winkler WE. (1998) Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. *Research in Official Statistics* 1: 87-104.

Winkler WE. (1999) Record Linkage Software and Methods for Merging Administrative Lists. *American Statistical Association, Proceedings of the Section on Social Statistics*. pp. 262-267.

Winkler WE. (2000) Machine Learning, Information Retrieval, and Record Linkage. *American Statistical Association, Proceedings of the Section on Survey Research Methods*. (to appear).

Witten IH, Moffat A, Bell TC. (1999) *Managing Gigabytes.* San Francisco: Morgan Kaufmann Publishers, Inc.

Woogh CM. (1988) An experience in psychiatric record linkage. *Can J Psychiatry* 33(2): 134-139.

Wu CFI. (1983) On the Convergence Properties of the EM Algorithm. *Annals of Statistics* II: 95-103.

Yu CT, Lam K, Salton G. (1982) Term Weighting in Information Retrieval Using the Term Precision Model. *Journal of the Association for Computing Machinery* 29: 152-170.

Zadeh LA. (1983) Commonsense knowledge representation based on fuzzy logic. *Computer* 16(10): 61-65.