Chapter 2

OX-LINK: The Oxford Medical Record Linkage System

Leicester E. Gill, University of Oxford

Abstract

This paper describes the major features of the Oxford record linkage system (OX-LINK), with its use of the Oxford name compression algorithm (ONCA), the calculation of the names weights, the use of orthogonal matrices to determine the threshold acceptance weights, and the use of combinational and heuristic algebraic algorithms to select the potential links between pairs of records.

The system was developed using the collection of linkable abstracts that comprise the Oxford Record Linkage Study (ORLS), which covers 10 million records for 5 million people and spans 1963 to date. The linked dataset is used for the preparation of health services statistics, and for epidemiological and health services research. The policy of the Oxford unit is to comprehensively link all the records rather than prepare links on an ad-hoc basis.

The OX-LINK system has been further developed and refined for internally cross matching the whole of the National Health Service Central Register (NHSCR) against itself (57.9 million records), and to detect and remove duplicate pairs; as a first step towards the issue of a new NHS number to everyone in England and Wales. A recent development is the matching of general practice (primary care) records with hospital and vital records to prepare a file for analyzing referral, prescribing and outcome measures.

Other uses of the system include ad hoc linkages for specific cohorts, academic support for the development of test programs and data for efficiently and accurately tracing people within the NHSCR, and developing methodologies for preparing registers containing a high proportion of ethnic names.

Medical Record Linkage

The term record linkage, first used by H. L. Dunn (1946; Gill and Baldwin, 1987), expresses the concept of collating health-care records into a cumulative personal file, starting with birth and ending with death. Dunn also emphasised the use of linked files to establish the accuracy or otherwise of the recorded data. Newcombe (Newcombe et al., 1959; and Newcombe, 1967, 1987, and 1988) undertook the pioneering work on medical record linkage in Canada in the 1950's and thereafter, Acheson (1967, 1968) established the first record linkage system in England in 1962.

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

the same individual. In this paper, I will confine my discussion to the linkage of records for different events which relate to the same person.

Matching and Linking

The fundamental requirement for correct matching is that there should be a means of uniquely identifying the person on every document to be linked. Matching may be *all-or-none*, or it may be *probabilistic*, i.e., based on a computed calculation of the probability that the records relate to the same person, as described below. In probability matching, a threshold of likelihood is set (which can be varied in different circumstances) above which a pair of records is accepted as a match, relating to the same person, and below which the match is rejected.

The main requirement for all-or-none matching is a unique identifier for the person which is fixed, easily recorded, verifiable, and available on every relevant record. Few, if any, identifiers meet all these specifications. However, systems of numbers or other ciphers can be generated which meet these criteria within an individual health care setting (e.g., within a hospital or district) or, in principle, more widely (e.g., the National Health Service number). In the past, the National Health Service number in England and Wales had serious limitations as a matching variable, and it was not widely used on health-care records. With the allocation of the new ten digit number throughout the NHS all this is being changed (Secretaries of State, 1989; National Health Service and Department of Health, 1990), and it will be incorporated in all health-care records from 1997.

Numbering systems, though simple in concept, are prone to errors of recording, transcription and keying. It is therefore essential to consider methods for reducing errors in their use. One such method is to incorporate a checking device such as the use of *check-digits* (Wild, 1968; Hamming, 1986; Gallian, 1989; Baldwin and Gill, 1982; and Holmes, 1975). In circumstances where unique numbers or ciphers are not universally used, obvious candidates for use as matching variables are the person's names, date of birth, sex and perhaps other supplementary variables such as the address or postcode and place of birth. These, considered individually, are partial identifiers and matching depends on their use in combination.

Unique Personal Identifiers

Personal identification, administrative and clinical data are gradually accumulated during a patient's spell in a hospital and finalized into a single record. This type of linkage is conducted as normal practice in hospital information systems, especially in those hospitals having Patient Administration Systems (PAS) and District Information Systems (DIS) which use a centrally allocated check-digited District Number as the unique identifier (Goldacre, 1986).

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

Difficulties arise, however, where the health event record does not include a unique identifier. In such cases, matching and linking depends on achieving the closest approach to unique identification by using several identifying variables each of which is only a partial identifier but which, in combination, provide a match which is sufficiently accurate for the intended uses of the linked data.

Personal Identifying Variables

The personal identifying variables that are normally used for person matching can be considered in five quite separate groups.

- **Group 1**.--Represents the persons proper names and with the exception of present surname when women adopt their husbands surname on marriage, rarely changes during a person's lifetime: birth surname; present surname; first forename or first initial; second forename or second initial; and, other forenames.

- **Group 2**.--Consists of the non-name personal characteristics that are fixed at birth and very rarely changes during the person's lifetime: gender (Sex at birth); date of birth; place of birth (address where parents living when person was born); NHS number (allocated at birth registration, both old and new formats); date of death; and ethnicity.

- **Group 3**.--Consists of socio-demographic variables that can change many times during the course of the person's lifetime: street address; post code; general practitioner; marital status; social class; number(s) allocated by a health district or special health-care register; number(s) allocated by a hospital or trust; number(s) allocated by a general practitioner's computing system; and, any other special hospital allocated numbers.

- **Group 4**.--Consists of other variables that could be used for the compilation of special registers: clinical specialty; diagnosis; cancer site; drug idiosyncrasy or therapy; occupation; date of death; and other dates (for example, LMP, etc.).

- **Group 5**.--Consists of variables that could be used for family record linkage: other surnames; mother's birth surname; father's surname; marital status; number of births; birth order; birth weight; date of marriage; and number of marriages.

File Ordering and Blocking

Matching and linkage in established datasets usually involves comparing each new record with a master file containing existing records. Files are ordered or *blocked* in particular ways to increase the efficiency of searching. In similar fashion to looking up a name in a telephone directory the matching algorithm must be able to generate the "*see also*" equivalent to this surname for variations in spelling (e.g., Stuart and Stewart, Mc, Mk, and Mac). Searching can be continued, if necessary, under the alternative surname.

Algorithmics that emulate the "see also" method are used for computer matching in record linkage. In this way, for example, Stuarts and Stewarts are collated into the same block. A match is determined by the amount of agreement and disagreement between the identifiers on the "incoming" record and those on the master file. The computer calculates the statistical probability that the person on the master file is the same as the person on the record with which it is compared.

File Blocking

The reliability and efficiency of matching is very dependent on the way in which the initial grouping or the "file-blocking" step is carried out. It is important to generate blocks of the right size. The balance between the number and size of blocks is particularly important when large files are being matched. The selection of variables to be used for file blocking is, therefore, critical and will be discussed before considering the comparison and decision-making stages of probability matching.

Any variable that is present on each and every record on the dataset to be matched could be used to divide or block the file, so enhancing the search and reducing the number of unproductive comparisons. Nevertheless,

if there is a risk that the items chosen are wrongly recorded -- which would result in the records being assigned to the wrong file block, then potential matches will be missed. Items that are likely to change their value from one record to another for the same person, such as home address, are not suitable for file blocking. The items used for file blocking must be universally available, reliably recorded and permanent. In practice, it is almost always necessary to use surnames, combined with one or two other ubiquitous items, such as sex and year of birth, to subdivide the file into blocks that are manageable in size and stable. Considerable attention has been given to the ways in which surnames are captured and algorithmic methods to reduce, or eliminate, the effects of variations in spelling and reporting, and which "compress" names into fixed-length codes.

Phonemic Name Compression

In record linkage, name compression codes are used for grouping together variants of surnames for the purposes of blocking and searching, so that effective match comparisons can be made using both the full name and other identifying data, despite misspelled or misreported names.

The first major advance in name compression was achieved by applying the principles of phonetics to group together classes of similar-sounding groups of letters, and thus similar-sounding names. The best known of these codes was devised in the 1920's by Odell and Russell (Knuth, 1973) and is known as the Soundex code. Other name compression algorithms are described by Dolby (1970) and elsewhere.

Soundex Code and the Oxford Name Compression Algorithm (ONCA)

The Soundex code has been widely used in medical record systems despite its disadvantages. Although the algorithm copes well with Anglo-Saxon and European names, it fails to bring together some common variants of names, such as Thomson/Thompson, Horton/Hawton, Goff/Gough, etc., and it does not perform well where the names are short, as is the case for the very common names, have a high percentage of vowels, or are of Oriental origin.

It is used principally, for the transformation of groups of consonants within names, to specific combinations of both vowels and consonants (Dolby, 1970). Among several algorithms of this type, that devised by the New York State Information and Intelligence System (NYSIIS) has been particularly successful, and has been used in a modified form by Statistics Canada and in the USA for an extensive series of record linkage studies (Lynch and Arends, 1977). A recent development in the Unit of Health-Care Epidemiology (UHCE) (Gill and Baldwin, 1987; Gill et al., 1993), 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 blocks produced using ONCA alone vary in size, from quite small and manageable for the less common surnames, to very large and uneconomic for the more common surnames. Further subdivision of the ONCA blocks on the file can be effected using sex, forename initial and date of birth either singly or in combination.

ORLS File Blocking Keys and Matching Variables

The file blocking keys used for the ORLS are generated in the following fashion:

- The primary key is generated using the ONCA of the present surname.

- The secondary key is generated from the initial letter of the first forename. Where this forename is a nickname or a known contraction of the "formal" forename, then 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 would be W. A further record is set up on the master file where a second forename or initial is present; the key is derived from this second initial.

- Where the birth surname is not the same as the present surname, as in the case of married women, a further record is set up on the master file under the ONCA code of birth surname and again subdivided by the initial. (This process is termed *exploding* the file.)

- Further keys based on the date of birth and other blocking variables are also generated.

In addition to the sorting header, four other variables are added to each record before sorting and matching is undertaken:

- **Accession Number**.--A unique number allocated from a pool of such numbers, and is absolutely unique to this record. The number is never changed and is used for identification of this record for correction and amendment. The number is check digited to modulus 97.

- **Person or System Number**.--A unique number allocated from a pool of such numbers. The number can be changed or replaced if this record matches with another record. The number is check digited to modulus 97.

- **Coding Editions**.--Indicators that record the various editions of the coding frames used in this record, for example the version of the ICD (International Classification of Diseases) or the surgical procedure codes. These indicators ensure that the correct coding edition is always recorded on each and every record and reliance is not placed on a vague range of dates.

- **Input and Output Stream Number**.--This variable is used for identifying a particular dataset during a matching run, and enables a number of matches to be undertaken independently at the same pass down the master file.

Generating Extra Records Where a Number of Name Variants Are Present

To ensure that the data record can match with the blocks containing all possible variants of the names information, multiple records are generated on the master file containing combinations of present and birth surnames, and forenames. To illustrate the generation of extra records where the identifying set for a person contains many variants of the names, consider the following example:

birth surname:	SMITH
present surname (married surname):	HALL
first forename:	LIZ (contraction of Elizabeth)
second forename:	PEGGY (contraction of Margaret)
year of birth:	1952 (old enough to be married).

Eight records would be generated on the master file and each record indexed under the various combinations of ONCA and initial, as follows:

Indexed 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);
Indexed under the birth surname SMITH: i.e., ONCA S530:	
S530L	for Liz
S530E	for Elizabeth (formal version of Liz)
S530P	for Peggy
S530M	for Margaret (formal version of Peggy).

Mrs. Hall would have her master file record included under each of the above 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 above, and would have a high expectation of matching to any of the variants during the matching phase.

To reduce the number of unproductive comparisons, a data record will only be matched with *an other record in the same block* provided that the year of birth on both records are within 16 years of each other. This constraint has been applied, firstly, to reduce the number of unproductive matches, and secondly to confine matching to persons born within the same generation, and in this way eliminate father/son, mother/daughter matches. Further constraints could be built into the matching software for example, matching only within the same sex, logically checking that the dates on the two records are in a particular sequence or range, or that the diagnoses on the two records are in a specified range, as required in the preparation of a cancer registry file.

Matching Methods

There are two methods of matching data records with a master file.

- The **two file method** is used to match a data record from a data file with a block on the master file, and in this way compare the data record with every record in the master file block.

- The **one file/single pass** method is used to combine the data file block and the master file block into one block, and to match each record with every other in the block in a triangular fashion, i.e., first with the rest, followed by second with the rest etc. In this way every record can be matched with every other record. Use of a stream number on each record enables selective matching to be undertaken, for example data records can be matched with the master file and with each other, but the master file records are not matched with themselves.

Match Weights

Considerable work has been undertaken to develop methods of calculating the probability that pairs of records, containing arrays of partial identifiers which may be subject to error or variation in recording do, or do not, relate to the same person. Decisions can then be made about the level of probability to accept. The issues are those of reducing false negatives (Type I errors) and false positives (Type II errors) in matching (Winkler, 1995; Scheuren and Winkler, 1996; and Belin and Rubin, 1995). A false negative error, or "missed match," occurs when records which relate to the same person are not drawn together (perhaps because of minor variations in spelling or a minor error in recorded dates of birth). Matches may also be missed if the two records fall into different blocks. This may happen if, for example, a surname is misspelled and the phonemic compression algorithm puts them into two different blocks.

Methods for probability matching depend on making comparisons between each of several items of identifying information. Computer-based calculations are then made which are based on the *discriminating power* of each item. For example, a comparison between two different records containing the same surname has greater discriminating power if the surnames are rare than if they are common. Higher scores are given for

agreement between identifiers (such as particular surnames) which are uncommon than for those which are common. The extent to which an identifier is uncommon or common can be determined empirically from its distribution in the population studied. Numerical values can then be calculated routinely in the process of matching for the amount of agreement or disagreement between the various identifying items on the records. In this way a composite score or *match weight* can be calculated for each pair of records, indicating the probability that they relate to the same person. In essence, these weights simulate the subjective judgement of a clerk. A detailed discussion of match weights and probability matching can be found in publications by Newcombe (Newcombe et al., 1959; and Newcombe, 1967, 1987, and 1988), and by Gill and Baldwin (1987) (See also Acheson, 1968.)

Calculating the Weights for the Names Items

Name identifiers are weighted in a different fashion to the non-name identifiers, because there are many more variations for correctly spelled names. Analysis of the NHS central register for England and Wales shows that there are:

57,963,992	records
1,071,603	surnames
15,143,043	surname/forename pairs.

The low frequency names were mainly non Anglo-Saxon names, hyphenated names and misspelled names. In general the misspellings were due to embedded vowel changes or to miss keying. A more detailed examination of the register showed that 954 different surnames covered about 50% of the population, with the following frequency distribution:

10% population	24 different surnames
20% population	84 different surnames
30% population	213 different surnames
40% population	460 different surnames
50% population	954 different surnames
60% population	1,908 different surnames
70% population	3,912 different surnames
80% population	10,214 different surnames
90% population	100,000 different surnames
100% population	1,071,603 different surnames.

Many spelling variations were detected for the common forenames. Using data from the NHSCR register, various forename directories and other sources of forenames, a formal forename lexicon was prepared that contained the well known contractions and nicknames. The problem in preparing the lexicon was whether to include forenames that had minor spelling errors, for example JOHN and JON. This lexicon is being used in the matching algorithm, to convert nicknames and contractions, for example LIZ, to the formal forename ELIZABETH, and both names are used as part of the search strategy.

Calculation of Weights for Surnames

The binit weight calculated from the frequency of the first letter in the surname (26 different values) was found to be too crude for matching files containing over 1 million records. The weights for Smith, Snaith, Sneath, Smoothey, Samuda, and Szabo would all have been set to some low value calculated from the frequency of Smith in the population, and ignoring the frequency of the much rarer Szabo. Using the frequencies of all of the 1 million or more different surnames on the master match file is too cumbersome, time consuming to keep up-to-date, and operationally difficult to store during the match run. The list would also have contained all of the one-off surnames generated by bad transcription and bad spelling. A compromise solution was de-

vised by calculating the weights based on the frequency of the ONCA block (8,000 values), with a cut-off value of 1 in a 1,000 in order to prevent the very rare and one-off names from carrying very high weights. Although this approach does not get round the problem of the very different names that can be found in the same ONCA block (Block S530: contains Smith, Smithies, Smoothey, Snaith, Sneath, Samuda, Szabo, etc.) it does provide a higher level of discrimination and, in part, accommodate the erroneous names.

The theoretical weight based on the frequency of the surname in the studied population is modified according to the algorithm devised by Knuth-Morris-Pratt (Stephen, 1994; Gonnet and Baeza-Yates, 1991; and Baeza-Yates, 1989), and takes into account 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. Where the two names are absolutely identical, the weight is set to +2N, but falls down to a lower bound of -2N where the amount of disagreement is quite large.

If the birth surname and present surname are swapped with each other, exploding the file as described previously enables the system to find and access the block containing the records for the appropriate surnames. The weights for the present and birth surname pairs are calculated, then the present surname/birth surname and birth surname/present surname pairs are also calculated. The highest of the two values is used in the subsequent calculations for the derivation of the match weight.

In cases where the marital status of the person is single, i.e., never married, or the sex is male, or the age is less that 16 years, it is normal practice in the UK for the present surname to be the same as the birth surname, and for this reason only the weight for the present surname is calculated and used for the determination of a match.

Forenames

The weights derived for the forenames are based on the frequency of the initial letter of the forename in the population. Since the distribution of male and female forenames are different, there are two sets of different weights, one for males and a second for females. Since the forenames can be recorded in any order, the weights for the two forenames are calculated and the highest value used for the match. Where there are wide variations in the spelling of the forenames, the Daitch-Motokoff version of Soundex ("Ask Glenda") is being evaluated for weighting the forenames in a fashion similar to that used for the surnames.

Calculating the Weights for the Non-Names Items

The weights for date of birth, sex, place of birth and NHS number are calculated using the frequency of the item on the ORLS and on the NHSCR file. The weight for the year of birth comparison has been extended to allow for known errors, for example, only a small deduction is made where the two years of birth differ by 1 or 10 years, but the weight is substantially reduced where the year of births differ by say, 7 years.

The weight for the street address is based on the first 8 characters of the full street address, where these characters signify a house number (31, High Street), or house name (High Trees), or indeed a public house name (THE RED LION). Terms like "Flat" or "Apartment" are ignored and other parts of the address are then used for the comparison. The postcode is treated and weighted as a single field although the inward and outward parts of the code can be weighted and used separately.

The range of binit weights used for the ORLS is shown in Table 1.

When the matching item is present on both the records, a weight is calculated expressing the amount of agreement or disagreement between the item on the data record and the corresponding item on the master file record.

Table 1. -- The Range of Binit Weights Used for Matching

Identifying Item		Score in Binits[1]		
		Exact Match	Partial Match	No Match
Surnames:	Birth	+2S	+2S to -2S	-2S
	Present[2]	+2S	+2S to -2S	-2S
	Mother's birth	+2S	+2S to -2S	-2S
	(where: common surname S = 6, rare surname S = 17)			
Forenames[3]		+2F	+2F to -2F	-2F
	(where: common forename F = 3, rare forename F = 12)			
NHS number		+7	NP[4]	0
Place of birth (code)		+4	+2	-4
Street address[5]		+7	NP	0
Post Code		+4	NP	0
GP (code)		+4	+2	0
Sex [6]		+1	NP	-10
Date of birth		+14	+13 -> -22	-23
Hospital and Hospital unit number		+7	NP	-9

[1]Where an item has been recorded as not known, the field 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.

[2]Where the surname is not known or has been entered as blank, the record can not be matched in the usual way, but 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 indices and the sex is corrected where it is missing or in error.)

It is possible for the calculated weight to become negative where there is extreme disagreement between the item on the data record and the corresponding item on the master file. In matching street address, postcode and general practitioner the score cannot go negative, although it can assume zero, because the individual may have changed their home address or their family doctor since they were last entered into the system, this is really a change in family circumstances and not errors in the data and so a negative weight is not justified.

Threshold Weighting

The procedure for deciding whether two records belong to the same person, was first developed by Newcombe, Kennedy, Axford, and James (1959), and rigorously examined by Copas and Hilton (1990), Belin and Rubin (1995), and Winkler (1995). The decision is based on the total binit weight, derived by summing algebraically the individual binit weights calculated from the comparisons of each identifying item on the master file and data file. The algebraic sum represents a measure of the probability that two records match. By comparing

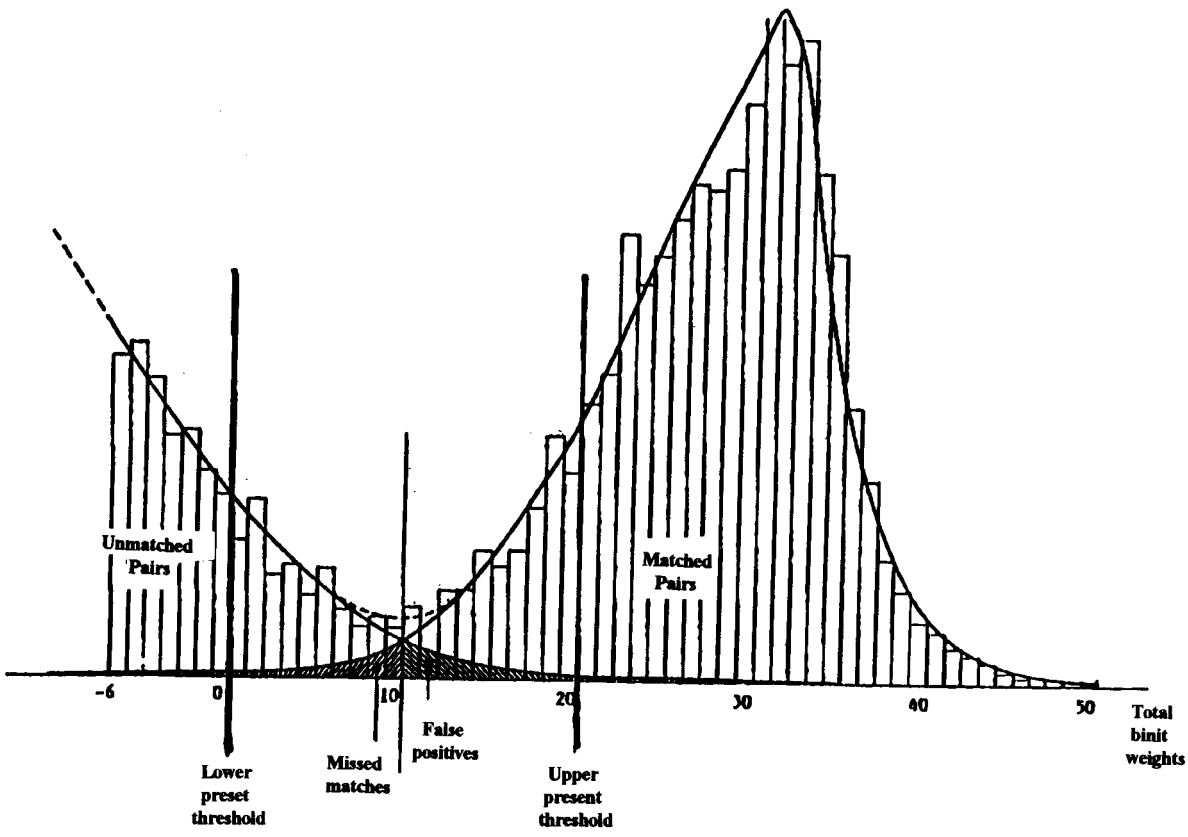
the total weight against a set of values determined empirically, it is possible to determine whether the two records being compared refer to the same person.

Two types of error can occur in record matching. The first, *false negative* matching or Type I error, is the more common and is a failure to collate records which refer to the same person and should have the same system number instead the person is assigned two or more person/system numbers and their records are not collated together. The second, *false positive* or Type II error, is less common but potentially more serious in allocating the same system number to two or more persons, where their records are wrongly collated together. The frequency of both types of error is a sound measure of the reliability of the record matching procedure.

In preparing earlier versions of the ORLS linked files, a range of binit weights was chosen and used to select records for clerical scrutiny. This range was delimited by the upper and lower pre-set thresholds, see Figure 1. The *false positive* and *false negatives* are very sensitive to the threshold cut-off weight: too low gives a very low *false positive* rate and a high *false negative* rate; too high 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:

- The minimisation of *false positives*, at the risk of increased missed matches;
- The minimisation of missed matches, at the risk of increased false positives; and
- The minimisation of the sum of *false positives* and *missed matches*.

Figure 1. — Frequency Distribution of the Binit Weights for Pairs of Records



The simple approach for the determination of a match based on the algebraic sum of the binit 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 items are based on distributions of those items 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 items in the algebraic total, and 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 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 grid approach developed by the UHCE and used for all its more recent matching and linking (Gill, et al., 1995; Vitter and Wen-Chin, 1987).

Algebraic Summation of the Individual Match Weights

In recent years we have, therefore, developed an approach in which a two dimensional orthogonal matrix is prepared, analogous to a spreadsheet, with the names scores forming one axis and the non-names scores the other axis. In the development of the method, sample runs are undertaken; pairs of records in cells in the matrix are checked clerically to determine whether they do or do not match; and the probability of matching is derived for each cell in the sample. These probabilities are stored in the cells of an orthogonal matrix designated by the coordinates (names score, non-names score). The empirical probabilities entered into the matrix are further interpolated and smoothed across the axes using linear regression methods.

Match runs using similar data types would access the matrix and extract the probability score from the cell designated by the coordinates. The array of probabilities can be amended after experience with further runs, although minor tinkering is discouraged. Precise scores and probabilities may vary according to the population and record pairs studied. A number of matrices have therefore been prepared for the different types of event pairs being matched, for example, hospital to hospital records, hospital to death records, birth to hospital records, hospital and District Health Authority (DHA) records, cancer registry and hospital records, and so on.

Over 200,000 matches were clerically scrutinized and the results recorded in the two axes of a orthogonal matrix, with the algebraic sum of the weights for the names items being X coordinate ("X" axis), and the algebraic sum of the fixed and variable statistics items plotted on the Y coordinate ("Y" axis). In each cell of the orthogonal matrix the results of the matches were recorded, with each cell holding the total number of matches, the number of good matches and the number of non-matches. A sample portion of the matrix is shown in Figure 2.

A graphical representation of the matrix is shown in Figure 3, where each cell contains the empirical decision about the likelihood of a match between a record pair. The good matches are shown as "Y," the non-matches as "N" and the doubtful matches that require clerical intervention as "Q." This graph is the positive quadrant where both the names and non-names weights are 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.

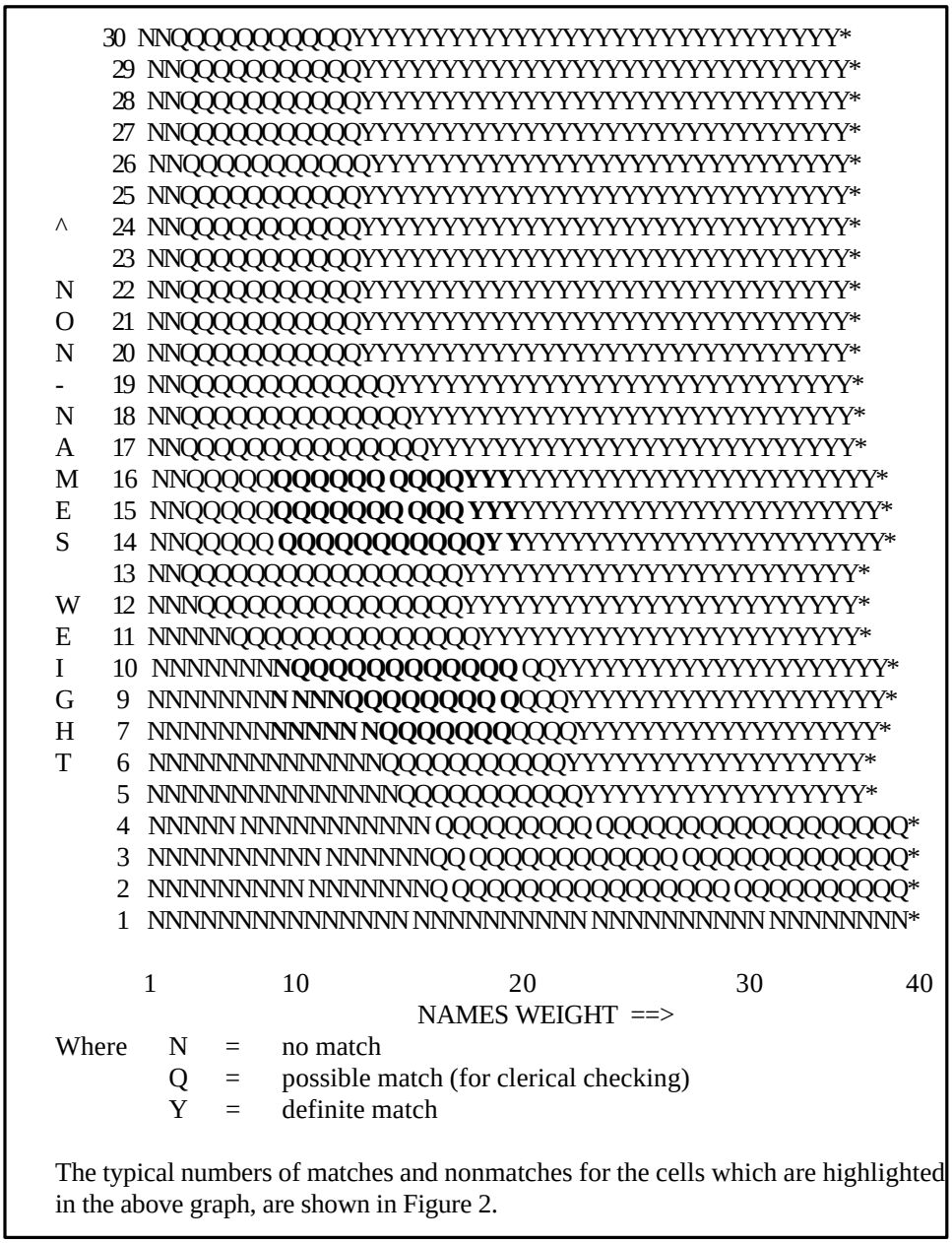
Figure 2. — Sample Portion of the Threshold Acceptance Matrix Showing the Number of Matches and Nonmatches, by Binit Weight for Names and Non-names Identifiers

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				

Record pairs with weights that fall in the upper right part of the matrix and shown in Figure 3 as "Y" are considered to be “good” matches and only a 1% random sample is printed out for clerical scrutiny. 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 is taken of record pairs that fall adjacent to N-Q boundary.

At the end of each computer run, the results of the clerical scrutiny are pooled with all the existing matching results and new matrices are prepared. The requirement is to reduce the “Q” zone to the minimum consistent with the constraints of minimum false positives and false negatives. Clerical intervention is invariably the most costly and rate determining stage.

Figure 3. -- A Sample Portion of the Matrix Used for Matching Hospital Records with Hospital Records



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

The typical numbers of matches and nonmatches for the cells which are highlighted in the above graph, are shown in Figure 2.

Separate matrices have been modelled for the different types of record pairs entering the system, for example:

hospital discharge / hospital discharge
hospital discharge / death record
birth record / hospital discharge
hospital discharge / primary care/FHSA record
hospital discharge / Cancer registry.

Further matrices have also been prepared that record the number of match items used in matching a record pair, for example, number of surnames, forenames and numbers of other matching variables. Since the number of matrices can become quite large, intelligent systems and neural net techniques are being developed for the interpretation of the N dimensional matrices and the determination of the match threshold (Kasabov, 1996; Bishop, 1995).

Special procedures have been developed for the correct matching of similarly-named same sex twins. Where the match weights fall within the clerical scrutiny area, the clerks are able to identify the two records involved and take the appropriate action.

The marked records are printed out for clerical scrutiny and the match amended where necessary. This situation also arises where older people are recorded in the information system under a given set of forenames but, on a subsequent hospital admission or when they die, a different set of forenames are reported by the patient or by the next of kin.

Linking

The output from the matching run, is a text file that contains details about each pair of records that were matched together. A sample portion of this file is shown in Figure 4, the layout of which is:

Details of data record	Person/system number Accession number Record type
Details of main file record	Person/system number Accession number Record type
Details about the match run	Output stream (good match or query match) Names weight Non-names weight Cross-reference to the clerical printout Matching probability/decision (either Y or N).

The number of records written to the output file for any one person can be very large, and is approximately the number of records on data file multiplied by the number of records on the master file. Using combinational and heuristic algebraic methods these records are reduced to a small number for each potential match pair, ideally one (Hu, 1982; Cameron, 1994; Lothaire, 1997; and Pidd, 1996).

Figure 4.—A Sample of the Typical Output from the Match Run

Example of OX-LINK System Number Output

389447756	860895558	GS	229800034	352-68394	GN	2	50	26	(GH1/ 500001)	Y	O
379194856	858751858	GS	233513082	369890337	GN	2	29	24	(GH1/ 500002)	Y	O
379194856	858751858	GS	233513082	911759078	TU	2	29	15	(GH1/ 500003)	Y	O
379194856	858751858	GS	233513082	911759078	TU	2	29	15	(GH1/ 500004)	Y	O
437096752	781384114	GS	323947927	524582350	BL	2	31	19	(GH1/ 500005)	Y	O
357816810	726892961	GS	249173530	472792138	GN	2	31	23	(GH1/ 500006)	Y	O
357816810	726892961	GS	249173530	343537893	GN	2	31	21	(GH1/ 500007)	Y	O
357816810	726892961	GS	249173530	406349427	GN	2	31	23	(GH1/ 500008)	Y	O
540814037	883641514	GS	210500551	448983383	GM	2	50	19	(GH1/ 500009)	Y	O
110463907	559719951	GN	408578989	738005030	GS	2	50	30	(GH1/ 500010)	Y	O
110463907	262969219	GH	408578989	738005030	GS	2	50	30	(GH1/ 500011)	Y	O
110463907	63685552	GH	408578989	738005030	GS	2	50	26	(GH1/ 500012)	Y	O
133714360	188729480	GH	414567239	748873845	GS	2	50	25	(GH1/ 500013)	Y	O
133714360	205039688	GH	414567239	748873845	GS	2	50	23	(GH1/ 500014)	Y	O

The rules for undertaking this reduction are:

- Ideally, all records for the same person will have the same person/system number.

- The records for a person who has only one set of identification details will be of the following type, where each record only carries one person/system number (A):

  A = A = A = A, etc. (= signifies matches with).

- Where a single woman gets married within the span of the file, records will be recorded under maiden name, person/system number (A) and also under her married name (B). Links will be effected between (A) and (B) and all the records will be converted to person/system number (A). The person/system number (B) will be lost to the system. Future matches will link to either her single or married records, both of which will carry the person/system number (A):

  A = A = B = B = A = B, etc.

  A being links under her maiden name
  B being links under her married name.

- Where there are records for a women recorded under her maiden name (A), and records that contain details of both her maiden and married name (B) and just her married name (C), these chains are will be made up of three types of links,

  A = A = B = B = C = B = C, etc.

  Successive matches will convert all the records to person/system number (A). If the linked file contains records type (A) and (C) only, linkage cannot be effected between (A) and (C) until records of type (B) are captured and linked into the system.

- Where the person has had many changes of name and marital status, the number of different types of links will increase. Over the 30 year span of the file, links up to 5 deep have been found.

Each record entering the system is given a new purely arbitrary person/system number from a pool of such numbers. Where the record on the data file matches with a record on the master file, the person/system number stored on the master file record is copied over the person/system number on the data record, over-writes it, and the original person/system number on the data record number is then lost from the system and cannot be re-issued.

Where two sets of records for the same person, but having two different person/system numbers are brought together during a subsequent matching run; all the records are given the lowest person/system number and any other person/system numbers are destroyed.

Results

When the matching, linking and clerical stages are completed, the file of linked records will contain two types of error. Firstly, the records that have matched together but do not belong to the same person, these are known as *false positives*. Secondly, records belonging to the same person that have not been brought together, i.e., reside on the file under two or more different person identifiers, these are known as "*false negatives or missed matches*."

The *false positive* rate was estimated using two different methods. Firstly, all the records for a random sample of 5,000 people having two or more records were extracted from the ORLS file and printed out for clerical scrutiny. Secondly, all the record pairs that matched together with high match weights but where the forenames differed, were printed out for clerical scrutiny.

The "*false negative or missed match*" rate was estimated, by extracting a subset of people who had continuing treatment, such as repeated admissions for diabetics, nephritics, etc., and for those patients who had died in hospital, where the linked file should contain both the hospital discharge record and the death record.

The latest results from the ORLS file and the Welsh and Oxfordshire Cancer registry files are very encouraging, with the *false positive* rate being below 0.25 percent of all people on the file, and the *missed match* rate varying between 1.0 percent and 3.0 percent according to the type of sample investigated. Recent works on matching 369,000 records from a health district with 71 million *exploded* records from NHS Central Register has given a *false positive* rate of between 0.2 and 0.3%; the higher figure is produced from records which have very common Anglo-Saxon or Asian names.

The worst *false negative* rate was found where hospital discharges were matched with the corresponding death record. The identifying information on the hospital discharge was drawn from the hospital master index supplemented by information supplied by the patient or immediate family. The identifying information on the death record is usually provided by the next of kin from memory and old documents.

The completed ORLS file is serial file that is indexed using the person/system number, and contains the partial identifiers, administrative and socio-demographic variables and clinical items. This file used for a wide range of epidemiological and health services research studies. For ease of manipulation and other operational reasons, subsets of the file are prepared for specific studies, usually by selecting specified records or record types, or by selecting on geographical area or span of years or on clinical specialty.

Acknowledgments

The Unit of Health Care Epidemiology and the work on medical record linkage is funded by the Research and Development Directorate of the Anglia and Oxford Regional Health Authority. The Office of Population Censuses and Surveys (now the Office of National Statistics) for permission to publish the frequencies of the surnames from the NHS central register.

References

Acheson E.D. (1967). *Medical Record Linkage*, Oxford: Oxford University Press.

Acheson E.D. (ed) (1968). Record Linkage in Medicine, *Proceedings of the International Symposium,Oxford, July 1967*, London: ES Livingstone Limited.

"Ask Glenda," Soundex History and Methods, World Wide Web: *http://roxy.sfo.com/~genealogysf/glenda.html* .

Baeza-Yates, R.A. (1989). Improved String Searching, *Software Practice and Experience*, 19, 257-271.

Baldwin, J.A. and Gill, L.E. (1982). The District Number: A Comparative Test of Some Record Matching Methods, *Community Medicine*, 4, 265-275.

Belin, T.R. and Rubin, D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 694-707.

Bishop, C.M. (1995). Three Layer Networks, in*: Neural Networks for Pattern Recognition*, United Kingdom: Oxford University Press, 128-129.

Cameron, P.J. (1994). Graphs, Trees and Forests, in: *Combinatorics*. United Kingdom: Cambridge University Press, 159-186.

Copas, J.R. and Hilton, F.J. (1990). Record Linkage: Statistical Models for Matching Computer Records, *Journal of the Royal Statistical Association*, Series A, 153, 287-320.

Dolby, J.L. (1970). An Algorithm for Variable Length Proper-Name Compression, *Journal of Library Automation*, 3/4, 257.

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

Gallian, J.A. (1989). Check Digit Methods, *International Journal of Applied Engineering Education*, 5, 503-505.

Gill, L.E. and Baldwin, J.A. (1987). Methods and Technology of Record Linkage: Some Practical Considerations, in: *Textbook of Medical Record Linkage* (Baldwin,J.A., Acheson, E.D., and Graham, W.J., eds). Oxford: Oxford University Press, 39-54.

Gill, L.E.; Goldacre, M.J.; Simmons, H.M.; Bettley, G.A.; and Griffith, M. (1993). Computerised Linkage of Medical Records: Methodological Guidelines, *Journal of Epidemiology and Community Health*, 47, 316-319.

Goldacre, M.J. (1986). The Oxford Record Linkage Study: Current Position and Future Prospects, *Proceedings of the Workshop on Computerised Record Linkage in Health Research* (Howe, G.R. and Spasoff, R.A., eds). Toronto: University of Toronto Press, 106-129.

Gonnet, G.H. and Baeza-Yates, R. (1991). Boyer-Moore Text Searching, *Handbook of Algorithms and Data Structure*, 2nd ed, United States: Addison-Wesley Publishing Co Inc, 256-259

Hamming, R.W. (1986). *Coding and Information Theory*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

Holmes, W.N. (1975). Identification Number Design, *The Computer Journal*, 14, 102-107.

Hu, T.C. (1982). Heuristic Algorithms, in: *Combinatorial Algorithms*., United States: Addison-Wesley Publishing Co. Inc, 202-239.

Kasabov, N.K. (1996). Kohonen Self-Organising Topological Maps, in*: Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, Cambridge, MA, USA: MIT Press, 293-298.

Knuth,D.E. (1973). Sorting and Searching, in: *The Art of Computer Programming*, 3, United States: Addison-Wesley Publishing Co. Inc., 391.

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

Lynch, B.T. and Arends, W.L. (1977). *Selection of a Surname Encoding Procedure for the Statistical Reporting Service Record Linkage System*, Washington, DC: United States Department of Agriculture.

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

Newcombe, H.B. (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. (1987). Record Linking: The Design of Efficiency Systems for Linking Records into Individual and Family Histories, in*: Textbook of Medical Record Linkage* (Baldwin, J.A.; Acheson, E.D.; and Graham, W.J., eds), Oxford: Oxford University Press, 39-54.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

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

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

Scheuren, F. and Winkler, W.E. (1996). Recursive Merging and Analysis of Administrative Lists and Data, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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

Stephen, G.A. (1994). Knuth-Morris-Pratt Algorithm, in: *String Searching Algorithms*, Singapore: World Scientific Publishing Co. Pte. Ltd, 6-25.

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

Wild, W.G. (1968). The Theory of Modulus *N* Check Digit Systems, *The Computer Bulletin*, 12, 308-311.

Winkler,W.E. (1995). Matching and Record Linkage, *Business Survey Methods* (Cox, Binder, Chinnappa, Christianson, Culledge, and Kott, eds.), New York: John Wiley and Sons, Inc., 355-384.