Journal of Epidemiology and Community Health 1993; **47**: 316–319

Computerised linking of medical records: methodological guidelines

Leicester Gill, Michael Goldacre, Hugh Simmons, Glenys Bettley, Myfanwy Griffith

Abstract

***Objectives*—To report on the development of computer assisted methods for linking medical records and record abstracts.**

***Design*—The methods include file blocking, to put records in an order which makes searching efficient; matching, which is the process of comparing records to determine whether they do or do not relate to the same person; linkage, which is the process of assembling correctly matched records into a time sequenced composite record for the individual; and validation checks and corrections, in which any inconsistencies between different records for the same person are identified and corrected.**

***Setting*—The dataset comprising the Oxford record linkage study which includes hospital inpatient records and vital records.**

***Results and conclusions*—Probability matching, using an array of identifiers, achieves much higher levels of correct matching than is generally achievable by exact character by character comparisons. The increasing use of information technology to store data about health and health care means that there is increasing scope to link records for research and for patient care. Sophisticated methods to achieve this on a large scale are now available.**

J Epidemiol Community Health 1993; **47**: 316–319

Record linkage is the process of bringing together related records, or abstracts of records, which have been compiled separately. The increasing use of computers to store information about health and health care, and the capacity and processing power of modern computing systems, means that there is increasing scope for linking records for epidemiology, health services research, and health service management. In the latter context, *Working for Patients*[1] and *The Next Steps*[2] established a requirement to link records of patient care in England in support of contracting and in the establishment of information systems relating to resident populations. Because of this widening interest, we summarise some of the basic theoretical and practical aspects of record linkage, drawing in particular on our experience with the Oxford record linkage study.[3–6]

Linking continuous and discontinuous health care events

Record linkage may be used to bring together different data about the same medical event. For example, a hospital admission may generate an admission record, a discharge record, clinical data, laboratory data, and a financial record; and it may also involve transfers between consultants and hospitals. The most common approach to the identification of different records relating to the same person, in the case of a continuous event, is the use of a database with a numbering system which is unique to the event, or to the person, or to both, and which is generated in the health care setting in which the event occurs.

The second circumstance where record linkage may be required is that when the records relate to different episodes of illness for the same person and may have been compiled in different places and at different times. In practice, most of the work on record linkage in medical research has been in this category. In 1946, Dunn[7] described the potential for linking records of "the principal events in life" into "a book of life" starting with birth record and ending with the death record. Newcombe undertook pioneering work on medical record linkage in Canada in the 1950s and thereafter.[8–10] Acheson established the first medical record linkage system in England in 1962.[3 4] When the requirement is to link records created at different times and in different places, it should, in principle be possible to link such records using a unique personal identification number. In practice, such a number has not generally been available on records 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. Record linkage can also be used to link family records: examples include linkage of records for a mother and her baby, records for siblings, and records for husband and wife, or for other family members. In this paper, we confine our discussion to the linkage of records for different events which relate to the same person.

The three main steps in linking records

The first step in creating a file of linked records is to search the file of potentially linkable records to identify different records that relate to the same person. The file is usually ordered, before searching, in ways which make the searching efficient. This ordering is often called *blocking* the file. If identifying numbers are used, the blocking may simply be a numerical sequence. Where surnames are used, the blocks may be alphabetical or, in most modern systems, they are aggregates of phonetically manipulated names (see below). The next step is *matching* which is the process by which potentially linkable records are systematically

Unit of Health-Care Epidemiology, Department of Public Health and Primary Care, University of Oxford, Oxford Regional Health Authority, Old Road, Headington Oxford OX3 7LF
L Gill
M Goldacre
H Simmons
G Bettley
M Griffith

Correspondence to:
Dr M Goldacre

Accepted for publication
February 1993

compared against all other "candidate" records (for example, those in the same block) to determine whether they relate to the same person or not. *Linking* is the process by which correctly matched records, identified as relating to the same person, are brought together and assembled in such a way that they can be analysed as a composite record for one individual. Methods used for blocking and searching depend on methods used for matching, and the latter will therefore be described first.

Identifiers and matching

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*—that is, computer-generated decisions are made that a pair of records either do or do not relate to the same person; or it may be *probabilistic*—that is, based on a computed calculation of the probability that the records relate to the same person, as described below. In probability matching, a threshold 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 an identifier for the person which is fixed, unique, easily recorded, verifiable, and available on every relevant record. Few, if any, identifiers meet all these specifications. However, systems of numbers or other cyphers can be generated which meet these criteria within an individual health care setting (for example, within a hospital or district) or, in principle, more widely (for example, the National Health Service number). In practice, the present National Health Service number in England and Wales has serious limitations as a matching variable. It has a cumbersome structure and its recording is prone to error. As is well known, it has also not been widely used on health care records in the past, and is therefore not widely available for linkage, although this will probably change.[2] 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*.[6 11 12] In circumstances where unique numbers or cyphers 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. One approach to matching such identifiers is automated comparison on a character by character basis to identify all-or-none matches. This approach is usually disappointing in practice. Error rates are often surprisingly high: it is not uncommon to get "failure to match" rates of 5% or even 10% by comparing records of names and dates of birth on a character by character basis. "Failure to match" on this basis occurs partly because there are fairly high levels of error in spelling and recording names; but it occurs particularly because the recording of names may vary (for example, one forename used on one occasion, two on another; a forename used on one occasion, an initial on another; the use of contractions such as Madge or Peggy for Margaret, Bill or Will for William; reversal of forenames when both are used; changes of name by women at marriage or divorce). Character by character matching is not recommended when precision of matching is required.

Match weights

Considerable work has therefore 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 and false positives in matching. 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 mis-spelt and the phonemic compression (see below) puts them into two different blocks. It is worth considering the use of at least two different methods of blocking to effect matches by one method which may have been missed by another. We routinely use the method of blocking on dates of birth, followed independently by blocking on names, to increase the likelihood of identifying true matches.

A false positive error or "mismatch" occurs when two records are brought together when they do not, in fact, relate to the same person. This may occur with very common names in large files of data (for example, two John Smiths with the same date of birth); or when, in order to avoid unduly high false negative rates, the matching criteria have been relaxed (for example, to accept minor errors). The latter can reach the point when two records which vary in minor respects are drawn together but, in fact, do not relate to the same person.

Methods for probability matching depend, first, 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 whole 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 and by Gill and Baldwin.[6 8–10]

The simple approach for calculating the composite match weight is to use the algebraic sum of the individual component scores. However, that ignores the fact that the distribution of variables such as names and dates of birth in the population have different characteristics. Names are propagated through families by procreation, marriage, and naming practices. This can give rise to name clumping in small geographical areas. Even now, in village communities in Britain, intermarriage between local families can produce generations of families with a very limited range of different surnames. By contrast, the frequency distribution of gender and dates of birth, within a given age range in a local population, is based on values which are all almost equally probable. This may present problems in large datasets. For example, a rare set of names would generate very high scores that could over-ride any scores derived from the non-names items; and, conversely scores derived from very common names could be overridden by a perfect set of non-names identifiers. In recent years we have therefore used an approach in which a two dimensional array 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 array 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 the array designated by the coordinates (names score, non-names score). The empirical probabilities entered into the array are further interpolated and smoothed across the axes using linear regression methods. Match runs using similar data types would access the array 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, at least a little, according to the population and record pairs studied. A number of arrays 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 family health service authority records, cancer registry and hospital records, and so on.

Blocking and searching

Matching and linkage in established datasets usually involves comparing each new record with a master file containing existing records. However, the principles and procedures described in this section also apply to the linkage of two different sets of records to be linked on an ad hoc basis. Files are ordered or *blocked* in particular ways to increase the efficiency of searching. This is analogous to finding a person in a telephone directory. To find a telephone number quickly and efficiently, the searcher skips to the page in the telephone book making intuitive mental use of (say) the first three letters of the surname, and then identifies the full surname, forename or initial, town and street address, by scanning systematically through the entries on the page until the correct "match" is found. Where there are spelling variants (for example, Stuart and Stewart, Mc and Mac), the telephone directory may remind the searcher that there is a *"see also"* equivalent to this surname. Searching can be continued, if necessary, under the alternative surname. Where the searcher is locating a number of different telephone entries, it is expedient to sort the list of people into the same order as that used by the telephone directory (that is, in alphabetical order) before searching. Similar algorithmic approaches are used for computer matching in record linkage. Both the new data file—perhaps containing hundreds or thousands of new records—and the master file are sorted into the same order, usually based on the surname, and the "see also" technique is accomplished by the use of phonetic conversion of the names (appendix). In this way, for example, Stuarts and Stewarts are collated into the same block. The new "incoming" record is compared with all the records in a block which contains all the names in the same phonetic group. 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 telephone-number searcher uses intuitive logic and other clues to effect the match. 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. When the new record matches with a record on the master file (that is, its match weight reaches the threshold for unconditional acceptance), the system number may be copied from the master file onto the new record. A record which did not match with the master file would be issued a new system number as a "new" individual.

Linking and corrections

Once the matching decisions have been made, a sequenced file can be built by linking the records relating to each individual. Linkage may reveal inconsistencies and errors across records which are not apparent from single records alone. There may therefore be a cycle of validation checks, error detection and correction. This cycle may, in practice, account for a significant part of the resource required to match and link records in an established linkage system.

Hardware and software in Oxford: a technical note

Record linkage in Oxford has been used mainly to link hospital records, birth records and death records. However, matching between many other types of health care records is possible and recently, for example, we have matched family health services authority records, laboratory records, and cancer registry records. Ad hoc linkage of smaller, research datasets is also an increasing requirement. There have been several generations of matching and linking systems used

in Oxford and modifications have been made as hardware and software have changed. The current software is written in IBM FORTRAN H EXTENDED specifically for use on IBM mainframe systems. The software is supported by the staff of the Unit of Health-Care Epidemiology and runs on the unit's IBM mainframe. We have recently rewritten the matching and linking software in a more modern version of FORTRAN. This means that our software could be transported across a wide range of mainframe and microcomputing systems. Our software has been further rewritten to permit the end-user to tailor the input/output file specifications to accept their own datasets, without recourse to expensive and time consuming reformatting.

APPENDIX

Name compression and equivalencing

It is common for information systems to incorporate *phonetic compression* of names into fixed length codes. Methods of name compression were developed partly to reduce the effects of variation in spelling and errors in recording of names; and partly to simplify the processes of blocking and searching. The main requirements of a name compression code are that all variations of a name are included in the same group, but widely dissimilar names are excluded. A typical approach to name compression is to reduce each name to a short fixed format code by, for example, eliminating vowels, regarding certain consonants as silent, regarding other as equivalent, and taking a set number of significant consonants as the compressed name. Well established compression codes are the Soundex Code, name compression methods devised by Dolby, those used in the New York State Information and Intelligence system (NYSIIS)[10] and the Oxford name compression algorithm.[5 6] To ensure that a record is given a high probability of matching to a block containing all possible variants of names like its own, one individual may appear in several blocks (*file explosion*). For example, the record of a married women would have an entry in one block under her maiden name and in another block under her present surname. If the woman's forename was given as Liz, she might have a record created corresponding to Liz and a second record created corresponding to Elizabeth under the entry for each surname. As an alternative design, forenames can be automatically referred, in the computer, to an *equivalencing dictionary* of recognised contractions and variants of the same forename.

The Unit of Health-Care Epidemiology and our work on record linkage is funded by the Department of Health and the Oxford Regional Health Authority.

1 Secretaries of State for Health, Wales, Northern Ireland, and Scotland. *Working for patients*. London: HMSO, Cmn 555.
2 National Health Service and Department of Health. *Working for patients: framework for information systems: the next steps*. London: HMSO, 1990.
3 Acheson ED. *Medical record linkage*. Oxford: Oxford University Press, 1967.
4 Acheson ED (ed). *Record linkage in medicine*. Proceedings of the International Symposium, Edinburgh, July 1967. London: ES Livingstone, 1968.
5 Baldwin JA, Gill LE. The district number: a comparative test of some record matching methods. *Community Medicine* 1982; 4: 265–75.
6 Gill LE, Baldwin JA. Methods and technology of record linkage: some practical considerations. In: Baldwin JA, Acheson ED, Graham WJ, eds. *Textbook of medical record linkage*. Oxford: Oxford University Press, 1987: 39–54.
7 Dunn HL. Record linkage. *Am J Public Health* 1946; **36**: 1412–6.
8 Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science* 1959; **130** (3381): 954–9.
9 Newcombe HB. The design of efficient systems for linking records into individual and family histories. *Am J Human Gen* 1967; **19**: 335–9.
10 Newcombe HB. *Handbook of record linkage: methods for health and statistical studies, administration, and business*. Oxford: Oxford University Press, 1988.
11 Wild WG. The theory of modulus N check digit systems. *The Computer Bulletin* 1968; **12**: 308–11.
12 Gallian JA, Check digit methods. *International Journal of Applied Engineering Education* 1989; **5**: 503–5.