Chapter 10

# The Development of Record Linkage in Scotland: The Responsive Application of Probability Matching

*Steve Kendrick, National Health Service, Scotland*

**Abstract**

*Since 1968, patient identifiable records of hospital discharges, cancer registrations and death records have been held centrally in Scotland in machine readable form. Patient details are held in order to enable record linkage using probability matching. In the 1970s and early 1980s over forty ad hoc linkages were carried out. Since the late 1980s, the records have been brought together into permanently linked data sets the largest of which now contains over 12 million records spanning the years 1981 to 1995.*

*These linked data sets have enabled a wide range of analyses to be carried out in response to demands from the health service and the medical research community. They have ranged from relatively simple aggregations of data at the patient level to complex studies of long term patient outcomes. Outcome indicators such as 30 day survival after acute myocardial infarction are now published at hospital level.*

*In addition to the main "internal" linkages over seventy linkages have been carried out between external data sets such as surveys (e.g., the West of Scotland Coronary Prevention Study), employee records and clinical audit records and the centrally held linked data sets.*

*The linkage techniques used have evolved to meet the challenges posed by a wide range of customer requirements and data sets. In particular there has been a shift from traditional sort-and-match methods to one pass techniques involving indexing in memory. This has been necessary to enable the linking of relatively small data sets to the main data set without multiple sorting of the data. The technique is currently being adapted to the main linkages to enable much more rapid incorporation of new data. Appropriate use of "best-link" principles has made possible either very high linkage accuracy (e.g., the linkage of Scotland's two main population registers) or reasonable accuracy in linking very poor quality data sets (e.g., linkage of records of victims of cardiac arrest to death records).*

*The paper will use the Scottish experience to illustrate how the application of probability matching needs to be closely attuned to the precise characteristics of and, in particular, the relationship between the data sets to be linked.*

## Introduction

Record linkage using probability matching, like many fields of human endeavour, has progressed as a highly fruitful interplay between theory and experiment, axioms and pragmatism. One viewpoint would see record linkage as primarily a highly practical enterprise based on common-sense and close attention to the empirical characteristics of the data sets involved in any linkage. Another would emphasize the rigorous grounding of record linkage practice in statistical theory and the theory of probability (Fellegi and Sunter, 1969; Newcombe et al., 1992; Arellano, 1992).

Howard Newcombe, pioneer and founder of probability matching techniques, recognises the value of both perspectives. His work has illustrated the continuing dialectic between the theory and the practical craft of linkage. From the point of view of the development of record linkage in Scotland however his most valuable contribution, beyond his initial formulation of the principles of probability matching, has been his emphasis on being guided by the characteristics and structure of the data sets in question and close empirical attention to the emergent qualities of each linkage (Newcombe et al., 1959; Newcombe, 1988). Particularly inspiring has been his insistence that probability matching is at heart a simple and intuitive process and should not be turned into a highly specialised procedure isolated from the day to day concerns of the organization in which it is carried out (Newcombe et al., 1986).

In this paper we wish to show how the development of the methods of record linkage used in the Scottish Health Service have been driven forward by concrete circumstances and in particular by the practical demands of our customers and the needs of the health service as a whole. Although almost no specifically "research and development" time has been devoted to the development of the Scottish system, our openness to the demands of customers and the sheer variety of linkages which this has engendered has in fact produced a rapid pace of development and change which shows no sign of abating.

Although we have pursued a highly pragmatic rather than a theoretical approach, the variety of linkages which have been undertaken has served to give shape to an overview of some of the main factors which need to be taken into account in designing linkages most effectively. The paper is thus in part the story of record linkage in Scotland, in part a concrete account of how our methods have evolved but also contains an overview of some the factors to do with data structure which may be relevant to linkage strategy. More than anything however the paper is an illustration of how the sensitive and flexible application of the very simple and basic principles outlined by Howard Newcombe can produce very powerful results.

## The Context

The current system of medical record linkage in Scotland was made possible by an extremely far sighted decision made as long ago as 1967 by the predecessor organisation to the Information and Statistics Division of the Scottish Health Service and by the Registrar General for Scotland. The decision was taken that from 1968 all hospital discharge records, cancer registrations and death records would be held centrally in machine readable form and would contain patient identifying information (names, dates of birth, area of residence etc.).

The decision to hold patient identifying information was taken with probability matching in mind and reflected familiarity with the early work of Howard Newcombe in Canada and close contact between Scotland and the early stages of the Oxford record linkage initiative. (Heasman, 1968; Heasman and Clarke, 1979).

In what can now be regarded as the first phase of medical record linkage in Scotland, over 40 often sizeable linkages were carried out between the late 1960s and the mid-1980s (for example, Hole et al., 1981; Kendell et al., 1987). The linkages were primarily for epidemiological purposes and each involved the rather laborious specification and development of a bespoke computer program, the whole process often taking over a year to complete. Although the system represented a considerable achievement, by the mid-

1980s it was acknowledged that it would be increasingly inadequate for the perceived future needs of the Scottish Health Service especially in terms of management information.

In the late 1980s the decision was taken to reconstitute the linkage system. Increased computing power and data storage capacity enhanced the feasibility of linking once and for all the set of records pertaining to a given patient. New enquiries, whether epidemiological or relating to service management, would increasingly involve analysis of already linked data rather than requiring fresh linkages.

The years since 1989 have seen the creation of such permanently linked data sets of Scottish health related data. (Kendrick and Clarke, 1993). The largest currently contains all hospital discharge data, cancer registrations and Registrar General's death records from 1981 to 1995 (over 14 million records relating to just over 4 million individuals). A maternity/neonatal data set contains all maternity admissions, neonatal records, Registrar General's birth records and stillbirth and infant death records from 1980 to 1995. Finally, the data set with the longest time span contains linked psychiatric inpatient records and Registrar General's death records from 1970 onwards.

It was envisioned that the creation of the national linked data sets would be carried out purely by automated algorithms with no clerical checking or intervention involved. After linkage of five years of data in the main linked data set it was found that the false positive rate in the larger groups of records was beginning to creep up beyond the 1% level felt to be acceptable for the statistical and management purposes for which the data sets are used. Limited clerical checking has been subsequently used to break up falsely linked groups. This has served to keep both the false positive and false negative rates at below one per cent. More extensive clerical checking is used for specialised purposes such as the linking of death records to the records of the Scottish cancer registry to enable accurate survival analysis for example.

The existence of the linked data sets has generated a high level of demand for analysis. Approaching a thousand analyses have been carried out ranging from simple patient based counts to complex epidemiological analyses. Among the major projects based on the linked data sets have been clinical outcome indicators (published at hospital level on a national basis), analyses of patterns of psychiatric inpatient readmissions and post-discharge mortality and analyses of trends and fluctuations in emergency admissions and the contribution of multiply admitted patients.

However, far from reducing the requirement for specialised data linkage, the existence of permanently linked national data and facilities for linkage has served to fuel the demand for new linkages. Over a hundred and fifty separate probability matching exercises have been carried out over the last five years. These have consisted primarily of linking external data sets of various forms -- survey data, clinical audit data sets -- to the central holdings. A particularly important linkage in the context of a major trial of cholesterol lowering drugs enabled comparison of the accuracy of follow-up using probability matching with reporting based on direct contact with patients. Automated linkage was found to be just as accurate for tracking hospital admissions (West of Scotland Coronary Prevention Study Group, 1995). Other specialised linkages have involved extending the linkage of subsets of the ISD data holdings back to 1968 for epidemiological purposes. (for example, Gillespie et al., 1996). These exercises have varied enormously in scale and complexity, from following up the patients of a particular consultant to linking what are virtually two different registers of the population of Scotland. Linkage proposals are subjected to close scrutiny in terms of the ethics of privacy and confidentiality by a Privacy Advisory Committee which oversees these issues for ISD Scotland and the Registrar General for Scotland.

The Scottish linkage project has been funded primarily as part of the normal operating budget of ISD Scotland. Relatively little time or resources have been available for general research into linkage methodology. Instead the development and refinement of linkage methods has taken place as a response to a wide variety of immediate operational demands. We have become to all intents and purposes a general purpose linkage facility at the heart of the Scottish Health Service operating to very tight deadlines often set in terms of weeks and in extreme cases, days. This has placed a high premium on developing quick, effective and accurate methods of linkage with an emphasis on fitness for purpose rather than straining for precision for its own sake.

Despite the lack of time and resources available for background research and development in linkage methodology, these conditions have in fact fostered, especially in recent years, a rapidly changing and developing approach to linkage.

Before describing the most significant developments involved, a brief overview of the main components will serve to set them in context.

## The Elements of Linkage

For the purposes of this discussion, record linkage using probability matching can be regarded as having three phases or elements each involving a key question.

- **Bringing pairs of records together for comparison**. **--** How do we bring the most effective subset of pairs of records together for comparison? It is usually impossible to carry out probability matching on all pairs of records involved in a linkage. Usually only a subset are compared, those which share a minimum level of identifying information. This has been traditionally achieved by sorting the files into “blocks” or “pockets” within which paired comparisons are carried out (Gill and Baldwin, 1987).

- **Calculating probability weights**. **--** How do we assess the relative likelihood that pairs of records belong to the same person? This lies at the heart of probability matching and has probably been the main focus of much of the record linkage literature. (Newcombe, 1988).

- **Making the linkage decision**. **--** How do we convert the probability weights representing relative odds into absolute odds which will support the linkage decision? The wide variety of linkages undertaken has been particularly important in moving forward understanding in this area.

It would probably be fair to say that of the three areas, it is the second, the calculation of probability weights which has received the most attention and is the best understood. Developments in Scotland over the last few years have occurred in the other two areas as the two subsequent sections will demonstrate.

Before moving on to these developments, our approach to the calculation of probability weights has been relatively conventional and can be quickly summarised. A concern has been to avoid overelaboration and over complexity in the algorithms which calculate the weights. Beyond a certain level increasing refinement of the weight calculation routines tends to involve diminishing returns. This relatively basic approach has been facilitated by the relative richness of the identifying information available on most health related records in Scotland. To take an example, for the internal linking of hospital discharge (SMR1) records across Scotland we have available the patient’s surname (plus sometimes maiden name), first initial, sex and date of birth. We also have postcode of residence. For records within the same hospital (or sometimes the same Health Board) the hospital assigned case reference number can be used. In addition positive weights can be assigned for correspondence of the date of discharge on one record with the date of admission on another. Surnames are compressed using the Soundex/NYSIIS name compression algorithms (Newcombe, 1988) with additional scoring assigned for more detailed levels of agreement and disagreement. Wherever possible specific weights relating to degrees of agreement and disagreement are used.

## Bringing the Pairs of Records Together: One Pass Linkage

## The Limitations of Sort-and-Match

By the time the largest linked data set covered several years of data and consisted of several millions of records, a particular challenge emerged. The linkage team began to be asked to link data sets consisting of relatively small numbers of "external" or "newcomer" records to the central catalog of identifiable records. The external or newcomer records might consist of respondents to a survey, a specialised disease register or a particular group of employees.

In all cases the aim was to link the newcomer data set to the central catalog of records so that the experience of the individuals involved could be traced forward from the date of survey, the date last known to the disease register or the date of employment.

As we have seen, in record linkage it is impossible to bring together and compare all the pairs of records involved in the linkage. The number of pairs which are brought together for comparison is normally reduced to manageable proportions by some form of blocking by which only those pairs of records which share common sets of attributes are compared. For example a common strategy is to compare only those pairs of records which share either the same first initial and NYSIIS/Soundex code or the same date of birth. The normal method of achieving such blocking is to sort the two files concerned on the basis of the blocking criteria. Thus, for a first pass of linkage, the files would be sorted by first initial and NYSIIS/Soundex code to bring together into the same "pocket" or "block" all records sharing the same NYSIIS/Soundex code and first initial. Records would only be compared within this block. Because a number of truly linked pairs of records would not be brought together on this basis (for example, because of a misrecording of first initial), a second pass could be carried out which blocks by date of birth. This second pass involves resorting the files on the basis of date of birth to create a second set of pockets or blocks within which comparison takes place. The results of the first and second passes need to be reconciled and this involves sorting the file yet again.

The key point is that standard methods of blocking involve sorting all the records involved in the linkage at least twice and usually more often. When faced with the kind of linkage mentioned above, involving linking a small number of newcomer records to a central catalog holding several millions of records such a procedure is at best immensely wasteful and at worst impossible. No matter how few newcomer records are involved, it is still necessary to sort all the central catalog records for the years of interest. If only a few years are involved, and especially if linkage is restricted to a subset of the central records e.g., cancer registrations, the exercise is feasible but immensely inefficient. If it is desired to link newcomer records to the entire data set, the exercise becomes, in reality, impossible.

## One Pass Linkage: Blocking Without Sorting

The question thus became: how can we link a relatively small number of newcomer records to the catalog without having to repeatedly sort the catalog? The solution adopted has been to store the newcomer records in memory and carry out blocking using indexes based on numerical elements of the blocking criteria. The catalog records can then be read in sequentially and compared with all the newcomer records which fit the chosen blocking criteria (Kendrick and McIlroy, 1996).

The linkage is thus carried out in the course of "one pass" through the catalog data set.

Before they are brought into contact with the catalog, all newcomer records are read into memory and stored in an array indexed by a unique numeric record identifier. Necessary pre-processing such as generation of NYSIIS/Soundex codes is also carried out.

The next step is the creation of blocking index arrays. In this description we assume that two sets of blocking criteria are being used: first initial and NYSIIS/Soundex code on the one hand, and date of birth on the other.

The blocking index arrays are indexed by numeric elements of the blocking criteria. Thus the first blocking array uses the numeric element of the NYSIIS/Soundex code as its index. All NYSIIS/Soundex codes consist of a letter followed by three figures e.g., A536 or B625. The first blocking index array has a row for each number from 001 to 999 which covers all possible numeric elements of NYSIIS/Soundex codes. In each row are stored the numeric identifiers of the newcomer records whose NYSIIS/Soundex code has the relevant numeric element. For example, the identifiers of newcomer records with surname FRAME (NYSIIS/Soundex code F650) and BROWN (B650) would be stored in the same row.

The second blocking index array has three indices: year, month and day of birth. Along the fourth dimension of the array are stored the numeric identifiers of the newcomer records sharing that date of birth.

Catalog records are then read in one by one. Suppose the first catalog record is for someone named BROWN (NYSIIS/Soundex code B650) with date of birth 1st March, 1922.

- Row 650 of the first blocking index array is inspected to see whether any newcomer records share the numeric element of the Soundex code. If any are found, then the first newcomer record is accessed via its numeric identifier in the newcomer record array. An immediate comparison is made of the first letter of the Soundex code and the first initial. If both match then we proceed to full probability matching between the catalog and newcomer records. If neither or only one match then no further action is taken. We then look at the next newcomer record (if any) indexed on the relevant row of the blocking index array.

- Blocking by date of birth is even easier to simulate in that the blocking criteria are entirely numeric. The catalog record can be directed to all newcomer records which share the same day (in this case 1); month (in this case 3); and year (in this case 22) by directly accessing the relevant array.

How the results of the ensuing pair comparisons are stored and implemented depends upon the structure and purpose of the linkage. Whenever links above a certain weight occur they can be output and stored for implementation in a provisional linkage file. This provisional linkage file can itself be flexibly interrogated to implement a given structure of linkage e.g., we may be only interested in the best link (the link with the highest weight) achieved by each newcomer record (see below).

## One Pass Linkage: Practical Considerations

The above strategy, whereby the newcomer records can be indexed only in terms of the numeric elements of any blocking criteria, is necessary when we are using a programming environment which only allows numeric indexing of arrays. If either the newcomer data or the catalog data is stored in a database which allows direct access by any type of key then the logic of the exercise would be simplified. The file could be flexibly indexed by whatever blocking keys are felt appropriate.

Our impression at present would be that using memory still has advantages in terms of speed of access. This of course is a practical issue and may well change quickly as relational databases and “search engines” improve in speed and efficiency.

The number of newcomer records which can be linked in one pass through the data is of course limited by the available memory. Memory is needed both for storing the elements of the newcomer records which are necessary for linkage and for storing the blocking index arrays. For most ad hoc linkages involving anything up to 15,000 newcomer records this has not tended to be a problem. Larger newcomer data sets can often be linked in sections without affecting the logic of the linkage.

Of the three elements of linkage it is this which is most dependent upon the capabilities of available hardware and software. The implication is that as these capabilities develop, there will be immense potential for moving beyond current limitations.

## The Linkage Decision: Relative and Absolute Odds, Structuring the Linkage and the Best Link Principle

At the heart of the record linkage enterprise is the decision as to whether two records are truly linked. Most often the question is one of whether the records involved relate to the same person. The calculation of probability weights aims to provide a mathematical grounding for this decision.

However, it is a fundamental characteristic of the odds represented by probability weights that they are relative odds rather than absolute odds. They only serve to rank the pairs of records involved in a given linkage in order of the probability that they are truly linked. The relative odds do not represent fixed absolute or betting odds such that a probability weight of 25, for example, would always representing absolute or betting odds of 50/50. The conversion factor will vary from linkage to linkage.

It is absolute odds which are needed to inform the linkage decision. In practical terms the issue of the determinants of the conversion from relative odds to absolute odds can often be bypassed in that the required threshold in terms of absolute odds can usually be identified empirically from inspection of a sample of pairs. However a broad understanding of the relationship between relative and absolute odds is useful in that it can help optimise the way a given linkage should be structured.

Not all linkages require the same absolute odds. The absolute odds required depend upon the purpose of the linkage. They depend upon the costs associated with missing a true link compared with making a false link. For statistical purposes, the required absolute odds may be 50/50. If the linkage is to be used for administrative or patient contact purposes where a false linkage may have extremely damaging consequences, very high absolute odds may be required.

### Relative to Absolute Odds: A Priori Factors

Two of the factors involving in converting relative to absolute odds take the form of relatively straightforward numerical principles. Newcombe has stated them in the context of a search file and a file being searched (Newcombe, 1988; Newcombe, 1995). The first principle is that the higher the proportion of records in the search file for which there exists a linked record in the file being searched, the more favorable will be the conversion factor between relative and absolute odds. The second proposition is that the larger the file being searched, the less favourable will be the conversion factor between relative and absolute odds.

These factors are given for any specific linkage.

### Relative to Absolute Odds: Structural Factors

However the conversion factor between relative and absolute odds can be influenced by how the linkage is implemented. It is important to design the linkage in a way which takes maximum advantage of the structures of the files involved and the relationships between the records in the files. For example, are the relationships between the records in two files one-to-one, one-to-many or many-to-many? How much confidence do we have in previous linkages which may have been carried out on the files involved? How confident are we that a file to be linked already contains only one record per person?

For example, if we want to link to each other a set of hospital discharge records, we have no a priori knowledge of how many records belong to each person. Our best bet is to do a conventional internal linkage and inspect all resulting pairs in setting a threshold. In this case we have relatively little leverage to improve the terms of conversion between relative odds and absolute odds.

If however we are linking a file of hospital discharge records to a file of death records we can obtain some “structural leverage.” Death only occurs once and assuming that this is reflected in there being only one death record per person in the file of death records, the linkage becomes many-to-one. Each hospital discharge record should link to only one death record. The terms of conversion from relative to absolute odds can be improved by only retaining, for each hospital discharge record, the best (highest weight) link which is achieved to a death record (see also Winkler, 1994).

Similarly, at the other end of the life cycle, if we are linking baby records to mothers records, assuming that the mothers records themselves have been correctly linked we should allow each baby to link to only one mother. This in fact was the first context in which the importance of structuring the linkage emerged in Scotland.

## Best Link and Structural Leverage: An Example

### The CHI/NHSCR Linkage

These rather abstract considerations can be best understood in the context of a particular example. In common with the rest of the United Kingdom, Scotland is committed to the development of a unique patient identifier to help streamline the management of all patient contacts with the health service. Historically, Scotland has possessed two health-related registers of the Scottish population. It was felt that the combined strengths of the two registers would provide a firm basis for a new patient identifier.

For the last twenty years the Community Health Index (CHI) has operated on a regional basis as a primary care patient register for such purposes as screening for breast and cervical cancer and childhood immunisation. It contains a wealth of operational information with high population coverage. However, the regional indexes were initially compiled on an opportunistic basis and there was a general perception that there were gaps in its coverage and that there was a high proportion of duplicate records for people who had moved from one area of Scotland to another.

Scotland also possesses a National Health Service Central Register (NHSCR) which has been carefully maintained to contain one record for each resident of Scotland. The NHSCR however contains relatively little operational information.

The initial plan was to carry out an internal linkage of the aggregated regional CHI indexes in order to remove duplicate records and then to link the resulting aggregated CHI data set to the NHSCR to form the basis of a national index.

However, based on our early experience of structuring linkages to maximise the power of the linkage, it was felt that linking the CHI databases to each other via a linkage to the NHSCR would provide more "leverage" in the linkage. (Kendrick et al., 1997)

The data which was available on both data sets to enable linkage was reasonable but not excessively rich. We had forenames, surnames, sex and date of birth but the only residential data was Health Board of residence (average size 300,000). A major bonus was that National Health Service number was available in a well formatted form on all NHSCR records. Because of its irregular format, the British NHS number has been notoriously difficult to use and was available on only a proportion of CHI records and with wide variations in accuracy and formatting.

Although the linkage was primarily concerned with "current" CHI records, those reflecting the current residence and GP registration of the Scottish population, "redundant" CHI records for people who had died or moved to a new Health Board were included in the linkage as a possible basis for constructing historical traces. In order to find a correct NHSCR "home" for as many CHI records as possible the NHSCR file also contained deaths from 1981 as well as known emigrants.

Since we were confident that the NHSCR did contain one record for every Scottish resident but there were suspicions that the CHI data set contained duplicate records (as well as legitimately multiple historical records), it was decided to structure the link as many-to-one. Each CHI record was allowed to link to only one NHSCR record -- the one with which it achieved the highest probability weight. Each NHSCR record on the other hand was allowed to link to as many CHI records as necessary.

## Relative to Absolute Odds: Conversion Factors

The linkage can be described in terms of the factors which were outlined in the previous section as determining the conversion factor between relative and absolute odds.

- **Purpose of the linkage**. -- Any links accepted from the linkage would form the basis of patient contact. A very high level of confidence in the validity of any links was required. Missed links were regarded as less of a problem in that they would normally be picked up in the course of the running of the new index. Thus very high absolute odds for linkage were required.

- **A priori probabilities**. -- Given that both sets of data represented a high level of coverage of the Scottish population, there was a very high probability that a person represented on the CHI file would also be represented on the NHSCR file. In terms of Newcombe's first rule, circumstances could not have been more favourable.

- **File sizes**. -- Reflecting as they did the entire Scottish population as well as deaths and transfers these were large files: approximately 6.3 million NHSCR records against 7.8 million CHI records. This gives a high coincidence factor and, according to Newcombe's second rule, would normally serve to push up the relative odds required for given absolute odds.

- **Structuring the file**. -- Given the knowledge that all Scottish residents were likely to be represented by one NHSCR record and one or more CHI records, it made sense to structure the linkage as a best link many-to-one linkage i.e. allowing each CHI record to link only to the NHSCR record with which it achieved best link would be the most effective route and would maximise the conversion factor between relative and absolute odds.

In broad terms then the linkage faced two difficult circumstances: the requirement for very high absolute odds and the large file sizes. These were more than outweighed however by the massive leverage contributed by the use of the best link principle in the context of a very high a priori probability that people were represented in both files.

## Linkage Results

Of approximately 5,360,000 current registered CHI records, 4,600,000 or 86% linked deterministically to an NHSCR record. There was a match between the Soundex/NYSIIS code of surname, first initial, date of birth, sex and NHS number. For the remaining 750,000 CHI records, probability matching was carried out.

Resources for clerical checking were limited and such checking was limited to a sample of best link pairs to determine a probability weight which would represent absolute odds for the correctness of a linkage which were sufficiently high for administrative purposes. Staff of Health Board Primary Care Teams and the National Health Service Central Register checked 2,500 pairs using existing search and confirmation systems. No incorrect links were found at a probability weight greater than 30 and this was chosen as the administratively acceptable threshold.

To put this outcome into a broad comparative perspective we can compare the CHI/NHSCR linkage with previous linkages in Scotland which did not use the best link principle but which linked similar types of record using virtually the same agreement and disagreement weights for the main identifying items such as name and date of birth.

In the linkage of the Scottish hospital discharge and death record data sets using probability matching, the fifty/fifty threshold (i.e., the weight at which it is equally likely that the two records belong or do not belong to the same person) has remained relatively constant at a probability weight of 25. The fifty/fifty threshold for the best links of CHI to NHSCR records is around 15. Similarly, the threshold below which links between Scottish Cancer Registrations and death records are clerically checked and above which they are accepted automatically is a weight of 40. In the CHI/NHSCR linkage as we have seen, this threshold is 30. In both cases the difference is ten units in the currency of binit weights or logs to the base 2. In terms of odds this is an improvement in the conversion factor from relative to absolute odds of $2^{10}$ or around a thousandfold.

Why the use of only best links in this context should contribute so much extra leverage compared with a pure threshold method is perhaps intuitively obvious but is much more difficult to explain in principle. The logic is perhaps best illustrated by a hypothetical example.

Let us suppose that a CHI record on which is recorded the name Angus MacAllan with date of birth 25/01/1952 has achieved its best link with an NHSCR record on which is recorded the name Angus McAllan born 24/01/1951. There is no NHS number on the CHI record and no other elements agree so that the link achieves what would be, in the context of an unstructured purely threshold linkage, only a moderate probability weight implying a less than fifty/fifty chance that the records belong to the same person. We can best assess the likelihood that these two records would belong to the same person in the CHI/NHSCR linkage context by an indirect route. Let us imagine what would have to be true for the two records not to belong to the same person. Either:

- there is no NHSCR record relating to the individual represented on the CHI record and in addition there exists on the NHSCR file a record relating to another Angus Mc/MacAllan with a highly similar date of birth; or

- there is an NHSCR record corresponding to the individual represented on the CHI record but there are sufficient discrepancies in the recording of the identifying information for this "true link" Angus MacAllan that an NHSCR record for another Angus MacAllan in fact achieves a higher probability weight with the CHI record.

Neither of these scenarios are impossible but they are highly improbable and it is much more likely that the two records really do belong to the same person.

The method used had two additional advantages. The file which was output from the linkage took the form of a copy of each CHI record to which was appended an extract from the NHSCR record to which it had achieved the best link and the weight at which the link was achieved. This file was used as a basis for generating pairs for inspection and links could be extracted at whatever weights were necessary. In essence this means that the threshold for linkage was set and could be varied retrospectively without having to rerun the linkage.

The problem of twins has always bedevilled record linkage. The CHI/NHSCR linkage was able to take advantage of the fact that the NHS numbers for most pairs of twins are consecutive and a high negative weight was given for pairs of records with consecutive NHS numbers. Linkages using best link are in normal circumstances better than linkages using only a numeric threshold. In the presence of consecutive NHS numbers for twins the linkage was very successful in correctly allocating the records for twins.

## One Pass Linkage and the Structuring of Linkages

Although one pass linkage and the structuring of linkages in terms of the best link have developed as separate responses to different challenges, they are not entirely independent.

Given that one of the main aims of one pass linkage is to avoid having to repeatedly sort or restructure the larger or target file, it is natural to implement one pass linkage as a best link procedure i.e., each newcomer record is allowed to link only to the catalog or target record with which it achieves the highest probability weight. Thus, it is not possible for the linkage to bring together records in the target file by "bridging" between them -- this would involve restructuring or resorting. As we saw earlier, as patient record sets in the main linked database grew larger, the false positive rate crept upwards, often because of illegitimate bridging by new records. As the main production linkages are adapted to one pass linkage, this problem will be minimised.

Although the affinity between one pass linkage and the best link principle is one of practical convenience, as we have seen, depending upon the circumstances of the linkage, the best link principle often has highly beneficial effects. Practicality and best practice often go hand in hand.

## Linkage in Scotland: A Possible Future

Another way of looking at the CHI/NHSCR linkage is to see the NHSCR file as a target file at which the regional CHI files were aimed for linkage. As we have seen finding the best link record in the target file for each CHI record proved to have a dramatic effect on the accuracy of the linkage.

The much richer "national CHI" file which has resulted from the linkage and the introduction of national search and enquiry facilities provides an even better target for the linkage of other data sets.

For example, in November 1996 Scotland experienced a severe outbreak of infection by the E-coli 0157 bacterium. Several different sets of records were generated in the course of the outbreak: a case register, community clinic contacts, laboratory records, known exposed cohorts and hospital patients. The quality of identifying information on many of these records was rather poor reflecting the circumstances in which they were collected. ISD Scotland was asked to link these records so that the records for each individual involved in the outbreak could be gathered together. Rather than attempt to link the different sets of records directly to each other, the records were "aimed" at the local Community Health Index and linked to it. Again this method paid off in terms of much more accurate linkage.

It is likely that more and more linkages in Scotland will take the form of aiming data sets at the target of the national CHI. Ultimately the objective is to use such linkages, whereby for example laboratory data sets or hospital Master Patient Indexes are linked to the national CHI , to populate an increasing proportion of Scotland's health records with a unique patient identifier. It is intended that this will eventually reduce the need to record patient identification details such as names and dates of birth on operational records and communications. Instead identification will be via the national CHI number. Such a system is already in place in Tayside Health Board where the CHI number is implemented on a wide range of primary and acute health care records.

In this context the role of probability matching in the Scottish Health Service and the methods used to carry it out are likely to change even more rapidly over the next few years than they have over the last ten years.

As we have emphasised it has been the openness of record linkage in the Scottish Health Service to the demands of a wide range of customers which has driven the rapid development in our methods and this is likely to continue.

In this context the common sense and pragmatic approach to record linkage championed by Howard Newcombe has been especially useful and appropriate as guidance. Working as we are in his footsteps we can summarise some of the most salient emphases.

Record linkage is about being guided by the data and staying as close to the data as possible at all stages. The people who know the data best must be involved. Linkage is an evolutionary and recursive process at all levels. Linkage is a continual learning process and linkage is about what works, not what ought to work.

Finally, record linkage is not about the mechanical application of complex and abstract rules. As circumstances change and data sets vary there is unlikely ever to be one definitive best method of carrying out record linkage using probability matching. Progress will come rather from the flexible and responsive application of what are, at heart, very simple principles.

## Acknowledgments

The following people have contributed over the years to the collective enterprise which is described. In ISD Scotland, James Boyd, Dorothy Gardner, Lena Henderson, Kevin McInneny, Margaret MacLeod, Fiona O'Brien, Chris Povey, Jack Vize, David Walsh, and Bruce Whyte have contributed their insight and programming skills. John Clarke and May Sleigh provided invaluable continuity and guidance. The expertise of Angela Bailey, Eileen Carmichael, Janey Read, and Maggi Reid in checking output has helped keep the system on course. In the Data Centre of the Scottish Health Service's Common Services Agency Gary Donaldson, Alison Jones, Ruth McIlroy, and Debbie McKenzie-Betts have laboured to put the linkage systems on a production basis.

## References

Arellano, M.G. (1992). Comment on Newcombe et al.1(992). *Journal of the American Statistical Association*, 87, 1204-1206.

Fellegi, I.P. and Sunter, A.B., (1969). A Theory of Record Linkage, *Journal of the American Statistical Association*, 40, 1183-1210.

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. et al. (eds), Oxford: Oxford University Press.

Gillespie, W.J.; Henry, D.A.; O'Connell, D.L.; Kendrick, S.W.; Juszczak, E.; McInneny, K.; and Derby, L. (1996). Development of Hematopoietic Cancers after Implantation of Total Joint Replacement, *Clinical Orthopaedics and Related Research*, 329S, S290-296.

Heasman, M.A. (1968). The Use of Record Linkage in Long-term Prospective Studies, in *Record Linkage in Medicine: Proceedings of the International Symposium, Oxford, July 1967,* Oxford: Oxford University Press.

Heasman, M.A. and Clarke, J.A. (1979). Medical Record Linkage in Scotland, *Health Bulletin (Edinburgh),* 37: 97-103.

Hole, D.J.; Clarke, J.A.; Hawthorne, V.M.; and Murdoch, R.M. (1981). Cohort Follow-Up Using Computer Linkage with Routinely Collected Data, *Journal of Chronic Disease*, 34, 291-297.

Kendell, R.E.; Rennie, D.; Clarke, J.A.; and Dean, C. (1987). The Social and Obstetric Correlates of Psychiatric Admission in the Puerperium., in *Textbook of Medical Record Linkage*, Baldwin J.A. et al. (eds), Oxford: Oxford University Press.

Kendrick, S.W. and Clarke, J.A. (1993). The Scottish Medical Record Linkage System., *Health Bulletin (Edinburgh),* 51, 72-79.

Kendrick, S.W. and McIlroy, R. (1996). One Pass Linkage: The Rapid Creation of Patient-Based Data, in *Proceedings of Healthcare Computing 1996: Current Perspectives in Healthcare Computing 1996,* Weybridge, Surrey: British Journal of Healthcare Computing Books.

Kendrick, S.W.; Douglas, M.M.; Gardner, D.; and Hucker, D. (1997). The Best-Link Principle in the Probability Matching of Population Data Sets: The Scottish Experience in Linking the Community Health Index to the National Health Service Central Register, *Methods of Information in Medicine* (in press).

Newcombe, H.B. (1988). *Handbook of Record Linkage,* Oxford: Oxford University Press.

Newcombe, H.B. (1995). Age-Related Bias in Probabilistic Death Searches Due to Neglect of the Prior Likelihoods, *Computers and Biomedical Research,* 28, 87-99.

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

Newcombe, H.B.; Smith, M.E.; and Lalonde, P. (1986). Computerised Record Linkage in Health Research: An Overview, in *Proceedings of the Workshop on Computerised Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*, Howe, G.R. and Spasoff, R.A. (eds), Toronto: University of Toronto Express.

Newcombe, H.B.; Fair, M.E.; and Lalonde, P. (1992). The Use of Names for Linking Personal Records, *Journal of the American Statistical Association*, 87, 1193-1204.

West of Scotland Coronary Prevention Study Group (1995). Computerised Record Linkage Compared with Traditional Patient Follow-up Methods in Clinical Trials and Illustrated in a Prospective Epidemiological Study, *Journal of Clinical Epidemiology*, 48, 1441-1452.

Winkler, W.E. (1994). *Advanced Methods for Record Linkage*, Statistical Research Division, Statistical Research Report Series No. RR94/05, Washington D.C.: U.S. Bureau of the Census.