The Scottish Record Linkage System

BACKGROUND

Howard Newcombe, pioneer and founder of probability matching techniques, 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 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 organisation 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 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 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, 1967; Heasman and Clarke, 1979).

The potential for bringing the records together on a patient basis was first outlined by Heasman in 1968 [1] [2]. Linkage was carried out afresh for each exercise and each linkage involved ad hoc programming by the Scottish Office Computer Service (the turnaround time for each project tended to lie between 6 months and a year).

In the late 1980's increases in Computing power and data storage capacity meant that for the first time it was possible to envision a system in which all the records for the individual could be linked once and held together on a data set. Such a system would enable linked data simply to be interrogated rather than re-linked for each enquiry. It was felt that increasing management and monitoring of health service activity would require a facility for the rapid generation and analysis of patient based data.

THE CURRENT PROJECT

Development of the current system began in May 1989 as a joint project between the Information and Statistics Division and the Computer Centre of the Scottish Health Service. The eventual plan for the new Scottish Record Linkage system was that all records centrally held at ISD would be brought together into the data set with all records pertaining to each patient grouped together.

At present the data set holds eighteen years (1981-1999) of hospital discharge records (SMR1) together with Scottish Cancer Registry records (SMR6/SOCRATES) and Registrar General's death records. Thus for example a cancer patient would have his or her cancer registration, any hospital admissions and any death record held together on the data set.

A maternity/neonatal data set holds maternity (SMR2), neonatal (SMR11) and infant deaths/stillbirths records for 1980-1995. All records pertaining to a mother and her births are held together.

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 three per cent. More extensive clerical checking is used for specialised purposes such as the linking of death records to the Scottish Cancer Registry to enable accurate survival analysis for example.

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. Other specialised linkages have involved extending the linkage of subsets of the ISD data holdings back to 1968 for epidemiological purposes. (for example, MIDSPAN). Linkage proposals are subjected to close scrutiny in terms of the ethics of privacy and confidentiality by a Privacy Advisory Committee, which oversees these cases for ISD Scotland and the Registrar General for Scotland.

Approaching a thousand linked 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.

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.

METHODS OF LINKING

In a world with perfect recording of identifying information and unchanging personal circumstances, all that would be necessary to link records would be the sorting of the records to be matched by personal identifiers. In the real world of data however, for each of the core items of identifying information used to link the records (surname, initial, year, month and day of birth), there may be a discrepancy rate of up to 3% in pairs of records belonging to the same person. Thus exact matching using these items could miss up to 15 % of true links.

To allow for the imperfections of the data, the system uses methods of probability matching which have been developed and refined in Canada [3] , Oxford [4] and Scotland [5] itself over the last thirty years. Despite the size of the data sets, linking the records consists of carrying out the same basic operation over and over again. This operation is the comparison of two records and the decision as to whether they belong to the same individual.

THE ELEMENTS OF LINKAGE.

1. 1. **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 e.g. soundex, date of birth etc. (Gill and Baldwin, 1987).

2. 2. **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 record linkage literature (Newcombe, 1988).

3. 3. **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.

1. Blocking

In an ideal world with infinite computing power we would carry out probability matching between every pair of records in order to determine whether they belong to the same person. At present this is realistically beyond current computing capacities and would be enormously wasteful even if it were possible. It is necessary to cut down in some way the number of pair comparisons which are made in a given linkage. Instead of comparing all pairs of records we compare only those records which have some minimum level of agreement in identifying items ('blocking' the records).

In the linkages carried out at ISD we tend to compare only those pairs of records between which there is agreement on:

	Soundex/NYSIIS code, first initial and sex	(Block A)
or	All elements of date of birth (day, month, year)	(Block B)

Thus records will not be compared if they disagree on one or more of the first set of blocking items and also disagree on one or more of the second set of blocking items. It is of course possible that two records belonging to the same person will disagree on for example, first initial and also date of birth. Experience shows that the proportion of true links thus lost because of blocking is less than 0.5%.

2. Probability Weights

Our approach to the calculation of probability weights has been relatively conventional and can be quickly summarised. A concern has been to avoid over-elaboration 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.

For the internal linking of hospital discharge (SMR1) records across Scotland we have available the patient's surname (plus sometimes maiden name), forename, 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. Soundex and related name compression algorithms overcome some of the problems associated with misspelling of names and variant spellings.

Blocking allows subsets of the records to be efficiently brought together for comparison. Finally and most importantly probability matching allows mathematically precise assessment of the implications of the levels of agreement and disagreement between records.

Probability matching

Two very simple and common sense principles underlie probability matching:

A. A. Every time an item of identifying information is the same on the two records, the probability that they apply to the same person is increased.

B. B. Every time that an item of identifying information differs between two records, the probability that they apply to the same person is usually decreased.

Whatever kind of matching we are doing, whether linking records within a file or linking records between files, we are looking at pairs of records and trying to decide whether they belong to the same person or don't belong to the same person. We are trying to divide the pairs into two classes - which are more generally referred to as 'truly linked' or 'truly unlinked', i.e. in our case belonging to the same person or not belonging to the same person.

The common core of identifying items are as follows:

1. 1. Surname
2. 2. First initial (also full forename and second initial if available)
3. 3. Sex
4. 4. Year, month and day of birth
5. 5. Postcode.

In principle, any items whose level of agreement or disagreement influences the probability that two records do or do not belong to the same person can be used by the computer algorithm. However, items should be statistically independent as far as possible.

Every time we compare an item of identifying information between two records we obtain what can be called an outcome. In the first instance this is either agreement or disagreement.

For every outcome we ask the same two questions.

1. 1. How often is this outcome likely to occur if the two records really do belong to the same person (are truly linked)?

2. 2. How often is this outcome likely to occur if the two records really don't belong the same person (are truly unlinked)?

The ratio between these two probabilities or odds is what is called an odds ratio - this is a measure of how much that particular outcome has increased or decreased the chances that the two records belong to the same individual. Odds can be awkward to handle so probability matching tends to use binit weights instead. The binit weight is the odds expressed as a logarithm to base 2.

The linkage methodology is aimed at squeezing the maximum amount of discrimination from the available identifying information. Thus the distribution of

probability scores differs for each kind of linkage. The threshold (or score at which the decision to link is made) is determined by clerical checking of a sample of pairs for each type of link.

The odds ratio: an example

Suppose we have two records, and we are comparing their first initials. We find that they both have first initial 'J'. We want to calculate an odds ratio which will tell us what effect this outcome - agreement of first initial 'J' - has on the chances that the records belong to the same person.

If both records belong to the same person how often will one record have the initial 'J'? In a perfect world with perfect data the answer would be always - the probability would be one, or in percentage terms, 100%. However, there are often going to be discrepancies in identifying information between records applying to the same person. If we estimate that the first initial is likely to disagree 3% of the time on records applying to the same person, then it will agree 97% of the time. So on the top line of our odds ratio we have a figure of 97%.

Next we look to the bottom line of the odds ratio. How often are we going to get agreement on the initial 'J' among pairs of records which do not belong to the same person? The answer quite simply depends upon how common that first initial is. If 20% of all first initials are 'J', then if we take any record with first initial 'J' and compare it with all the other records, then 20% of the time the record it is compared with will have first initial 'J'. So the bottom line of the odds ratio is 20%. The odds ratio then is 97%/20% or 4.85.

So agreement of first initial 'J' has improved our chances that the records belong to the same person by 4.85 to one.

What if the first initial disagrees? Again we compare the outcome among pairs of records, which do belong to the same person against pairs of records which do not.

The top line of the odds ratio is 3% (if you take all records with initial 'J', then 3% of the time - even among records belonging to the same person - the other record will have a different initial.) For the bottom line, we want to know how often the first initial disagrees when the records do not belong to the same person. For illustration we can take the initial as disagreeing 92.5% of the time among records not belonging to the same person. So for disagreement of first initial we have an odds ration of 3%/92.5% or 1 to 32. So disagreement of first initial has reduced the chances that the records belong to the same person by 32 to 1.

So we now have a quantitative estimate of how much an agreement on first initial 'J' has improved our chances that we are looking at records belonging to the same person. Similarly we have a quantitative estimate of how much a disagreement on first initial has reduced the chances that the records relate to the same person.

We can now give an example of how the odds ratios deriving from comparison of individual identifying items can be combined to give odds for the overall comparison of the two records.

Suppose we have two records each with the identifying information:

Male J Thompson born 15 05 1932
Male J Thompson born 05 05 1932

The odds associated with these comparisons are as follows:

				Binit
Sex				
Agreement: odds ratio	99.5%/50%	=	1.99	+0.99
First initial				
Agreement: odds ratio	97%/20%	=	4.85	+2.28
Surname				
Agreement: odds ratio	97%/0.8%	=	121.25	+6.92
Day of birth				
Agreement: odds ratio	3%/92%	=	0.0326	-4.94
Month of birth				
Agreement: odds ratio	97%/8.3%	=	11.7	+3.55
Year of birth				
Agreement: odds ratio	97%/1.4%	=	70.0	+6.13

How much have all these comparisons of identifying information improved the chances that these two records really apply to the same person? You combine odds by multiplying them:

1.99 x 4.85 x 121.25 x 0.0326 x 11.7 x 70 = 31,245 to 1.

So the comparisons have increased the likelihood that the two records belong to the same person by 31,245 to 1. However, that does not mean that it is a certainty. Our files have millions of records on millions of individuals. It is not inconceivable that there is more than one male J. Thompson born on the 14th or 15th of May 1932. That is why the procedure is known as probability matching - there are never any certainties. And since there are no certainties, we still have to make a decision as to whether or not the records do apply to the same person.

Binit weights

Odds like 31,245 to 1 are rather awkward to handle. Probability matching tends to use instead what are called binit weights. So far we have talked about odds ratios e.g. the odds ratio for agreement on initial ‘j’ is 4.85 to 1. The binit weight is this number expressed as a logarithm to base 2.

In this context, the most useful thing about logarithms in general, or binit weights in particular, is that they can be added together. Adding together the binit weights is the same as multiplying the odds ratios. So our overall improvement in the chances that

the records belong to the same person of 31,245 to 1 is equivalent to a binit weight of 14.93.

The essence of record linkage is to calculate the overall binit weight for each pair of records. High binit weights mean that the records are likely to belong to the same person. Low binit weights (which reflect odds against) mean that the records are unlikely to belong to the same person.

Soundex/NYSIIS codes

Surnames are changed to a coded format in order to overcome the effects of most discrepancies in the spelling. First the NYSIIS (New York State Intelligence Information System) name compression algorithm is applied. This carries out such tasks as bringing together commonly confused letter groups like 'ch' and 'gh' or 'sh' and 'sch' as well as removing vowels. The surnames are then Soundexed [6] , which involves giving the same code to similar sounding non-initial constants. The resulting compression and soundex codes are assigned different weights for agreement depending upon their frequency in the population.

3. Decision-making

Binit weights present us with a mathematical expression of the extent to which the available identifying information increases or decreases the chances that two records belong together. These however are only relative odds. They allow us to rank order the pairs of records in order of likelihood. They are not absolute odds. Such absolute odds depend upon various factors such as the size of the data sets involved. Methods of calculating such absolute odds are available but they are usually based on rather speculative assumptions. It is much safer to base the decision on which records belong together on a match weight threshold based on empirical inspection. In other words we compare records, calculate relative odds for each pair and look at a selection of odds before deciding on the cut off point for accepting matches.

When the frequencies of pairs of records with given values of the binit weight are graphed, a *bimetal* pattern usually emerges (see IARC report No.32 - Automated Data Collection in Cancer Registrations). The group of pairs of records with high binit weights can be taken as matches (as belonging to the same person). The group with low binit weights can be regarded as non-matches. It is the group in between which cause problems.

The crucial step is to identify a threshold above which pairs will be taken as linking, and below which the pairs will not be accepted as linking.

This threshold is usually determined by clerical inspection of a sample of pairs of records. The threshold is usually set at the 50/50 point. In other words, at the threshold it is a fifty-fifty bet as to whether the pair of records belongs to the same person. Above the threshold it is more likely than not that they do belong to the same person. Below the threshold it is more likely than not that they do not belong to the same person.

Once a threshold in terms of the binit weight has been set, the computer can be allowed to make the decisions as to whether records belong together. In practice the development of match weights and the setting of the threshold is an iterative process with results depending on the precise characteristics of the data sets involved.

Tuning the linkages

Tuning the linkages, either in terms of adjusting the weights for particular comparisons or by adjusting the match threshold is an iterative process.

All linkages are different and the quality of the linkage is best ensured by taking careful account of the precise properties of the data sets involved and the different problems, which emerge in linking two particular data sets.

The linkage threshold is established or confirmed by inspecting the pairs output file and thereby establishing the weight above which it is more likely than not that a pair of records belong to the same person and below which it is more likely than not that the pair of records do not belong to the same person. This threshold is often confirmed in terms of the graph of the frequency of the outcome weights for a particular linkage. The 50/50 threshold (the weight at which it is evens whether records do or do not belong together) often corresponds to the low point of the trough in the frequency counts.

ONE PASS LINKAGE

The limitations of sort-and-match

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. The traditional method of achieving such blocking is to sort the two files concerned on the basis of the blocking criteria. The files would be first sorted by first initial and NYSIIS/Soundex code to bring together into the same '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 mis-recording of first initial), a second pass could be carried out which blocks by date of birth. This involves resorting the files on the basis of date of birth to create a second set of '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 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.

One pass linkage in principle

The major problem with traditional methods of record linkage when faced with the necessity of linking relatively small sets of 'newcomer' records with large national data sets, is that they involve sorting the rather large combined file at least twice as well as copying and reformatting the national data sets.

One pass linkage avoids these problems by storing the newcomer records in the computer memory so that they can be accessed directly rather than sequentially, efficiently and at high speed.

Thus instead of combining the newcomer records with the national data set and sorting the two files together in order to bring together the right pairs for comparison, the entire set of newcomer records can be brought into comparison with each of the records in the national data set.

Thus by reading the newcomer records into memory we have solved the problem of bringing pairs of newcomer and national data set records together. However we have not solved the problem of limiting the number of comparisons - the task which is carried out by the process of blocking in traditional methods.

Essentially we solve the problem by simulating the blocking process using blocking index arrays. Each newcomer record has a unique identifier. These identifiers are

stored in the blocking index arrays to allow the national data set records to be directed towards comparison with simulated blocks of newcomer records.

In principle the process of one pass linkage consists of:

a) a) storing all the newcomer records in memory
b) b) reading in the national data set records sequentially
c) c) directing each national data set record towards comparison with the appropriate subsets(blocks) of newcomer records by using the blocking index arrays

Onepass Example

<<Flow Chart Example Required for this explanation>>

When a record in the national data set of someone called Anderson, with Soundex code A536 and date of birth 12th December 1942 is read into the matching program, the following happens.

First the program uses the Soundex blocking index array to find the identifiers of all newcomer records with numeric component of the Soundex 536. These candidate newcomer records are then checked to see whether they have the same first initial and the same sex. If they do not they are excluded from full comparison

Secondly the program uses the date of birth blocking array to find the identifiers of all the newcomer records with date of birth 12th December 1942. These identifiers are in array cells with first three subscripts 12, 12, 42. These newcomer records are accepted for comparison unless they have already been accepted according to the first set of blocking criteria i.e. to avoid comparing the same records more than once. If accepted they are added to the list of newcomer records to be compared to the national data set record belonging to Mrs Felicity Anderson born on the 12th of December 1942.

QUALITY OF LINKAGE

The linkage system has been automated as much as possible. The probability matching algorithm alone makes the decision as to whether records belong together. Clerical monitoring shows that on a pair-wise basis, both the false positive rate (the proportion of pairs which are incorrectly linked) and the false negative rate (the proportion of pairs which the system fails to link) are around three per cent.

As the data set has expanded the number of patient record sets with large numbers of records has grown. In order to construct a patient record set with 10 records, up to 45 pair comparisons will have been carried out, each comparison contributing its own possibility of a false positive link. Thus larger groups of records are more likely to be false positive. Some of the more important groups moreover tend to be the larger groups. Patient record sets containing cancer registrations tend to have more records than average have thus have a relatively high error rate. For this reason, groups of records where there is an obvious error such as two death records or a hospital admission following death have been targeted for clerical correction. Such errors will help to keep the overall false positive and false negative rates close to one percent.

By using such a focused approach to clerical checking we are intending to achieve the advantages of the quality of a fully clerically checked system without the massive investment of time and expense which such a system would involve.

CONFIDENTIALITY

Medical Record Linkage poses particular problems of data confidentiality both because of the comprehensive nature of the linked data and the necessity to use personal identifying information in order to carry out the linkage. All data involved in the linkage both as input and output is handled under the same strict regulations as any other data containing personal identifying information, which is held centrally by the Scottish Health Service.

ORGANISATION OF THE DATA

At present the linked data is stored as a conventional flat ASCII file of records with the records for each individual stored adjacently in chronological order with a unique personal identifier attached to each. The different types of records are stored in their original unlinked format preceded by several fields of linkage information.

This means that any information contained in the unlinked records is available for the analysis of the linked data. The main advantage is that the system exhibits the maximum degree of flexibility in terms of the range of analyses possible. The down side is that the data set is relatively complex to work with requiring the use of bespoke FORTRAN programs to access the data.

LINKAGE BEYOND THE LINKED DATA SETS

Although an increasing proportion of enquiries requiring linked data can be satisfied on the basis of the linked data sets, there is a continuing demand for linkage of data beyond their limits.

Much of ISD's data holds sufficient identifying information for linkage as far back as 1968. Subsets of this data can be linked on an ad hoc basis for the entire period from 1968 to the present. However the data from 1968 to 1974 does not contain many of the identifying items required to provide robust matches. There are also facilities for linking external data to ISD's own data holdings.

ANALYSIS

Episodes, stays and patients.

Three levels of data can be seen as defining the basic building blocks of the linked data set.

a) a) At the lowest level is the individual record - whether SMR1 in-patient, SMR6 cancer registration or Registrar General's death record.
b) b) An intermediate level is defined primarily in terms of SMR1 records. This is the continuous inpatient stay. A continuous inpatient stay is defined as all the SMR1 records referring to the same continuous spell of inpatient treatment - whether or not this involves transfers between hospitals or even between health boards. All records belonging to the same continuous inpatient stay are marked as such on the linked data set.
c) c) Finally, all the records deemed to refer to the same patient form the patient record set.

Monitoring of levels of activity in the National Health Service has tended to be carried out in terms of patient episodes corresponding to the SMR1 record. Despite the fact that statistics are usually based on such counts of patient episodes, they are often loosely referred as 'numbers of patients being treated'. The linked data set allows us to be more precise in referring to counts of episodes, inpatient stays and numbers of different patients treated in a given time period.

Any analysis of levels of activity can be carried out at episode, stay or patient level. Great care is needed however when assigning stays or patients to, for example, Health Board of Treatment in that the episodes which make up an inpatient stay or a patient record set can belong to more than one Health Board.

Mortality analysis

A crucial element of many analyses involving linked data is the ability to link to death records. The Registrar General has allowed ISD to carry out linkages with death records on the understanding that permission is sought and that the Registrar General is kept fully informed of such linkages.

The records of the Scottish Cancer Registry from 1968 onwards have been linked with Registrar General's death records in order to supplement the flagging carried out at National Health Service Central Register. From 1992 onwards, record linkage will replace flagging at NHSCR except for deaths in England and Wales. A similar exercise has been undertaken for the Scottish Cardiac Surgery Register.

Statistical Modelling

Although counts in terms of linked blocks of data such as inpatient stays and patient record sets are an important aspect of the output of the system, the bulk of the added value of the linked data set comes from analysing the temporal relationships between the records in the system.

Such 'relationship' information is more difficult to present than normal record based data. The Record Linkage team is investing time in working out the best ways of presenting the data.

Beyond the basic presentational level, and especially for purposes of epidemiological research, such topics call for the use of types of statistical modelling, which have only emerged over the last couple of decades. For example, the Cox's proportional hazard model [7] allows the disentanglement of the effects of such variables of age of patient, date of operation and type of operation on the survival rate after a particular type of operation such as prostatectomy. Logistic regression can be used to explore the extent to which variation in readmission rates is independent of changes in the age and sex distribution of patients (A model of the kind of analysis involved is provided in a different context by Leyland et al. [8]).

References

1. 1. Heasman, M. A. The use of Record Linkage in long-term prospective studies in Acheson, E. D. (Ed) **Record Linkage in Medicine,** proceedings of the actual symposium, Oxford, July 1967. OUP: Oxford 1968.
2. 2. Heasman, M. A. and Clarke, J. A. Medical Record Linkage in Scotland. **Health Bulletin (Edinburgh)** 1979; 37; 97-103.
3. 3. Newcombe, H. B. Handbook of Record Linkage, OUP: New York, 1988.
4. 4. Baldwin, J. A., Acheson, E. D. and Graham, W. J. (eds) **Textbook of Record Linkage** OUP: Oxford 1968.
5. 5. Heasman, M. A. The use of Record Linkage in long-term prospective studies in Acheson, E. D. (Ed) **Record Linkage in Medicine,** proceedings of the international symposium, Oxford, July 1967. OUP: Oxford 1968.
6. 6. Newcombe, H. B. Handbook of Record Linkage, OUP: New York, 1988.
7. 7. Cox, D. R. and Oakes, D. **Analysis of Survival Data** Chapman and Hall: London, 1984.
8. 8. Leyland, A. H., Pritchard, C. W., McLoone, P. and Boddy, F. A. Measures of performance in Scottish maternity hospitals. **British Medical Journal** 303: 389-93 1991.

References

Arellano, M. G. (1992) "Comment on Newcombe et al., 1992" *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. "Development of Hematopoietic Cencers after Implantation of Total Joint Replacement" *Clinical Orthopaedics and Related Research.* 329S, S290-296.

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., 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. (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", *Bureau of The Census Statistical Research Division Statistical Research Report Series No. RR94/05. U. S. Bureau of the Census, Statistical Research Division, Washington D. C.*