© 2005 Schattauer GmbH

The EpiLink Record Linkage Software

Presentation and Results of Linkage Test on Cancer Registry Files

P. Contiero[1], A. Tittarelli[1], G. Tagliabue[1], A. Maghini[1], S. Fabiano[1], P. Crosignani[1], R. Tessandori[2]

[1]Cancer Registry Division, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy
[2]Azienda Sanitaria Locale della Provincia di Sondrio, Sondrio, Italy

Summary

Objectives: Record linkage, the process of bringing together separately compiled but related records from different databases, is essential in many areas of biomedical research. We developed a record linkage program (EpiLink), which employs a simple mathematical approach. We describe the program and present results obtained testing it in a linkage task.

Methods: EpiLink was designed to be flexible with user-friendly settings to tailor linkage and operating parameters to specific linkage tasks, and employ deterministic, probabilistic or sequential deterministic-probabilistic linkage strategies as required. The user can also standardize data format, examine linkage results and accept or discard them. We used EpiLink to link a subset of cases of the Lombardy Cancer Registry (20,724 records) with the Social Security file of the population (1,021,846 records) covered by the registry. The linkage strategy was deterministic, followed by several probabilistic linkage steps.

Results: Manual inspection of the results showed that EpiLink achieved 98.8% specificity and 96.5% sensitivity.

Conclusions: EpiLink is a practical and accurate means of linking records from different databases that can be used by non-statisticians and is efficient in terms of human and financial resources.

Keywords

Cancer registry, record linkage, computer program, follow-up methods

Methods Inf Med 2005; 44: 66–71

Introduction

Record linkage is the process of bringing together separately-compiled but related records from different databases [1] and is essential in cohort follow-up, clinical trials, epidemiological studies, health service research and health service management [2-9]. The increasing use of large electronic health databases has accentuated the requirement for automated systems of record linkage, since it is impractical to carry out record linkage by hand. In fact, essential or desirable record linkage projects have not been actuated because, in the absence of automated procedures, the human and financial resources required to complete such projects are prohibitively large [10].

The difficulties of record linkage vary with the characteristics of the data files being linked. When a unique identifier is present (e.g. social security number), the linkage is straightforward (if the numbers are accurate) and consists of matching the records in the two data files by means of the common identifier [1]. However, unique identifiers are not always present, and this is typically the case when the data set is assembled retrospectively. Moreover the unique identifier may be subject to error. In such cases it is necessary to use other identifiers such as surname, name, date of birth, etc. However this often leads to other difficulties, the main ones being that personal identifiers are not unique to an individual and even a combination may not uniquely identify an individual; they are also subject to errors, can be missing, or may change with time (e.g. address).

To surmount these problems two approaches are possible: deterministic linkage and probabilistic linkage [10, 11]. The deterministic approach only links records that perfectly match in terms of the linkage items chosen. This is a simple low-cost procedure, however it does not match records with errors or repetitions of individual records. The probabilistic approach links records according to probabilistic rules designed to overcome problems due to errors, including multiple occurrences of supposedly unique identifiers. However, the use of probabilistic methods introduces considerable complexity into the linkage process and requires the involvement of personnel with competence in statistics.

The most practicable approach at present is to implement a tailored strategy for each record linkage task, deciding whether it should be deterministic or probabilistic, what fields should be involved in the linking process, the rank of their importance, and the statistical model used to identify the links between records. The choice of strategy will therefore depend on the structure and quality of the database, user competence, and the aims of the linkage. If the quality of data is high or a certain percentage of errors is acceptable, the best linkage approach may be deterministic; otherwise probabilistic record linking is necessary. It should be evident therefore that a probabilistic record linkage program should be highly flexible in that its settings should be modifiable to suit the linkage project being undertaken.

We have developed a relatively simple probabilistic data linkage program, called EpiLink, using a straightforward mathematical approach that can be understood by non-expert users. Our aim was to produce a program that is cost-effective in terms of computer and human resources, and that can be used as an alternative to more complex probabilistic record linkage programs [12-15]. The program was developed using database files of the Lombardy Cancer Registry (LCR) and was intended to be:

- flexible, permitting users to impose linkage settings according to the characteristics of the records being linked;
- user-friendly, allowing operators unfamiliar with computer languages or statistical methods to use the program efficiently via a familiar graphic user interface (GUI);
- portable: implementable on various computer configurations and operating systems with easy installation and maintenance procedures.

Record linkage is a central concern of cancer registries, whose principal mission is to record all cancer cases in a defined population and collect and integrate information on those cases from various sources including hospital records, mortality files and pathology reports [16]. The specific aim of the work reported here was to test the program in linking a subset of the cases in the LCR [16, 17] to the Lombardy region's social security file in order to check the life status of all cases not known to be dead, and to assess cancer site-specific survival.

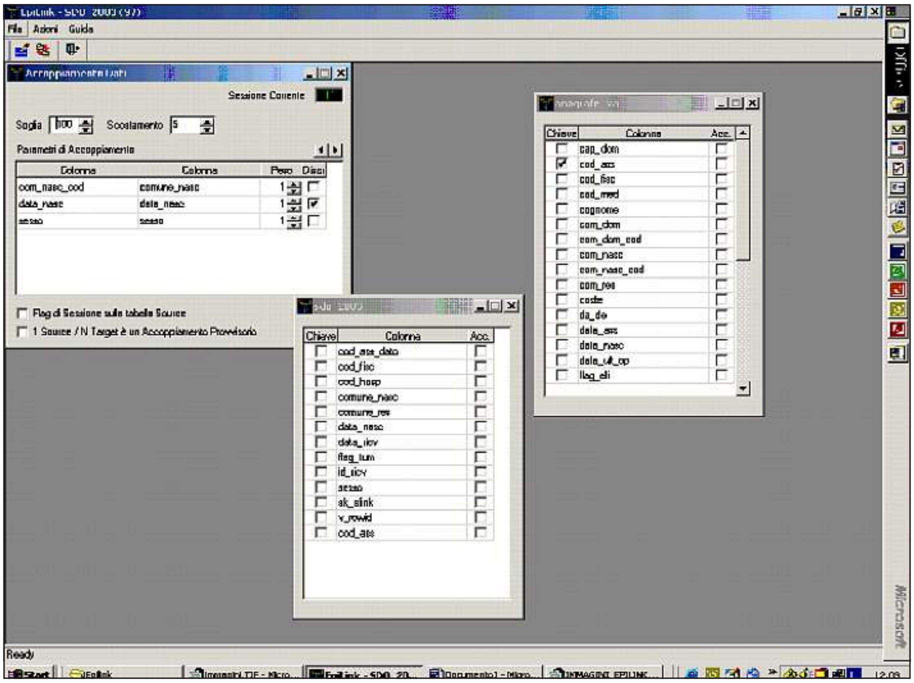
Materials and Methods

Program Features

To enhance the portability of the EpiLink record linkage computer program it was implemented on client server architecture, and is capable of running on a stand-alone PC or on more complex configurations, on top of Unix or various Windows operating systems. The program has a simple installation procedure. When installed on a server more than one client can use it simultaneously. It can perform multiple linkage projects simultaneously following definition of multiple profiles. It works on existing databases, and is also capable of accessing external tables and inserting them to an existing database. The program can accept several types of data structure and working environments so the data do not require modification before the program can use them. To enhance ease of use the program was designed so that all options and operating parameters can be set without changing the code.

The record linkage process takes place in a series of steps, each characterized by specific user-defined options and parameters.

The choice of information items and statistical parameters to use is critical to the success of probabilistic linkage. Running the linkage without a preliminary analysis of the information in the databases is likely



Sessione Corrente

Soglia 100 Scostamento 5

Parametri di Accoppiamento

Colonna	Colonna	Peso	Disci
com_nasc_cod	comune_nasc	1	□
data_nasc	data_nasc	1	☑
sesso	sesso	1	□

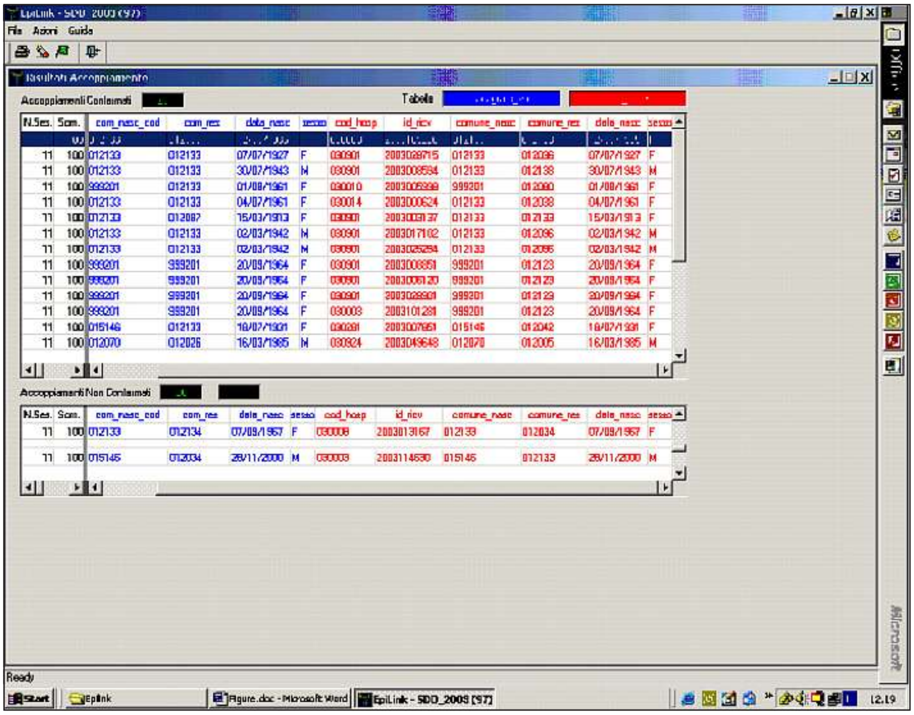
□ Flag di Sessione sulle tabelle Source

□ 1 Source / N Target è un Accoppiamento Provvisorio

Chiave	Colonna	Acc.
□	cod_ass_dato	□
□	cod_fisc	□
□	cod_hosp	□
□	comune_nasc	□
□	comune_res	□
□	data_nasc	□
□	data_ricv	□
□	flag_tum	□
□	id_ricv	□
□	sesso	□
□	sk_elink	□
□	v_rowid	□
□	cod_ass	□

Chiave	Colonna	Acc.
□	cap_dom	□
☑	cod_ass	□
□	cod_fisc	□
□	cod_med	□
□	cognome	□
□	com_dom	□
□	com_dom_cod	□
□	com_nasc	□
□	com_nasc_cod	□
□	com_res	□
□	coste	□
□	da_de	□
□	data_ass	□
□	data_nasc	□
□	data_ult_op	□
□	flag_eli	□

Fig. 1 Form to select items to link, weightings and thresholds



Accoppiamenti Confermati

Tabella

N.Ses.	Som.	com_nasc_cod	com_res	data_nasc	sesso	cod_hosp	id_ricv	comune_nasc	comune_res	data_nasc	sesso
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	100	012133	012133	07/07/1927	F	030901	2003028715	012133	012096	07/07/1927	F
11	100	012133	012133	30/07/1943	M	030901	2003008594	012133	012138	30/07/1943	M
11	100	999201	012133	01/08/1961	F	030010	2003005999	999201	012080	01/08/1961	F
11	100	012133	012133	04/07/1961	F	030014	2003000624	012133	012038	04/07/1961	F
11	100	012133	012087	15/03/1913	F	030901	2003003137	012133	012133	15/03/1913	F
11	100	012133	012133	02/03/1942	M	030901	2003017102	012133	012096	02/03/1942	M
11	100	012133	012133	02/03/1942	M	030901	2003025254	012133	012096	02/03/1942	M
11	100	999201	999201	20/09/1964	F	030901	2003006851	999201	012123	20/09/1964	F
11	100	999201	999201	20/09/1964	F	030901	2003006120	999201	012123	20/09/1964	F
11	100	999201	999201	20/09/1964	F	030901	2003028901	999201	012123	20/09/1964	F
11	100	999201	999201	20/09/1964	F	030003	2003101281	999201	012123	20/09/1964	F
11	100	015146	012133	18/07/1931	F	030281	2003007951	015146	012042	18/07/1931	F
11	100	012070	012026	16/03/1985	M	030824	2003049648	012070	012005	16/03/1985	M

Accoppiamenti Non Confermati

N.Ses.	Som.	com_nasc_cod	com_res	data_nasc	sesso	cod_hosp	id_ricv	comune_nasc	comune_res	data_nasc	sesso
11	100	012133	012134	07/09/1967	F	030008	2003013167	012133	012034	07/09/1967	F
11	100	015145	012034	28/11/2000	M	030003	2003114630	015145	012133	28/11/2000	M

Fig. 2 Form to look at linkage results

to result in a large number of erroneous linkages. We resolved this problem by incorporating a procedure that allows trial linking on small subsets of data. This allows the testing of hypotheses regarding the linkage, the heuristic discovery of improvements, and provides information to decide which items are best involved in the linkage process.

Options incorporated into the user interface allow the user to:

- load external tables into the working database;
- standardize data, for example by specifying date format;
- select the items to be used in the linkage, and choose the parameters for comparison (Fig. 1);
- execute the linkage;
- check all linked records by a customizable interface (Fig. 2);
- examine record linkage statistics;
- discard trial linkage results.

Statistical Approach to Linking

We used a simple mathematical approach based on a family of functions called similarity indexes which are widely used in taxonomy, ecology, genetics and text analysis [18-21].

Similarity indexes are intuitive and simple to handle from the mathematical point of view. The linkage is performed between two tables, the target table and the source table. The target table has records which are to be matched with records in the source table. The target table can have record duplications, the source cannot. A field $\underline{x}$ in the source is matched to a field $\underline{t}$ in the target by means of the following similarity function which affords a similarity value $s(\underline{x}, \underline{t})$ that expresses the percentage correspondence between the values of the source and target fields:

$$s(\underline{x}, \underline{t}) = 2 \mid \underline{x} \wedge \underline{t} \mid / (lx + lt) \ (\times 100)$$

where lx and lt are the lengths of the compared strings and $\mid \underline{x} \wedge \underline{t} \mid$ are the characters in common, taken in sequence.

Numerous string manipulators more or less similar to the s function are described in the literature [22-24]. Some of these are more efficient than the s function in that they are able to recognize probable misspellings for example, but introduce a degree of complexity that is contrary to the aim of the present study.

Before computing the s function, EpiLink cleans the strings being considered of special characters such as accents, apostrophes, commas, full stops (periods), and multiple spaces between words, etc.

EpiLink does not specifically incorporate any search code algorithms such as Soundex or Davidson that use various criteria (for example words pronounced alike) to suggest matches. However, such algorithms can be applied to databases before using EpiLink.

The process of matching a source record to a target record involves assigning a weighting to each considered field, and summing them as shown in the general formula below.

The need for weightings derives from the fact that the fields will have different error rates and discriminating powers. Thus, if the error rate is high, a low weighting would be assigned and vice versa. The discriminating power of a field is inversely proportional to the average recurrence frequency of values in that field. Thus, the less often values recur in a field, the more useful the field is for identifying a record, consequently it is assigned a higher weighting [8].

The general formula for comparing records is:

$$S(\underline{x}, \underline{t}) = \Sigma_i\, w_i\, s(\underline{x}_i/\underline{t}_i)/ \Sigma_i\, w_i \ (\times 100)$$

where w_i is the weighting assigned to the i^{th} field. For each target record, the S value is calculated for each record of the source and takes values from 0 (no similarity) to 1 (all field values equal between the two records compared).

EpiLink leaves the user free to choose the weightings for each field, but the program suggests weightings using the following formula derived from information theory:

$$w_i = \log_2(1 - e_i)/f_i$$

where f_i is the average frequency of values in the field, which can be obtained by analysis of the source file or can be input by the user; and e_i, is the error rate for a field which can be set by the user after a trial linkage or based on previous knowledge of the data set.

Consider for example the variable sex: if f_i is 1/2 and the error rate is 1/20, w_i is

$$\log_2 [(1.0 - 0.05)/ 0.5] = 0.93$$

To clarify the use of the S function, consider Table 1, which shows a sample of records to be linked.

Comparing the target record with the first record in the source file, the s function gives the following values for each field: Surname = 1, Name = 1, Date of birth = 7/8, Sex = 1.

If we assign the weightings 4, 1, 4, and 1, to the fields, respectively, the value of the S function for the whole target record is 95%. Comparison of the target record with the other three records of the source file gives values for the S function of 93%, 87%, and 67%.

This algorithm only identifies similarities between the two records under comparison; it does not calculate the frequency of a given value in a field in the files to be linked. Only the average frequency of the values in the fields used for the linkage is

Surname	Name	Date of birth (dd-mm-yyyy)	Sex
Source file:			
Rossi	Paolo	10–10–1950	M
Rossi	Paoli	21–10–1950	M
Rosso	Paolo	21–10–1950	M
Fossa	Paola	10–10–1950	F
Target file:			
Rossi	Paolo	11–10–1950	M

Table 1
Example to illustrate use of the S function. Records in the source file that are candidates to be linked with the Paolo Rossi record in the target file

used, to set the weighting before starting the linkage process. In the above example the average frequency of surnames was used, not the frequency of the surname Rossi.

The user must impose a threshold (acceptability threshold) for the percentage correspondence between the source and target records used in the linking. If the acceptability threshold is set at 80%, then records of the target table that are 80% or more similar to a given source record are linked. An acceptability threshold of 100% is equivalent to fully deterministic linkage; lower thresholds are probabilistic linkages. The linkage procedure selects the record with the highest similarity exceeding the defined acceptability threshold; if none are found at or above the threshold, no linkage is made.

Using this approach, however, erroneous linkages can be made. For example if the program finds two possible links that differ by only a few percentage points, automatically linking the one with the highest percentage similarity carries a fairly high risk of error. To overcome this problem the program presents not only the most similar record, but also those that are slightly less similar, the percentage range (tolerance threshold) being user-selectable.

If more than one record has a similarity above the acceptability threshold but the percentage differences fall within the tolerance threshold, these records are placed in a temporary link list, so that the user can choose the correct one by visual inspection. If the record with the highest similarity falls above the acceptability threshold and there are no other records within the tolerance set by the tolerance threshold, the link is considered ‘stable’.

In the above example, if the acceptability threshold is set at 80% then the program potentially flags three records (those with S function values of 95%, 93%, and 87%). If the tolerance threshold is also set at 5 percentage points, the program chooses the records with S function values of 95% and 93%, and places them in the temporary link list, as the difference between them is only two percentage points.

As noted, the linkage process takes place in several sessions. The first session is typically a deterministic matching (100% similarity). The matches obtained in this session are flagged and not considered in future sessions. In subsequent sessions probable links are identified, progressively lowering the acceptability thresholds for items, and possibly adjusting the weightings for different linkage items. Each session results in flagged links. By opportunely setting the weightings, a session can select putative links in which the error is in one or more items only, and this facilitates manual checking of putative links.

The efficacy of linkage is measured by the number of false negative (records left unlinked) and false positive (records linked erroneously) [1]. In the EpiLink system the way to reduce false negatives is to lower the acceptability threshold, although this must increase the number of false positives. To decrease the false positives the user can increase the tolerance threshold, but this increases the amount of manual checking required to validate the putative links placed in a temporary list.

The basic way that EpiLink operates is to compare each record from the target table with all the records of the source table, selecting the most similar. However this is a time-consuming operation when the databases are large, and to increase speed it is possible for the user to set a particular field as a deterministic one. This means that a given record in the target table is only compared with those records in the source table for which the deterministically flagged field matches exactly. This operation greatly decreases the number of comparisons that have to be made and reduces the computer resources and time required to complete the linkage process. When setting this flag the string cleaning function noted above is not applied.

Table 2 Fields chosen for linkage, with error rate and average frequency

Field	Initial error rate	Average frequency
Surname	0.02	1/23310
Name	0.05	1/5951
Date of birth	0.03	1/23976
Sex	0.001	1/2

The Lombardy Cancer Registry

The LCR is a population-based cancer registry [16, 17], established in 1976, that covers the population of the Province of Varese, Region of Lombardy, northern Italy. The population is about 800,000. The information items collected by the registry for each patient (cancer case) are general demographic characteristics, cancer site, tumor histotype according to the Standard International Classification of Disease (WHO, 1990), and date of first diagnosis. The total number of personal data records contained in LCR is 88,410, covering cancer incidence from 1976 to 1996.

The LCR is now archived on the Oracle 8.0 database, running on an IBM RISC 6000 machine with 128M RAM. The operating system is Unix, but clients run on Windows NT. The completeness of the LCR personal data used for linkage in the present study is 100% for name, surname, sex and date of birth, 62% for place of birth.

The Social Security List

All those entitled to receive public health services in Italy (practically the whole population) are listed in Social Security files. A computerized Social Security List (SSL) is managed separately by every regional Health Authority; that for Lombardy was set up in 1987. The main uses of the SSL are to manage payments to general practitioners arising from consultation by patients [25] and to identify patients admitted to hospitals.

Hospitals and local health authorities have direct connections to the server containing the SSL database. We worked with a copy of the SSL file pertaining to all residents in the province of Varese during the period 1/1/1987 to 1/1/1997. The SSL copy was provided by the Lombardy Health authority and loaded into our server. A total of 1,021,846 records were present in the Varese SSL file. The completeness of the SSL personal data fields used in linkage procedure is 100% for name, surname, sex and date of birth, and 57% for place of birth.

Session number	Automated stable links	Cumulative % of stable links	Stable links checked manually and unlinked	Temp links	Temp links manually unlinked
1	14500	70.0	0	56	30
2–5	4946	93.8	6	34	23
Total	19446	93.8	6	90	53

Table 3 Cancer Registry linkage results by session number. Temp links are uncertain links that have to be checked manually. The first session was deterministic; the others were probabilistic

Record Linkage

All LCR records of cancer cases incident between 1/1/1988 and 31/12/1996, and ascertained as alive on 1/7/1999 were linked with the SSL. These records (total 20,724) formed the target table, whereas the SSL was the source table. Preliminary analysis of the data permitted us to decide which fields to use, the error rate, the discriminating power of each field and the linkage strategy. Table 2 shows the fields chosen for linkage (surname, name, date of birth, place of birth and sex) with source table error rates and average frequencies of specific values; we used these to estimate field weightings for use in the S function.

Test linkages performed during the preliminary analysis suggested an optimal linkage strategy as follows:

a) Run a deterministic linkage session to identify most links.
b) Run four probabilistic sessions with (i) acceptability threshold first set at 95% and progressively lowered to 90%; (ii) weightings progressively changed according to the error rate and discriminating power of the records left unlinked; and (iii) tolerance threshold always set at 5% to detect false positives.
c) Run additional sessions (in the present case six sessions) to assist manual linking of the small percentage of records not linked by the previous sessions.

To test the procedure, links produced by the first two sessions, where the acceptability threshold was ≥95%, were checked randomly; all links produced in subsequent sessions were also checked manually. Records not linked at the end of these sessions were searched manually.

Results

The results are shown in Table 3, according to session, session 1 being deterministic and sessions 2–5 probabilistic. Average linkage speed was about a thousand records per hour. A total of 19,446 (93.8%) of the 20,724 records were linked after the first five sessions, with 90 records placed in a temporary list; 53 of the latter were unlinked and 37 confirmed following visual inspection. Six stable links were also discarded after manual inspection of the results. Of the 1188 records left unlinked by the linkage process, 688 were linked by manual inspection, and 500 were left unlinked.

Table 4 shows data pertaining to the first five sessions. T+ refers to true links: records linked by the deterministic linkage and checked manually in a sample of random cases, or records linked probabilistic sessions and checked manually in each case. T- are the unlinked records. E+ refers to the records linked by EpiLink, E- refers to the records not linked by EpiLink. From these data the specificity of the process was 98.8% and the sensitivity 96.5%.

The 1188 records left unlinked by the first five sessions were linked manually with the aid of six additional linkage sessions. These results are shown in table 5.

Table 4 Results of the first five record linkage sessions. T– = unlinked records or false links, T+ = automatically or manually linked true links, E– = not automatically linked records, E+ = automatically linked records

	T–	T+	Totals
E–	500	688	1188
E+	6	19440	19446
Totals	506	20128	20634

Because we worked with a low acceptability threshold, some of the putative linkages made in these extra sessions were discarded following manual checking. However the process assisted the manual identification of true links. The unlinked records left over at the end of the process were persons (cancer cases) not present in the SSL.

Discussion

Our aim was to develop a record linkage system that was easy to use, portable and flexible. Ease of use was achieved by using a simple mathematical approach and designing a user-friendly GUI that allowed the user complete freedom to impose the parameters of linkage process required without the need to modify the source code or the statistical formulae. Flexibility was achieved by making it possible to define a linkage plan with a variable number of sessions, with the additional possibility of imposing different weightings and thresholds in each session. Our experience is that about three to four days are required to train a novice operator to use EpiLink, although this does not include deciding what linkage strategy to use. The program is portable, in that it can be installed on Windows-based PCs and servers, and also on Unix servers.

In the present study we investigated whether the simple mathematical approach implemented in EpiLink produced acceptable linkage in terms of sensitivity and specificity. The study was divided into two parts. In the first part (first five EpiLink sessions) the linkages were automatic, although links flagged as temporary by the program were checked manually. We also manually checked all links provided by the third to fifth sessions, to verify the accuracy of the linking process. As a result of this check, six false positives were found. This low figure suggests it is possible to link without manual checking, while still maintaining high specificity. If we had stopped after the fifth session we would have successfully linked 96.5% of the records automatically.

The remaining 3.5% of records were linked manually in the second part of the

Table 5
Results of additional linkage sessions used to assist manual linking

Session number	Stable links	Stable links checked manually and unlinked	Temp links	Temp links unlinked manually
6–11	691	50	33	25

study, although we still used EpiLink to suggest possible links, thereby making the manual work much easier.

The results of any record linkage depend on the quality of information on the databases. Independent assessment shows that the LCR is a high quality database and the overall success of the linkage process must in part be due to this.

EpiLink has been used in two other linkage projects. The first of these was to link hospital discharge data, pathology reports and mortality files (information sources for the LCR) with SSL records. These data sources are more heterogeneous and not as complete and accurate as the LCR database, having error rates on fields of up to 9%. Nevertheless we found that EpiLink had only slightly inferior sensitivity and specificity for these sources than in the present study [26]. This first linkage allowed us to perform various tests on weighting definition that led to the formalization expressed in the section 'Statistical Approach to Linking'.

The second was to link a follow-up cohort with LCR files [27-29]. In this case sensitivity and specificity were closely similar to those presented in this paper.

The nature of similarity function used has an important influence on linkage efficiency. We designed EpiLink so that it will be easy, in future versions, to change the similarity function, and to introduce a feature whereby the user can redefine the similarity function. Whether or not this will be necessary will depend on the results on ongoing tests using other databases.

Acknowledgments

The authors are grateful to D. Ward for his editorial assistance and to G. Grimoldi for his contribution in computer program development.

References

1. Leicester G, Goldacre M, Simmons H, Bettley G, Griffith M. Computerized linking of medical records: methodological guidelines. Journal of Epidemiology and Community Health 1993; 47: 316-9.
2. Howe GR. Use of computerized record linkage in cohort studies. Epidemiol Rev 1998; 20: 112-21.
3. Alsop JC, Langley JD. Determining first admissions in a hospital discharge file via record linkage. Meth Inform Med 1998; 37: 32-7.
4. The West of Scotland Coronary Prevention Study Group. Computerized record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. J Clin Epidemiol 1995; 48: 12: 1441-52.
5. Hole DJ, Clarke JA, Hawthorne VM, Murdoch RM. Cohort follow-up using computer linkage with routinely collected data. J Chronic Dis 1981; 34: 291-7.
6. Kato I, Toniolo P, Koenig KL, Kahn A, Schymura M, Zeleniuch-Jacquotte A. Comparison of active and cancer registry-based follow-up for breast cancer in a prospective cohort study. Am J Epidemiol 1999; 149: 372-8.
7. Van den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Hunen PMH. Development of a record linkage protocol for use in the Dutch cancer registry for epidemiological research. Int J Epidemiol 1990; 19: 553-8.
8. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D and the clinical epidemiology group from centres d'information et de soins de l'immunodeficience humaine (CISIH). Record linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990–1993. Int J Epidemiol 2000; 29: 168-74.
9. Camargo KR Jr, Coeli CM. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cad Saude Publica 2000; 16 (2): 439-47.
10. Roos LL, Wajda A. Record linkage strategies: Part 1. Estimating information and evaluating approaches. Meth Inform Med 1991; 30: 117-23.
11. Wajda A, Roos LL, Layefsky M, Singleton JA. Record linkage strategies: Part 2. Portable software and deterministic software. Meth Inform Med 1991; 30: 210-4.
12. Howe GR, Lindsay J. A generalized iterative record linkage computer system for use in medical follow-up studies. Comput Biomed Res 1981; 30: 327-40.
13. Fellegi I, Sunter A. A theory for record linkage. JASA 1969; 64: 1183-210.
14. MacLeod MCM, Bray CA, Kendrick SW, Cobbe SM. Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods. Comput Biomed Res 1998; 31: 257-70.
15. Newcombe HB, Fair ME, Lalonde P. The use of names for linking personal records. JASA 1992; 87: 1193-208.
16. Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J. Cancer incidence in five continents, Vol VII. IARC Sci Pub 1997; 566-7
17. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB. Cancer incidence in five continents, Vol VIII. IARC Sci Pub 2002; 386-7
18. Tao YC, Leibel RL. Identifying functional relationships among human genes by systematic analysis of biological literature. BMC Bioinformatics. 2002; 3 (1): 16.
19. Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proc AMIA Symp 2001; 319-23.
20. Lynch M. The similarity index and DNA fingerprinting. Mol Biol Evol 1990; 7 (5): 478-84.
21. Carranza L, Feoli E, Ganis P. Analysis of vegetation structural diversity by Burnaby's similarity index. Plant Ecol 1998; 138: 77-87.
22. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. J Am Stat Ass 1989; 84: 414-20.
23. Friedman C, Sideli R. Tolerating spelling errors during patient validation. Comput Biomed Res 1992 ; 25 (5): 486-509.
24. Sideli RV, Friedman C. Validating patient names in an integrated clinical information system. Proc Annu Symp Comput Appl Med Care. 1991; 588-92.
25. http://www.lispa.it/mercati/r_s_anagrafe.htm
26. Tittarelli A et al. Epilink, sistema di linkage del registro tumori di Varese. VII Riunione dell'Associazione Italiana registri Tumori. April 3-4, 2003.
27. Contiero P, Evangelista A, Tittarelli A, Del Sette D, Krogh V, Berrino F, Tagliabue G. Benign neoplasms: a follow-up study in Italy, 1993-1998. IARC Sci Publ 2002; 156: 537-9.
28. Evangelista A, Tagliabue G, Del Sette D, Tittarelli A, Contiero P, Krogh V, Crosignani P, Berrino F: Malignant tumour follow-up in Italy, 1993-1998. IARC Sci Publ 2002; 156: 535-6.
29. Tagliabue G, Evangelista A, Tittarelli A, Del Sette D, Contiero P, Crosignani P, Berrino F, Micheli A. Follow-up of the ORDET cohort, Lombardy Cancer Registry, 1987-1997. IARC Sci Publ 2002; 156: 67-8.

Correspondence to:
Paolo Contiero
Cancer Registry Division
Istituto Nazionale per lo Studio e la Cura dei Tumori
Via Venezian 1, 20133 Milan
Italy
E-mail: biomol@istitutotumori.mi.it