METHODOLOGIC ISSUES

Practical introduction to record linkage for injury research

D E Clark

Injury Prevention 2004;**10**:186–191. doi: 10.1136/ip.2003.004580

See end of article for author's affiliation

Correspondence to: Dr David E Clark, 887 Congress Street, Portland, ME 04102, USA; clarkd@mmc.org

The frequency of early fatality and the transient nature of emergency medical care mean that a single database will rarely suffice for population based injury research. Linking records from multiple data sources is therefore a promising method for injury surveillance or trauma system evaluation. The purpose of this article is to review the historical development of record linkage, provide a basic mathematical foundation, discuss some practical issues, and consider some ethical concerns.

Clerical or computer assisted deterministic record linkage methods may suffice for some applications, but probabilistic methods are particularly useful for larger studies. The probabilistic method attempts to simulate human reasoning by comparing each of several elements from the two records. The basic mathematical specifications are derived algebraically from fundamental concepts of probability, although the theory can be extended to include more advanced mathematics.

Probabilistic, deterministic, and clerical techniques may be combined in different ways depending upon the goal of the record linkage project. If a population parameter is being estimated for a purely statistical study, a completely probabilistic approach may be most efficient; for other applications, where the purpose is to make inferences about specific individuals based upon their data contained in two or more files, the need for a high positive predictive value would favor a deterministic method or a probabilistic method with careful clerical review. Whatever techniques are used, researchers must realize that the combination of data sources entails additional ethical obligations beyond the use of each source alone.

The frequency of early fatality and transient nature of trauma care mean that a single database will rarely suffice for population based injury research. Emergency Medical Services (EMS) and vital statistics data have been combined to determine outcomes of cardiac arrest,[1] and a similar approach is warranted for victims of severe injuries, who often die without entering an EMS system or require transfer from one hospital to another. Record linkage methods have therefore been advocated for studies of injury outcomes.[2–4]

For small applications, enough information is usually present to allow an accurate human judgment about whether a record from one source refers to the same case as a record from another source. However, this ''manual'' or ''clerical'' method becomes impractical with large numbers. A natural solution is to use a computer for ''matching'' or ''linking'' records; for simplicity, these terms will be used interchangeably, although some have reserved the former for the true relationship and the latter for the decision to accept that two records from different sources refer to the same case.[5]

The easiest computer assisted method is to link cases that have the same identification number, or some other element or group of elements that uniquely identify a given person or episode. This approach may be referred to as deterministic (or ''exact'' or ''all-or-none'') matching, and is effective in many cases. However, the necessary information may be absent, may have different formats or variations in different sources, or may be inaccurately entered or missing. Most of the interest in large scale record linkage research has therefore focused on probabilistic methods that simulate human pattern recognition when deciding that a record from one source refers to the same person or event as a record from another source. Despite the sophistication of some computer methods, the only ''gold standard'' for whether two records truly match is still the judgment of a human reviewer, and a combination of deterministic and probabilistic computer methods, along with human judgment, will often be the best approach.[6 7]

Much information about record linkage is available, and there have been previous reviews of the subject,[8–10] but references are in diverse locations mostly irrelevant to the field of injury control. The purpose of this article is to review the historical development of record linkage, provide a basic mathematical foundation, discuss some practical issues, and consider some ethical concerns arising from linking multiple databases. This is not an exhaustive review, but an outline of the main principles. More detailed information is available in the references, including proceedings of the United States Federal Committee on Statistical Methodology workshops from 1985 and 1997, which contain reprints of some classic articles.[11 12]

HISTORICAL BACKGROUND

The potential benefits of linking medical and vital statistics records were recognized even before computers became widely available.[13] By 1959, Newcombe and colleagues in Canada reported the ability to link such records contained on punch cards at a rate of about 10 per minute, and hoped that technology would increase this rate by a factor of at least 20.[14] Twenty years later, Newcombe was able to demonstrate the superiority of his computer methods over clerical methods for a large record linkage project, and the processing rate had increased to about 14 000 records per minute.[15] This processing speed has now also been vastly exceeded, along with further improvements in programming and data storage, and reductions in the size and cost of computer hardware.

Increasingly practical applications have been developed, largely based on Newcombe's methodology (see table 1).[6 12 14 16–30] Advances in probabilistic record linkage methodology have begun to see applications in injury epidemiology, which had previously been limited to cumbersome manual methods[31–34] or *ad hoc* deterministic procedures.[35–38] A ''road injury database'' was constructed in Western Australia by linking medical, police, and traffic data using the Canadian *Generalized Iterative Record Linkage System*[39]; later the methods were adapted to use *Automatch* and hospital

Abbreviations: CODES, Crash Outcome Data Evaluation System; EMS, Emergency Medical Services; NPV, negative predictive value; PPV, positive predictive value

trauma registries were added.[40 41] Hospital trauma registries in Maine were linked to death certificates, hospital discharge abstracts, and EMS data initially using a deterministic computer program,[37] but when *Automatch* became available, the latter was found easier to specify and generalize.[42] Hospital readmissions for injury in New Zealand were more easily identified by probabilistic methods, which allowed for more variables to be used for linkage than a deterministic method, even when some values were missing or erroneous.[43]

The Crash Outcome Data Evaluation System (CODES) project has been carried out in the past decade under the direction of the United States National Highway Traffic Safety Administration. This project has used probabilistic methods to link crash data with EMS, hospitalization, and death certificate data in several states. Many of the results from this project are available only as government documents,[28 44 45] although limited results from some states are accessible in the medical literature.[46–51] Despite some criticism,[52] CODES has produced a major increase in the experience and understanding of record linkage methods within the injury control community. Building on this experience, probabilistic linkage of other injury data has been successfully accomplished in Maine[53] and Utah.[54] The latest CODES projects have used new software, with an easier user interface.

PRACTICAL AND THEORETICAL ISSUES

Preprocessing

Although the mathematics and computer matching procedures are very interesting (see Appendix), the most difficult and time consuming part of a record linkage project is the preprocessing.[14 20 28] Missing or miscoded data, duplicate records, etc must be dealt with, and files must be put into standard formats for dates, locations, etc. Indeed, the success of record linkage is much more dependent on data quality than on software.

Special problems arise if names are available for linkage.[16] Although this may allow greater accuracy, the relative frequency or infrequency of different names, changes due to marriage, potential variations in spelling, nicknames, abbreviations, etc, greatly increase the complexity of matching. Numerous clever approaches can be programmed,[55] but human pattern recognition is particularly hard to replicate in this area.[5] In practice, confidentiality restrictions usually do not allow the use of names in large medical databases.

Table 1 Some historically notable software applications for probabilistic record linkage, along with published information about commercial availability

Probabilistic record linkage software system (year of initial publication)	Commercial availability
Oxford Record Linkage Study, *OX-LINK* (1967)[16–18]	Available (1997)[12]
Generalized Iterative Record Linkage System, *GIRLS*, *GRLS* (1981)[6 19 20]	Canadian $30,000 (2001)[20]
California Automated Mortality Linkage System, *CAMLIS* (1984)[21]	
LINKS, LinkPro (1991)[22 23]	
Leicester University Record Linkage System, *LYNX* (1994)[24]	
Automatch (1989)[25 26]	US $ 1995 (1994) US $195000 (2001)[20]
United States Census ''Winkler system'' (1997)[27]	Available, no cost (1997)[27]
CODES 2000, LinkSolv (2000)[28 29]	US $ 4,500 (2001)

Cost information not referenced is based upon the author's personal experience. Absence of an entry does not necessarily mean that these or other software systems are not available. Information about currently available record linkage software may be best obtained from internet queries, ideally including objective reviews.[30] *Caveat emptor*.

Stratification

With the probabilistic approach, the number of possible comparisons increases with the product of the file sizes, which becomes impractical when the files are large. The usual remedy is to stratify the procedure by restricting the comparisons to ''blocks'' or ''pockets'' of cases where one or more variables match exactly. This essentially utilizes a deterministic approach to assist the probabilistic method, but can be further modified by ''blocking'' sequentially using different variables.

Error rates

In epidemiologic studies using record linkage, the probability of falsely matching records that should not have been matched must be balanced against the probability of failing to match records that should have been matched. Records that are falsely matched (''mismatches'' or ''homonym errors'') will lead to misidentification of the outcome for specific cases as well as underestimation of the total number of cases; records that are falsely unmatched (''false non-matches'', ''erroneous non-matches'', ''failures to match'', or ''synonym errors'') will lead to missing data from one or the other source and overestimation of the total number of cases. The theoretical magnitude of these errors can be estimated algebraically after certain assumptions.[56]

The frequency of false positives and false negatives can be expressed in familiar terms of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV),[57] as depicted in table 2. In practice, the number of records truly unmatched is generally so large that specificity and NPV are not useful measurements. Furthermore, for any real application, it may be difficult to specify the ''gold standard'' against which matched or unmatched records are considered ''true'' or ''false''; because of this, a method to estimate the PPV based on the frequency of duplicate links has been proposed.[58]

Advanced mathematics

The basic mathematical concepts are described in the appendix; more advanced mathematical implications of automated record linkage have attracted the interest of some famous statisticians over the past half century,[59–63] and the status of current research in this area has been summarized by Winkler.[63] Fellegi and Sunter presented the formal theoretical structure for record linkage most often cited today,[64] and showed that the approach based on likelihood ratios (developed empirically by Newcombe[14]) was in accordance with classical hypothesis testing theory. Newcombe, in one of his last publications,[65] acknowledged that this approach can also be derived from Bayes' Theorem as in the Appendix.

Table 2 Possible outcomes for two records from different files

	Records truly are from the same person	Records truly are *not* from the same person
Records matched	Truly matched (TM)	Falsely matched (FM)
Records *not* matched	Falsely unmatched (FU)	Truly unmatched (TU)

At least theoretically, these lead to the concepts of sensitivity = TM/(TM+FU), specificity = TU/(TU+FM), PPV = TM/(TM+FM), NPV = TU/(TU+FU).

The mathematical approach to record linkage theory becomes more complicated when allowing for blocking or missing data.[61 64] Other theoretical complications result if one allows partial credit for ''near matches''.[16] For very large samples, sophisticated mathematical research has gone into the problem of minimizing the need for human review by estimating error using models based upon past human experience.[62]

Combining methods

A certain degree of ''art''[6 66] or ''fiddling around''[67] with the linkages will be necessary despite mathematical and technological advances. As mentioned above, the ''blocking'' strategy essentially combines deterministic and probabilistic approaches, and human review of preliminary results is certainly part of the validation of any computer program.

The best method for a given linkage project depends in part on its purpose.[68] If a population parameter is being estimated for a purely statistical study (for example, the effect of wearing safety belts on mortality), a completely probabilistic approach may be most efficient. To some extent, the numbers of records falsely matched and records falsely unmatched will cancel each other as match cut offs are varied.[56 57 69] The sensitivity and PPV can be estimated and used to develop confidence limits on the parameter estimate.[62]

For other applications, where the purpose is to make inferences about specific individuals based upon their data contained in two or more files (for example, flow of patients through multiple phases of care in a trauma system), a completely probabilistic approach would not be likely to give acceptable results. In this case, we must be quite sure that records from different sources truly refer to the same person (high PPV), and might favor a deterministic method.[70] However, probabilistic methods with careful clerical review may also be useful.[41]

A few studies have compared deterministic and probabilistic methods, using human review or artificially withheld identifying information as a ''gold standard''. Roos *et al*[23] and Jamieson *et al*[7] both found that a probabilistic method identified more matches; however the latter study found that only their deterministic method was free from falsely matched records and suggested that a combination of methods might be valuable. Gomatam *et al* have compared *Automatch* and a ''stepwise deterministic strategy'' using two files for which the true relationships were known from other data[71]; they also found that the sensitivity of the probabilistic method was better, but the PPV for the deterministic method was nearly 100%.

ETHICS OF RECORD LINKAGE

In 1946, the Chief of the United States Public Health Service's Office of Vital Statistics proposed that hospital, insurance, and other records for an individual be linked to provide statistical information for research.[13] Noting that registration systems developed in Europe under police authority ''will find disfavor in the United States'', he admired the decentralized Canadian system in which vital records were kept ''in their proper place, i.e., under the control of public health and statistical agencies'', but linked to a federal index with a personal identification number. As predicted, American concerns for privacy have led to a more cautious approach to record linkage than in Canada.[72 73]

Ethical issues were not on the program of a symposium on record linkage techniques held in 1985, but the editors recognized that this was an important area for further research.[11] Privacy issues were prominently addressed at a subsequent symposium held in 1997,[12] where some of the leading theoreticians carefully analyzed the social implications of their scientific work.[74 75] Citizens in the United States have a healthy mistrust of government, especially the huge federal bureaucracy. While there is broad support for the use of statistical information in public health research, this support depends upon the trust of the public that information accumulated for the general good will not be used against individual citizens.[68]

While the risk to patients may seem small, linkage of one database to another does create not only new generalizable knowledge about cause-and-effect relationships but also more specific knowledge about some individuals. Even if permission has been obtained to use separate databases, combining them adds a new level of obligation to the researcher and should only be done with the approval of the owners of the original data sets and an institutional review board; this does not necessarily mean that informed consent has to be obtained from each person whose records may be included (which would generally be impractical), but an impartial evaluation should show that the research is of good quality, that the risks are minimal, and that confidentiality of individual information will be maintained.[10 70 76]

In the United States, privacy considerations are even more important since the Health Insurance Portability and Accountability Act took effect in 2003; these regulations specifically prohibit the use of names, social security numbers, or vehicle identification numbers, and mandate informed consent for research using medical records unless waived by an institutional review board. The effect of this new legislation on clinical research is still being debated,[77 78] although it should be noted that special provisions are made for public health authorities, including ''an individual or entity acting under a grant of authority from or contract with such public agency''.[79]

ACKNOWLEDGEMENTS

Supported by Grant #R49/CCR119798-01 from the National Center for Injury Prevention and Control.

Author's affiliation

D E Clark, Center for Outcomes Research and Evaluation, Maine Medical Center and the Harvard Injury Control Research Center, Harvard School of Public Health

APPENDIX: MATHEMATICAL BACKGROUND

PROBABILITY AND ODDS

Let us define the *probability* of A, signified by P(A), to mean your degree of belief that A is true, expressed as a fraction ranging from slightly more than 0 (impossible) to slightly less than 1 (certain). We can define:

$$Odds\ (A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)} \qquad (1)$$

where $P(\bar{A})$ means the probability that A is not true. Note that when P(A) is very small, there is not much difference between the probability and the odds. Also, equation 1 can be rewritten as:

$$\mathrm{P(A)} = \frac{\mathrm{Odds\ (A)}}{1 + \mathrm{Odds\ (A)}}$$

JOINT AND CONDITIONAL PROBABILITY; INDEPENDENCE

Let us define the *joint probability* of A and B to be the probability that both A and B are true, written symbolically as P(A,B). Let us also define the *conditional probability* that A is true, given that B is true, written symbolically as:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

so also $$P(B|A) = \frac{P(B,A)}{P(A)}$$

and $$P(B|A)P(A) = P(A|B)P(B) = P(A,B) = P(B,A)$$

A and B may be defined as *independent* if:

$$P(A,B) = P(A)P(B) \quad (2)$$

Record linkage theory uses *mutual information* between two variables to assess independence.[80 81] If A and B are independent, their mutual information should be near zero.

BAYES' THEOREM; WEIGHTS

From the definition of conditional probability, we get:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

and $$P(\overline{B}|A) = \frac{P(A|\overline{B})P(\overline{B})}{P(A)}$$

Further algebra gives us:

$$\frac{P(B|A)}{P(\overline{B}|A)} = \frac{P(B)}{P(\overline{B})}\frac{P(A|B)}{P(A|\overline{B})} \quad (3)$$

which is the *odds ratio form of Bayes' Theorem*.[82 83] In equation 3,

$\frac{P(B|A)}{P(\overline{B}|A)}$ is called the *posterior odds*,

$\frac{P(B)}{P(\overline{B})}$ is called the *prior odds*, and

$\frac{P(A|B)}{P(A|\overline{B})}$ is called the *likelihood ratio*.

If we assume $A_1|B \ldots A_n|B$ are independent, then with repeated applications of Bayes' Theorem we get:

$$\frac{P(B|A_1,A_2\ldots A_n)}{P(\overline{B}|A_1,A_2\ldots A_n)} = \frac{P(B)}{P(\overline{B})}\frac{P(A_1|B)}{P(A_1|\overline{B})}\cdots\frac{P(A_n|B)}{P(A_n|\overline{B})} \quad (4)$$

Now, consider *P(B)* to mean "the probability that two records on different lists refer to the same person" and A_1 (for example) meaning "element 1 (age, sex, or whatever) is the same on both lists". Record linkage terminology refers to $P(A_1|B)$ as an *M probability* (the probability that element 1 is the same if the records truly match), and refers to $P(A_1|\overline{B})$ as a *U probability* (the probability that element 1 is the same, just by chance, when the records truly should be unmatched). If a given element is not the same on both lists, the likelihood ratio becomes (1−M)/(1−U).

Newcombe introduced logarithms in his explanation of record linkage methods, but later was concerned that they might be more confusing than helpful.[8] If we take the logarithm of both sides of equation 4, we obtain:

$$\log\frac{P(B|A_1,A_2\ldots A_n)}{P(\overline{B}|A_1,A_2\ldots A_n)} = \log\frac{P(B)}{P(\overline{B})} + \log\frac{P(A_1|B)}{P(A_1|\overline{B})} + \cdots + \log\frac{P(A_n|B)}{P(A_n|\overline{B})} \quad (5)$$

In other words, the *posterior log odds* (or *overall weight*) that the two records refer to the same person equals some constant (the *prior log odds*) plus the sum of the log likelihood ratios (*agreement* or *disagreement weights*) for each element.

ESTIMATING POSTERIOR PROBABILITIES

If we can demonstrate that our linking variables are nearly independent, then equation 4 will be approximately valid. If you have reason to believe (from other knowledge) that the number of matching records is about N_X, the number of records in file A is N_A, and the number of records in file B is N_B, then you can estimate the prior probability that a randomly selected record from file A matches with a randomly selected record from file B as:

$$P(match) = \frac{N_X}{N_A} \times \frac{1}{N_B} \quad (6)$$

This will generally be a very small number, so the prior odds will be similar. If you choose to work with logarithms, the log odds will be a very negative number, to which the agreement weights (minus the disagreement weights) will be added to obtain the posterior log odds (equation 5). With or without logarithms, by reversing our previous transformations (equation 4 and equation 1) you can obtain a posterior probability (or absolute probability) that two records match.

This approach can also evaluate the feasibility of a proposed record linkage project.[8 81 84] If the file sizes are known, and the number of expected links between them can be estimated, and the M and U probabilities can be approximated as described earlier, then equations 6, 4, and 1 can be used to see whether two truly matching records will be assigned a very high (for example, 95% or 99%) posterior probability of being correct.[84] If not, the project may be impractical.

HYPOTHETICAL EXAMPLE

Suppose you have the data presented in table 3, and need to decide which ambulance cases correspond to which emergency department cases. For this small number, you could match them using your own inspection and judgment (based on past experience with these kinds of patients and records),

Key points

- Record linkage methods are important for injury research or surveillance, because any single database is often inadequate.
- Computer assisted methods, simulating the human judgment that two records from different sources actually refer to the same event or person, are only superior when a large number of records must be processed.
- The basic mathematical theory behind probabilistic record linkage is not difficult to explain, and accords well with human intuition.
- Despite the speed and sophistication of modern record linkage software, deficiencies in data quality are the greatest obstacles to successful record linkage.
- Deterministic (exact) methods or careful human review of probabilistic results are required if record linkage is used to make inferences about individual cases.
- Linking two or more databases entails ethical obligations beyond the use of each separate database.

Table 3 Hypothetical data from 10 ambulance records and 20 emergency department records

Ambulance data							Emergency department data						
Case	Year	Day	Hosp	Birth year	Birthday	Sex	Case	Year	Day	Hosp	Birth year	Birthday	Sex
A01	01	Jan01	X	1950	Jan21	M	E01	01	Jan01	X	1950	Jan21	M
A02	01	Jan01	X	1950	May01	F	E02	01	Jan10	Z	1987	Jul17	M
A03	01	Jan10	Y	1975	Dec27		E03	01	Feb23	X	1992	Oct19	M
A04	01	Aug13	X	1977	Apr29	F	E04	01	Apr22	Y	1979	May09	M
A05	01	Sep12	Y	1980	Feb16	F	E05	01	May02	X	1929	Nov12	F
A06	01	Dec31	Z	1919	Sep16	M	E06	01	May23	Y	1964	Dec01	M
A07	02	Feb02	X	1924	Mar26	F	E07	01	Jun01	X	1950	May01	F
A08	02	Jun10	Y	1951	Mar29	M	E08	01	Aug14	X	1977	Apr29	F
A09	02	Aug06	Y	1953	Apr17		E09	01	Sep12	Y	1980	Feb16	F
A10	02	Sep21	Z	1956	Jun03	F	E10	01	Oct21	Y	1985	Mar12	M
							E11	02	Jan01	Z	1919	Sep16	M
							E12	02	Jan10	Y	1975	Dec27	F
							E13	02	Feb02	X	1924	Mar26	F
							E14	02	May16	X	1924	Oct12	M
							E15	02	Jun10	Y	1951	Mar29	M
							E16	02	Jul04	Z	1982	Jun12	M
							E17	02	Aug05	Y	1953	Apr17	M
							E18	02	Aug06	Y	2002	Apr17	F
							E19	02	Sep21	Z	1956	Jun03	F
							E20	02	Nov22	X	1917	May29	M

From past experience or calculations using the larger list, we might assume that: (1) About 90% of injured patients brought by ambulance will generate an emergency department record, so the prior probability of a match between records on the two lists might be about 0.045 (equation 6), and the prior odds (equation 1) about 0.047. (2) The chances that records truly from the same event will have the same admission year, admission date, hospital, birth year, birth date, and sex (the M probabilities) might be 0.99, 0.95, 0.99, 0.95, 0.99, and 0.95 respectively. (3) The chances that variables would match for randomly selected records (the U probabilities) might be 0.50 for admission year, 0.0027 (1/365) for admission date, 0.4 for hospital X or Y, 0.2 for hospital Z, 0.01 for birth year, 0.0027 for birth date, 0.60 for males, and 0.40 for females.

but let us employ the probabilistic method (and the assumptions given in table 3) to simulate this reasoning.

We can calculate posterior probabilities for each pair of records from the ambulance list and the emergency department list. The highest score would be for the pair A10-E19, with posterior odds (equation 4) of about:

$$0.047 * \frac{0.99}{0.50} * \frac{0.95}{0.0027} * \frac{0.99}{0.20} * \frac{0.95}{0.01} * \frac{0.99}{0.0027} * \frac{0.95}{0.40} = 1\,340\,000$$

and therefore a posterior probability (from equation 1) much greater than 0.9999. Also scoring very high would be the other exact matches A01-E01, A05-E09, A07-E13, and A08-E15. These are easily identified in a sorted list (like table 3), and would also be found by deterministic computer methods.

Scoring not quite so high would be those pairs where one or more elements did not match, for example A03-E12, with posterior odds calculated as about:

$$0.047 * \frac{1-0.99}{1-0.50} * \frac{0.95}{0.0027} * \frac{0.99}{0.40} * \frac{0.95}{0.01} * \frac{0.99}{0.0027} * \frac{1-0.95}{1-0.40} = 2376$$

and a posterior probability of 0.9996. Here, the admission year and sex were different, so that the likelihood ratio for these terms is (1−M)/(1−U). Our assumption that nearly all the ambulance records should match to an emergency department record resulted in a relatively large prior probability; the relatively large posterior probability thus reflects our judgment that the discrepancies are likely due to data entry errors. We would probably also accept the pairs A02-E07 and A04-E08, with posterior probabilities of 0.9990. Notice that it would be difficult for a human to find these probable matches on a longer list, and not simple to develop a deterministic computer strategy to identify them.

A09 presents a problem because it might be matched either to E17 or to E18, with posterior probabilities of 0.9805 or 0.9921, respectively. The pair A06-E19, with posterior probability of 0.9507, is also uncertain. Human judgment might help resolve such cases, but error is still possible. All other pairs of records not yet mentioned have much lower posterior probabilities, and would probably not be considered as potential matches.

This process could be made more sophisticated by allowing dates to differ by one day, separating month from day, penalizing missing data less than erroneous data, etc. We might also modify the M or U probabilities, or the prior odds, after reviewing initial results. This human/machine interaction should produce results that accord with human intuition, but can be expanded to manage thousands of records in each file.

REFERENCES

1 **Waien SA**. Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data. *Acad Emerg Med* 1997;**4**:1087–95.
2 **Weiss HB**, Dill SM, Garrison HG, *et al*. The potential of using billing data for emergency department injury surveillance. *Acad Emerg Med* 1997;**4**:282–7.
3 **Institute of Medicine Committee on Injury Prevention and Control**. *Reducing the burden of injury: advancing prevention and treatment*. Washington, DC: National Academy Press, 1999.
4 **Runge JW**. Linking data for injury control research. *Ann Emerg Med* 2000;**35**:613–15.
5 **Bell RM**, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. *Med Care* 1994;**32**:1004–18.
6 **Roos LL Jr**, Wajda A, Nicol JP. The art and science of record linkage: methods that work with few identifiers. *Comput Biol Med* 1986;**16**:45–57.
7 **Jamieson E**, Roberts J, Browne G. The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. *Methods Inf Med* 1995;**34**:371–7.
8 **Newcombe HB**. *Handbook of record linkage*. Oxford: Oxford University Press, 1988.
9 **Howe GR**. Use of computerized record linkage in cohort studies. *Epidemiol Rev* 1998;**20**:112–21.
10 **Neutel CI**, Johansen HL, Walop W. ''New data from old'': epidemiology and record-linkage. *Prog Food Nutr Sci* 1991;**15**:85–116.
11 **Federal Committee on Statistical Methodology**. *Record linkage techniques—1985: Proceedings of the workshop on exact matching methodologies*. Arlington, Virginia, 1985. Available at: www.fcsm.gov/working-papers/RLT_1985 (accessed on 30 May 2002).
12 **Federal Committee on Statistical Methodology**. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
13 **Dunn HL**. Record linkage. *Am J Public Health* 1946;**36**:1412–16.

14 **Newcombe HB**, Kennedy JM, Axford SJ, *et al*. Automatic linkage of vital records. *Science* 1959;**130**:954–9.
15 **Smith ME**, Newcombe HB. Accuracies of computer versus manual linkages of routine health records. *Methods Inf Med* 1979;**18**:89–97.
16 **Acheson ED**. *Medical record linkage*. London: Oxford University Press, 1967.
17 **Gill L**, Goldacre M, Simmons H, *et al*. Computerised linking of medical records: methodological guidelines. *J Epidemiol Community Health* 1993;**47**:316–19.
18 **Gill LE**. *OX-LINK: The Oxford medical record linkage system—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:15–33. Available at: www.fcsm.gov/working-papers/RTL_1997 (accessed on 30 May 2002).
19 **Howe GR**, Lindsay J. A generalized iterative record linkage computer system for use in medical follow-up studies. *Comput Biomed Res* 1981;**14**:327–340.
20 **Winkler WE**. *Record linkage software and methods for merging administrative lists*. Bureau of the Census Statistical Research Report Series No RR2001/03. Available at: www.census.gov/srd/papers/pdf/rr2001-03.pdf (accessed on 31 May 2002).
21 **Arellano MG**, Petersen GR, Petitti DB, *et al*. The California Automated Mortality Linkage System (CAMLIS). *Am J Public Health* 1984;**74**:1324–30.
22 **Wajda A**, Roos LL, Layefsky M, *et al*. Record linkage strategies: part II. Portable software and deterministic matching. *Methods Inf Med* 1991;**30**:210–14.
23 **Roos LL**, Walld R, Wajda A, *et al*. Record linkage strategies, outpatient procedures, and administrative data. *Med Care* 1996;**34**:570–82.
24 **Langley JD**, Botha JL. Use of record linkage techniques to maintain the Leicestershire Diabetes Register. *Comput Methods Programs Biomed* 1994;**41**:287–95.
25 **Jaro MA**. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J Am Statist Assoc* 1989;**84**:414–20.
26 **Jaro MA**. Probabilistic linkage of large public health data files. *Stat Med* 1995;**14**:491–8.
27 **Winkler WE**. Matching and record linkage. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:374–403. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
28 **US Department of Transportation, National Highway Traffic Safety Administration**. *Problems, solutions and recommendations for implementing CODES*. Washington, DC: NHTSA, 2001. Technical report DOT HS 809 200.
29 **Strategic Matching, Inc**. Available at: www.linksolv.org (accessed on 16 May 2003).
30 **Australian National University Data Mining Group**. Available at: http://datamining.anu.edu.au (accessed on 20 January 2004).
31 **Clark DE**, Katz MS, Campbell SM. Decreasing mortality and morbidity rates after the institution of a statewide burn program. *J Burn Care Rehabil* 1992;**13**:261–70.
32 **Copes WS**, Stark MM, Lawnick MM, *et al*. Linking data from national trauma and rehabilitation registries. *J Trauma* 1996;**40**:428–36.
33 **Esposito TJ**, Nania J, Maier RV. State trauma system evaluation: a unique and comprehensive approach. *Ann Emerg Med* 1992;**21**:351–7.
34 **Russell J**, Conroy C. Representativeness of deaths identified through the injury-at-work item on the death certificate: implications for surveillance. *Am J Public Health* 1991;**81**:1613–18.
35 **James HF**. Under-reporting of road traffic accidents. *Traffic Engineering+Control* 1991;**32**:574–73.
36 **Fife D**. Matching fatal accident reporting system cases with National Center for Health Statistics motor vehicle deaths. *Accid Anal Prev* 1989;**21**:79–83.
37 **Clark DE**. Development of a statewide trauma registry using multiple linked sources of data. *Proc Annu Symp Comput Appl Med Care* 1993:654–8.
38 **Meyer S**. Using Microsoft Access to perform exact record linkages. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:280–6. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
39 **Ferrante AM**, Rosman DL, Knuiman MW. The construction of a road injury database. *Accid Anal Prev* 1993;**25**:659–65.
40 **Lopez DG**, Rosman DL, Jelinek GA, *et al*. Complementing police road-crash records with trauma registry data—an initial evaluation. *Accid Anal Prev* 2000;**32**:771–7.
41 **Rosman DL**. The western Australian road injury database (1987–1996): Ten years of linked police, hospital and death records of road crashes and injuries. *Accid Anal Prev* 2001;**33**:81–8.
42 **Clark DE**, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proc Annu Symp Comput Appl Med Care* 1995:397–401.
43 **Alsop JC**, Langley JD. Determining first admissions in a hospital discharge file via record linkage. *Methods Inf Med* 1998;**37**:32–7.
44 **Johnson S**. Technical issues related to the probabilistic linkage of population-based crash and injury data. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:222–6. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
45 **Utter D**. Use of probabilistic linkage for an analysis of the effectiveness of safety belts and helmets. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:67–70. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
46 **Van Tuinen M**. Unsafe driving behaviors and hospitalization. *Mo Med* 1994;**91**:172–5.
47 **Farrell TM**, Sutton JE, Clark DE, *et al*. Moose-motor vehicle collisions: an increasing hazard in northern New England. *Arch Surg* 1996;**131**:377–81.
48 **Karlson TA**, Quade C, Florey M. Nonfatal motor vehicle crash injuries: Wisconsin's experience with linked data systems. *Wis Med J* 1996;**95**:301–4.
49 **Patterson L**, Weiss H, Schano P. Combining multiple data bases for outcomes assessment. *Am J Med Qual* 1996;**11**:S73–7.
50 **Moore M**. Comparison of young and adult driver crashes in Alaska using linked traffic crash and hospital data. *Alaska Med* 1997;**39**:95–102.
51 **Cook LJ**, Knight S, Olson LM, *et al*. Motor vehicle crash characteristics and medical outcomes among older drivers in Utah, 1992–1995. *Ann Emerg Med* 2000;**35**:585–91.
52 **Robertson LS**. *Injury epidemiology*. New York: Oxford University Press, 1998.
53 **Clark DE**, Hahn DR. Hospital trauma registries linked with population-based data. *J Trauma* 1999;**47**:448–54.
54 **Dean JM**, Vernon DD, Cook L, *et al*. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. *Ann Emerg Med* 2001;**37**:616–26.
55 **Fair ME**, Lalonde P, Newcombe HB. Application of exact ODDS for partial agreements of names in record linkage. *Comput Biomed Res* 1991;**24**:58–71.
56 **Brenner H**, Schmidtmann I. Determinants of homonym and synonym rates of record linkage in disease registration. *Methods Inf Med* 1996;**35**:19–24.
57 **Brenner H**, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med* 1997;**16**:2633–43.
58 **Blakely T**, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol* 2002;**31**:1246–52.
59 **Deming WE**, Glasser GJ. On the problem of matching lists by samples. *J Am Statist Assoc* 1959;**54**:403–15.
60 **DeGroot MH**, Feder PI, Goel PK. Matchmaking. *Ann Math Statist* 1971;**42**:578–93.
61 **Copas JB**, Hilton FJ. Record linkage: statistical models for matching computer records. *J R Statist Soc A* 1990;**153**:287–320.
62 **Belin TR**, Rubin DB. A method for calibrating false-match rates in record linkage. *J Am Statist Assoc* 1995;**90**:694–707.
63 **Winkler WE**. The state of record linkage and current research problems. Available atwww.census.gov/srd/papers/pdf/rs99-04.pdf (accessed on 12 February 2002).
64 **Fellegi IP**, Sunter AB. A theory for record linkage. *J Am Statist Assoc* 1969;**64**:1183–210.
65 **Newcombe HB**. Age-related bias in probabilistic death searches due to neglect of the ''prior likelihoods''. *Comput Biomed Res* 1995;**28**:87–99.
66 **Newcombe HB**. Strategy and art in automated death searches. *Am J Public Health* 1984;**74**:1302–3.
67 **Scheuren F**, Oh HL. Fiddling around with nonmatches and mismatches. *Record linkage techniques—1985: Proceedings of the workshop on exact matching methodologies*. Arlington, Virginia, 1985:79–88. Available at: www.fcsm.gov/working-papers/RLT_1985 (accessed on 30 May 2002).
68 **Kelman CW**, Bass AJ, Holman CD. Research use of linked health data—a best practice protocol. *Aust N Z J Public Health* 2002;**26**:251–5.
69 **Brenner H**, Schmidtmann I. Effects of record linkage errors on disease registration. *Methods Inf Med* 1998;**37**:69–74.
70 **Muse AG**, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. *Stat Med* 1995;**14**:499–509.
71 **Gomatam S**, Carter R, Ariet M, *et al*. An empirical comparison of record linkage procedures. *Stat Med* 2002;**21**:1485–96.
72 **Wallman KK**, Coffey JL. Sharing statistical information for statistical purposes. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:268–78. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
73 **Beebe GW**. Record linkage systems—Canada vs the United States. *Am J Public Health* 1980;**70**:1246–8.
74 **Fellegi IP**. Record linkage and public policy—a dynamic evolution. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:3–14. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
75 **Scheuren F**. Linking health records: human rights concerns. *Record linkage techniques—1997: Proceedings of an international workshop and exposition*. Arlington, Virginia, 1997:404–26. Available at: www.fcsm.gov/working-papers/RLT_1997 (accessed on 30 May 2002).
76 **Breen KJ**. Consent for the linkage of data for public health research: is it (or should it be) an absolute pre-requisite? *Aust N Z J Public Health* 2001;**25**:423–5.
77 **Annas GJ**. Medical privacy and medical research—judging the new federal regulations. *N Engl J Med* 2002;**346**:216–20.
78 **Califf RM**, Muhlbaier LH. Health Insurance Portability and Accountability Act (HIPAA): must there be a trade-off between privacy and quality of health care, or can we advance both? *Circulation* 2003;**108**:915–18.
79 **Guidance from CDC and the US Department of Health and Human Services**. HIPAA privacy rule and public health. *MMWR Morb Mortal Wkly Rep* 2003;**52**(suppl May 2):1–20.
80 **Cover TM**, Thomas JA. *Elements of information theory*. New York: John Wiley, 1991.
81 **Roos LL**, Wajda A. Record linkage strategies. Part I: estimating information and evaluating approaches. *Methods Inf Med* 1991;**30**:117–23.
82 **O'Hagan A**. *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*. New York: Halsted Press, 1994.
83 **Clarke JR**. A scientific approach to surgical reasoning: II. Probability revision—odds ratios, likelihood ratios, and Bayes' theorem. *Theor Surg* 1990;**5**:206–10.
84 **Cook LJ**, Olson LM, Dean JM. Probabilistic record linkage: relationships between file sizes, identifiers and match weights. *Methods Inf Med* 2001;**40**:196–203.