METHODOLOGIC ISSUES

Practical introduction to record linkage for injury research

D E Clark

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 the transient nature of emergency medical 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, 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.

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.

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.

Much information about record linkage is available, and there have been previous reviews of the subject, 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.

HISTORICAL BACKGROUND

The potential benefits of linking medical and vital statistics records were recognized even before computers became widely available. 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. 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. 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). Advances in probabilistic record linkage methodology have begun to see applications in injury epidemiology, which had previously been limited to cumbersome manual methods or ad hoc deterministic procedures. A “road injury database” was constructed in Western Australia by linking medical, police, and traffic data using the Canadian Generalized Iterative Record Linkage System; later the methods were adapted to use Automatch and hospital...
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,46–51 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 Maine53 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.44 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.14 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,53 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.

### 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.54

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),55 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.56

### 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.59–61 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)65 was in accordance with classical hypothesis testing theory. Newcombe, in one of his last publications,66 acknowledged that this approach can also be derived from Bayes’ Theorem as in the Appendix.

### Table 1

<table>
<thead>
<tr>
<th>Probabilistic record linkage software system (year of initial publication)</th>
<th>Commercial availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINKS, LinkPro (1991)</td>
<td></td>
</tr>
<tr>
<td>Leicester University Record Linkage System, LYNX (1994)</td>
<td></td>
</tr>
</tbody>
</table>

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.54 Caveat emptor.

### 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) | False positives (FP) |
| Records not matched | False negatives (FN) | Truly unmatched (TU) |

At least theoretically, these lead to the concepts of sensitivity = TM/(TM+FN), specificity = TU/(TU+FP), PPV = TM/(TM+FP), NPV = TU/(TU+FN).
The mathematical approach to record linkage theory becomes more complicated when allowing for blocking or missing data. Other theoretical complications result if one allows partial credit for “near matches”. 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.

Combining methods
A certain degree of “art” or “fiddling around” 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. 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. The sensitivity and PPV can be estimated and used to develop confidence limits on the parameter estimate.

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. However, probabilistic methods with careful clerical review may also be useful.

A few studies have compared deterministic and probabilistic methods, using human review or artificially withheld identifying information as a “gold standard”. Roos et al and Jamieson et al 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; 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. 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.

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. Privacy issues were prominently addressed at a subsequent symposium held in 1997, where some of the leading theoreticians carefully analyzed the social implications of their scientific work. 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.

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.

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, 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.”

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:

\[
\text{Odds}(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)} \tag{1}
\]

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:

\[
P(A) = \frac{\text{Odds}(A)}{1 + \text{Odds}(A)}
\]

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:
A and B may be defined as independent if:

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

so also

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

and

\[ P(A|B)P(B) + P(B|A)P(A) = P(A,B) = P(A,B) \]

A and B may be defined as independent if:

\[ P(A,B) = P(A)P(B) \] (2)

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

BAYES’ THEOREM; WEIGHTS

From the definition of conditional probability, we get:

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

and

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

Further algebra gives us:

\[ \frac{P(B|A)}{P(A|B)} = \frac{P(A,B)}{P(B)} \cdot \frac{P(B)}{P(A,B)} \] (3)

which is the odds ratio form of Bayes’ Theorem. In equation 3,

\[ \frac{P(B|A)}{P(A|B)} \]

is called the posterior odds,

\[ \frac{P(B)}{P(A)} \]

is called the prior odds, and

\[ \frac{P(A|B)}{P(A)} \]

is called the likelihood ratio.

If we assume \( A_1|B...A_n|B \) are independent, then with repeated applications of Bayes’ Theorem we get:

\[ \frac{P(B|A_1,A_2...A_n)}{P(B|A_1,A_2...A_n)} = \frac{P(A_1|B)P(A_2|B)...P(A_n|B)}{P(A_1|B)P(A_2|B)...P(A_n|B)} \]

(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 \( P(A_1|B) \) 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. If we take the logarithm of both sides of equation 4, we obtain:

\[ \log \frac{P(B|A_1,A_2...A_n)}{P(B|A_1,A_2...A_n)} = \log \frac{P(B)}{P(B)} + \log \frac{P(A_1|B)}{P(A_1|B)} + \cdots + \log \frac{P(A_n|B)}{P(A_n|B)} \]

(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:

\[ P(match) = \frac{N_X}{N_A} \times \frac{1}{N_B} \] (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. 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. 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.
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:

\[
\begin{align*}
0.047 & \times \frac{0.99}{0.50} \times \frac{0.95}{0.0027} \times \frac{0.99}{0.20} \times \frac{0.95}{0.01} \times \frac{0.99}{0.0027} \\
& \times \frac{0.95}{0.40} = 1340000
\end{align*}
\]

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:

\[
\begin{align*}
0.047 & \times \frac{1 - 0.99}{0.95} \times \frac{0.95}{0.0027} \times \frac{0.99}{0.20} \times \frac{1 - 0.95}{0.01} \\
& \times \frac{0.99}{1 - 0.40} = 2376
\end{align*}
\]

and a posterior probability of 0.9996. Here, the admission year and sex were different, so that the probability for a match would be less than 1. Our assumption that nearly all 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.

The potential of using billing data for emergency medical services cohort studies using existing data. Acad Emerg Med 1997;4:1087–95.


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

<table>
<thead>
<tr>
<th>Ambulance data</th>
<th>Emergency department data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>Year</td>
</tr>
<tr>
<td>A01</td>
<td>01</td>
</tr>
<tr>
<td>A02</td>
<td>01</td>
</tr>
<tr>
<td>A03</td>
<td>01</td>
</tr>
<tr>
<td>A04</td>
<td>01</td>
</tr>
<tr>
<td>A05</td>
<td>01</td>
</tr>
<tr>
<td>A06</td>
<td>01</td>
</tr>
<tr>
<td>A07</td>
<td>02</td>
</tr>
<tr>
<td>A08</td>
<td>02</td>
</tr>
<tr>
<td>A09</td>
<td>02</td>
</tr>
<tr>
<td>A10</td>
<td>02</td>
</tr>
</tbody>
</table>

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 4), 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.

REFERENCES


63. 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.


74. Langley JD, Botha JL. Use of record linkage techniques to maintain the Leicestershire Diabetes Register. Comput Methods Programs Biomed 1994;41:287–95.


Practical introduction to record linkage for injury research

D E Clark

*Inj Prev* 2004 10: 186-191
doi: 10.1136/ip.2003.004580

Updated information and services can be found at:
http://injuryprevention.bmj.com/content/10/3/186

These include:

**References**

This article cites 58 articles, 5 of which you can access for free at:
http://injuryprevention.bmj.com/content/10/3/186#BIBL

**Email alerting service**

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Notes**

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/