Probabilistic Record Linkage of Infection Records and Death Registrations: A Tool to Strengthen Surveillance

Nicola Potz, Health Protection Agency
David Powell, Health Protection Agency
Theresa L. Lamagni, Health Protection Agency
Richard Pebody, Health Protection Agency
David Bridger, Health Protection Agency
Georgia Duckworth, Health Protection Agency

Recommended Citation:
Potz, Nicola; Powell, David; Lamagni, Theresa L.; Pebody, Richard; Bridger, David; and Duckworth, Georgia (2010) "Probabilistic Record Linkage of Infection Records and Death Registrations: A Tool to Strengthen Surveillance," Statistical Communications in Infectious Diseases: Vol. 2: Iss. 1, Article 6.
DOI: 10.2202/1948-4690.1015
Available at: http://www.bepress.com/scid/vol2/iss1/art6

©2010 Berkeley Electronic Press. All rights reserved.
Probabilistic Record Linkage of Infection Records and Death Registrations: A Tool to Strengthen Surveillance

Nicola Potz, David Powell, Theresa L. Lamagni, Richard Pebody, David Bridger, and Georgia Duckworth

Abstract

An important element for many infectious disease surveillance programmes is their capacity to monitor not only the incidence of infection, but also the associated mortality. The ability to monitor post-infection mortality is dependent on outcome information being collected through the surveillance reports, or on infections being precisely specified on death certificates. For many infectious diseases, neither of these sources provides a reliable source of this information, so a method for linking infection and death registration data is needed. Given that surveillance data often lacks a unique patient identifier, a probabilistic record linkage method was developed to reliably bring together large-scale data sources to identify deaths following infection. The method was developed using Streptococcus pneumonia infection records but with wider applicability to other infectious disease surveillance programmes. Evaluation of the mechanism was undertaken by tracing patients through a central health service database. Results of the evaluation showed a positive predictive value of 97.7-99.8% for correctly identifying deaths following infection, and a negative predictive value of 90.2-98.0%. The successful application of probabilistic matching to link infections and death registrations paves the way for a new era in infectious disease surveillance in the UK, with its potential application to augment a wide array of ongoing surveillance programmes with information on patient outcome.

KEYWORDS: disease surveillance, probability, infection, mortality

Author Notes: We extend our thanks to the following additional members of the Project Board: Andrew Chronias, HPA; Clare Griffiths, Office for National Statistics (ONS); Nourieh Hoveyda, ONS; Cleo Rooney, ONS; Levin Wheller, ONS; Jennie Wilson, HPA and the following members of the Steering Group: Joy Dobbs, ONS; Peter Goldblatt, ONS; Andrew Phillips, University College London; Sarah Scobie, National Patient Safety Agency; Robert Spencer, Hospital Infection Society. We also thank Robert George (HPA Respiratory and Systemic Infection Laboratory), Elizabeth Miller, Pauline Kaye and Usha Gungabissoon (HPA Immunisation Department) for use of the enhanced S pneumonia surveillance dataset, Natasha Crowcroft and Julia Granerod (HPA Immunisation Department) for their conceptual input into the original project proposal, Nick Andrews (HPA Statistics, Modelling and Bioinformatics Department) for the sample size calculations and our microbiology colleagues in laboratories across England and
Wales for their continued reporting of infection data. The study was funded by the Department of Health for England.
INTRODUCTION

Information on outcome in patients diagnosed with a serious infection, or other life-threatening event, is an important element of any analysis estimating disease burden and its associated cost. The quality of information collected by routine national infectious disease surveillance programmes on mortality following infection is variable; many important programmes do not routinely collect information concerning mortality. Use of death registrations for such monitoring is also limited, with the quality of reporting of infections on death certificates varying between certifiers (Crowcroft and Catchpole, 2002; Health Protection Agency, 2007). One possible means to circumvent this problem would be to directly link infection and mortality datasets to identify individuals who have died following an infection, increasing the interpretive value of both sets of data. For the UK, this would be consistent with the current shift towards increased data sharing by public agencies (Department for Constitutional Affairs, 2003; UK Clinical Research Collaboration and Wellcome Trust, 2007). The method chosen to link records should depend on the quality and availability of personal identifiers within the records.

Record linkage work is often undertaken in an ad hoc manner, with the researcher making an empirical decision as to which combination of matched variables constitute matched records (Seagroatt and Goldacre, 1994; Wyllie et al., 2005; Wyllie et al., 2006). Records may be linked in two ways: ‘deterministically’ or ‘probabilistically’ (Newcombe, 1988). Deterministic or exact record linkage requires a universally available, unique variable (or combination of variables) to be available in both files being linked. Probabilistic record linkage, however, is suitable when no unique identifier is available, data are ‘noisy’ (i.e. contain random errors), and/or key variables have missing values. The method compares records from two files in pairs and applies probability-based criteria to determine the strength of the match between the record pairs.

The probabilistic record linkage method detailed in this paper was specifically developed to link infection records to death registrations in fulfilment of a Department of Health funded project examining mortality post meticillin-resistant Staphylococcus aureus (MRSA) infection diagnosed in patients in England (HPA, 2005). Matches between two data records are based on a comparison of multiple data fields within the two files. Identifiers need not match exactly; each identifier is assigned a weight (which is derived from the data itself) and the cumulative weights for all identifiers being compared yielding a score which can be used to classify records as matches, non-matches or possible matches. The successful linkage of infection and mortality records described here has since allowed an analysis of death following MRSA bacteraemia to be undertaken, which could not have been previously undertaken using either dataset alone (Lamagni et al., 2010). The details of the linkage method and its evaluation

Published by Berkeley Electronic Press, 2010
are described in this paper, with discussion of the possible limitations and applications for public health.

METHODS

Data sources

The probabilistic record linkage method was developed using records of invasive (bloodstream and cerebrospinal fluid) Streptococcus pneumoniae infections collected as part of the national surveillance of laboratory-confirmed infections collated by the Health Protection Agency (HPA), a system which also captures data on laboratory diagnoses of MRSA infection. S pneumoniae records were used as they were similar in all respects to the MRSA records in terms of the fields available for record linkage (unique patient id - National Health Service number, date of birth, sex and postcode) but additionally contained the patient’s name as opposed to an anonymised ‘soundex’ code derived from the patient’s surname which the majority of infectious diseases under surveillance are restricted to. Collection of patient identifiers by the HPA, including surnames, is regulated by the National Information Governance Board Ethics and Confidentiality Committee (formerly the Patient Information Advisory Group) and is restricted to specified infections where this information is needed for prevention and control purposes. Availability of the patient surname within the S pneumoniae records was essential to the evaluation of the linkage method as it allowed these records to be processed by the NHS Central Register (NHSCR) Tracing Service as an alternative method of obtaining vital outcome, the results of which were used as a comparison with the probabilistic linkage method for evaluative purposes.

All 3256 records of invasive S pneumoniae infection diagnoses reported by English laboratories in the period 1 July 2003 – 30 June 2004 were extracted and stratified into four age groups (18-64 years, 65-74 years, 75-79 years, 80+ years). From these, 1252 records were randomly sampled within the strata based on a sample size calculation that indicated a sample size of 1250 was needed to yield a 95% confidence interval of 8.3-11.7% around an estimated case fatality of 10%. All surnames within the S pneumoniae dataset were soundexed prior to record linkage to reproduce the fields available for the majority of infectious disease surveillance data, allowing the method developed to have wide applicability. These records were then paired using the method detailed below with all death registrations in England and Wales (obtained from the Office for National Statistics) during the period 1 July 2003 – 31 March 2005. In parallel, the same set of records was sent to the National Health Service Central Register (NHSCR) for tracing.
Probabilistic record linkage

Prior to record linkage, coding of all variables was standardized across the infection and death record datasets (Table 1). NHS numbers were checked for their validity (Department of Health, 2002). Where any data items were missing or invalid, the field entry was left null. The variables used for record are given in Table 1, with date of birth split into three variables - day, month, year - giving a total of eight matching variables.

Table 1 Availability of variables for record linkage in the invasive Streptococcus pneumoniae infection and death registrations datasets and their coding formats

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coding format</th>
<th>Availability of variable (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Infection records</td>
</tr>
<tr>
<td>NHS number</td>
<td>10-digit (validity checked)</td>
<td>29.6</td>
</tr>
<tr>
<td>Forename initial</td>
<td>Single letter (A–Z)</td>
<td>96.8</td>
</tr>
<tr>
<td>Surname soundex</td>
<td>Letter + 3 digits</td>
<td>97.7</td>
</tr>
<tr>
<td>Sex</td>
<td>1 (male), 2 (female)</td>
<td>97.9</td>
</tr>
<tr>
<td>Date of birth</td>
<td>DD/MM/YYYY</td>
<td>99.0</td>
</tr>
<tr>
<td>Postcode</td>
<td>Letter prefix only</td>
<td>51.4</td>
</tr>
</tbody>
</table>

Blocking of datasets Both files were broken down into smaller manageable ‘blocks’, with corresponding blocks within each file then linked to each other, rather than all records to all records. This reduced the number of pairs of records being compared, thereby reducing the amount of computational power and time for linkage required. Blocks were created using two different variables, soundex and date of birth. Both variables were well and reliably completed and had good discriminatory power, containing a large number of different values. Results from the two independent rounds of record linkage were then combined by merging together and then removing all duplicate records. Undertaking blocking using two different variables meant that records which were incorrectly blocked by one variable due to data entry error could still be linked to any potential matching record using the other variable. Records lacking both soundex and date of birth, and therefore not blocked by either, were not linked and would have insufficient identifiers to be reliably linked by any method; no records without soundex and date of birth contained a unique identifier such as NHS number.
**Weighting of record pairs** ‘Weight scores’ were assigned to all variables based on their frequencies within the death registration file and the matched NHSCR traced dataset (Blakely and Salmond, 2002). The frequencies of the variables within the death registration file were used to calculate ‘u’ probabilities – the probability that a variable will agree purely by chance within a pair of records not belonging to the same individual. u probabilities were calculated for all values within each variable. The ‘m’ probability, the probability that a given variable will agree when the pair of records is a match, was calculated for all variables using the matched NHSCR traced dataset, and was calculated for each variable as a whole, rather than each value within each variable, due to the smaller size of the dataset and consequent lack of representation of all variables. These u and m probabilities were then used to calculate frequency ratios and (dis)agreement weight scores, with agreement between variables resulting in a positive weight score and disagreement resulting in a negative weight score.

An example of the calculation of weight scores is shown in Table 2 using an initial of ‘A’ as an example, where the probability of the initial being correct in the matched dataset is 0.99 and the proportion of death registrations where the forename initial is ‘A’ is 0.079. The total weight score for a linked pair of records is the sum of the weight scores of (dis)agreement for all variables. If either record has a blank field for a variable, score that variable as ‘0’. The total weight score for a pair of records under comparison will be a large positive number if all of variables agree or a large negative number if all of the weights disagree. Total weight scores can be placed on a scale of weight distribution and subsequent match likelihood.

**Table 2** Calculation of agreement and disagreement frequency ratios and weight scores for the matching variable value ‘forename initial A’

<table>
<thead>
<tr>
<th>Comparison outcome</th>
<th>Proportion</th>
<th>Frequency ratio</th>
<th>Weight score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Matched pairs</td>
<td>Unmatched pairs</td>
<td></td>
</tr>
<tr>
<td>Agreement</td>
<td>0.993 (m)</td>
<td>0.079 (u)</td>
<td>12.57 (m/u)</td>
</tr>
<tr>
<td>Disagreement</td>
<td>0.007 (1-m)</td>
<td>0.921 (1-u)</td>
<td>0.008 [(1-m)/(1-u)]</td>
</tr>
</tbody>
</table>

http://www.bepress.com/scid/vol2/iss1/art6
DOI: 10.2202/1948-4690.1015
Threshold setting Upper and lower thresholds for the total weight scores were determined above and below which record pairs were accepted as matches and mismatches, respectively. Rounds of threshold rule setting and evaluation were undertaken to minimize missed matches (Type I errors) and false positives (Type II errors), and the following rules derived:

- The upper threshold was set at the lowest total weight score at which two records matched on seven of the eight matching variables, in this case a score of $\geq 31$ indicates a matched record pair.

- In order to be able to distinguish whether a comparison pair was a match or not, it was considered that a minimum of five variables e.g. full date of birth, soundex and sex would need to match. The lower threshold was therefore set at a score of one below the lowest total weight score at which five variables matched, with or without a mismatch in a maximum of one variable; in this case a score of $\leq 8$ indicates a mismatched record pair. Allowing a mismatch in one variable allows for a low level of data error.

Manual checking The record pairs falling between the two thresholds were manually checked. Record pairs for checking were firstly restricted to those matching on at least five variables with only one mismatching variable permitted, with all others regarded as mismatches. The mismatching variable (if present) was then manually checked, and the location of the hospital/laboratory reporting the infection compared to the place of death. A decision as to whether the record pair were a matched pair was then made, taking into account such factors as possible typographical errors such as swapping of digits, alternative forenames and surnames given within the death data, and mismatching postcodes in high population density areas where areas are small and movement into a nursing home, for example, may change someone’s postcode prefix. Additionally all records with multiple matches where one of the matched records fell beyond the lower threshold were subject to a manual check. Matched pairs identified through manual checking were then added to those already accepted due to falling above the upper threshold to produce the final matched infection-death dataset.

Method evaluation The resultant dataset from the probabilistic record linkage of invasive S. pneumoniae infection records and death registrations was compared to that obtained when the exact same set of infection records were traced for vital outcome through a search of the national patient database of individuals receiving medical care by the NHSCR Tracing Service. The sensitivity, specificity, positive predictive value and negative predictive value of the probabilistic record linkage method were calculated using the NHSCR tracing results as a ‘gold standard’.
RESULTS

When plotted, the distribution of record pairs by total weight score fell into three main groups. The three populations evident were non-matches, matches where at least one of the records lacked NHS number and matches where both records had matching NHS numbers, as a matching NHS number contributed a score of approximately 20 to the total weight score (Figure 1). When probabilistic record linkage results were overlaid with the results obtained when the same *Streptococcus pneumoniae* infection records were traced for vital outcome through the NHSCR Tracing Service, the upper end of the distribution was an exact match with most discrepancies between the two methods occurring in the manual checking area (total weight scores 9-30).

**Fig. 1** Distribution of total weight scores of record pairs and comparison with deaths identified through NHSCR tracing

◊ Probabilistically-linked *Streptococcus pneumoniae* record pairs
+ Matched infection and death records identified by the NHSCR Tracing Service

The probability of a probabilistically matched record being a true match as assessed by the tracing results increased steeply beyond a total weight of 19, fluctuating around certainty from 21 to 38 after which all linked records were a true match (Figure 2).
Fig. 2 Probability of true match according to distribution of total weight scores derived from probabilistic matching

A total of 476 deaths were identified by probabilistic record linkage; 327 (69%) through automatic acceptance and 149 through manual acceptance. NHSCR Tracing identified 480 deaths. Seventy records could not be traced through NHSCR (Table 3).

Errors in the identification of individuals who had died following an infection occurred by both methods. Of the 15 individuals found to have died by NHSCR Tracing but not by probabilistic record linkage (“false negative”), ten were individuals whose records held too few data variables to be linked by record linkage but had very distinctive names of value to the tracing process. Two NHSCR matches were missed by probabilistic record linkage due to more than one mismatching variable. The three remaining record pairs appear to have been matched in error by NHSCR Tracing for the following reasons: 1) an identical NHS number was wrongly given for two individuals, possibly a grandfather and grandson, 2) a 01/01/yyyy date of birth was given in the death record (given if the exact date of birth is not known) and 3) two individuals with differing days of birth and middle initial were accepted as a match by NHSCR, but upon evaluation the matched individual was shown to have died four years prior to the date of the infection record.
### Table 3 Comparison of the developed probabilistic record linkage method with NHSCR Tracing in the identification of deaths following invasive *S pneumoniae* infection

<table>
<thead>
<tr>
<th>Probabilistic record linkage</th>
<th>Matched to a death record</th>
<th>Not matched to a death record</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traced - dead</strong></td>
<td>465</td>
<td>15</td>
<td>480</td>
</tr>
<tr>
<td><strong>NHSCR Tracing</strong></td>
<td>464</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td><strong>Not traced</strong></td>
<td>10</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>476</td>
<td>767</td>
<td>1243</td>
</tr>
</tbody>
</table>

- **Sensitivity** 96.9% (465/480)
- **Specificity** 99.9% (692/693)

One record pair (false positive) was matched in error by probabilistic record linkage, relating to two individuals with the same date of birth, soundex, initial, sex and location (surnames in fact differed). Ten record matches were made by probabilistic record linkage but the individuals could not be traced by the NHSCR. None of these records contained an NHS number, however, all contained a full date of birth, soundex and/or forename initial and sex.

When evaluating the probabilistic record linkage method against the NHSCR Tracing, it is not known whether the ‘not traced’ NHSCR cases are alive or dead. The best-case scenario would be that they are all allocated correctly by probabilistic record linkage. The worst-case scenario would be that record linkage allocated them all incorrectly. The positive predictive value, i.e. the probability that someone identified as dead by probabilistic record linkage had truly died, therefore lies in the range 97.7-99.8% (Table 3). Likewise, the negative predictive value, i.e. the probability that a person not identified as dead by probabilistic record linkage (either alive or dead but unlinked) was truly alive,
was 90.2-98.0%. The sensitivity and specificity of the probabilistic record linkage method were shown to be 96.9% and 99.9%, respectively.

**DISCUSSION**

The probabilistic record linkage method compared favourably to the best available alternative - identification of patients using the National Health Service Central Register Tracing service - for identifying individuals who had died following an invasive *S. pneumoniae* infection using personal identifiers available on routine infection surveillance data. Although some errors were made by the probabilistic method, these were few resulting in a level of accuracy acceptable for surveillance purposes. Furthermore, the method developed holds many advantages over the NHSCR tracing service for identification of mortality following infection, namely its wider applicability, increased accuracy where errors may be present within the data and lower cost.

The probabilistic record linkage method was developed to utilise variables commonly available in surveillance datasets in the UK, allowing for the use of anonymised/truncated fields such as surname soundex and forename initial, which reduces the levels of patient identifiable information being required. As is the case in other countries, flow of patient identifiers for routine surveillance purposes is restricted to infectious diseases where a vital public health function requires this to happen, currently standing at 28 infectious pathogens in England. This means that the method has wider applicability than NHSCR for identifying patient outcome as NHSCR cannot use soundex and would only be of routine use where surname data or NHS number are available. This is supported by the findings that even after removal of records lacking key variable combinations such as NHS number/postcode/date of birth and NHS number/postcode/forename initial, only 91% of MRSA records, which lack surname data, submitted for NHSCR tracing were traceable, with the majority (65%) being exact matched on NHS number alone (data not shown). This highlights the dependence of the NHSCR Tracing on highly discriminatory variables in order to identify individuals. Likewise, results from the method evaluation showed a number of record matches missed by NHSCR Tracing due to the patient’s surname being common (e.g. Smith) whereas additional identifiers could be utilised in combination with (soundexed) surname by the probabilistic method to find a matching record.

Aside from the wider applicability of probabilistic methods for linking such datasets, these methods hold one other important advantage over deterministic methods in that they are not immediately thrown by errors in the data which would cause mismatches or failures to match in a deterministic system. Probabilistic methods are highly intuitive in that they make cumulative use of a range of variables for identifying matches, a process akin to that which
individuals would typically adopt in evaluating two records as possible matches. For this reason they can substantially improve on the accuracy of deterministic methods.

The method was developed to enable the researcher to have complete control over all data manipulations and to limit processing capacity and time. Where different datasets are used, the same methods could be applied, the only requirement being the recalculation of the frequencies of the variables used to link the datasets. Although manual checking can be time-consuming, the need for manual checking is dependent on the intended use of the data. If for ongoing monitoring purposes, then simply accepting matches above a specified upper threshold will be sufficient to monitor changes in mortality over time.

The successful linkage of infection and mortality records described here has since allowed an analysis of death following MRSA bacteraemia and invasive group A streptococcal infection to be undertaken, which could not have been previously done using either dataset alone (Lamagni et al., 2009; Lamagni et al., 2010). The development of this method lays the groundwork for linkage of a wide range of health-related datasets, as it has been developed with the knowledge of the limitations of the availability of variables in such datasets. This will substantially augment existing surveillance programmes by allowing the routine monitoring of mortality following infection. Application of such a method could be similarly attractive to other countries with comparable surveillance systems, and could provide a means for the rapid investigation of sudden changes flagged by real-time mortality monitoring systems currently under development (Mazick, 2007).

REFERENCES


