THEORY AND METHODS

Probabilistic record linkage and a method to calculate the positive predictive value

Tony Blakely and Clare Salmond

Background Computerized record linkage is commonly used in cohort studies to ascertain the study outcome, and as such its accuracy classifying the outcome can be described using the standard epidemiological terms of sensitivity and positive predictive value (PPV).

Method We describe a 'duplicate method' to calculate the PPV of record linkage when each record can only be involved in one match (e.g. linking population files to death files). The method does not require a validation subset of records from both files with detailed personal information (e.g. name and address), and is therefore ideal for linkage projects using anonymous data. The duplicate method assumes that the number of records from one file with zero, one, two, etc., links from the other file is distributed in a manner predicted by combinatorial probabilities. Having made this assumption, the number of false positive links, and hence the PPV, are estimable. We demonstrate this duplicate method using output from anonymous and probabilistic record linkage of census and mortality records in New Zealand.

Results The PPV estimates conform to the pattern expected based on the underlying theory of probabilistic record linkage, and were robust to sensitivity analyses. We encourage other researchers to further assess the accuracy of this method.

Keywords Medical record linkage, predictive value of tests, sensitivity and specificity, epidemiological methods, censuses, mortality

Accepted 12 August 2002

Computerized record linkage is commonly used in cohort studies to ascertain the study outcome,[1,2] often using probabilistic record linkage methods.[3,4] This paper serves three purposes. First, we briefly review record linkage methodology. Second, we briefly describe the record linkage process in the epidemiological terms of a screening test (e.g. sensitivity and positive predictive value [PPV]). Third, we describe a method to calculate the PPV when each record can only be involved in one match (e.g. linking population files to death files) and there is no 'gold-standard' data-set against which to validate the record linkage (i.e. there is no subset of records with complete data for, say, names and addresses against which to validate the record linkage).

Record linkage methodology

Detailed descriptions of record linkage methodology can be found elsewhere.[3–5] In this section, we provide a brief overview. Table 1 is a glossary of record linkage terms. The first use in the text of this paper of any term in this glossary is in bold.

Record linkage involves searching files for records that belong to the same individual. For example, we might be conducting a cohort study, and use record linkage of our cohort data set with mortality data set(s) to determine who has (or has not) died.

Deterministic record linkage

Deterministic record linkage is where we look for exact (dis)agreement on one or more **matching variables** between files. For example, we might simply use a social security number common to two files. However, coding errors of the social security number on one file mean that some true **matches** (a **comparison pair** of two records from different files *for the same person*) will be missed.

Probabilistic record linkage

Probabilistic record linkage uses information on a greater number of matching variables, and allows for the amount of information provided by any (dis)agreement on matching variables. For example, agreement on social security number is

Department of Public Health, Wellington School of Medicine, University of Otago, PO Box 7343, Wellington, New Zealand. E-mail: tblakely@wnmeds.ac.nz

Table 1 Glossary of record linkage terms

Term	**Definition**
Probabilistic record linkage	Record linkage of two (or more) files that utilizes the probabilities of agreement and disagreement between a range of matching variables.
Deterministic record linkage	Record linkage of two (or more) files based on exact agreement of matching variables.
Comparison pair	Any possible comparison of a record from one file with a record from another file.
Match	A comparison pair of records that are for the same person.
Non-match	A comparison pair of records that are not for the same person.
Link	A comparison pair that is accepted as being highly likely for the same individual.
Non-link	A comparison pair that is not accepted as being highly likely for the same individual.
False negative link	A comparison pair that is not accepted as a link when it actually was a match.
False positive link	A comparison pair that is accepted as a link when it actually was not a match.
True positive link	A comparison pair that is accepted as a link when it actually was a match.
True negative link	A comparison pair that is not accepted as a link when it actually was not a match.
Sensitivity	The proportion of all records on one file that have a match in the other file that were correctly accepted as a link.
Specificity	The proportion of all records on one file that have no match in the other file that were correctly not accepted as a link.
Matching variable	Variable common to the two files that is used for comparing records.
Blocking variable	Variable common to the two files that is used to 'block' (or partition) the two files. Only within these blocks are matching variables compared between the records. Blocking greatly reduces the number of comparisons.
u probability	The probability that a matching variable agrees given that the comparison pair being examined is as a non-match (i.e. the probability that variables agree purely by chance among non-matches).
m probability	The probability that a matching variable agrees given that the comparison pair being examined is a match.
Agreement weight	The weight assigned for an agreement on a given matching variable: $[\ln(m/u)/\ln(2)]$ where m and u are short for [m probability] and [u probability].
Disagreement weight	The weight assigned for a (dis)agreement on a given matching variable: $[\ln(1-m/1-u)/\ln(2)]$, where m and u are short for [m probability] and [u probability].
Total weight	The sum of the agreement weights for all matching variables that agree (positive values) and the disagreement weights for all matching variables that disagree (negative values).
Cut-off weight	The total weight above which comparison pairs are accepted as links.
Duplicate link(s)	A record on one file that has two or more links with records on the other file for which the total weight was above the cut-off.
Automatch®	A probabilistic record linkage software package.

more suggestive of a match than is agreement on sex. Also, agreements on rare values of a given matching variable (e.g. surname Blakely) are more suggestive than agreements on common values (e.g. Smith).

At the heart of probabilistic record linkage are ***u* probabilities** and ***m* probabilities**. Consider the matching variable 'month of birth'. The probability of this variable agreeing purely by chance for a comparison pair of two records not belonging to the same individual (i.e. a **non-match**) is about 1/12 = 0.083. This value is the *u* probability. (For a matching variable that has an uneven distribution of values in the files [e.g. country of birth], the *u* probability will vary by value.) The *m* probability is the probability of agreement for a given matching variable when the comparison pair is a match. As all matching variables are prone to mis-coding, the *m* probability is less than 1.0. The value of the *m* probability is estimated (sometimes iteratively) during the specification of the record linkage strategy based upon prior information and the proportion of agreements among the comparison pairs accepted as **links**. (As we never know which comparison pairs are actually the matches, we use the links we accept during the record linkage process to iteratively estimate the *m* probability.) In this example, assume the *m* probability was 0.95. These *u* and *m* probabilities are then used to determine frequency ratios or (**dis)agreement weights** (Table 2). In this example, a comparison pair that agreed on month of birth would be assigned a weight of 3.51 and a comparison pair that

Table 2 Example of agreement and disagreement frequency ratios and weights for the matching variable 'month of birth'

	Proportion			
Comparison outcome	Links	Non-links	**Frequency ratio**	**Weight**
Agreement	0.95 (m)	0.083 (u)	11/1 (m/u)	3.51 $[\ln(m/u)/\ln(2)]$[a]
Disagreement	0.05 ($1-m$)	0.917 ($1-u$)	1/18 ($1-m/1-u$)	–4.20 $[\ln(1-m/1-u)/\ln(2)]$[a]

[a] The divisor, ln(2), transforms the natural logarithm to a base 2 logarithm. It is conventional to use base 2 logarithms in record linkage. Accordingly, each 1-unit increase in the weight corresponds to a doubling of the relative likelihood of the comparison being a match.

disagreed on month of birth would be assigned a weight of –4.20. The setting of *u* and *m* probabilities and the corresponding weights is repeated for all matching variables, and possibly additionally for all values of each/some of the matching variables. The **total weight** for a given comparison pair is simply the sum of the (dis)agreement weights for each matching variable. The total weight will be a large positive number if all/most matching variables agree, or a large negative number if all/most matching variables disagree.

Record linkage from an epidemiological perspective

The objective of record linkage is to find matches. Figure 1 schematically shows the bimodal distribution of total weight scores for matches and non-matches in a record linkage project. Note that in reality it is not possible to determine exactly which comparison pairs are matches and non-matches, rather we just observe the combined (matches and non-matches) number of comparison pairs at any given total weight score. The task in record linkage is to set a **cut-off weight** (of the total weight) above which comparison pairs are categorized as links and below which the comparison pairs are categorized as **non-links**. Hopefully the (vast) majority of links are matches (**true positives**), and few matches are missed (**false negatives**). The vertical dotted line in Figure 1 is a possible cut-off score. A two-by-two table of link/non-link status by match/non-match status is shown below.

	Matches	**Non-matches**
Linked	**a** (true positives)	**b** (false positives)
Unlinked	**c** (false negatives)	**d** (true negatives)

As being a match in an epidemiological study is often equivalent to having the outcome of interest (e.g. death), the performance of the record linkage in *classifying the outcome* can be quantified with the familiar terms:

Sensitivity = a/(a + c)
Specificity = d/(b + d)

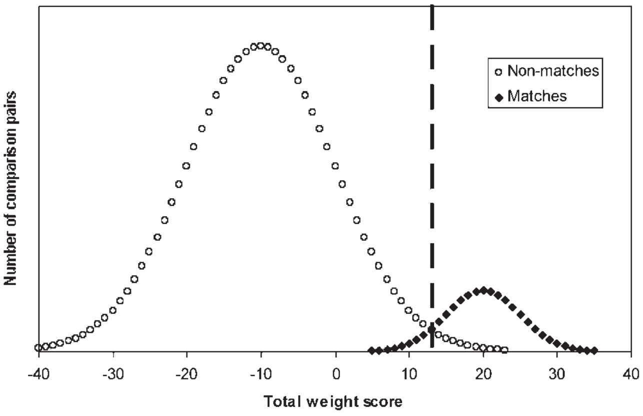

Figure 1 Number of comparison pairs for matches and non-matches by total weight score in a probabilistic record linkage project

Positive predictive value = a/(a + b)
Negative predictive value = d/(c + d)

These parameters will vary depending on the cut-off weight: moving it to the left in Figure 1 will increase the sensitivity, but also increase the number of false positives; moving it to the right will decrease the sensitivity, but also decrease the number of false positives.

When record linkage is used to determine the outcome in a cohort study, what effect do errors in the record linkage have on subsequent analyses of the association of exposure with the outcome? False positives incurred during the record linkage will bias both the risk ratios and risk differences to the null, so long as the specificity is non-differential by the exposure variable(s) measured for the cohort study-base (i.e. a non-differential misclassification bias of the mortality outcome).[1,6,7] However, the effect of false negatives incurred during the record linkage (i.e. imperfect sensitivity) is to cause an underestimate of the risk difference only—the risk ratio remains unaffected so long as the sensitivity is non-differential by the exposure variable(s).[1,8] Thus, when trade-offs are required between the number of false positives and false negatives incurred in a record linkage project a sensible strategy is to sacrifice the sensitivity (and incur many false negatives or missed matches) but maintain a high specificity (and incur few false positives or incorrect links). With this strategy the measured risk ratio in subsequent cohort analyses should be unbiased, although statistical power will be somewhat reduced.[1] (An additional strategy is to actually adjust the observed risk ratios and risk differences for misclassification bias of the outcome incurred during the record linkage process. A description of these adjustment procedures using estimates of the sensitivity and specificity or positive predictive value is beyond the scope of this paper, but are well described elsewhere.[6,9–11])

Minimizing the number of false positive links requires first quantifying their number by values of the total weight score to permit an informed decision about what value to set the final cut-off weight. There are several examples in the published literature where the cut-off was determined by manual inspection of a subset of the comparison pairs that had matching variables which were not available for all the records.[12–18] For example, Muse *et al.* linked anonymous human immunodeficiency virus data but for a sub-sample of records had names allowing a validation of the larger anonymous record linkage project.[18] In the absence of such a 'gold-standard' practitioners are forced to rely more on the 'art' of record linkage.[19] For example, comparison pairs in the grey-zone (i.e. the zone either side of the dotted line in Figure 1) are manually reviewed and a decision on linkage status made on the basis of what looks 'alright'. In probabilistic record linkage, it is also possible to estimate the absolute odds (and thereby the PPV) of a comparison pair being a match for a given weight score.[3,19–21] However, this method is prone to bias due to correlated agreements and disagreements between matching variables for a given comparison pair. For example, if sex was coded incorrectly for a given record the chance of another coding error for that particular record is probably greater than for any randomly selected record. Also, age-related bias due to the alteration in prior probability of death for any cohort followed over time may bias the absolute odds method for calculating the PPV.[3,20]

Duplicate method for determining false positives

In the remainder of this paper we describe an empirical method for estimating the number of false positive links. This method is only applicable when there can be no more than one match for a given record—a common situation in epidemiology (e.g. linking mortality files to other files). We describe and illustrate this duplicate method based on our experience linking census and mortality records in the New Zealand Census-Mortality Study (NZCMS).[22] In this study, a combination of large file sizes and a limited number of matching variables meant that even at high total weight scores there were instances of a mortality record agreeing exactly with two (or more) census record(s). The duplicate method described in this paper quantifies the false positive rate above a given total weight by using the number of observed duplicate links above that total weight score. As the number of census records far outweighed the number of mortality records in the NZCMS, we describe the duplicate method from the standpoint of mortality records linked to one, two, or more census records.

The duplicate method involves simultaneously solving the combinatorial probabilities for zero, one, or two census links for a given mortality record. Assume that above a given total weight score, there is a uniform probability, p, that any one mortality record will have a purely chance link with any one census record. Let t be the probability that a mortality record has a true link or match, and n be the number of census records (trials) compared to each mortality record. Thus:

P_1 = Pr (no match and 0 false positives)
$= [1-t] \quad [\quad (1-p)^{n} \quad]$
P_3 = Pr (no match and 1 false positive)
$= [1-t] \quad [n \quad p \quad (1-p)^{n-1}]$
P_5 = Pr (no match and 2 false positives)
$= [1-t] \quad [n(n-1)/2) \quad p^{2} \quad (1-p)^{n-2}]$
P_7 = Pr (no match and 3 false positives)
$= [1-t] \quad [n(n-1)(n-2)/6) \quad p^{3} \quad (1-p)^{n-3}]$
etc.
P_2 = Pr (1 match and 0 false positives)
$= [t] \quad [\quad (1-p)^{n-1}]$
P_4 = Pr (1 match and 1 false positive)
$= [t] \quad [(n-1) \quad p \quad (1-p)^{n-2}]$
P_6 = Pr (1 match and 2 false positives)
$= [t] \quad [(n-1)(n-2)/2) \quad p^{2} \quad (1-p)^{n-3}]$
etc.

Note that the sum of the odd-numbered probabilities is just $(1-t)$ since the terms in the second brackets are the binomial probabilities of observing 0, 1, 2,... n false links in n comparisons and thus sum to unity. Similarly, the even-numbered probabilities sum to t. Thus the sum of all possible probabilities is $(1-t) + t = 1$.

In practice, at and above a given total weight score we may observe the proportion of mortality records with zero, one, and two census record links at the specified weight cut-off in the linkage as X, Y, and Z, where:

$$X = P_1$$
$$Y = P_2 + P_3$$
$$Z = P_4 + P_5$$

Multiplying the equation for Y by $(n-1)(1-(1-p))/(1-p)$, subtracting the equation for Z, and then substituting $X/(1-p)^{n}$ for $(1-t)$ (from the equation for X), we get a quadratic in $(1-p)$:

$$[n(n-1)X + 2(n-1)Y + 2Z]\ (1-p)^{2} - [2n(n-1)X + 2(n-1)Y]\ (1-p) + [n(n-1)X] = 0 \qquad (1)$$

where n is the number of census records that can possibly be compared to each mortality record. The equation has two roots. Back substitution gives values for p and t. The correct one of these two roots will give $t < 1$ and $0 < (1-p) < 1$.

When a mortality record agreed exactly with two or more census records (therefore each link scores exactly the same total weight), one of these duplicate links was almost certainly the match and the other(s) a false-positive link. As they were indistinguishable we discarded both links to prevent false positive links. When the duplicate links had different total weight scores we assumed the highest scoring link was the match (a reasonable assumption when the majority of matches [if present] agree on all matching variables as was the case in this study), and rejected the remaining lower scoring duplicate links. Given these two decision rules, none of the even number probabilities above contribute false positive links. The proportion of all mortality records involved in false positive links can thus be approximated from the odd numbered probabilities in $\{P_i, i \geqslant 3\}$, where each P_i is estimated by substitution of the derived values for p and t.

Two refinements may be used with this duplicate method, first to improve efficiency, and second to recognize that not all mortality records are eligible to have a comparison pair as the cut-off becomes very high.

Efficiency is improved by '**blocking**', that is by comparing records on the two files only when a highly discriminating variable already agrees. For example, we might block the census and mortality files by geocode and thus only compare census and mortality records when they come from the same neighbourhood. This blocking dramatically reduces the number of comparisons between the two files, but also reduces the sensitivity (a match with disagreeing geocode would be missed or 'skipped') and increases the PPV (the number of false positives is a function of how many census records are compared to any given mortality record). In the above equations, n becomes the average number of census records in each block—not the total number of census records in the file. (The effect of using an average n is explored below.)

Second, very high total weight scores will only be possible for *exact* agreements between records with *uncommon* values of the matching variables (e.g. born in Asia). In order for the duplicate method to work at these very high total weights, allowance must be made for the decreasing number of records able to score this high (a method for which is presented below). However, as most record linkage projects will accept all exact agreements this problem is not critical.

Illustrating the duplicate method in the New Zealand Census-Mortality Study (NZCMS)

The NZCMS study involves linking census records to mortality records.[22,23] A limited range of matching variables are available in the NZCMS: geocodes, sex, date of birth (disaggregated to day, month, and year of birth), ethnicity and country of birth. Thus there was the potential for false positive links that we wanted to minimize to preserve the validity of the risk ratios in subsequent cohort analyses.

The linkage of the 1986 census and 1986–1989 mortality records in the NZCMS involved eight passes using Automatch®.[24] In the first pass the census and mortality records were blocked into approximately 32 000 meshblocks, the smallest administrative geographical area in New Zealand with an average of around 100 people. In all, 39 515 mortality records and 3 131 176 census records were submitted to the first pass. Among other things, the output from Automatch® includes the number of 'highest-scoring' pairs and 'duplicate' pairs (i.e. MP and DA Pairs, respectively, in Automatch® jargon). (Automatch® does not produce values for X, Y and Z directly.) A 'highest-scoring' pair is the highest total weight scoring comparison pair for a given mortality record. A 'duplicate' pair is any other comparison pair involving a mortality record that is already involved in a highest-scoring pair. Thus, above any given cut-off:

- A mortality record linked to only one census record results in just one highest-scoring pair. The proportion of mortality records with this outcome is equivalent to 'Y' above.
- A mortality record linked to two census records results in one highest-scoring pair and one duplicate pair. The proportion of mortality records with this outcome equivalent to 'Z' above.
- A mortality record linked to three census records results in one highest-scoring pair and two duplicate pairs; and so on.

Note that:

[No. duplicate pairs] =
[No. mortality records linked to two census records] +
2 × [No. mortality records linked to three census records] +
3 × [No. mortality records linked to four census records] +
etc …

We used an iterative process to estimate X, Y, and Z. Equation (1) was first solved using the number of 'highest-scoring' for X, and the number of duplicate pairs for Y (and consequently Z was initially set at zero). Next, P_1, P_2, P_3, P_4, and P_5 were calculated using the p and t estimates from the first iteration, and then revised estimates of X (P_1), Y ($P_2 + P_3$), and Z ($P_4 + P_5$) were made and used in the second iteration. This process was repeated until convergence was achieved.

The number of highest weight-scoring pairs and duplicate pairs above varying cut-off weights is shown in the first two columns of Table 3. In this project the majority of comparison pairs above a total weight of 14 (calculated probabilistically by Automatch®) agreed exactly on all matching variables. For any cut-off below 14 we assume that all 39 515 submitted mortality records had a chance of being involved in a false positive link. However, for any cut-off above 14 we adjusted downwards the number of submitted mortality records to approximate the number that could have actually had a link above the given weight. We used the distribution of highest-scoring pairs by weight score to approximate that number. For example, above a cut-off of 17 there were 7205 highest-scoring pairs, or 29.6% of all the 24 352 highest-scoring pairs above 14. Thus we assumed that the number of mortality records with values of their matching variables that permitted a weight score above 17 was 29.6% of 39 515, i.e. 11 691. This adjusted number of mortality records was used in combination with the number of highest weight-scoring pair and duplicate pairs to calculate X, Y and Z.

The fourth column of Table 3 presents the estimated number of false positive links calculated by solving equation (1) and then calculating the number of false positive links. Note that as we used blocking by geocode in the record linkage, n is 100 (the average number of census records in each block) not 3 131 176 (the total number of census records). The PPV was then calculated as [1–([estimated number of false positives]/[number of highest-scoring pairs])].

The calculations so far determine the PPV *above* different total weights. Of more relevance in setting the cut-off weight is the PPV at the margin, i.e. at or about the potential cut-off weight. We estimated this 'marginal PPV' by determining the number of highest-scoring pairs and estimated false positives for each 1-point range of the total weight score. Results are shown in the final columns on Table 3. For example, we estimated that 70.9% of links with a total weight-score between 7 and 8 were matches, i.e. the PPV was 70.9% for this narrow range of total weight scores. The marginal PPV increased rapidly from close to 0% at a weight score of about 3.5 to 90% for a weight score of about 9.5. Thus, to ensure that the marginal false positive percentage was always greater than 90%, a cut-off score of 9 was indicated in this project.

Whilst we were unable to validate our duplicate method for calculating the PPV against a gold-standard sub-sample of comparison pairs with more discriminating matching variables (e.g. names and text addresses), two additional methods provided reassuringly similar patterns of results. (See ref. 22 for details). First, for each 1-point increase in the weight score the *odds* of being a false positive link approximately halves exactly as would be predicted by the absolute odds method.[3,19–21] Second, PPV calculations using the duplicate method for very high total weight scores (i.e. where most comparison pairs were exact agreements) were similar to calculations using a method based on the probability of any one mortality record agreeing exactly with a census record by purely chance. However, there are two advantages of the duplicate method compared to the absolute odds method and the latter chance method. Unlike the absolute odds method the duplicate method is not prone to bias from correlated coding errors; and unlike the 'chance method' it is applicable to weight scores for non-exact agreements.

We conducted sensitivity analyses of the effect of variations about the average block size (i.e. n), assuming that false positive links only arose for P_3, P_5, and P_7, and assuming that p was constant for all mortality records. For the situation encountered in the NZCMS, it appeared that the duplicate method was not particularly sensitive to moderate violations of these assumptions described above. (See reference 22 for details.)

Conclusion

There is both an art and a science to computerized record linkage.[3,19] In this paper, we have attempted to introduce a little more science by describing a method to calculate the PPV when only one match per record is possible, and it is not possible to validate the record linkage against a gold-standard sub-sample with more discriminating matching variables. We encourage other researchers to further assess this duplicate method in two ways. First, its performance should be assessed against PPV estimates obtained in linkage projects where a gold-standard

Table 3 Calculations of the positive predictive value (PPV) above varying total weight scores in a probabilistic linkage project of 39 515 mortality records and 3 131 176 census records

	PPV calculations *above* a given total weight score					**PPV calculations at the *margin***			
Total weight score	Mortality records linked to ⩾1 census record(s)	Duplicates pairs	Adjusted no. of submitted mortality records	Estimated no. of false positives	Estimated PPV	Total weight score range	Mortality records linked to ⩾1 census record(s)	Estimated no. of false positives	Estimated 'marginal' PPV
⩾3	30 123	5887		2027.0	93.3%	3–4	327	320.6	2.0%
⩾4	29 796	4813		1706.1	94.3%	4–5	1292	958.4	25.8%
⩾5	28 504	1861		747.7	97.4%	5–6	313	113.3	63.8%
⩾6	28 191	1526		634.4	97.7%	6–7	1415	363.2	74.3%
⩾7	26 776	559		271.2	99.0%	7–8	317	92.2	70.9%
⩾8	26 459	357		178.9	99.3%	8–9	158	22.4	85.8%
⩾9	26 301	307		156.5	99.4%	9–10	364	31.8	91.3%
⩾10	25 937	235		124.7	99.5%	10–11	948	43.5	95.4%
⩾11	24 989	138		81.2	99.7%	11–12	62	5.4	91.3%
⩾12	24 927	128		75.8	99.7%	12–13	166	5.3	96.8%
⩾13	24 761	117		70.5	99.7%	13–14	409	8.8	97.8%
⩾14	24 352	98	39 515	61.7	99.7%	14–15	350	5.0	98.6%
⩾15	24 002	90	38 947	56.6	99.8%	15–16	1741	15.1	99.1%
⩾16	22 261	66	36 122	41.5	99.8%	16–17	15 056	30.2	99.8%
⩾17	7205	18	11 691	11.3	99.8%	17–18	1162	0.6	99.9%
⩾18	6043	17	9806	10.7	99.8%	18–19	319	2.5	99.2%
⩾19	5724	13	9288	8.2	99.9%	19–20	2410	2.5	99.9%
⩾20	3314	9	5377	5.7	99.8%	20–21	1851	3.0	99.8%
⩾21	1463	6	2374	2.7	99.8%	21–22	148	-1.1	100%
⩾22	1315	6	2134	3.8	99.7%	22–23	875	1.9	99.8%
⩾23	440	3	714	1.9	99.6%	23+	440	1.9	99.6%

[a] Duplicate (DA) pairs in Automatch® output are comparison pairs involving a mortality record that is already included in a highest-scoring (MP) pair. Thus, the number of duplicate pairs = [mortality records linked to two census records] + 2 × [mortality records linked to three census records] + 3 × [mortality records linked to four census records] + …

sub-sample is available. Second, we described the method in the context of one file greatly outnumbering the other. Whilst we believe the underlying principle and assumptions are applicable to projects with similar sized files (but still only one possible match per record), this needs scrutinizing.

Acknowledgements

The NZCMS is conducted in collaboration with Statistics New Zealand. The NZCMS is primarily funded by the Health Research Council of New Zealand, with co-funding from the Ministry of Health. We thank Alistair Woodward for useful comments on a draft of this paper. The (New Zealand Census Mortality Study) NZCMS is a study of the relationship between socioeconomic factors and mortality in New Zealand, based on the integration of anonymized population census data from Statistics New Zealand and mortality data from the New Zealand Health Information Service. The project was approved by Statistics New Zealand as a Data Laboratory project under the Microdata Access Protocols in 1997. The data-sets created by the integration process are covered by the Statistics Act and can be used for statistical purposes only. Only approved researchers who have signed Statistics New Zealand's declaration of secrecy can access the integrated data in the Data Laboratory. For further information about confidentiality matters in regard to this study please contact Statistics New Zealand.

KEY MESSAGES

- Record linkage is commonly used to determine the occurrence of the outcome (e.g. mortality) in cohort studies.
- Errors in the record linkage, therefore, manifest as misclassification bias of the study outcome.
- The accuracy of record linkage can be quantified in terms of sensitivity, specificity and positive predictive value.
- The occurrence of duplicate links (e.g. one mortality record linked to two census records) can be used to quantify the positive predictive value of the outcome (mis)classification. This quantification allows an informed decision about where to set the cut-off weight above which links are accepted.

References

[1] Howe G. Use of computerized record linkage in cohort studies. *Epidemiol Rev* 1998;**20:**112–21.

[2] Gill L, Goldacre M, Simmons H, Bettley G, Griffith M. Computerised linking of medical records: methodological guidelines. *J Epidemiol Community Health* 1993;**47:**316–19.

[3] Newcombe H. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press, 1988.

[4] Jaro M. Probabilistic linkage of large public health data files. *Stat Med* 1995;**14:**491–98.

[5] Baldwin J, Acheson E, Graham W. *Textbook of Medical Record Linkage*. Oxford: Oxford University Press, 1987.

[6] Copeland K, Checkoway H, McMichael A, Holbrook R. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol* 1977;**105:**488–95.

[7] Rothman K, Greenland S. *Modern Epidemiology*. *2nd Edn*. Philadelphia: Lippincott-Raven, 1998.

[8] Rodgers A, McMahon S. Systematic underestimation of treatment effects as a result of diagnostic test inaccuracy: implications for the interpretation and design of thromboprophylaxis trials. *Thromb Haemost* 1995;**73:**167–71.

[9] Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. *Am J Epidemiol* 1993;**138:**1007–15.

[10] Green M. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. *Am J Epidemiol* 1983; **117:**98–105.

[11] Blakely T. Socio-economic factors and mortality among 25–64 year olds: The New Zealand Census-Mortality Study. (Also at http://www.wnmeds.ac.nz/nzcms-info.html) [Doctorate]. University of Otago, 2001.

[12] Muse A, Mikl J, Smith P. 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.

[13] van den Brandt P, Schouten L, Goldbohm R, Dorant E, Hunen P. Development of a record linkage protocol for use in the Dutch cancer registry for epidemiological research. *Int J Epidemiol* 1990;**19:** 553–58.

[14] 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. *Meth Inform Med* 1995;**34:** 371–77.

[15] Goldberg M, Carpenter M, Theriault G, Fair M. The accuracy of ascertaining vital status in a historical cohort study of synthetic textiles workers using computerised record linkage to the Canadian mortality data base. *Canadian J Public Health* 1993;**84:**201–04.

[16] Mi M, Kagawa J, Earle M. An operational approach to record linkage. *Meth Inform Med* 1983;**22:**77–82.

[17] Calle E, Terrell D. Utility of the National Death Index for ascertainment of mortality among Cancer Prevention Study II Participants. *Am J Epidemiol* 1993;**137:**235–41.

[18] Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. *Meth Inf Med* 1998;**37:**69–74.

[19] Roos LJ, Wajda A, Nicol J. The art and science of record linkage: methods that work with few identifiers. *Comput Biol Med* 1986;**16:**45–57.

[20] Newcombe H. Age-related bias in probabilistic death searches due to neglect of the 'Prior Likelihoods'. *Computers and Biomedical Research* 1995;**28:**87–99.

[21] Newcombe H, Smith M, Howe G, Mingay J, Strugnell A, Abbatt J. Reliability of computerized versus manual death searches in a study of the health of Eldarado uranium workers. *Comput Biol Med* 1983; **13:**157–69.

[22] Blakely T, Salmond C, Woodward A. Anonymous record linkage of 1991 census records and 1991–94 mortality records: The New Zealand Census-Mortality Study (Also at http://www.wnmeds.ac.nz/nzcms-info.html). Wellington: Department of Public Health, Wellington School of Medicine, University of Otago, 1999.

[23] Blakely T, Salmond C, Woodward A. Anonymous linkage of New Zealand mortality and Census data. *Aust NZ J Public Health* 2000;**24:**92–95.

[24] MatchWare Technologies I. *Automatch Generalised Record Linkage System, Version 4.2: User's Manual*. Kennebunk, Maine: MatchWare Technologies, Inc, 1998.