Record linkage to enhance data from perinatal registries

Tromp, M.

Publication date
2009
Document Version
Final published version

Chapter 1

Understanding the differences in performance between deterministic and probabilistic medical record linkage

M Tromp, ACJ Ravelli, GJ Bonsel, A Hasman, JB Reitsma

Submitted

Abstract

Introduction

The goal in medical record linkage is to combine data on the same patient stored in different databases. Deterministic record linkage uses a pre-defined rule to classify record pairs as links and non-links whereas probabilistic record linkage uses the observed data patterns to find the optimal classification. Our aim was to compare the performance of both strategies under different conditions by varying the frequency of registration errors and the amount of discriminating power.

Methods

We performed a simulation study in which we varied key data characteristics to create a range of realistic record linkage scenarios. For each scenario we compared the number of misclassifications (number of false links and false non-links) made by the different linking strategies: deterministic full, deterministic N-1 (all but one have to agree) and probabilistic.

Results

The full deterministic strategy produced the lowest number of false positive links, but at the expense of missing considerable numbers of matches depending on error rate of the linking variables. The probabilistic strategy outperformed the deterministic strategy (full or N-1) across all scenarios. A deterministic strategy will only produce comparable results in those situations where the deterministic rule closely matched the patterns identified by a probabilistic strategy. However, the information to guide the selection of patterns is not present in a deterministic strategy.

Conclusion

Probabilistic record linkage is the preferred method for record linkage as deterministic strategies can lead to many linking errors if the predefined rule is not correctly specified. Furthermore, the probabilistic strategy is a more flexible approach that provides additional information about the quality of the linkage process.

1.1 Introduction

The growing number of electronically available databases with patient information offers the possibility to reuse existing databases for medical research as an alternative for setting up clinical trials or prospective cohort studies. To answer clinical questions often requires that information on the same patient residing in different databases needs to be combined. Because of privacy concerns, only a few countries make use of a unique national patient identifier to combine stored patient data.[1-3] Medical record linkage (MRL) is a tool to combine data on the same patient stored in different databases in the absence of a unique identifier.[4;5] Record linkage is possible when combining a set of partially identifying variables generates a powerful discriminating system. Examples of such linking variables are first/last name, date of birth, gender, city of residence and postal code.[6-10]

Two frequently applied strategies in record linkage are deterministic (DRL) and probabilistic (PRL) record linkage that differ fundamentally in their approach. In deterministic record linkage all or a predefined subset of linking variables have to agree (corresponding values on a linking variable are the same within a pair) to consider a pair as a link.[7;11-13] In probabilistic linkage, weights for agreement (reward) or disagreement (penalty) are estimated for each variable based on the difference in probability that a variable agrees among matches and non-matches.[14;15] The term match refers to the situation that two records in reality belong to the same person, while a link refers to the outcome of the record linkage procedure. The first probability reflects the reliability of the variable (1 - error rate) and the second probability the discriminating power of the variable (1 - chance agreement). If the total sum of weights is above a certain threshold value, the pair is considered a link.

Two types of errors can occur in record linkage: the failure to link two records that belong to the same person (false non-link) and the linking of two records that belong to different persons (false link). False non-links occur when there is disagreement on linking variables while the records belong to the same person (a match), which can be caused by data entry errors. False links occur when two different persons share the same value on several linking variables just by chance. The discriminating power of a linking key is considered high if this probability that two different persons will have the same value of their linking key is low. The (theoretical) number of possible values of the linking variables (more values – higher discriminating power), and their distribution (more uniformly distributed – higher discriminating power) determines this probability. Error rate and discriminative power are the two fundamental concepts in record linkage. Together they influence the number of incorrect decisions in a linking procedure; a high error rate increases the number of false non-links and a low discriminating power increases the number of false links.

Although probabilistic record linkage incorporates the concepts of discriminating power and error rate in a natural way, the method is argued to be more complex and less transparent than the deterministic approach. It is unclear whether and when the results of these two approaches are comparable and when relevant differences occur. The objective of this study is to compare the performance of deterministic and probabilistic record linkage under a range of different linking conditions.

1.2 Methods

Overall approach

We performed a simulation study in which we created different datasets by varying the amount of registration errors and the discriminating power of the linking variables, thereby mimicking a range of realistic linking 'scenarios' (figure 1.1). Each dataset consisted of 4 linking variables (the linking key) and a unique identifier for the record. Each linking variable was specified by a certain number of possible values, an underlying distribution of these values, and a proportion of registration errors. Each linking variable reflected a specific type of variable that can be encountered in real linkage situations. By systematically varying these data-generating parameters (number and distribution of values and error rate), we created a total of 10 scenarios. For each scenario we compared the results (number of false links and false non-links) of the different linking strategies. This approach allowed us to study the effects of increasing error rate and decreasing discriminating power on the difference in performance between DRL and PRL.

Simulation of datasets

The four linking variables and their characteristics in the basic scenario are displayed in table 1.1. Variable 1 (V1) represents a date of birth, variable 2 (V2) models postal code, variable 3 (V3) models gender and variable 4 (V4) represents a hospital code. Two datasets (dataset A and B, with size N_A and N_B respectively) were created having these four variables in common. We chose N_A = 10,000 and N_B = 6,000. Records were generated by randomly drawing values for each variable based on the specified distribution (same for A and B). A unique number was assigned to each record in both datasets. In the next step, a random sample S of records of size N_s = 4,000 was drawn from dataset A (see Figure 1.1). Errors were randomly introduced in the linking variables of this subset and subsequently this subset was added to dataset B. Errors were introduced by randomly drawing a new value from the same distribution as the original variable. This sample S of records present in both dataset A and B are the matches that need to be identified by the record linkage strategy without using the unique record identifier (see Figure 1.1).

In the basic scenario, the date variable (V1) is normally distributed with a mean of 30 years and SD of 1825 days (5 years), postal code (V2) consists of a stepwise distribution with 400 values per stratum and a different probability per stratum (representing five different address density areas with probabilities 0.4, 0.25, 0.2, 0.1 and 0.05), gender is uniformly distributed with two values (V3) and hospital number is uniformly distributed with 120 values (V4) (see Table 1.1). Given the distribution of the four variables, the probability of agreement on all four variables by chance is 6.10^{-10}. If we compare all records of dataset A with all records of dataset B under basic conditions (this implies $N_A * (N_B + N_S)$ comparisons), the expected number of pairs agreeing on all four variables just by chance is (10,000 * (6,000 + 4,000)) * 6.10^{-10} = 0.06 for the basic scenario. The amount of error introduced in the basic scenario was 1% for date of birth (V1), 5% for postal code (V2), 0.5% for gender (V3) and 2% for hospital number (V4).

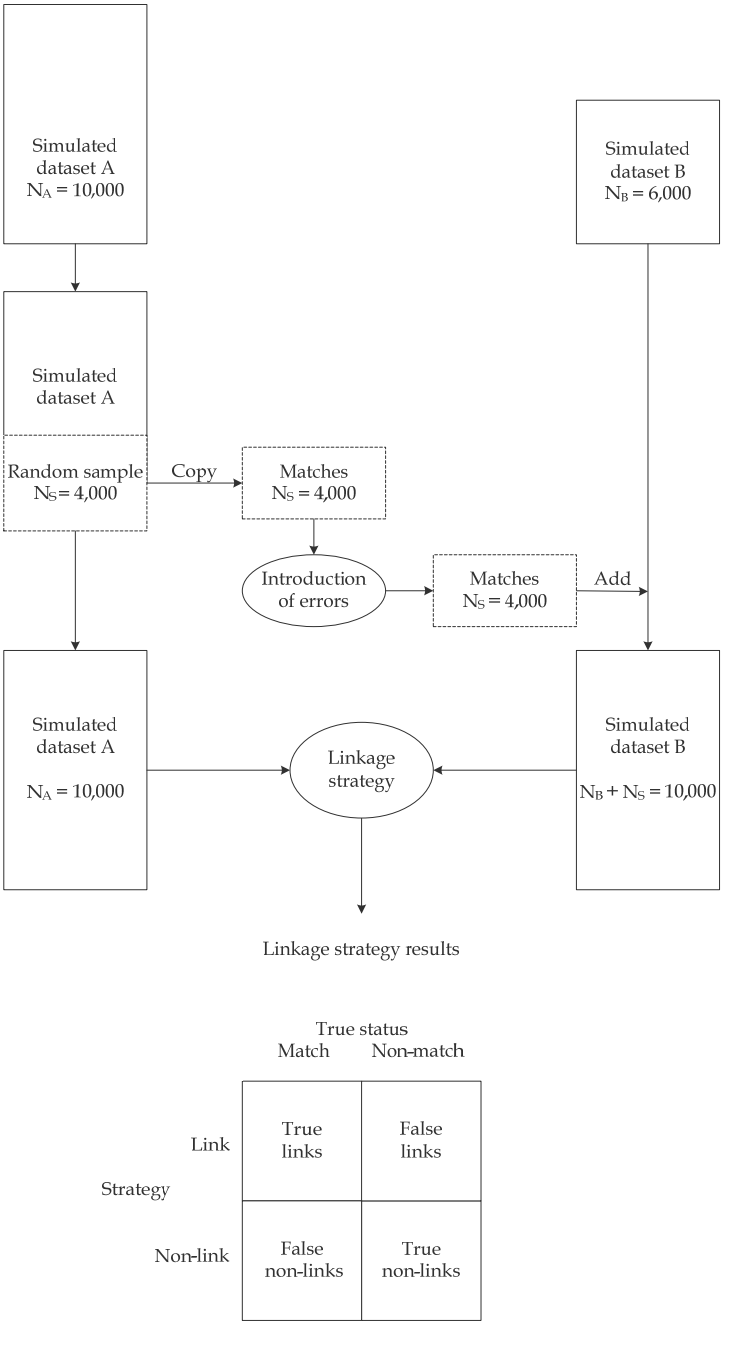

Figure 1.1 Outline of simulation study.

In the 9 alternative scenarios, the chance agreement of the linking key was varied (0.1x, 0.5x, 2x, 8x and 80 times the chance agreement in the basic scenario) and the amount of error (0.25x, 0.5x, 2x and 4 times the amount of error in the basic scenario). Each scenario was repeated 100 times, so the results presented per scenario reflect the average result of 100 simulations ('runs').

Table 1.1 Characteristics of the linking variables in the basic scenario.

Variable		Discriminating power		Error rate	
V1	Date of birth	High	Normal (30,1825)	Low	1%
V2	Postal code	High	Stepwise*	High	5%
V3	Gender	Low	Uniform (2)	Low	0.5%
V4	Hospital	Medium	Uniform (120)	Medium	2%

* Five strata with a uniform distribution of 400 values with different probability (0.4, 0.25, 0.2, 0.1, 0.05).

Linking strategies

In the deterministic strategy, all records of dataset A and B were compared and within each pair the number of linking variables that had the same value was determined. Two different DRL decision rules were applied: the DRL full match strategy (all four variables have to agree) and the DRL N-1 match strategy (one linking variable may disagree).

The probabilistic strategy starts with estimating the so called m_i and u_i probabilities for each linking variable. These probabilities of agreement among matches and non-matches are estimated in a latent class model by using the observed frequency of patterns of agreement and disagreement among all pairs. In this latent class model the unknown status of a pair (match or not) is mathematically linked to the individual m and u probabilities of each linking variable and a prevalence parameter. Maximum likelihood methods are used to estimate the parameters of this model (see Appendix and reference [16] for more details).

A weight for agreement and a weight for disagreement were calculated using the estimated m_i and u_i probabilities:

Agreement weight of the i^{th} variable is given by: $\log_2 \frac{m_i}{u_i}$,

Disagreement weight of the i^{th} variable is defined by: $\log_2 \frac{1-m_i}{1-u_i}$.

Using these linking weights a total linking weight for each pair was calculated by summing up the individual linking weights. The threshold value above which pairs were considered a link was based on the estimated prevalence of matches provided by the latent class model.

Data analyses

By design, the prevalence of linked records was N_S / (N_A * N_B) = 4,000 / (10,000 * 10,000) = 4.10^{-5}. Regardless of the linking strategy the true status of each pair (match or not) could be determined by the added unique number. The linking status as result of applying either DRL or PRL was then compared to the true status (true link/ non-link, false link or false non-link). The number of false links and false non-links per strategy was our main outcome

measure. Since for each scenario 100 datasets were simulated, the mean value of these 100 runs was used.
The simulation syntax was created in SAS version 9.1.

1.3 Results

Figure 1.2 shows the performance of the full deterministic strategy under different linking conditions. Because all four linking variables have to agree before a record pair is considered a link, the main problem is missing a match (false non-link) because one (or more) linking variables contain an error. The number of false non-links with the full deterministic approach in the basic scenario was on average 330, so 330 out of the 4,000 matches could not be identified due to errors in the linking variables. If the amount of error is increased, the number of false non-links increased in parallel (Figure 1.2). The likelihood of chance agreement leading to false links increased when the discriminative power of the linking key was lowered.

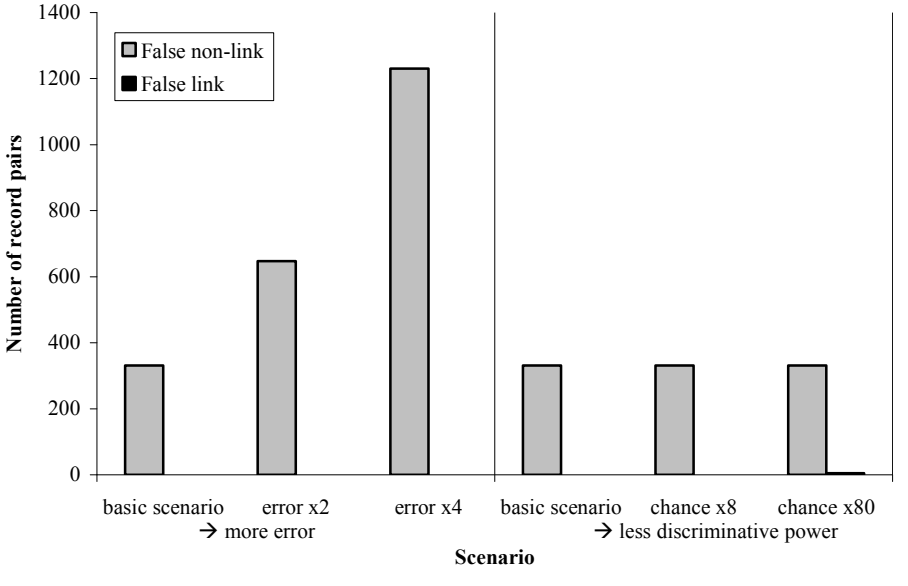

Figure 1.2 Linkage errors by the full deterministic strategy under conditions which vary by error rate or discriminative power.

A logical approach to overcome false non-links as a result of errors in linking variables is to accept differences in linking values within a pair (e.g. a pair is still classified as link if only one variable disagrees). This is the deterministic N-1 strategy. The rationale is that discriminative power may be sacrificed to compensate for errors in linking variables. Figure 1.3 shows the results of the individual N-1 variants (one specific variable is allowed to disagree) and the overall N-1 deterministic strategy.

Allowing only disagreement on the variable that has a high discriminating power and a low error rate (V1), produced poorer results than the deterministic full strategy. Only a few false non-links were prevented, but many false links were introduced by reducing the amount of discriminating power. The N-1 variant were V2 is allowed to differ (high discrimination, high error) repaired about half of the false non-links at the cost of introducing some false links. For a variable with low discrimination and low error rate (V3), a few false non-links were repaired and the loss of power by allowing V3 to differ did not (directly) lead to the occurrence of false links. For the variable with a medium error rate and medium discriminating power (V4), some false non-links were now correctly classified at the cost of only a few false links. Compared to the full deterministic strategy, the overall N-1 deterministic strategy introduced more false links than it repaired false non-links. The probabilistic strategy led to the lowest total number of linking errors.

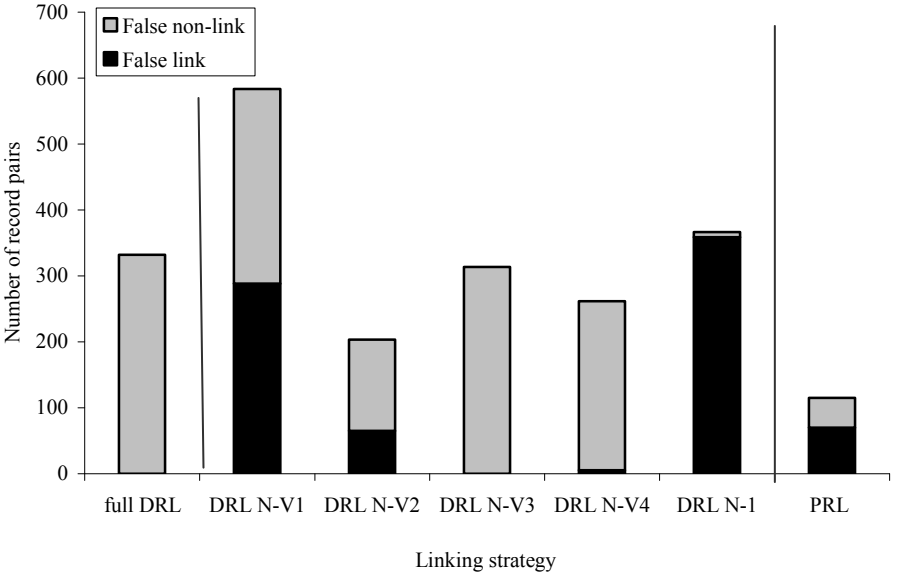

Figure 1.3 Performance of the full deterministic (DRL full), individual N-1 deterministic (DRL N-Vx), overall N-1 deterministic (DRL N-1) and the probabilistic strategy (PRL) for the basic scenario.

The advantage of a probabilistic strategy is that it provides a ranking of all possible patterns of agreement based on their total linking weight. This total weight is a reflection of the likelihood of a record pair belonging to the same individual (match). Table 1.2 shows the agreement and disagreement weight for the individual linking variables from a run. Table 1.3 shows the ranking of record pairs by the probabilistic approach for the basic scenario. The probabilistic strategy correctly indicates that one N-1 deterministic strategy (allowing

V1 to differ – pattern '0111') should not be considered as a link as this variant will increase the total number of misclassifications.

Table 1.2 Agreement and disagreement weight for the basic scenario obtained from the probabilistic linking procedure.

Variable		m_i	u_i	weight agree	weight disagree
V1	Date of birth	0.9953	0.0002	12.6	-7.72
V2	Postal code	0.9486	0.0007	10.4	-4.28
V3	Gender	0.9948	0.5000	0.99	-6.60
V4	Hospital	0.9793	0.0083	6.88	-5.58
Total				30.9	-24.2

Table 1.3 Ranking of all possible patterns of agreement and disagreement for the basic scenario based on the total weight derived from the probabilistic record linkage strategy.

(Dis-)agreement pattern (V1,V2,V3,V4)	Number of record pairs	Total linking weight	Matches
0000	49536622	-24.2	0
0010	49540550	-16.6	0
0001	416319	-11.7	0
0100	34812	-9.5	0
0011	416702	-4.1	4
1000	7761	-3.8	0
0110	34775	-1.9	0
0101	291	3.0	1
1010	7753	3.8	2
1001	64	8.6	0
0111	310	10.5	35
1100	8	10.9	0
Threshold value			
1011	264	16.2	194
1110	83	18.4	78
1101	19	23.3	19
1111	3667	30.9	3667
Total	10000000		4000

Figure 1.4 shows the performance of the deterministic N-1 and probabilistic strategy for scenarios with varying error rates. In the basic scenario, the total number of linking errors for the deterministic N-1 strategy was about 3 times higher than for the probabilistic strategy and mainly consisted of false links. With increasing error rates, more false non-links started to occur for the deterministic N-1 strategy. It reflects the increase in likelihood that a true match will have registration errors in two or more linking variables, while the deterministic N-1 strategy only compensates for an single error within a match. The number

of false links in the deterministic N-1 strategy occurred due to a lack of discriminating power and was not influenced by variations in error rate. For the probabilistic approach, the increasing error rate gave more linking errors; both false links and false non-links. This reflects the general point that more registration errors produce less favourable linking conditions hampering the discrimination between matches and non-matches by the algorithm. In all scenarios, the number of linking errors with the probabilistic strategy was lower than with the deterministic N-1 strategy.
Figure 1.5 compares the performance of the deterministic N-1 and probabilistic strategy for scenarios in which the discriminative power was varied. In a situation with very high discriminative power, the performance of both strategies becomes comparable. With decreasing discriminative power, the number of false links increased for the deterministic N-1 strategy reflecting the higher chance agreement. The number of false non-links with the deterministic N-1 strategy was not influenced by variations in discriminative power. The performance of the probabilistic approach also decreased with lower discriminative power, but the reduction in performance was much slower than with the deterministic N-1 strategy.

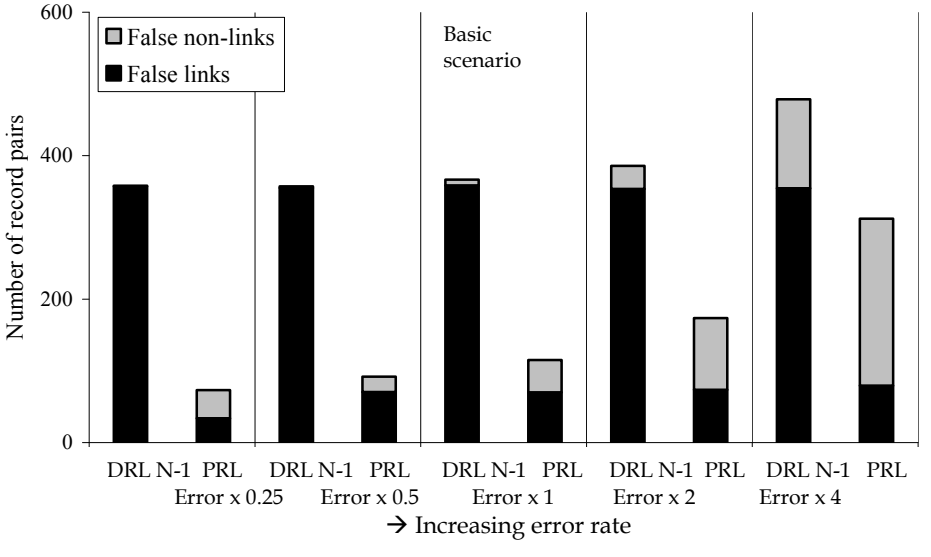

Figure 1.4 Performance of the overall N-1 deterministic (DRL N-1) and the probabilistic (PRL) strategy as a function of error rate of the linking key.

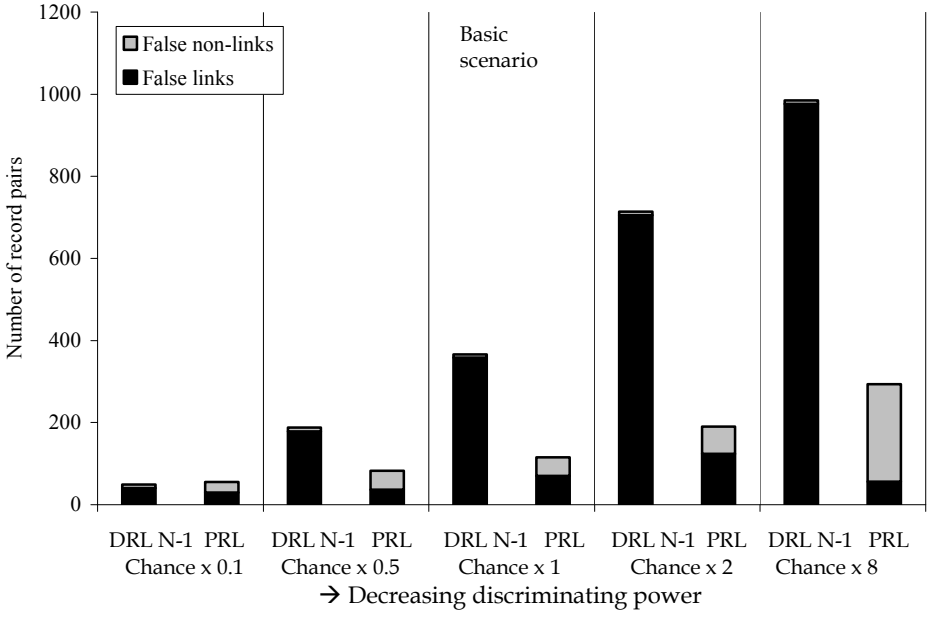

Figure 1.5 Performance of the overall N-1 deterministic (DRL N-1) and the probabilistic (PRL) strategy as a function of discriminative power of the linking key.

1.4 Discussion

In linking situations with few to no error and a powerful discriminating linking key, a simple deterministic full linking strategy performs adequately. Unfortunately, such situations are very rare. A deterministic N-1 strategy can compensate for errors in linking variables by sacrificing discriminating power, but it treats all N-1 variants in the same way which may not be correct, thereby introducing unnecessary errors (too many false links). A probabilistic record linkage strategy is superior in finding the optimal balance between false non-links and false links which results in fewer linking errors overall. In addition to the linkage outcome, the probabilistic strategy provides for each observed pattern the likelihood that a record pair belongs to the same or to different individuals. This information can be used to adjust the threshold in situation where false links (then increase the threshold) or false non-links (lower threshold) have more negative consequences.

The parameters that influence the outcome of a linking procedure can be varied in an infinite manner. An illustrative basic scenario with four common available linking variables was chosen from which variations were made to demonstrate the impact on the performance of the different linking strategies. We studied scenarios with only four linking variables, where more candidate linking variables might be available. Adding linking variables to the linking key, adds both discriminative power and potential errors. More

linking variables will increase the number of possible patterns and defining a deterministic rule based on common sense becomes even more complex. Moreover, the meta-information on quality of a linkage strategy should in our view be preferred to the 'black box' of deterministic linkage which rests on assumptions which are often not made explicit (e.g. the unrealistic assumption that all N-1 have a similar probability of being a match).
In the current study we did not include string variables such as first and last name, which can be available for record linkage. Characteristics of a string variable are that they have a large number of possible values (high discriminating power) but they are also prone to error (high error rate). Using Soundex[17] or another phonetic algorithm[18;19] can reduce errors and make a string variable more suitable for record linkage. In anonymous datasets however, names are usually absent.
In the current study we did not incorporate missing values. Missing values can be handled in several ways. If a variable is truly unknown, agreement (missing in both records) can be informative. Generally however, a missing value will be non-informative. In deterministic linkage a missing value in one or both variables in the variable comparison has to be considered as either agreement or disagreement. In probabilistic linkage it is also possible to assign a fixed weight of zero (non-informative) in case of missing values.[6;16] When applying a fixed weight for a missing value in variable comparison, missing values should be excluded from the weight estimation. The method of handling missing values will determine the difference between a deterministic and probabilistic approach, but the probabilistic approach can be more refined in handling missing values, for instance by considering agreement on missing values as a special outcome (see next paragraph).
We only considered agreement or disagreement as the outcome of variable comparison, while partial agreement is another possible outcome category. Partial or 'close' agreement is used to correct for data entry or transcription errors or for small differences in values due to rounding off or different measurements instruments. Also a certain range around a variable can be considered as 'close' agreement. The advantage of the probabilistic strategy is that an actual weight can be estimated for close agreement as 'close' agreement can still add some evidence in favour of a link.[20]
A key assumption in most probabilistic algorithms including ours is that conditional on the true status agreement on one variable does not affect the probability of agreement on another variable (conditional independence assumption). The same is true for the occurrence of error. In our simulations these independence assumptions were – by design - satisfied. Dependencies in values can result in incorrect linking weight estimations with the probabilistic strategy.[21] Dependency among variables can be addressed by combining two correlated variables into a single variable[21] or by extending the latent class model with additional parameters addressing dependency[22]. We did not study the effect of correlation among errors or values on the outcome of different linking strategies.

The advantages of a probabilistic strategy over a deterministic strategy have been shown by other studies. However, most of these studies have compared the performance of both strategies for one particular real life situation, where the match status was known or based on a unique identification number.[16;23-26] In our simulations we were able to vary the key parameters that influence the quality of a record linkage procedure and show the impact of these variations for the different linking strategies. If the deterministic strategy closely matches the patterns selected by the probabilistic strategy differences can be small [23], but differences easily become large beyond [16;25]. Gomatam et al.[24] found that a probabilistic

strategy as implemented in Automatch[27] gave higher sensitivity, but a lower positive predictive value as a deterministic strategy. However, by adjusting the threshold of the probabilistic strategy sensitivity can be sacrificed to obtain a higher positive predictive value. In the current study we only applied a deterministic full and deterministic N-1 strategy, while in literature stepwise deterministic linkage is also reported.[24;28] Stepwise deterministic linkage is equal to considering different deterministic N-1 or N-2 variants one at a time. The order in which the different variants are considered is guided by common sense. A probabilistic strategy is more versatile because it provides the ordering of all possible patterns of agreements and disagreements.

The results of our study show that the probabilistic strategy is superior and more flexible than the deterministic strategy in finding the optimal balance between keeping sufficient discriminative power and allowing disagreements to overcome registration errors. The final outcome of a probabilistic strategy is a set of patterns indicating (dis-)agreement on the linking variables that are accepted as link. This outcome can be implemented as a deterministic strategy. However, the advantage of the probabilistic strategy is its flexibility to adapt to changes in the source files over time. A drawback of the probabilistic strategy is that the algorithm requires fitting of statistical models. However, the latent class models in record linkage can be fitted with many standard statistical packages. For the current study we used a standard SAS statistical procedure to estimate the m_i and u_i probabilities. Also, commercial and freely available software is available to perform probabilistic record linkage.[29] In our view, the probabilistic strategy should be the standard approach for linking data from registries unless there is reliable (external) evidence which patterns of agreement and disagreement should be considered as links and non-links.
Future research should focus on how to examine and incorporate dependency among variables into the probabilistic linkage strategy. More research is also needed on how the additional information from the probabilistic strategy about the quality of a link (e.g. total weight) could be used in the subsequent analysis of the linked datasets, for instance by using weighted regression techniques.

Appendix

For the probabilistic strategy, the m_i and u_i probabilities and the prevalence of matches were estimated in a latent class model by using the observed frequency of patterns of agreement and disagreement on the linking variables among all pairs.[16] If the outcomes of the comparisons are independent between variables, the total log likelihood can be written as:

$$\sum_p n(\gamma^p)\left\{\log\left(\pi\prod_{i=1}^{k} m_i^{\gamma_i^p}(1-m_i)^{1-\gamma_i^p}+(1-\pi)\prod_{i=1}^{k} u_i^{\gamma_i^p}(1-u_i)^{1-\gamma_i^p}\right)\right\}$$

where m_i is the probability of agreement of the i^{th} variable among matches, u_i is the probability of agreement among non-matches, π is the proportion of true matches among all possible record combinations, $n(\gamma^p)$ the number of record pairs with pattern γ, γ_i^p is the outcome of the comparison of variable i (0,1) in the pattern p, for i = 1,…,k and p = 1,…,2^k. The number of parameters to be estimated equals 2xk+1, namely k m_i parameters and k u_i parameters and one prevalence parameter (π). For a dataset with k variables per record, there are 2^k unique agree/disagree comparison vectors. Maximum likelihood methods have been used to estimate the parameters of the equation.

References

1 Libby G, Smith A, McEwan NF, Chien PF, Greene SA, Forsyth JS, et al. The Walker Project: a longitudinal study of 48,000 children born 1952-1966 (aged 36-50 years in 2002) and their families. Paediatr Perinat Epidemiol 2004;18(4):302-12.

2 Nyren O, Yin L, Josefsson S, McLaughlin JK, Blot WJ, Engqvist M, et al. Risk of connective tissue disease and related disorders among women with breast implants: a nation-wide retrospective cohort study in Sweden. BMJ 1998;316(7129):417-22.

3 Westergaard T, Wohlfahrt J, Aaby P, Melbye M. Population based study of rates of multiple pregnancies in Denmark, 1980-94. British Medical Journal 1997;314(7083):775-9.

4 Dunn HL. Record Linkage. Am J Public Health Nations Health 1946;36(12):1412-6.

5 Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959;130:954-9.

6 Bell RM, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. Med Care 1994;32(10):1004-18.

7 Maizlish NA, Herrera L. A record linkage protocol for a diabetes registry at ethnically diverse community health centers. J Am Med Inform Assoc 2005;12(3):331-7.

8 Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. J Am Med Inform Assoc 1997;4(3):233-7.

9 Quantin C, Binquet C, Bourquard K, Pattisina R, Gouyon-Cornet B, Ferdynus C, et al. Which are the best identifiers for record linkage? Med Inform Internet Med 2004;29(3-4):221-7.

10 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(1):45-57.

11 Buescher PA. Method of linking Medicaid records to birth certificates may affect infant outcome statistics. Am J Public Health 1999;89(4):564-6.

12 Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA Symp 2002;305-9.

13 O'Reilly D, Rosato M, Connolly S. Unlinked vital events in census-based longitudinal studies can bias subsequent analysis. J Clin Epidemiol 2008;61(4):380-5.

14 Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association 1969;64(328):1183.

15 Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14(5-7):491-8.

16 Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidemiol 2007;60(9):883-91.

17 Newcombe HB. Handbook of record linkage: Methods for Health and Statistical Studies, Administration and Business. Oxford: Oxford University Press; 1988.

18 Newcombe HB, Fair ME, Lalonde P. Discriminating powers of partial agreements of names for linking personal records. Part I: The logical basis. Methods Inf Med 1989;28(2):86-91.

19 Oberaigner W, Stuhlinger W. Record linkage in the Cancer Registry of Tyrol, Austria. Methods Inf Med 2005;44(5):626-30.

20 Tromp M, Reitsma JB, Ravelli AC, Meray N, Bonsel GJ. Record Linkage: Making the most out of errors in linking variables. AMIA Annu Symp Proc 2006;779-83.

21 Tromp M, Meray N, Ravelli AC, Reitsma JB, Bonsel GJ. Ignoring Dependency between Linking Variables and Its Impact on the Outcome of Probabilistic Record Linkage Studies. J Am Med Inform Assoc 2008;15(5):654-60.

22 Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 1996;52(3):797-810.

23 Campbell KM. Impact of record-linkage methodology on performance indicators and multivariate relationships. J Subst Abuse Treat 2008.

24 Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med 2002;21(10):1485-96.

25 Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Annu Symp Proc 2003;259-63.
26 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(4):371-7.
27 Jaro MA. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association 1989;84(406):414-20.
28 Adams MM, Wilson HG, Casto DL, Berg CJ, McDermott JM, Gaudino JA, et al. Constructing reproductive histories by linking vital records. Am J Epidemiol 1997;145(4):339-48.
29 Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. Health Informatics J 2008;14(1):5-15.