Probabilistic Medical Record Linkage incorporating close agreement

M Tromp, N Méray, ACJ Ravelli, JB Reitsma, GJ Bonsel
Department of Medical Informatics; Department of Public Health Methods
University of Amsterdam
P.O. Box 22700, 1100 DE Amsterdam, The Netherlands
m.tromp@amc.uva.nl

Abstract

Probabilistic close medical record linkage techniques have been used to link the Dutch perinatal registries. Close agreement weights further enhance the probabilistic procedure. External validation showed that the developed procedure is highly reliable.

1 Introduction

In the Netherlands, four different caregivers (midwifes, general practitioners, obstetricians and paediatricians) are involved in perinatal care. Each caregiver collects data into their own national registry. These registries are anonymous and because of privacy laws in the Netherlands, no unique personal identifier exists.

A combined dataset of the separate registries is needed to produce valid reports on outcome measures of perinatal healthcare and to enable further data analysis. Medical record linkage is a technique to identify records belonging to the same individual in the absence of a unique identifier [Newcombe, 1988]. Partly identifying variables present in both datasets are combined to create a powerful discriminating linking key.

There are two different approaches to medical record linkage; deterministic medical record linkage and probabilistic medical record linkage. Deterministic strategies only look at (dis)agreement on the linking variables. For full deterministic linkage all linking variables have to agree, while N-1 deterministic linkage allows one of the linking variables to disagree [Newcombe, 1988]. Probabilistic medical record linkage strategies use the information value of the different linking variables by assigning a weight for agreement and disagreement for every variable. This weight is calculated using two probabilities: the probability that a variable agrees among matches (m_i) and the probability that a variable agrees among non-matches (u_i). The m_i value reflects the reliability of the variable, while the u_i value reflects the discriminating power of the variable [Newcombe, 1988; Bell et al.,1994].

This paper in short describes the applied probabilistic medical record linkage algorithm incorporating close agreement to link the registry of midwifes ("MR") and the registry of obstetricians ("OR").

2 Methods

The datasets of the perinatal registries were linked for the years 2001 - 2003. First, both datasets for one year were internally linked to detect administrative doubles using a deterministic N-1 approach. In the next step, the MR and OR were linked using a probabilistic medical record linkage algorithm. The two datasets were separated for singletons and twins. Records of multiple births require a different, stricter, approach because these records have a lot of variables in common. U_i probabilities were calculated from the marginal distribution in the two files as true non-matches make up the largest part of the total number of pairs. Because the matching status is unknown, m_i values were estimated using the Expectation Maximization (EM) algorithm with the observed patterns of agreement and disagreements of the singleton files without missing values [Felligi and Sunter, 1969; Reitsma, 1999]. Missing values influence the calculation of weights in an undesirable way because a missing value on a variable in both records is seen as agreement by the EM algorithm. Besides full agreement, close agreement was defined for certain variables. Even in case of agreement the value of a variable in two records can differ because of different calculation methods or rounding off of figures. Identification of these variables, and the definition of the close range was established with help of caregivers combined with information from the data (Table 1).

Because of large file sizes (about 160.000 records for MR, 125.000 for OR) blocking was applied in two steps. Records were only compared if they agreed on the blocking variable *date of birth from the mother* in the first step and *ZIP code of the mother* in the second step.

Table 2 list the variables there were compared in the linking procedure. For *date of birth* (DOB), *expected date of birth* and *birth weight* close ranges were defined.

Table 1 Effect of choice of close range on linking weight for the variable *birth weight*

Close range	Weight agree	Weight disagree
Birth weight (full)	7,99	
Birth weight (± 5g)	1,44	-4,16
Birth weight (± 10g)	**0,91**	**-4,44**
Birth weight (± 25g)	0,17	-4,70
Birth weight (± 50g)	-0,45	-5,17
Birth weight (± 100g)	-1,12	-5,76

Table 2 Linking variables with m_i and u_i value and linking weight for 2002 data

	m_i value	u_i value	weight agree	weight close agree	weight dis-agree
DOB mother	Blocking				
ZIP code mother	0,9573	0,0006	10,74		-4.55
DOB child	0,9780	0,0028	8,47	1,50	-7,28
(± 1 day)	0,0156	0,0055			
Exp. DOB child	0,8877	0,0027	8,36	1,35	-5,79
(± 7 day)	0,0949	0,0371			
Birth weight	0,9356	0,0037	7,98	0,91	-4,44
(± 10 gram)	0,0191	0,0102			
Place of birth	0,8818	0,0064	7,11		-3,07
Minute of birth	0,9173	0,0180	5,67		-3,57
Hour of birth	0,9701	0,9701	4,50		-5,00
Gravidity	0,9457	0,3016	1,65		-3,69
Gender	0,9918	0,5006	0,99		-5,93

Linkage weights of the different variables were calculated using the m_i and u_i values (Table 2):
Full agreement weight of the i^{th} variable: $\log_2 (m_{if}/u_{if})$
Close agreement weight of the i^{th} variable: $\log_2 (m_{ic}/u_{ic})$
Disagreement weight of the i^{th} variable:
$\log_2 (1-(m_{if}+m_{ic})/ 1-(u_{if}+u_{ic}))$

For every record pair a total linkage weight was calculated by adding up the individual variable weights. Linking weight was set to zero if a variable was missing in one or both records compared. All pairs were sorted by this total weight and a threshold value was determined separating links from non-links based on the estimated match rate by the EM algorithm and by reviewing pairs around this estimated threshold value.

A double blind external validation with the medical record as gold standard was carried out for the MR^OR linkage, focussing mainly on the uncertain area of the linkage (around the threshold value).

3 Results

Duplicates were removed from the separate registry files; 0.5% in the MR and 0.05% in the OR. Linking of the MR with the OR showed that 41% of all pregnancies were present in both files. Figure 1 shows all pairs sorted by total linkage weight together with the threshold value (15.4) above which value pairs are considered a link. External validation revealed no errors outside the uncertain region (weight of ±5 around threshold value) and a false margin of 13% in the uncertain region. Only 0.35% of the linked pairs are in the uncertain region and only 0.055 % of the non-linked pairs, which means that our total linkage procedure has a margin of error of less than 1 %.

4 Discussion

A perinatal healthcare data file now exists with combined data from the different involved disciplines, which can be used to produce valid tables on outcome measures and offers new possibilities for further data analysis. As external validation showed, the developed probabilistic close medical record linkage procedure is highly reliable.

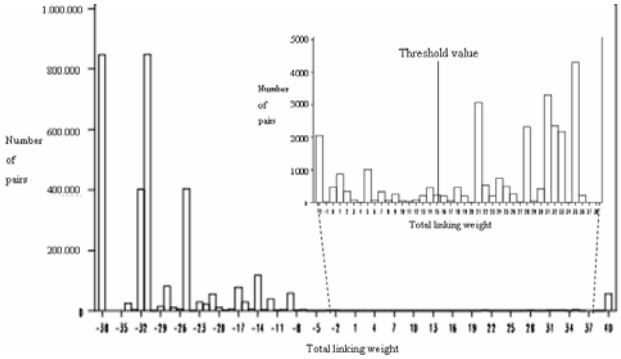

Figure 1 MR^OR singleton pairs sorted by total linking weight

The weights calculated for the linking variables proved to be very stable for variations in the number of linking variables and close ranges. Blocking had to be applied because of large file sizes, but we believe two step blocking minimizes the number of false negative links. Our decision the separate the linkage of singleton and multiple births worked well, although multiple births remain difficult to link. Choices made on handling missing values and defining close ranges should be further founded by simulation studies. SAS was used for all linking procedures and proved to be a flexible and powerful tool.

5 Further research

Additional simulation studies will further ground choices made so far in particular the dependency of their optimality on dataset characteristics (number of records, the ratio of possible links to file sizes and error rates of variables). Simulation studies conducted to refine the linking strategy will focus on the range of close agreement, the handling of missing values and dependencies between linking variables by using a simulated dataset with known match status. Yet a valid probabilistic linking procedure is now available and can be used for similar problems.

Acknowledgments

We gratefully acknowledge the support and funding of the SPRN (Foundation of the Dutch Perinatal Registry) and the investment of numerous caregivers.

References

[Bell *et al*., 1994] Bell RM, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. *Medical Care*, 32: 1004-1018, 1994.

[Felligi and Sunter, 1969] Felligi IP, Sunter AB. A theory for record linkage. *Journal of the Americal Statistical Association*, 64: 1183-1210, 1969.

[Newcombe, 1988] Newcombe HB. *Handbook of record linkage: methods for health and statistical studies, administration, and business.* Oxford: Oxford University Press, 1988.

[Reitsma, 1999] Reitsma JB. *Registers in Cardiovascular Epidemiology*. PhD thesis Academic Medical Center, University of Amsterdam, 1999. ISBN 90-9013206-6.