Record linkage to enhance data from perinatal registries
Tromp, M.

Citation for published version (APA):
Tromp, M. (2009). Record linkage to enhance data from perinatal registries

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

Record linkage:
Making the most out of errors in linking variables

M Tromp, JB Reitsma, ACJ Ravelli, N Méray, GJ Bonsel

American Medical Informatics Association Annual Symposium Proceedings 2006; 779-783.
Abstract

This paper presents a refinement of the probabilistic medical record linking algorithm. We introduced ‘close agreement’ to account for typical errors in administrative variables used for record linkage. Linking data on early pregnancy determinants with data on late child outcomes was used as a case study. We analyzed whether the addition of close agreement resulted in a higher discriminating power of the linking key reflected in a reduction of the number of links with an uncertain linking status. Incorporating close agreement for postal code and date of birth in the record linking algorithm resulted in a reduction of 95% of the number of pairs in the uncertain region. We showed that the extension of a third outcome ‘close’ when comparing values of corresponding linking variables led to a major improvement in our probabilistic record linkage study. Similar improvements are likely in other studies because the frequency, nature, and type of errors in other large databases will not be substantially different.
Medical record linkage techniques are commonly applied when data from different sources, if combined, provide an answer to a clinical or public health question. For instance, linking electronic health records of patients stored in multiple databases and registries. In the absence of a unique identifier in these data sources, medical record linkage techniques are used to identify records belonging to the same individual.\textsuperscript{1} Partly identifying variables present in both datasets are combined to create a powerful discriminating, yet anonymous, linking key. Partially identifying variables are combined to create a powerful discriminating, yet anonymous, linking key. Probabilistic medical record linkage strategies use the information value of the linking variables by assigning a weight for agreement and a weight for disagreement for every variable separately. These weights are based on 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 (1-error rate), while the $u_i$ value reflects the discriminating power of the variable (1-chance agreement).\textsuperscript{1,2} These are the two fundamental concepts in record linkage: error rate and discriminating power. Together they influence the number of incorrect decisions in a linking procedure: a low discriminating power increases the number of false links and a high error rate increases the number of false non-links. The discriminating power of the linking key is directly correlated with the probability that two different persons have the same values on the linking key. The (theoretical) number of possible values of the linking variables, and their (empirical) distribution determines this probability. A variable with a large number of possible values and a uniform distribution has a high discriminating power, reflected by a low $u_i$ probability. Errors in variables will lead to a difference in the value of a corresponding variable among matches thereby lowering the $m_i$ probability.

Differences in values between corresponding variables can have many underlying causes: errors created during collection of the data, during the data entry or true changes of a value. Making errors depends on the type of variable, the type of data collection and the way of data entry. If the occurrence or the origin of errors has a certain pattern and is therefore predictable, this knowledge can be incorporated in the record linking algorithm to improve the outcome of the linkage. For example: a date variable is more prone to error than a dichotomous variable (predictability based on type of variable). Data entry errors might occur more often at the end of the day than at the beginning of the day (predictability in occurrence). In case of error, the new value is often dependent on the original value (predictability in origin). This last type of predictability of errors in linking variables is the topic of this paper. This dependency can exist in many different forms: the literal distance (birth weight weighted on two scales), but also a figurative distance (different spelling of the same name, keyboard errors, transposition of day and month in a date). Soundex algorithms have been used to account for differences in spelling of names.\textsuperscript{3} Partial agreement on linking variables has been used in the manual review process of a linkage procedure or as a further refinement to resolve ties among links.\textsuperscript{4,5} A typical example of a relevant clinical question asking for combined data is to investigate the relation between early determinants and late outcomes. Profession based medical registries do not always cover this wide horizon, because they are limited to the care provided. We use linking data on pregnancy determinants and child outcome (neonatal and infant mortality) as a case study to account for errors in linking variables.
Chapter 2

In the Netherlands data on perinatal care are registered in four profession-based anonymous national registries: the LVR1-registry (midwives), the LVRh-registry (GP’s) the LVR2-registry (gynaecologists) and the LNR-registry (paediatricians and neonatologists). Low level care is provided by midwives/general practitioners and high level care by gynaecologists. The level of care might change (in two directions) during pregnancy, delivery and/or postpartum period. Admissions of newborns are recorded until 28 days after term date in the registry of paediatricians and neonatologists. The perinatal registries have been combined using probabilistic medical record linkage strategies to form one perinatal registry with records containing combined information on mother, pregnancy, birth, the postpartum period and the child.

The Dutch Population Register is a national assembly of all municipal registers and contains information on address and family relations of Dutch residents. A live born is registered along with its parents. The birth and mortality statistics of Statistics Netherlands are based on this Population Register. The causes of death register is legally held by Statistics Netherlands. In addition, a separate national register exists of stillborns of ≥ 24 weeks. Statistics Netherlands combines this register with information from the causes of death register for the statistics on stillborns. Because of privacy laws in the Netherlands the population and stillborn registers are anonymous on a national level.

In a pilot study (2005-2006), we tried to combine the perinatal registry data and the population register data (including mortality statistics) in order to produce valid figures on long term outcomes neonatal (1st month) and infant (1st year mortality) by early determinants (gestational age and birth weight) for the Netherlands. To overcome the limited discriminating power of our linking key, we introduced ‘close agreement’ to detect and account for typical errors in administrative variables. ‘Close agreement’ is the situation where two values are not in perfect agreement, but are close to each other. The definition of ‘close’ depends on the nature of the involved variable (for instance comparing dates of birth versus comparing birth weights) and on knowledge about the characteristics of the entry procedure, e.g. the type of error checks that are in place. We then investigated if the discriminating power of the linking key increased by adding close agreement, reflected by a reduction of the number of links with an uncertain linking status.

2.2 Methods

We used the data files of the year 2001 for this study as this was the first available linked perinatal registry file (indicated by date of birth child). Names were not included. The perinatal registry contained 202,904 records on 188,628 liveborns and 1,545 stillborns for 2001. The remaining records did not include pregnancy outcome. The data files of the national population register and the stillborn register that we used contained records of 205,525 liveborns and 1,261 stillborns for 2001.

We started by building a basic probabilistic record linkage algorithm that only includes two options with respect to comparing corresponding linking variables: variables either agree or disagree. Because linking variables differed considerably with respect to their discriminatory value and their likelihood of containing error, we calculated different \( m_i \) and \( u_i \) values for each variable. \( u_i \) 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 disagreement of the singleton files.\(^7\)\(^8\) We excluded records with missing values on one of the linking variables when estimating weights, because otherwise pairs with missing values in both records will be included in the weight estimation of agreement. If the outcomes of the comparisons are independent between variables, the total log likelihood can be written as:

\[
\sum_{p} n(p) \left\{ \log \left( \pi \prod_{i=1}^{k} m_i^{p_i}(1-m_i)^{(1-p_i)} + (1-\pi) \prod_{i=1}^{k} u_i^{p_i}(1-u_i)^{(1-p_i)} \right) \right\}
\]

(Equation 1)

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, \( \pi \) is the proportion of true matches among all possible record combinations, \( n(p) \) the number of record pairs with pattern \( p \), \( \gamma_i \) is the outcome of the comparison of variable \( i \) in the pattern \( p \), for \( i = 1, \ldots, k \) and \( p = 1, \ldots, 2^k \).

Using the \( m_i \) and \( u_i \) probabilities calculated in equation 1, we calculate the weights for agreement and disagreement as:

- Agreement weight of the \( i \)th variable = \( \log_2 \frac{m_i}{u_i} \).
- Disagreement weight of the \( i \)th variable = \( \log_2 \frac{1-m_i}{1-u_i} \).

To formally incorporate our prior knowledge on errors we defined a so called 'close' agreement. The underlying assumption is that close agreement is more likely to occur among matches if an error is being made than can be expected by chance alone. Thus, close agreement is a third possible outcome of variable comparison.

Incorporating this third outcome option into equation 1, leads to the following definition of the total log likelihood:

\[
\sum_{p} n(p) \left\{ \log \left( \pi \prod_{i=1}^{k} m_i^{p_i} \gamma_i^{p_i} (1-m_i)^{(1-p_i)} (1-\pi) \prod_{i=1}^{k} u_i^{p_i} \gamma_i^{p_i} (1-u_i)^{(1-p_i)} \right) \right\}
\]

where \( I(\varphi) \) is the indicator function, i.e. \( I(\varphi)=0 \) if \( \varphi \) is false and \( I(\varphi)=1 \) if \( \varphi \) is true, \( m_i \) is the probability of full agreement of the \( i \)th variable among matches, \( mc_i \) is the probability of close agreement among matches, \( uf_i \) is the probability of full agreement among non-matches and \( uc_i \) is the probability of close agreement among non-matches. The weight for close agreement is calculated the same way as the weight for full agreement: \( \log_2 \frac{mc_i}{uc_i} \).

The weight for disagreement can now be calculated as: \( \log_2 \frac{1-mf_i - mc_i}{1-uf_i - uc_i} \).

For every record pair a total linkage weight is calculated by adding up all 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 their 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. The small area around the threshold is called the ‘grey area’ because of the uncertainty of linking status in this area (figure 2.1). Pairs with a weight higher than the upper boundary of the grey area are certain links and under the lower boundary are certain non-links. If adding close agreement for linking variables increases the discriminating power of the linking key and thus better separates the matches from non-matches, the number of pairs within the grey area should decrease (depicted by the dotted line in figure 2.1).

The six available linking variables were: date of birth mother, postal code mother (4 digits), date of birth child, gender child, multiple birth status (number of children per pregnancy) and mortality. Multiple birth status was used to split both files as we conducted the linkage separately for singletons and multiple births. Only singleton records were selected for this paper on the potential utility of close. Death is rare and therefore we could not use this variable as a linking variable, but instead we used it afterwards as a control. Because of large file sizes, blocking was mandatory (variable date of birth mother). In a second step we used postal code as blocking variable to account for errors in date of birth mother (results not shown). We estimated and compared the $m_i$ and $u_i$ probabilities for the linking variables in the situation without and with close agreement. Within the close approach, we compared several options and the absolute and relative differences in $m_i$ and $u_i$ values between exact agreement, close, and disagreement. Of all these differences, we examined which outcome had the biggest impact on the size of the grey area.

### 2.3 Results

Table 2.1 shows the $m_i$ and $u_i$ probabilities and the associated linking weights for agreement and disagreement for postal code (4 digits), date of birth child and gender. Postal code has the highest discriminating power resulting in a weight for agreement of 10.76. The weight

![Figure 2.1 Number of pairs by their total linking weight for matches and non-matches with and without close agreement (fictional data).](image-url)
for agreement for gender is only 0.99 because of the low discriminating power of gender. The maximum linking weight is 20.23 if all variables agree. Based on the estimated prevalence of matches we set the threshold value at a total weight of 7, grey area ranging from 4 to 10. The total number of pairs we assigned as matches was 173,875 pairs.

Table 2.1 Reliability (m<sub>i</sub> probability) and discriminating power (u<sub>i</sub> probability) with associated weights, for linking variables to link the perinatal registry with the population register for the year 2001.

<table>
<thead>
<tr>
<th>Variable</th>
<th>m&lt;sub&gt;i&lt;/sub&gt;</th>
<th>u&lt;sub&gt;i&lt;/sub&gt;</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of birth mother</td>
<td>0.9592</td>
<td>0.00055</td>
<td>10.76</td>
<td>-4.61</td>
</tr>
<tr>
<td>Date of birth child</td>
<td>0.9893</td>
<td>0.00278</td>
<td>8.48</td>
<td>-6.54</td>
</tr>
<tr>
<td>Gender child</td>
<td>0.9960</td>
<td>0.00010</td>
<td>0.99</td>
<td>-6.93</td>
</tr>
</tbody>
</table>

Table 2.2 shows three different types of close agreement for postal code with their u<sub>i</sub> and m<sub>i</sub> probabilities. The first type of close agreement allows one of the four digits to differ (typing error). There are now three possible outcomes of variable comparison: full agreement, close agreement and disagreement. This close value shows a change in linking weight for values being close from disagreement weight of -4.61 to close agreement weight of +1.50 (increase in linking weight of +6.11). The weight for disagreement changes from -4.61 to -5.56, as disagreement is now everything besides full agreement and close agreement (decrease in linking weight of -0.95). The second type of close agreement allows only the fourth digit to differ, with the assumption that errors in the last digit are less apparent than errors in one of the other digits. This close agreement shows a change in linking weight for values being close of +7.59 and for disagreement of -0.7. The third type of close agreement allows the reversal of two digits within the four digit postal code.

Table 2.2 Reliability (m<sub>i</sub> probability) and discriminating power (u<sub>i</sub> probability) of three types of close agreement for postal code.

<table>
<thead>
<tr>
<th>Variable</th>
<th>m&lt;sub&gt;i&lt;/sub&gt;</th>
<th>u&lt;sub&gt;i&lt;/sub&gt;</th>
<th>(close) Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postal code (4 digits)</td>
<td>0.95920</td>
<td>0.00055</td>
<td>10.76</td>
<td>-4.61</td>
</tr>
<tr>
<td>1)</td>
<td>0.95920</td>
<td>0.00059</td>
<td>10.66</td>
<td>-5.56</td>
</tr>
<tr>
<td>2) Postal code</td>
<td>0.95910</td>
<td>0.00059</td>
<td>10.66</td>
<td>-5.31</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt; digit differs</td>
<td>0.01580</td>
<td>0.00200</td>
<td>2.98*</td>
<td></td>
</tr>
<tr>
<td>3)</td>
<td>0.95920</td>
<td>0.00059</td>
<td>10.66</td>
<td>-4.65</td>
</tr>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt; and 2&lt;sup&gt;nd&lt;/sup&gt; digit reversed</td>
<td>0.00009</td>
<td>0.00013</td>
<td>-0.56*</td>
<td></td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt; and 3&lt;sup&gt;rd&lt;/sup&gt; digit reversed</td>
<td>0.00008</td>
<td>0.00019</td>
<td>-1.29*</td>
<td></td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt; and 4&lt;sup&gt;th&lt;/sup&gt; digit reversed</td>
<td>0.00086</td>
<td>0.00013</td>
<td>2.76*</td>
<td></td>
</tr>
</tbody>
</table>

* Close agreement weights.

Reversal of the third and fourth digit shows the largest shift of +7.37 for values being close and -0.04 for disagreement. This can be explained by the fact that an error in the last two digits (indicating the neighborhood) is less apparent than an error in the first two digits (indicating the region in the Netherlands).
Table 2.3 shows close agreement types for the variable date of birth child: a difference of 1 day, 2 days and 1 month, with the $u_i$ and $m_i$ probabilities. The close agreement of 1-day difference is the only close agreement with a high discriminating power: the weight for being close is changed by +6.94 and the punishment for disagreement by -1.57.

Table 2.3 Reliability ($m_i$ probability) and discriminating power ($u_i$ probability) of three types of close agreement for date of birth child.

<table>
<thead>
<tr>
<th></th>
<th>$m_i$</th>
<th>$u_i$</th>
<th>(close)Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Date of birth child</td>
<td>0.98930</td>
<td>0.00278</td>
<td>8.47</td>
<td>-6.84</td>
</tr>
<tr>
<td>± 1 day</td>
<td>0.98920</td>
<td>0.00278</td>
<td>8.48</td>
<td>-8.11</td>
</tr>
<tr>
<td>2) Date of birth child</td>
<td>0.98930</td>
<td>0.00278</td>
<td>8.48</td>
<td>-6.62</td>
</tr>
<tr>
<td>± 2 days</td>
<td>0.00064</td>
<td>0.00541</td>
<td>-3.08*</td>
<td>-3.08*</td>
</tr>
<tr>
<td>3) Date of birth child</td>
<td>0.98930</td>
<td>0.00278</td>
<td>8.48</td>
<td>-6.63</td>
</tr>
<tr>
<td>± 1 month</td>
<td>0.00069</td>
<td>0.00485</td>
<td>-2.82*</td>
<td>-2.82*</td>
</tr>
</tbody>
</table>

* Close agreement weights.

Table 2.4 finally demonstrates the benefit in terms of grey area reduction of the best close agreement of Table 2.2 and 2.3. In this example we fixed the threshold at a total weight of 7, with a grey area range of 4-10. Adding close agreement for date of birth child of a 1-day difference reduces the grey area by 24%. Adding close agreement for postal code of a difference in the fourth digit or a reversal of the third and fourth digit reduces the grey area by 71%. When combined, the addition of these close agreements reduces the number of pairs in the grey area by 95%. The addition of close agreement for date of birth child and postal code lifted 1,660 links from the uncertain linked area over the upper boundary of the grey area (10%).
Table 2.4 Grey area reduction after adding close agreement for date of birth child, postal code and for date of birth and postal code for the linking of singleton files.

<table>
<thead>
<tr>
<th>Linkage strategy</th>
<th>From uncertain not linked (weight 4-7) to certain not linked (weight &lt;4)</th>
<th>Grey area (uncertain region) weight 4-10*</th>
<th>From uncertain linked (weight 7-10) to certain linked (weight &gt;10)</th>
<th>Number of links weight &gt;7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic probabilistic linkage</td>
<td>NA</td>
<td>16,465</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Close agreement for date of birth child (± 1 day)</td>
<td>3,504</td>
<td>12,523</td>
<td>76%</td>
<td>438</td>
</tr>
<tr>
<td>Close agreement for postal code (error 4th digit, reversal of 3rd and 4th digit)</td>
<td>10,417</td>
<td>4,826</td>
<td>29%</td>
<td>1,222</td>
</tr>
<tr>
<td>Close agreement for date of birth and postal code</td>
<td>13,908</td>
<td>897</td>
<td>5%</td>
<td>1,660</td>
</tr>
</tbody>
</table>

NA = not applicable.
* The number of record pairs in the grey area with basic probabilistic linkage was set to 100%.
2.4 Discussion

This paper shows, within the context of linking perinatal data, that the addition of ‘close agreement’ as a third outcome in a probabilistic linkage algorithm substantially improves the performance. Incorporating close agreement in the linkage algorithm corrects for typical errors in linking variables. Similar improvements are likely to be expected in other studies because the frequency, nature, and type of errors in other large databases will not be substantially different. Although prior knowledge is required to define close, the testing of the added value (if any) could be executed along existing statistical procedures. The same formal criteria could be applied to estimate the \( m \) and \( u \) probabilities for close agreement. The gain of close could be expressed as a grey area reduction which of course involves an arbitrarily decision regarding its boundaries. The surplus value of incorporating close agreement will be most apparent in situations with few linking variables that are sensitive for errors.

So far, no manual validation has been conducted for this pilot study. For previous conducted linking studies using a similar approach, validation showed good results. The linking weights were very stable regardless of perfect or close agreement and misclassification of links will be about the same (< 1%). Adding close agreement requires prior knowledge. For content variables if used as linking variable, \( \text{e.g. birth weight}, \) this includes knowledge on habits and customs as close is not random. It is still arbitrarily when to include additional types of close agreement, therefore further research will focus on deciding for optimal close values for linking variables.

The grey area reduction is a simple device to demonstrate the added value of close agreement. Application of close agreement can be anticipated in many circumstances replacing manual or semi-automatic record by record reviewing.

2.5 Conclusion

The extension of existing dichotomous perfect agreement/disagreement with a third category ‘close’ and incorporation in the record linking algorithm represents an improvement in probabilistic medical record linking expressed in a grey area reduction of 95% in our example.

Acknowledgements

We gratefully acknowledge the support and funding of the SPRN (Foundation of the Dutch Perinatal Registry), the partners in the pilot study of Statistics Netherlands (W.P. Schaasberg, A. de Bruin and M. Berger-van Sijl) and the investment of numerous caregivers in the Netherlands.
References


