

NEONATAL MORTALITY BY GESTATIONAL AGE AND BIRTH WEIGHT IN ITALY, 1998-2003: A RECORD LINKAGE STUDY

Cristiano Marini¹, Alessandra Nuccitelli²

ABSTRACT

Neonatal mortality rates by gestational age and birth weight category are important indicators of maternal and child health and care quality. However, due to recent laws on administrative simplification and privacy, these specific rates have not been calculated in Italy since 1999. The main aim of this work is to assess the possibility of retrieving information on neonatal mortality by the linkage between records related to live births and records related to infant deaths within the first month of life, with reference to 2003 and 2004 birth cohorts. From a strict methodological point of view, some critical aspects of the most used record linkage approach are highlighted: specific problems may arise from the choice of records to be linked if there are consistency constraints between pairs (in this context, one death record can be linked to at most one birth record). In the light of considerations on the quality of the starting data, the retrieval of information on neonatal mortality by gestational age and birth weight is restricted to Northern Italy. Specific neonatal mortality rates are provided with reference to 2003 and discussed with particular emphasis on quality issues in the data collection processes.

Key words: administrative data, Boolean linear programming, data quality, greedy algorithm

1. Introduction

Neonatal mortality rates by gestational age and birth weight category are important indicators of maternal and child health (Zeitlin *et al.*, 2003) and care quality (Horbar, 1999)³. However, due to recent laws on administrative

¹ University of Rome “La Sapienza”, Italy, e-mail: cristiano.marini@uniroma1.it.

² National Statistical Institute, Italy, e-mail: nuccitel@istat.it.

³ A neonatal death is the loss of a live-born infant within the first month of life. The neonatal mortality rate is the number of deaths within the first month of life per 1,000 live births.

simplification and privacy, these specific rates have not been calculated in Italy since 1999.

The time-honoured system of birth registration managed by Istat, the National Statistical Institute, was dismantled in 1998 and later rebuilt entrusting it to the Ministry of Health. Istat has remained in charge of the register of deaths; the transfer of birth certificates from the Ministry to Istat has been only permitted after deletion of personal identifiers. Thus, the linkage of birth certificates - containing crucial information, such as birth weight and gestational age - to the corresponding infant death records (if any) has turned more difficult.

The main aim of this work is to assess the possibility of retrieving information on neonatal mortality by a linkage between records related to live births and records related to neonatal deaths, with reference to 2003 and 2004 birth cohorts.

The paper is organized as follows. Some quality issues regarding the data used for the linkage are discussed in the next section, with special emphasis on the coverage for the new medical birth register entrusted to the Ministry of Health. The coverage of neonatal deaths can be assumed to be total. An overview of record linkage problems is provided in section 3; some critical aspects of the most used approach to choose records to be linked, when there are consistency constraints between record pairs, are highlighted. The strategy adopted for the matching is described in section 4. Some results from a provisional linkage at national level are reported in section 5; in the light of such preliminary results, the retrieval of information on neonatal mortality is restricted to Northern Italy. In an attempt to give an idea of the trend of the phenomenon at least for this area, specific neonatal mortality rates are provided with reference to 2003, the first year from which the time series calculation can be resumed (section 6). Finally, some conclusions are drawn.

2. Data quality issues

2.1. Coverage of births

Data on births are collected through a medical register entrusted in 2002 to the Ministry of Health, rather than to Istat, as it had been up to 1998.

This change has caused some organizational problems, and in 2004 the coverage⁴ for the new registration system was still about 86 percent at national level (Ministero della Salute, 2007). Over the period 2002-2004 the coverage showed a strong heterogeneity across regions, with a tendency to decrease from north to south.

⁴ The coverage is measured by the percent ratio of number of birth certificates to number of deliveries registered by the hospital discharge system.

2.2. Quality of information used for the linkage

For the sake of simplicity, let A and B be two data files containing n_A and n_B records, respectively. Each record corresponds to an entity (or individual) and consists of several fields (or variables).

The aim of record linkage is to partition the set of all record pairs $A \times B = \{(a, b); a \in A, b \in B\}$ into two disjoint subsets: M , the set of matches, and U , the set of non-matches. Generally, a record linkage procedure classifies the record pairs on the basis of results from the comparison between corresponding fields (or matching variables) common to both files.

Matching variables may be subject to errors and omissions and usually have a different identifying power. For example, a field such as *sex* only has two value states and consequently could not impart enough information to identify the records related to the same individual. Conversely, a field such as *municipality of birth* imparts much more information, but it may frequently be reported incorrectly. Besides, some variables may change their values during the course of a person's lifetime, making more difficult to recognize records related to the same individual (e.g. *municipality of residence*).

Rather than considering all pairs $A \times B$, comparisons are sometimes restricted to those records that agree on some matching variables (blocking variables). Clearly, in this case the reliability and efficiency of a record linkage procedure is highly dependent upon the way in which blocking is carried out. Errors in blocking variables can result in a failure to compare records corresponding to the same entity.

The matching variables used in retrieving information on neonatal mortality are:

- a. *infant's sex*;
- b. *plurality of pregnancy*;
- c. *infant's day of birth*;
- d. *infant's month of birth*;
- e. *infant's municipality of birth*;
- f. *infant's province of birth*;
- g. *mother's day of birth*;
- h. *mother's month of birth*;
- i. *mother's year of birth*;
- j. *mother's municipality of residence*;
- k. *mother's province of residence*;
- l. *mother's nationality*.

In such a framework, the variables concerning mother's residence and nationality can be considered virtually unchanging over the lifetime of the infants of interest (at most, one month). The quality of the data to be linked can be

assessed in terms of incidence of invalid or missing values for each matching variable, with reference to 2003 and 2004 live birth cohorts⁵ (Tables 1 and 2).

As displayed in Table 1, the accuracy of the neonatal death data is not high. The quality is poor for the matching variables concerning mother's date of birth and *plurality of pregnancy*. In particular, at national level unacceptable values for these variables increase respectively from 37 and 0 percent, for 2003 cohort, to 42 and 30 percent, for 2004 cohort. The incidence of invalid or missing values for Northern Italy is lower overall than for the whole country.

Table 1. Invalid or missing values (absolute and percent frequencies) in the neonatal death data for Italy and Northern Italy, by matching variable and infant's year of birth

matching variable	invalid or missing values							
	2003				2004			
	Italy		Northern Italy		Italy		Northern Italy	
	abs. fr.	%	abs. fr.	%	abs. fr.	%	abs. fr.	%
a	-	-	-	-	-	-	-	-
b	2	0.13	1	0.18	465	30.45	129	24.57
c	-	-	-	-	-	-	-	-
d	-	-	-	-	-	-	-	-
e	23	1.51	5	0.88	-	-	-	-
f	14	0.92	3	0.53	-	-	-	-
g	563	36.85	171	30.00	640	41.91	163	31.04
h	563	36.85	171	30.00	640	41.91	163	31.04
i	558	36.52	169	29.65	640	41.91	163	31.04
j	92	6.02	34	5.96	2	0.13	-	-
k	84	5.50	34	5.96	2	0.13	-	-
l	60	3.93	18	3.16	14	0.92	6	1.14

The birth data appear more accurate than the neonatal death ones (Table 2). In particular, with regard to *plurality of pregnancy* and *mother's nationality*, invalid or missing values decrease respectively from 12 and 3 percent, for 2003 cohort, to 1 and 2 percent, for 2004 cohort, while unacceptable values for the variables concerning the place of birth tend to disappear. However, the quality of *infant's month of birth* and of almost all other information on the infant's mother declines for the 2004 cohort. Compared to the whole country, for Northern Italy the infant data are more accurate, while the incidence of unacceptable values for the variables regarding the infant's mother is slightly higher.

⁵ Other types of mistake in the data are not considered, since no subset of records for which the matching variables are observed without errors is available.

Table 2. Invalid or missing values (absolute and percent frequencies) in the birth data for Italy and Northern Italy, by matching variable and infant's year of birth

matching variable	invalid or missing values							
	2003				2004			
	Italy		Northern Italy		Italy		Northern Italy	
	abs. fr.	%	abs. fr.	%	abs. fr.	%	abs. fr.	%
a	598	0.13	263	0.12	530	0.11	143	0.06
b	55,138	12.17	39	0.02	4,447	0.94	324	0.14
c	558	0.12	153	0.07	313	0.07	147	0.07
d	738	0.16	157	0.07	12,464	2.62	147	0.07
e	1,019	0.22	36	0.02	44	0.01	21	0.01
f	1,022	0.23	39	0.02	44	0.01	20	0.01
g	2,466	0.54	1,450	0.66	2,663	0.56	1,997	0.89
h	2,743	0.61	1,484	0.67	2,892	0.61	2,015	0.90
i	3,157	0.70	1,629	0.74	6,602	1.39	2,177	0.97
j	4,925	1.09	1,586	0.72	7,077	1.49	3,565	1.59
k	6,203	1.37	2,227	1.01	8,128	1.71	4,040	1.80
l	14,710	3.25	2,812	1.27	8,002	1.69	4,790	2.14

3. Constrained record linkage problems

Specifying a record linkage procedure requires:

- a measure of similarity between records;
- a decision rule using this measure for deciding when to classify a record pair as a link.

If k matching variables $X_1, X_2, \dots, X_j, \dots, X_k$ are compared for the record pair (a, b) , the results can be expressed by a k -dimensional comparison vector representing the level of agreement between the records $a \in A$ and $b \in B$:

$$\mathbf{y}_{ab} = (y_{ab,1}, y_{ab,2}, \dots, y_{ab,j}, \dots, y_{ab,k}).$$

The easiest and most used way of defining the component j ($j=1, 2, \dots, k$) of $\mathbf{y}_{ab}$ is to assume that $y_{ab,j} = 1$, if a and b agree on the field j , and $y_{ab,j} = 0$, if a and b disagree on j or at least one value for j is missing.

Once the comparison vector is defined, the second step is to decide how $\mathbf{y}_{ab}$ can be used to classify the record pairs into M or U . According to the most usual approach, $\mathbf{y}_{ab}$ can be given a weight w_{ab} on which to base the decision rule for

(a, b) :

$$\begin{aligned}
 &\text{if } w_{ab} \geq t_u && (a, b) \text{ is classified as a link;} \\
 &\text{if } t_l \leq w_{ab} < t_u && (a, b) \text{ is classified as a potential link and the final} \\
 & && \text{decision will be only taken after manual review} \\
 &\text{if } w_{ab} < t_l && (a, b) \text{ is classified as a non-link.}
 \end{aligned} \tag{1}$$

The estimation of w_{ab} , $(a, b) \in A \times B$, and the selection of the threshold values t_l and t_u ($t_l \leq t_u$) are crucial steps in defining a record linkage procedure. In a very straightforward way, weights and thresholds can be specified according to a deterministic criterion, with a view to the specific goals of the matching⁶.

In the decision rule proposed by Fellegi e Sunter (1969) the weights are defined in a probabilistic way by means of the log-likelihood ratio:

$$w_{ab} = \log \left(\frac{P(y_{ab} | (a, b) \in M)}{P(y_{ab} | (a, b) \in U)} \right) \quad (a, b) \in A \times B \tag{2}$$

Most software for probabilistic record linkage uses the Expectation-Maximization (EM) algorithm to estimate (2), while t_l and t_u are usually selected in an empirical way, by inspection of the distribution of the estimated weights (Jaro, 1989; Winkler, 2000). Besides, some software enables to adjust such estimates for relative frequencies of specific values observed for the matching variables, in order to better take into account of their different identifying power⁷.

In most applications, a one-to-many (or one-to-one) matching is required, depending on whether each entity is represented at most once in one file (or in each file) to be matched. When the decision rule is defined separately for each pair, as in (1), this requirement leads to one of the following methodological questions. First, there is the issue of how to modify the weight estimation method to incorporate the requirement. The second issue is how to solve the assignment problem arising in a one-to-many (or one-to-one) matching when a record in one file (or in each file) is included in more than one record pair classified as a link or a potential link by the decision rule.

⁶ For instance, a very simple deterministic criterion may consist in setting the weights equal to the number of agreements on the matching variables and both t_l and t_u equal to $k - 1$ (i.e. only the record pairs with at least $k - 1$ agreements are classified as links).

⁷ For example, with regard to *infant's municipality of birth*, an agreement observed on a small municipality (in terms of population) enables to identify the records related to the same individual more easily than an agreement observed on a larger municipality.

Since current implementations of the EM algorithm usually do not incorporate any explicit mechanism to force the matching requirement, additional procedures based on operational research techniques are needed to solve the assignment problem.

Relatively early software for record linkage introduced greedy algorithms to solve the assignment problem (Hill and Pring-Mill, 1985). With reference to the one-to-one case, a greedy algorithm usually works by ordering the record pairs by matching weight. At each step, the record pair with the highest weight is retained and included in the final assignment scheme. All other record pairs that conflict with this best pair, meaning that they include records that are also in the best pair, are discarded. This is repeated for all pairs whose weights are not less than t_l or t_u .

Some recent computer software uses the core of the algorithm proposed by Jaro (1989) to force the matching requirement. Specifically, the optimal assignment scheme for the one-to-one case can be obtained as the solution of the following Boolean linear programming problem⁸:

$$\begin{aligned}
 &\text{maximize} && \sum_{a=1}^{n_A} \sum_{b=1}^{n_B} w_{ab} z_{ab} \\
 &\text{subject to:} && \sum_{a=1}^{n_A} z_{ab} \leq 1 \\
 &&& b=1, \quad 2, \quad \dots, \quad n_B \\
 &&& \sum_{a=1}^{n_B} z_{ab} \leq 1 && a=1, \quad 2, \quad \dots, \quad n_A,
 \end{aligned} \tag{3}$$

where $z_{ab}=1$, if a is assigned to b , and $z_{ab}=0$, if a is not assigned to b , for $(a,b) \in A \times B$. From now on, this approach will be also called global. According to Jaro, once the optimal assignment scheme is obtained, an assigned pair can be classified as a link if its weight is not less than t_u .

Armstrong e Saleh (2000) found the algorithm introduced by Jaro to produce incorrect links in practice. In order to give here a general explanation for that, consider the records $a, a' \in A$ and $b, b' \in B$ and all the possible record pairs (a,b) , (a,b') , (a',b) , (a',b') whose weights are related as follows:

$$\begin{aligned}
 w_{ab} + w_{a'b'} &> w_{ab'} + w_{a'b}, \\
 w_{ab'} &> w_{ab}, w_{a'b'}, w_{a'b}.
 \end{aligned}$$

⁸ With reference to the one-to-many case (one record in A can be linked to several records in B , but one record in B can be linked to at most one record in A), the number of constraints decreases:

$$\begin{aligned}
 &\text{maximize} && \sum_{a=1}^{n_A} \sum_{b=1}^{n_B} w_{ab} z_{ab} \\
 &\text{subject to:} && \sum_{a=1}^{n_A} z_{ab} \leq 1 && b=1, \quad 2, \quad \dots, \quad n_B.
 \end{aligned}$$

Suppose that all the weights are not less than t_u and each entity can be represented at most once in each file to be matched. According to the global approach, (a, b') is not labelled as a link, despite its highest weight. Such a final decision may turn out to be unreliable, especially if a and b' agree perfectly on all matching variables.

It is worth noting that the assignment scheme obtained by maximizing the sum of weights depends upon the numerical values of the weights involved and not only on their order.

As a consequence, this scheme proves to be even less robust to possible biases in estimates of weights⁹ than that resulting from a greedy strategy. Besides, the widespread use of logarithms in defining weights (2) can affect final results in a global approach.

This approach may perform well in record linkage applications characterized by very reliable matching variables with a high identifying power (e.g. *surname* or *address*), as in the case study reported by Jaro (1989). However, in other situations in which most of matching variables are subject to frequent errors or missing values and have a low identifying power, it is advisable to ensure the consistency between links by a more suitable approach to the problem.

4. The matching strategy

The possibility of retrieving information on neonatal mortality by gestational age and birth weight is explored by a deterministic linkage. The files to be linked consist of the following records:

file *A1*. 1,528 records related to neonatal deaths in Italy¹⁰, with reference to 2003 birth cohort;

file *A2*. 1,527 records related to neonatal deaths in Italy, with reference to 2004 birth cohort;

file *B1*. 452,984 records related to live births in Italy in 2003;

file *B2*. 474,893 records related to live births in Italy in 2004.

⁹ For instance, consider the records $a, a' \in A$ and $b, b' \in B$ and all the possible record pairs with the following weights: $w_{ab} = 15 > w_{ab'} = 12 > w_{a'b} = 10 > w_{a'b'} = 8$. Suppose that all the weights are not less than t_u and each entity can be represented at most once in each file to be matched. By adopting either a global or a greedy approach, both (a, b) and (a', b') are classified as links. If the weights are modified, for example, by setting $w_{ab} = 13 > w_{ab'} = 12 > w_{a'b} = 10 > w_{a'b'} = 8$, the global solution changes as well - (a, b') and (a', b) are now classified as links - while the greedy one is unchanged.

¹⁰ The death records from the province of Bolzano are excluded, as the birth certificates for 2003 and 2004 are not available.

The matching is carried out between files related to the same birth cohort (*A1* and *B1*, *A2* and *B2*); therefore, *infant's year of birth* - which can be considered error-free - represents an implicit blocking variable. As only a low percentage of infants die within the first month of life, the size of the files to be linked is very different and the identification of records corresponding to the same individual is particularly difficult.

The matching weights are defined as a measure of similarity between records; this measure is based on the frequencies of the values observed for each matching variable in the larger file to be linked, which is related to live births. Let *A* and *B* represent two files related to the same birth cohort, *A1* and *B1* or *A2* and *B2*. The weight for (*a*, *b*) is:

$$w_{ab} = \frac{\sum_{j=1}^k [y_{ab,j} \times (n_B - fr_{md(x_{b,j}^B)})]}{k \times n_B - \sum_{j=1}^k fr_{md(x_{r,j}^B)}}, \quad (4)$$

where:

- n_B is the number of records in *B*;
- $y_{ab,j} = 1$, if *a* and *b* agree on *j*, and $y_{ab,j} = 0$, if *a* and *b* disagree on *j* or at least one value for *j* is missing;
- $fr_{md(x_{b,j}^B)}$ is the frequency (in *B*) of the value observed for *j* on the record $b \in B$;
- *r* is the record in *B* corresponding to the rarest perfect link, i.e. the pair whose records agree on all matching variables and for which $\sum_{j=1}^k fr_{md(x_{b,j}^B)}$ is minimum.

The weights (4) usually take values between 0 and 1. In particular, the pairs whose records disagree on all matching variables have a weight equal to 0; the maximum weight is usually attained at the rarest perfect link. On the basis of some tests on data from Italian regions characterized by different levels of coverage of births, such weights lead to the same final assignment scheme attainable by adopting the weights (2).

The threshold values t_l and t_u are selected by inspection of the distribution of the estimated weights. In this context, t_l and t_u are set to take values in the ranges 0.45-0.55 and 0.55-0.65, respectively (with $t_u - t_l \leq 0.1$).

As one death record can be linked to at most one birth record, a greedy algorithm is used to obtain the final assignment scheme. In case of one-to-many assignments¹¹ with the same matching weight, only the pair related to the infant with the lowest birth weight is labelled as a link, exploiting the findings from recent studies on neonatal mortality (Branum and Schoendorf, 2003).

¹¹ These pairs are related to cases in which one of the infants born from the same pregnancy died.

5. Results of linkage

5.1. Results from provisional linkage at national level

In this subsection the results from a provisional linkage of national data are reported. The weights are set equal to the number of agreements on the matching variables and $t_l = t_u = k$, *i.e.* only the record pairs with k agreements are classified as links. In case of one-to-many assignments with the same matching weight, only the pair related to the infant with the lowest birth weight is labelled as a link.

As displayed in Table 3, this provisional linkage does not produce a high number of links. The ratio of number of links to number of deaths is 29.2 and 23.6 percent for 2003 and 2004 cohorts, respectively. The largest decrease in the percent ratio can be observed for Central Italy (from 25.0 to 10.3 percent). With regard to Southern Italy and Islands, about one out of five death records is matched. Northern Italy shows the highest percentage of links, in the face of a decrease from 42.5 to 37.5 over the two-year period. The links coming from one-to-many assignments with equal matching weight - which are related to infants born from the same pregnancy - are 22 and 20 for 2003 and 2004 cohorts, respectively (for the sake of brevity, these data are not reported in Table 3).

The links achieved by perfect agreement on the matching variables are not many; so, a more sophisticated strategy for linkage is required. Besides, in the light of the considerations on coverage of births and quality of data used for the linkage (section 2), the retrieval of information on neonatal mortality by gestational age and birth weight is restricted to Northern Italy.

Table 3. Results from provisional linkage between records related to neonatal deaths and records related to live births, by geographical area of death and infant's year of birth

<i>geographical area of death</i>	<i>number of deaths</i>		<i>number of links</i>		<i>n. of links / n. of deaths (%)</i>	
	<i>2003</i>	<i>2004</i>	<i>2003</i>	<i>2004</i>	<i>2003</i>	<i>2004</i>
<i>Northern Italy</i>	570	525	242	197	42.5	37.5
<i>Central Italy</i>	288	302	72	31	25.0	10.3
<i>Southern Italy and Islands</i>	670	700	135	135	20.1	19.3
<i>Italy</i>	1,528	1,527	449	363	29.2	23.6

5.2. Results from linkage for Northern Italy

With regard to Northern Italy, the proposed strategy for linkage leads to a high number of final links (Table 4). The links coming from one-to-many assignments with equal matching weight are related to infants of the same gender, born from the same pregnancy (for the sake of brevity, such data are not reported in Table 4), and are about 4.8 and 5.0 percent of the total links for 2003 and 2004 cohorts, respectively. Besides, 91.4 and 95.7 percent of the links (for 2003 and 2004 cohorts, respectively) are labelled automatically, as their weights are above t_u , while 8.6 and 4.3 percent are designated manually, as their weights are between t_l and t_u .

Table 4. Results from final linkage between records related to neonatal deaths and records related to live births, by infant's year of birth - Northern Italy

<i>infant's year of birth</i>	<i>n. of deaths</i>	<i>n. of links</i>	<i>n. of links with weight $\geq t_u$</i>	<i>n. of links with weight between t_l and t_u</i>	<i>n. of links / n. of deaths (%)</i>
2003	570	526	481	45	92.3
2004	525	460	440	20	87.6

The high number of links that agree on all or most of the matching variables - at least 8 agreements occur for over 96 percent of links (Table 5) - does not rule out, however, the presence of errors in the final results (false matches or missing matches), mainly due to some critical issues concerning the data: the coverage of births is heterogeneous across regions; the matching variables do not have a great identifying power; the information on mother's date of birth in the death records is not very accurate. On the other hand, matching errors cannot be easily estimated, as no subset of pairs for which the linkage status is known is available; besides, model-based estimates of matching errors, according to other approaches proposed in literature (e.g. Armstrong and Mayda, 1993), could result inaccurate, due to the above mentioned critical aspects.

Table 5. Percentage distribution of links by number of agreements on the matching variables and infant's year of birth - Northern Italy

<i>number of agreements</i>	<i>links (%)</i>	
	<i>2003</i>	<i>2004</i>
12	46.01	42.83
11	12.17	18.26
10	8.17	6.30
9	19.96	15.22
8	9.70	14.57
7	2.47	2.61
below 7	1.52	0.22
<i>total</i>	100.00	100.00

6. Specific neonatal mortality rates

Neonatal mortality rates by gestational age and birth weight category have not been calculated in Italy since 1999. On the basis of the results from record linkage, estimates of these specific rates are provided for Northern Italy with reference to 2003, the first year from which the calculation of the time series can be resumed and for which the quality of the starting data appears to be slightly higher.

The neonatal mortality rates presented here, related to 2003 birth cohort, are calculated by using:

live births in Northern Italy by gestational age and birth weight category, at the denominator;

deaths within the first month of life, in Northern Italy, by gestational age and birth weight category, at the numerator¹².

The 570 neonatal deaths are assumed to have the same distribution by gestational age and birth weight category as the 481 links with weight not less than t_u ¹³.

The results in Figures 1 and 2 appear to be in line with those reported in recent studies on the survival of preterm and low birth weight infants (Demissie *et al.*, 2001; Horbar *et al.*, 2002) and with the progressive downward trend in neonatal mortality observed in the nineties - especially for preterm and low birth weight infants. Compared to 1998, in 2003 the neonatal mortality rate declined:

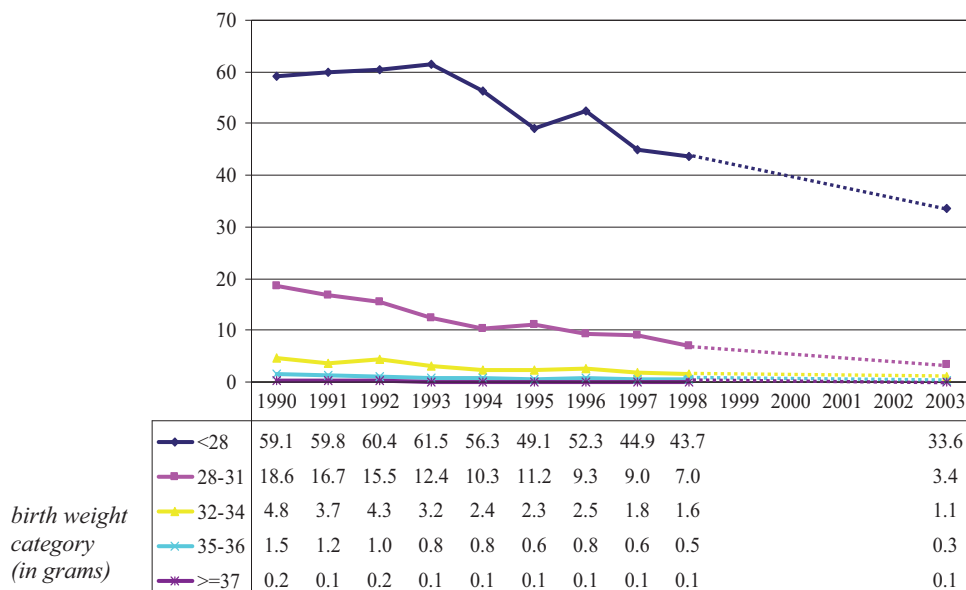
from 43.7 to 33.6 and from 36.7 to 30.8, respectively for the lowest categories of gestational age (below 28 weeks) and birth weight (below 1,000 grams);

by half, for gestations between 28 and 31 weeks (from 7.0 to 3.4) and for infants with birth weight between 1,000 and 1,499 grams (from 6.1 to 3.1).

¹² For 2003, neonatal mortality rates are calculated by cohort, while in the past (up to 1998) they were cross-sectional. However, if neonatal deaths are assumed to have a uniform distribution over the year, the differences between these types of rate can be negligible.

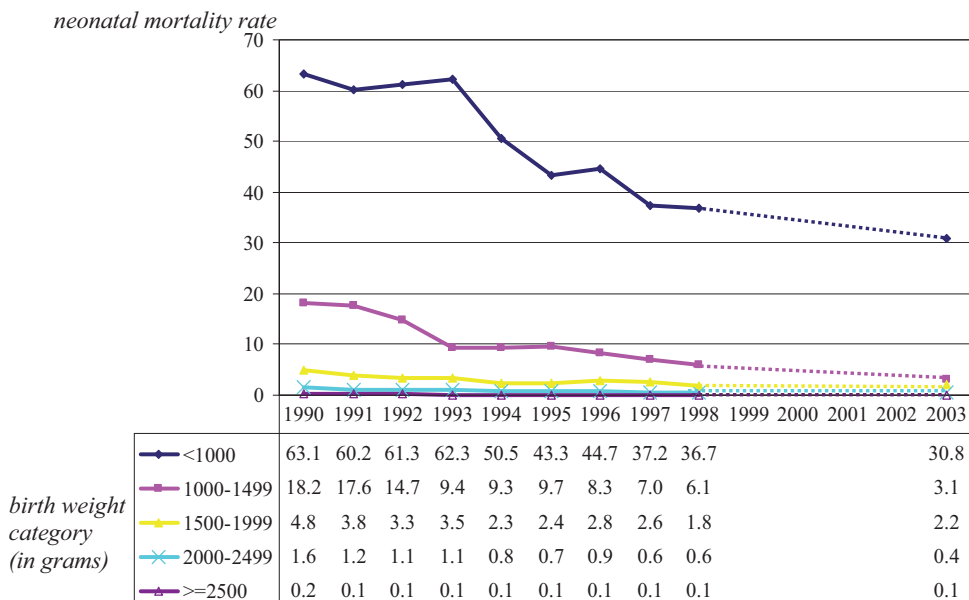
¹³ Of the total 526 links, only those with weight not less than t_u (column 4 of table 4) are considered, in order to contain the introduction of bias due to possible false matches.

Figure 1. Neonatal mortality rates by gestational age category and year - Northern Italy, years 1990-98 and 2003 (values per 100 live births in the same gestational age category)



Besides, compared to the past, in 2003 there was a higher concentration of neonatal deaths in the lowest categories of gestational age and birth weight: 42 and 55 percent of neonatal deaths were related to gestations below 28 and 32 weeks, while 40 and 48 percent concerned infants with birth weight below 1,000 and 1,500 grams.

Figure 2. Neonatal mortality rates by birth weight category and year - Northern Italy, years 1990-98 and 2003 (values per 100 live births in the same birth weight category)



7. Concluding remarks

In the light of the considerations on coverage of births and quality of data used for the linkage, the information on neonatal mortality by gestational age and birth weight is only retrieved for Northern Italy. The percentage of cases for which the retrieval is achieved by perfect agreement on the matching variables is low; so, a more sophisticated strategy for linkage is used. In this work, the possibility of retrieving crucial information on neonatal mortality is explored by using a deterministic record linkage. The adopted weights are based on the frequencies of the values observed for each matching variable in the larger file to be linked, which is related to live births. On the basis of some tests on data from Italian regions characterized by different levels of coverage of births, this choice - even though not very sophisticated - leads to the same links attainable by adopting a probabilistic strategy.

In this framework, one death record can be linked to at most one birth record. This requirement is the starting point for a critical review of the most used approach to choose records to be linked, when the decision rule is defined separately for each pair. Maximizing the sum of matching weights, incorrect links can be produced, as the solution depends upon the numerical values of the

weights involved and not only on their order. A greedy approach is believed to be more suitable, apart from the specific context.

However, the high number of final links does not rule out the presence of errors in the results and, as a consequence, the mentioned findings for neonatal mortality should be treated with caution. The reliability of neonatal mortality estimates depends on the accuracy and completeness of reporting and recording of births and deaths. In the retrieval of the information needed to compute specific rates, further efforts should be made to control matching errors, and this cannot be accomplished without improving the data collection processes over the whole country.

REFERENCES

- ARMSTRONG J. B., MAYDA J. E. (1993), Model-based estimation of record linkage error rates, *Survey Methodology*, 19, 137-147.
- ARMSTRONG J., SALEH M. (2000), Weight estimation for large scale record linkage applications, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1-10.
- BRANUM A. M., SCHOENDORF K. C. (2003), The effect of birth weight discordance on twin neonatal mortality, *Obstetrics & Gynecology*, 101, 570-574.
- DEMISSIE K., RHOADS G. G., ANANTH C. V., ALEXANDER G. R., KRAMER M. S., KOGAN M. D., JOSEPH K. S. (2001), Trends in preterm birth and neonatal mortality among blacks and whites in the United States from 1989 to 1997, *American Journal of Epidemiology*, 154, 307-315.
- DEMPSTER A. P., LAIRD N. M., RUBIN D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, B*, 39, 1-38.
- FELLEGI I. P., SUNTER A. B. (1969), A theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.
- HERZOG T. N., SCHEUREN F. J., WINKLER W. E. (2007), *Data quality and record linkage techniques*, Springer, New York.
- HILL T., PRING-MILL F. (1985), Generalized iterative record linkage system, in: KILSS. B., ALVEY W. (eds.), *Record linkage techniques - 1985, Proceedings of the Workshop in Exact Matching Methodologies*, Arlington, Virginia, May 9-10, 1985, 327-333.
- HORBAR J. D. (1999), The Vermont Oxford network: evidence-based quality improvement for neonatology, *Pediatrics*, 103, 350-359.

- HORBAR J. D., BADGERS G. J., CARPENTER J. H., FANAROFF A. A., KILPATRICK S., LACORTE M., PHIBBS R., SOLL R. F. (2002), Trends in mortality and morbidity for very low birth weight infants 1991-1999, *Pediatrics*, 110, 143-151.
- JARO M. A. (1989), Advances in record-linkage methodology as applied to matching the Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414-420.
- MINISTERO DELLA SALUTE (2007), *Certificato di assistenza al parto (CeDAP). Analisi dell'evento nascita - Anno 2004*, Dipartimento della Qualità, Direzione Generale del Sistema Informativo, Ufficio di Direzione Statistica.
- WINKLER W. E. (2000), Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage, *Statistical Research Report Series RR2000/05*, U. S. Bureau of the Census, Washington.
- ZEITLIN J., WILDMAN K., BRÉART G., ALEXANDER S., BARROS H., BLONDEL B., BUITENDIJK S., GISSLER M., MACFARLANE A. (2003), Selecting an indicator set for monitoring and evaluating perinatal health in Europe: criteria, methods and results from the PERISTAT project, *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 111, 5-14.