TEXTO PARA DISCUSSÃO Nº 198

PROBABILISTIC RECORD LINKAGE AND AN AUTOMATED PROCEDURE TO MINIMIZE THE UNDECIDED-MATCHED PAIR PROBLEM

Carla Jorge Machado
Kenneth Hill

Maio de 2003

Ficha catalográfica

314.01 M149p 2003	Machado, Carla Jorge. Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem. - / por Carla Jorge Machado, Kenneth Hill. Belo Horizonte: UFMG/Cedeplar, 2003. 18p. (Texto para discussão ; 198) I 1. Demografia – Técnica. 2. Levantamentos demográficos. I. Hill, Kenneth. II.Universidade Federal de Minas Gerais. Centro de Desenvolvimento e Planejamento Regional. III. Título. IV. Série.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE CIÊNCIAS ECONÔMICAS
CENTRO DE DESENVOLVIMENTO E PLANEJAMENTO REGIONAL

PROBABILISTIC RECORD LINKAGE AND AN AUTOMATED PROCEDURE TO MINIMIZE THE UNDECIDED-MATCHED PAIR PROBLEM

Carla Jorge Machado

Professora do Departamento de Demografia do Cedeplar/UFMG
carla@cedeplar.ufmg.br / cjmachado@terra.com.br

Kenneth Hill

Professor Department of Population and Family Health Sciences,
The Bloomberg School of Public Health, Johns Hopkins University
khill@jhsph.edu

CEDEPLAR/FACE/UFMG
BELO HORIZONTE
2003

SUMÁRIO

ABSTRACT

Probabilistic record linkage allows the assembling of information from different data sources. In this article, we present a procedure in case a one-to-one relationship between records in different files is expected but is not found only by applying the probabilistic record linkage methodology. Our data were births and infant deaths from the 1998-birth cohort whose mother's place of residence was the City of São Paulo at the time of birth. Our assumption was that pairs for which a one-to-one relationship was obtained, and a best-link was found with the highest combined weight would be considered as uniquevocally matched pairs or gold-standard and should then provide information in order to decide about pairs in which such a relationship could not be established. For example, we observed that the for the unequivocally matched pairs a clear and expected relationship between differences in dates of death and birth registration could be assessed. As a result, such a relationship was used to help solving the remaining pairs for which a one-to-one relationship could not be found. Indeed, we reduced the number of non-uniquely matched records and even though we could not establish a one-to-one relationship for every single death we reduced the number of uncertain. We suggest that future research using record linkage should use combined strategies from results from first record linkage runs before a full clerical review (the standard procedure under uncertainty) in order to most efficiently (and less costly), retrieve record matches.

Key words: probabilistic record linkage, best-link, birth-cohort, one-to-one match.

RESUMO

O relacionamento probabilístico de dados permite que fontes de informações relativas ao mesmo registro e encontradas em bancos de dados distintos sejam unificadas. Neste artigo apresenta-se um procedimento utilizado quando se espera que um registro de um banco de dados corresponda a apenas um outro registro num segundo banco de dados, ou seja, quando a relação é unívoca contudo a aplicação da metodologia de relacionamento probabilístico é insuficientee para a obtenção desta relação unívoca. As fontes de dados foram os registros de nascimento e óbito infantis da coorte de nascimentos de 1998, cuja residência da mãe era a cidade de São Paulo quando do nascimento, relacionados probabilisticamente. Partiu-se do princípio de que os dados relacionados probabilisticamente com o mais alto escore possível e relação unívoca, seriam utilizados como padrão-ouro e forneceriam subsídio para a decisão sobre pares obtidos a partir do relacionamento probabilístico não foi obtida relação unívoca. Por exemplo, uma vez observado que os dados univocamente relacionados obedeciam um comportamento esperado em termos da diferença nas datas de registro de óbito e de nascimento, aplicou-se esta relação aos dados cuja relação unívoca não havia sido inicialmente estabelecida. Como resultado, o número de pares com relação unívoca aumentou substancialmente e mesmo nos casos de ausência desta relação, diminui-se substancialmente o número de registros de nascimento ligados a um registro de óbito. Sugere-se que este procedimento deve ser associado à revisão manual de registros (o procedimento padrão na presença de incerteza) a fim de conseguir um pareamento o mais eficiente possível.

Palavras-chave: relacionamento probabilístico de dados, melhor par encontrado, coorte de nascimento, pareamento unívoco de registros.

INTRODUCTION

Record linkage is the methodology of finding a unified record from two or more records that are in different files and belong to the same entity. Record linkage methods can be deterministic or probabilistic or a combination of both. Deterministic linkage is used when there is a unique identifier or if variables used for comparison are error-free and highly discriminatory, whereas probabilistic linkage takes into account the uncertainty that can exist in comparing variables used for comparison in both files. The uncertainty is related to the 'rareness' of the characteristic used for comparison and on how much confidence we place in such characteristic. Sex, for example, induces a twofold partition of a file: males and females, and if records agree on sex, we cannot say with a high degree of confidence that they belong to the same person. On the other hand, since it is very easy to code, if records disagree on sex we can almost surely state that the linked records do not belong to the same person.

Probabilistic record linkage have been used in the Public Health field in the last fifty-years, since the Seminal work of Newcombe et al (1959). Sometimes, such methods are not sufficient in providing the basis for the decision about whether a pair is a true-link (matched pair) or not, and other information rather than the one provided by the matching variables – or variables common to both files used to identify matches – is needed. Clerical review is the most common option, which is considered the gold-standard, but sometimes the size of the file makes such a task prohibitively expensive or highly time-consuming.

We probabilistically linked data from the 1998-birth cohort of the City of São Paulo and our attempt was to match 3842 infant deaths from this birth cohort to their corresponding live birth. The size of the live birth file was 209628. Our data came from two sources: from DATASUS and from the SEADE Foundation and a description and a full review of data sources and quality can be found in Machado (2002). We aimed to obtain the corresponding death record to each birth record, assuming that a logical one-to-one relationship should hold. Using probabilistic methods, in a first pass, we obtained a one-to-one match for 2249 deaths (59% of the deaths). In this article, rather than describing the methodology of probabilistic record linkage itself, the aim is to describe a method to get around the undecided-matched pair problem – which happens here whenever a one-to-one relationship does not hold – by using information from a first matched file in order to help solving undecided links. Before that, however, we briefly review the results obtained from the probabilistic methodology used in order to familiarize the reader with our procedure and classification rules.

PROBABILISTIC RECORD LINKAGE PROCEDURE

For any probabilistic record linkage procedure, two steps are crucial: searching out the potential linked pairs for further comparison, and deciding whether a record pair is correctly matched. In the process of searching out the pairs we required that in order for records to be suitable for comparison, they had to agree exactly on a given variable selected to be mother's district of residence in the City of São Paulo. This variable is called a *blocking variable*. For any given block, all pairwise combinations between births and deaths were obtained. Therefore, we first generated 13680789 comparison pairs, using the *Reclink© program* (Camargo Jr. and Coeli, 2000). In the process of

deciding about matched pairs, the variables used as matching variables were birth date, birth weight, maternal age, delivery mode, sex, and plurality. A full description and estimation procedures of the matching weights – a value assigned to a linked pair that summarizes the comparison results of the two variables – for each matching variable is in Machado (2002). In Table 1 the estimated weights for each matching variable are displayed.

TABLE 1
Estimated weights for matching variables

Comparison results between two variables	Estimated Weights
Agreement on date of birth	19.52
Agreement on birth weight	15.93
Agreement on maternal age	11.69
Dates of Birth off by one day	9.56
Birth weights off by 100 grams	8.23
Maternal Ages off by one year	5.78
Agreement on Sex	2.82
Agreement on delivery mode	2.68
Birth Weights off by 200 grams	0.52
Agreement on plurality	0.10
Maternal ages are off by two years	-0.13
Dates of birth are off by two days	-0.41
Plurality is missing in either record	-1.67
Disagreement on plurality	-1.84
Maternal age is missing in either record	-5.87
Disagreement on maternal age	-6.04
Birth weight is missing in either record	-7.01
Disagreement on birth weight	-7.19
Delivery mode is missing in either record	-7.51
Disagreement on delivery mode	-7.68
Disagreement on dates of birth	-10.38
Disagreement on Sex	-10.93

Source: DATASUS/2000 and SEADE Foundation (2001).

As an example, it is clear that if a death record was linked to a birth record and the records agreed exactly on birth weight, birth date and mother's age, there was a very high chance that the pair belonged to the same infant, i.e., was a match. On the other hand, if records disagreed on sex and on dates of birth, the chance was very small. It is also noticeable that an agreement on plurality for example was not very informative and this is quite intuitive: a pair of singular infants were very likely to be linked by chance only since the vast majority of infants were singleton. Therefore, different combinations of comparisons for different variables can yield a range of combined weights, where combined weights are the linear sum of the each estimated weight for each matching variable. Indeed, we had 1800 possibilities of combined weights.

BEST LINKS

The next step was to select best link(s), defined as the linked pair with the highest combined weight, achieved by each record (MacLeod et al, 1998). One problem is the failure to match a death record with its corresponding record, which yields non-matched records. If just by chance there is another birth record within the same block that links to this death record with a higher weight, we will make the wrong decision. There is no way to avoid this kind of mistake, but an erroneous link due to this source of error is unlikely (Kendrick et al, 1998). In this case, the deceased infant would have to have recorded values more similar to those on the 'wrongly matched' birth record than to values recorded on its own corresponding birth record. We expect the degree of similarity to be higher between records that belong to the same infant, which is the fundamental assumption of the record linkage theory. A more frequent problem is the coincidental-match problem, which relates to the presence of missing values in birth and death records or to the case where the matching variables are not highly discriminatory (such as sex or plurality, for example). The issue of missing information, however, is more serious. In this context, we would expect that a given death record would be linked and would achieve a best-link with more than one death record. Another possibility, less likely, happens when one birth record is linked to more than one death record, and the two pairs formed have the same combined weight and we have no way to decide which infant represented in each death record is more likely to be also represented by the given birth record. In both situations, ties are generated, were ties are matches that can not be considered as definitively relating to the same infant.

'Ties' were solved using the basic principle that, in searching for matches, other than the evidence provided by the combined weights for or against a match, only one birth record should correspond to a given death record. Once this one-to-one relationship is established, the birth record is not allowed to link to any other death record. Pairs of records in which ties were identified were classified as 'temporary matches' (Tepping, 1968). We sought other information in order to resolve those ties and classify pairs as matches or non-matches.

From 13680789 pairs we selected for each death record its respective best-link(s) and went down to 17764 pairs. Then we kept the links in which the birth record achieved its best-link. The assumption is that in case a given birth record is involved in more than one pair, we should keep the link with the highest combined weight. We went down from 17764 best links to 16278 best links (a reduction of 8.4%). Examples of pairs obtained are in Table 2 and a summary of these results are in Table 3.

TABLE 2

Examples of best-links and second best-links

Identification Number		Combined Weight	Best link achieve by…		Pair selected as a temporary (or definitive) match?
Death	Birth		…Death Record?	…Birth Record?	
200	***4709***	*0.28*	*Yes*	*No*	*No*
200	11863	*0.28*	Yes	Yes	***Yes***
200	14232	*0.28*	Yes	Yes	***Yes***
200	10516	-9.68	No	No	No
200	15052	-9.68	No	No	No
200	15095	-9.68	No	No	No
200	145459	-9.68	No	No	No
235	***4709***	*52.76*	*Yes*	*Yes*	***Yes***
235	11863	16.96	No	No	No
131	***3362***	*9.96*	*Yes*	*No*	*No*
131	12970	3.57	No	No	**Yes** (later on it will be selected)
132	***3362***	*10.13*	*Yes*	*Yes*	***Yes***
132	3335	0.16	No	No	No
132	13134	0.16	No	No	No
132	20641	0.16	No	No	No

Source: DATASUS/2000 and SEADE Foundation (2001).

TABLE 3

Results obtained after selecting first best-links

Characteristics of pairs for which best links were found	Pairs		Death Records		Average number of pairs per death record
	N	%	N	%	
Tie due to a death record with multiple links; death records linked to a birth record involved in one link only.	13443	82.6	1466	38.6	9.1
Tie due to a death record with multiple links; death record linked to a birth record also involved in at least another link.	572	3.6	71	1.9	7.9
Tie due to a death record linked to more than one birth record; death record not involved in any other link and birth record involved in multiple links.	14	0.1	14	0.4	1
No tie (a one-to-one relationship established between a birth and a death record)	2249	13.8	2249	58.3	1
Death records with no best links			42[(1)]	0.9	
TOTAL	16278	100.0	3842	100.0	4.3

Source: DATASUS/2000 and SEADE Foundation (2001)

Notes: [(1)]At later stages in selecting the pairs, we defined as matches best links for 3 out of 42 death records and remained with 39 death records whose matched pairs were found among second best-links.

We kept as potential matches four pairs, the 'best-linked' ones. The ordered pair (235; 4709) is a definitive match and (200; 11863) and (200; 14232) were considered temporary matches.

Death record '131' achieved a unique best-link with birth record '3362'. However, birth record '3362' was involved in another pair (with 132) in which it achieved a higher composite weight *and* this pair was considered a best link from the standpoint of the death record. We kept (132; 3362) as a definitive match and refuted the pair (131; 3362) as so. We then look for another match for death record '131' among the second best links and in later stages, selected (131, 12970) as a definitive match, since birth record '12970' did not achieve a best link with a composite weight higher than 3.57 with any other death record. However, selecting a death record among 'second best links' was a rare event, that happened to only 39 death records (1% of the death records).

On average, there were 4.3 birth records linked to each death record and selected as best links. Therefore, for a typical death record, ties do exist. But, indeed, for the majority of ties the task is to decide among two to four birth records per death record, as we see in Table 4.

Ties arose for more than one reason. Most often, a tie was formed because one death record linked to multiple birth records that were not involved in another link. This was the most common situation here since the birth file was so much larger than the death file. In fact, the *a priori* probability that any birth record would link to any death record was very small, about 1.8 percent.

TABLE 4

Distribution of death records by number of linked birth records
Records with more than one best link and birth record not involved in any other pair

Birth records per death record	Number of Pairs	Number of death records	Percentage of death records	Cumulative Percent of death records
2	660	330	22.5	
3	663	221	15.1	37.6
4	588	147	10.0	47.6
5 to 10	4461	642	43.8	91.4
11 to 21	1589	124	8.5	99.9
100 +	5482	2	0.1	100.0
TOTAL	13443	1466	100.0	

Source: DATASUS/2000 and SEADE Foundation (2001)

For 572 pairs, corresponding to 71 death records, ties were formed due to the linking of a death record that achieved its best link with more than one birth record; these birth records also achieved best links with other death records. A number of them were allocated consecutively or very closely in the death record file and linked best to the same birth record, such as death records '431' and '432' that achieved their two best links with birth records '26355' and '26359' each of them. We speculate that those death records belong to non-singleton infants. Because so much identifying information was likely to be shared, the correct matching of records belonging to twins has been recognized as a major problem (Kendrick et al, 1998). For 14 death records, a tie was formed because

the death record achieved its unique best link with a birth record that was also involved in another unique best link from the standpoint of the death record. Since the deaths are allocated consecutively or very closely for ten out of the 14 records we invoke the same reason as before. The difference is that only one death record seems to have been issued. Examples are death records '344' and '345' that achieved a best-link with birth record '27657'.

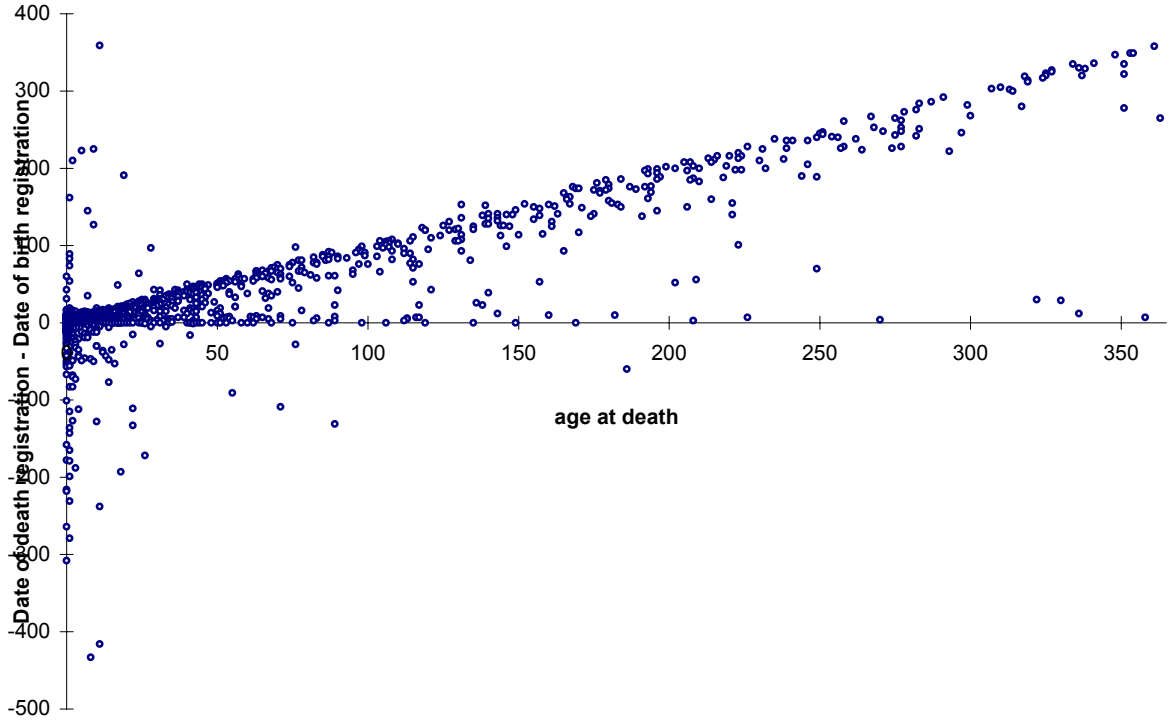
Finally, for 2249 death records it was possible to find a unique best link and these matches are called *unequivocally matched records* – and no further efforts will be implemented in making sure that these links are in fact matched pairs and they will be considered our *gold-standard*, since the expected one-to-one relationship was established.

RESOLUTION OF TIES

After a record linkage operation the researcher should seek other information in order to decide the matches and non-matches among ties (Waien, 1997). Clerical review has been extensively used and considered to be the standard method. However, given the size of the file to be reviewed, this option was not considered. We considered checking the agreement between comparison variables other than the ones used in the record linkage, such as comparing values for maternal education or gestational age category. However, this alternative is clearly not fruitful: above 80 percent of the death records (or 3093 records) do not posses information on maternal education and above 55 percent of the death records (or 2128 records) do not do so on gestational age. Indeed, had we believed that these variables were of value for matching we would have included them as comparison variables in the first place. Another idea was to inspect those pairs selected as unequivocal matches. Wadja and Ross (1987) suggest that results from a record linkage operation obtained from an initial run through the data generally suggest opportunities for improving the linkage. Winkler and Scheuren (1996) suggested a recursive method were firstly matched files provide information to a subsequent matching. Assuming that each of these 2249 pairs formed truly belongs to the same infant, we inspected information on date of birth registration combined with date of death registration. We expected that births would be registered around the time of birth and that deaths would be registered around the time of death. Therefore the difference between date of death registration and date of birth registration would be very close to the age at death of the infant. However births may go unregistered for some time, and in this case the time elapsed between births and deaths to be shorter compared to the time between the birth and death of the infant. Therefore, the idea was to use these 2249 pairs as a 'learning set of pairs' in order to calculated the ranges of birth and death registration for each of the 2249 matched pairs. First we plotted the difference in dates of registration against the age the reported age at death of the child in number of days. Results are shown in Figure 1:

FIGURE 1

Scatterplot of the differece in dates of registration against the reported age at death - Deceased infants with unequivocally matched birth records



Most points follow on a diagonal or slightly below the diagonal line indicating also that the time elapsed between dates of registration may be slightly inferior than the time elapsed between the birth and the death of an infant. In fact, by law, the live birth should be notified and registered within 15 days. The death is likely to be registered as soon as it happens, in order to obtain a death certifcate for the burial. Therefore, it would be reasonable to expect that the time between the registration of the two events would fall in between the age at death of the infant minus 15 days and the age at death of the infant. This is a reasonable assumption, corroborated by the observations.

Figure 1 also shows that a significant number of observations fall on a horizontal line where the difference between date of death registration and date of birth registration equals to zero, which means that the birth and the death of the infant were registered in the same day. Lastly, 216 deaths were registered before the birth had been registered.

In Brazil, for infant deaths, if the birth has not yet been registered at the time of death registration, this has to be done, by law, at the same time and at the same registrar. However, in some situations a birth might have been registered after the death, due to a misunderstanding of the law by the registrar, for example. Or, simply, a mistake might have occurred in the recording date of registration of either event.

In light of these findings, according to the age of death of an infant we defined acceptable ranges in which we could expect the differences in dates of registration to lie (Table 5)

TABLE 5

Acceptable ranges of time intervals between birth and death registration for resolution of temporary pairs, by age at death of the infant.

Age at death (in number of days)	Acceptable range that includes number of days between birth and death registration (inclusive time intervals)
0	0
1	(0, 1)
2	(0, 2)
(…)	(…)
16	(0, 16)
17	(2, 17) or events registered at the same day (i.e., 0)
(…)	(…)
363	(348, 363) or same day (i.e., 0)

Source: DATASUS/2000 and SEADE Foundation (2001)

Note: Date of birth registration, in number of days, starting with January 1st, 1998 as day '1', is "X" ; Date of death registration, in number of days, starting with January 1st, 1998 as day '1', is "Y"; Time elapsed between registrations in the second column relates to "Y-X"

We also hypothesized that the earlier the death, the higher the chance that the birth and the death would have taken place in the same hospital (or facility). Therefore, the chance that the birth and the death would have been registered in the same registrar's office would also be higher for earlier deaths. For the infants unequivocally matched, the earlier the death, the higher the proportion of deaths registered in the same place the birth was registered. For neonatal deaths, 77% of the infants were registered in the same registrar's office whereas for post-neonatal deaths, only 43% were registered in the same registrar. Therefore, to solve ties we assumed that deaths during the neonatal period were more likely to be registered in the same registrar and used a score system, to be applied in all temporary matches, as follows:

1. For each temporary matched pair, if a death occurred at any age and the number of days elapsed between the registration of birth and registration of death fell within the proposed ranges in Table 5, we gave the pair a score of one point (+1). A minus one point (-1) was given, otherwise. If either the birth or the death did not possess information on date of registration we gave the pair a score of zero (0).
2. For each temporary matched pair, if a death occurred during the neonatal period and the registrar's office of birth and death registration was the same, we gave the pair a score of one point (+1). A minus one point (-1) was given in case of discordant registrar's. In the absence of information on registrar for either the birth or the death, a null score of zero (0) was given. For post-neonatal deaths, we gave a full score of one point (+1) for all pairs, given the information on registrar was considered of no use for later deaths.

3. For each tied pair of records, we added the first to the second score. The range of possible scores is from (-2) to (+2).
4. For each death record for which tied pairs existed, we selected the pair with the highest score, if it existed.

As an example, we revisit pairs considered temporary in Table 2. Death record '200' belonged to an infant who died at 18 days of life.

TABLE 6
Example pair of resolved tie – death record '200'

Identification Number		Registrar's Office	Time between registration	SCORE		
Death	Birth			Registrar	Time between dates	TOTAL
200	11863	Same	4	+1	+1	+2
200	14232	Not Same	7	-1	+1	0

Source: DATASUS/2000 and SEADE Foundation (2001)

Pair (200, 11863) was selected as an uniquevocally matched pair. But, sometimes, we were unable to select only one pair, as for death record '277'. The infant died at 13 days (Table 7):

TABLE 7
Example pair of unresolved tie – death record '277'

Identification Number		Registrar's Office	Time between birth and death registration	SCORE		
Death	Birth			Registrar	Time between dates	TOTAL
277	16558	Same	2	+1	+1	+2
277	16567	Same	0	+1	+1	+2
277	16575	Same	1	+1	+1	+2
277	192334	Not Same	-262	-1	-1	0
277	32173	Same	-10	+1	-1	0
277	17649	Not Same	17	-1	-1	0
277	16491	Same	14	+1	-1	0

Source: DATASUS/2000 and SEADE Foundation (2001)

We were still uncertain about which birth record truly represented the death record and we kept the first three pairs in the absence of any other information to solve the tie.

Overall, after we used dates of birth registration and death registration combined and the information on registrar, we reduced the number of temporary matched pairs from 13443 to 3917, a 71 % reduction.

We further reduced the number of temporary pairs stating that a one-to-one relationship also provided evidence that the pair belonged to the same infant. When a one-to-one relationship was found, the birth record involved in that relationship should not be allowed to be involved in any other match, in case the death record in this later match was also involved in another match with other birth record (or records). For example, death record '196' belonged to an infant who died in the first day of life and death record '1447' belongs to an infant who died at the age of three months of life (119 days). These death records achieved a best link with other birth records at the combined weight of 0.288 (Table 8).

TABLE 8
Example pairs – death records 196 and 1447 – Evidence provided by one-to-one relationship in resolving temporary matched pairs

Identification Number		Registrar's Office	Time between registration	SCORE		
Death	Birth			Registrar	Time between dates	TOTAL
196	**11179**	**Same**	**0**	**+1**	**+1**	**+2**
196	16181	Not same	-2	-1	-1	-2
196	19630	Not same	-11	-1	-1	-2
196	20610	Not same	7	-1	-1	-2
196	27113	Not same	-14	-1	-1	-2
196	27133	Not same	-15	-1	-1	-2
196	27588	Not same	-37	-1	-1	-2
196	57358	Not same	-59	-1	-1	-2
196	73832	Not same	-88	-1	-1	-2
1447	**11179**	**----**	**119**	**+1**	**+1**	**+2**
1447	16181	----	117	+1	+1	+2
1447	19630	----	108	+1	+1	+2
1447	20610	----	126	+1	-1	0
1447	27113	----	105	+1	+1	+2
1447	27133	----	104	+1	+1	+2
1447	27588	----	82	+1	-1	0
1447	57358	----	60	+1	-1	0
1447	73832	----	31	+1	-1	0

Source: DATASUS/2000 and SEADE Foundation (2001).

Since the only birth record considered to be matched to death record '196' was birth record '11179', a one-to-one relationship was established. We then ruled out birth record '11179' as an option for death record '1447' because a one-to-one relationship was established between death record '196' and '11179', but not between '1447' and '11179'. We notice, however, that if the only birth record left for '1447' was '11179' we would be unable to proceed in this way. Indeed, the reduction in the number of temporary pairs by following this procedure existed, but was very small: only 6 pairs.

We have ruled out a number of pairs after implementing the scoring system in which we considered the consistency in dates of registration and the information on registrar's office for earlier deaths and by the later procedure described. We were left with several birth records that were now allowed to match with death records that did not achieve a best link in the first pass. We present the case of the death record '3410', that belongs to an infant who died at the age of twenty-five days (Table 9):

TABLE 9

Example pairs of resolution of ties after a first pass, based on remaining birth records

Identification Number		Combined Weight	Best link achieved by …		Pair selected as a temporary (or definitive) match
Death	Birth		…death record?	…birth record?	
3121	42431	17.85	Yes	Yes	Yes
3121	43442	17.85	Yes	Yes	Yes
3121	200212	17.85	Yes	Yes	Yes
3121	203961	17.85	Yes	Yes	Yes
3410	42431	11.94	Yes	No	No
3410	43442	11.94	Yes	No	No
3410	200212	11.94	Yes	No	No
3410	203961	11.94	Yes	No	No

Source: DATASUS/2000 and SEADE Foundation (2001)

All birth records that best-linked to death record ‘3410’ also best-linked to death record ‘3121’ with a higher composite weight. Death record ‘3121’ belonged to an infant that died in the first day of life. We thought at that time that it would be appropriate to search for the second best-link(s) for death record ‘3410’. However, not all birth records linked to death record ‘3121’ were kept after we checked on information about registration dates and registrar’s office. We ended up selecting only pairs (3121; 42431) and (3121; 200212) as definitive matches and the remaining birth records were allowed to be an option for the death record ‘3410’. Finally, we evaluated the consistency between dates of registration for each pair and also the information on registrar office and chose pair (3410; 42431) as the most likely to belong to the same infant.

A further 592 temporary matched were eliminated. The total reduction in the number of temporary matches was of 76 % (14029 to 3319 pairs). Final results are in Table 10.

TABLE 10

Final Results of the Record Linkage

Birth record per death record	Number		Percentage		Cumulative percentage	
	Pairs	Death records	Pairs	Death records	Pairs	Death records
1	2847	2847	46.2	74.1	46.2	74.1
2	854	427	13.9	11.1	60.0	85.2
3	687	229	11.1	6.0	71.2	91.2
4	536	134	8.7	3.5	79.9	94.7
5	475	95	7.7	2.5	87.6	97.1
6	318	53	5.2	1.4	92.7	98.5
7	203	29	3.3	0.8	96.0	99.3
8	152	19	2.5	0.5	98.5	99.8
9	45	5	0.7	0.1	99.2	99.9
11 +	49	4	0.8	0.1	100.0	100.0
TOTAL	6166	3842	100.0	100.0		

Source: DATASUS/2000 and SEADE Foundation (2001)

According to these results, 2827 death records were unequivocally matched (74 % of the death records), since that for 20 death records, one birth record was linked to more than one death record. Indeed, from the standpoint of the birth record, we eventually obtained that for 96 % of the pairs, the birth record involved in a match was best-linked to only one death record; for 150 pairs, to two death records. And for 94 pairs, to at least four death records.

CONCLUSION AND DISCUSSION

In this article we have described a procedure to circumvent the 'undecided-matched pair problem', when a one-to-one relationship is expected to hold, avoiding the need to undergo a full clerical review before considering first results from the record linkage. As a result, we increased the number of uniquely matched pairs from 2249 to 2827, which corresponds and increase from 59 to 74 % of the 3842 matched death records. We also reduced the number of death records best-linked to at least 4 records from 915 to 339 death records. Therefore, even though we could not find a one-to-one match for every single death record, we surely diminished the number of uncertain matches.

At least two limitations can be pointed out in this research though. First, the assumption that the district of residence of the mother at the time of the infant's birth was the same district at the time of the infant's death may not hold, especially for later deaths. Two records belonging to the same infant death might have genuinely different places of residence stated on those, since the mother may have changed district of residence between these two events. However, we believe that the failure to match records due to this reason is probably negligible, since 67% of all deaths took place in the neonatal period; 75%, before two months of life; and only 10% after 6 months of life.

Second, even though we have reduced the number of non-uniquely matched records from 1593 to 1015 death records, we recognize that this number of records with uncertain matched is far from satisfactory and clerical review might have eliminated a number of temporary matches. Our aim, however, was to show that before undergoing a full clerical review, information from first correctly matched pairs should be considered. By following the approach proposed here, we have reduced the number of uncertain matched from 14029 to 3319 pairs. Future research using record linkage should consider the combined strategies from results from first record linkage runs (such as we described here) before a full clerical review in order to most efficiently (and less costly), retrieve record matches.

ACKNOWLEDGMENTS

The first author's Doctoral studies in The Bloomberg School of Public Health, Johns Hopkins University, were fully funded by The Brazilian Agency for Post-Graduate Education (CAPES) – process number 2166/97-6 – and a preliminary version of this article was developed as part of the first author's Ph.D. dissertation submitted in November, 2002. The author's also would like to thank Dr Carlos E. C. Ferreira for thoughtful comments while the development of the first version of this manuscript as part of the dissertation thesis.

REFERENCES

CAMARGO JR. K. R. & COELI C. M., 2000. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cadernos de Saúde Pública 16: 439-447

KENDRICK S. W.; DOUGLAS M. M.; GARDNER D.& HUCKER D., 1998. Best-link matching of Scottish health data sets. Methods of Information in Medicine. 37: 64-68

MACHADO, C. J., 2002. Early Infant Morbidity and Infant Mortality in Brazil: A Probabilistic Record Linkage Approach. Tese de Doutorado, Baltimore: Bloomberg School of Public Health, The Johns Hopkins University.

MACLEOD M. C.; BRAY C. A.; KENDRICK S. W.& COBBE S. M., 1998. Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods. Computers and Biomedical Research. 31: 257-270.

NEWCOMBE H. B.; KENNEDY J. M.; AXFORD S. J. & JAMES A. P., 1959. Automatic linkage of vital records. Science. 30: 954-959

TEPPING B. J., 1968. A model for optimum linkage of records. Journal of American Statistical Association. 63: 1321-1332.

WADJA A. & ROSS L. L., 1987. Simplifying record linkage: software and strategy. Computers in Biology and Medicine. 17: 239-248

WAIEN S. A., 1997. Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data. Academic Emergency Medicine. 4: 1087-1095

WINKLER W.& SHEUREN S., 1996. Recursive analysis of linked files that are computer matched. 30 may 2002.<http://www.census.gov/srd/papers/pdf/rr96-8.pdf>