Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem

Relacionamento probabilístico de dados e um procedimento automático para minimizar o problema da incerteza no pareamento de registros

Carla Jorge Machado [1]
Kenneth Hill [2]

Abstract

Probabilistic record linkage allows the assembling of information from different data sources. We present a procedure when a one-to-one relationship between records in different files is expected but not found. Data were births and infant deaths, 1998-birth cohort, city of São Paulo, Brazil. Pairs for which a one-to-one relationship was obtained and a best-link was found with the highest weight were taken as unequivocally matched pairs and provided information to decide on the remaining pairs. For these, an expected relationship between differences in dates of death and birth registration was found; and places of birth and death registration for neonatal deaths were likely to be the same. Such evidence was used to solve for the remaining pairs. We reduced the number of non-uniquely matched records and of uncertain matches, and increased the number of uniquely matched pairs from 2,249 to 2,827. Future research using record linkage should use strategies from first record linkage runs before a full clerical review (the standard procedure under uncertainty) to efficiently retrieve matches.

Probability; Records; Cohort Studies

[1] *Departamento de Demografia, Centro de Desenvolvimento e Planejamento Regional, Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.*
[2] *Department of Population and Family Health Sciences, The Bloomberg School of Public Health, Johns Hopkins University, Baltimore, U.S.A.*

Correspondence
C. J. Machado
Departamento de Demografia, Centro de Desenvolvimento e Planejamento Regional, Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais.
Av. Augusto de Lima 1376/908, Belo Horizonte, MG 30190-003, Brasil.
carla@cedeplar.ufmg.br

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 has been used in the public health field in the last fifty years, since the seminal work of Newcombe et al. [1]. 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 information other than that provided by the matching variables – or variables com-

mon 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.

Data

We probabilistically linked data from the 1998-birth cohort of the city of São Paulo, Brazil, and our attempt was to match 3,842 infant deaths from this birth cohort to their corresponding live birth. The size of the live birth file was 209,628. Our data came from two sources: the Information Department of the Unified National Health System (SUS) [2] and the Foundation for the State Data Analysis System [3]; a description and full review of data sources and quality can be found in Machado [4]. 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 2,249 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 solve 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 pair-wise combinations between births and deaths were obtained. Therefore, we first generated 13,680,789 comparison pairs, using the *Reclink program* [5]. In the process of deciding about matched pairs, the matching variables used were birth date, birth weight, maternal age, delivery mode, sex, and plurality. Briefly, each matching weight – a value assigned to a linked pair that summarizes the comparison results of the two variables – is a result of the logarithm to the base two of a ratio between two probabilities, the *m* probability and the *u* probability ($\log_2(m/u)$). The *m* probability is the conditional probability of an agreement on a given variable if the pair is a true match; the *u* probability is the conditional probability of an agreement on a given variable if the pair is not a match. A description and estimation procedures of the matching weights for each matching variable can be found in Camargo Jr. & Coeli [5] and in Machado [4]. In Table 1 the estimated weights for each matching variable are displayed.

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 was 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 each estimated weight for each matching variable. Indeed, we had 1,800 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 [6]. One problem is the failure to match a death record with its corresponding birth 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 [7]. 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 birth 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 also be represented by the given birth record. In both situations, "ties" are generated, that is, matches that cannot be considered as definitively relating to the same infant. Another possible explanation would be a differential under-registration of births or infant deaths. Ferreira & Flores [8] pointed out that the under-recording of deaths is believed to be no greater than that of births in absolute numbers. It may be restricted to the very small percentage of live births that takes place at home, and these births are not likely to be registered either, for many of the same reasons. Indeed, in the city of São Paulo in the late 1960s, 91% of the infants who died had their birth registered and, to date, this percentage is considered to be even higher, close to 99% [8].

"Ties" were solved using the basic principle that, in searching for matches, other than 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" [9]. We sought other information in order to resolve those ties and classify pairs as matches or non-matches.

From 13,680,789 pairs, we selected for each death record its respective best-link(s) and decreased to 17,764 pairs. We then 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 decreased from 17,764 best links to 16,278 best links (a reduction of 8.4%). Examples of pairs obtained are in Table 2, and Table 3 provides a summary of these.

We kept as potential matches four pairs, the "best-linked" ones. The ordered pair ("235"; "4,709") is a definitive match, and ("200"; "11,863") and ("200"; "14,232") were considered temporary matches.

Death record "131" achieved a unique best-link with birth record "3,362". However, birth record "3,362" 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"; "3,362") as a definitive match and refuted the pair ("131"; "3,362") as such. We then searched for another match for death record "131" among the second-best links and in later stages selected ("131", "12,970") as a definitive match, since birth record "12,970" 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 which happened to only 39 death records (1% of the death records).

On average, there were 4.30 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.

Table 1

Estimated weights for matching variables.

Comparison of 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 off by two years	-0.13
Date of birth 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: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

Table 2

Examples of best-links and second best-links.

Identification number		Combined weight	Best link achieved by		Pair selected as a temporary (or definitive) match?
Death	Birth		Death record?	Birth record?	
200	4,709	0.28	Yes	No	No
200	11,863	0.28	Yes	Yes	Yes
200	14,232	0.28	Yes	Yes	Yes
200	10,516	-9.68	No	No	No
200	15,052	-9.68	No	No	No
200	15,095	-9.68	No	No	No
200	145,459	-9.68	No	No	No
235	4,709	52.76	Yes	Yes	Yes
235	11,863	16.96	No	No	No
131	3,362	9.96	Yes	No	No
131	12,970	3.57	No	No	Yes (later on it will be selected)
132	3,362	10.13	Yes	Yes	Yes
132	3,335	0.16	No	No	No
132	13,134	0.16	No	No	No
132	20,641	0.16	No	No	No

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

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	13,443	82.6	1,466	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	1.0
No tie (a one-to-one relationship established between a birth and a death record)	2,249	13.8	2,249	58.3	1.0
Death records with no best links			42*	0.9	
Total	16,278	100.0	3,842	100.0	4.3

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].
* At later stages in selecting the pairs, we defined as matches the best links for 3 out of 42 death records and remained with 39 death records whose matched pairs were found among second best-links.

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%.

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 "26,355" and "26,359". We speculate that those death records belong to non-singleton infants. Because so much identifying information was

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
5-10	4,461	642	43.8	91.4
11-21	1,589	124	8.5	99.9
100 +	5,482	2	0.1	100.0
Total	1,3443	1,466	100.0	–

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

likely to be shared, the correct matching of records belonging to twins has been recognized as a major problem [7]. 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", which achieved a best-link with birth record "27,657".

Finally, for 2,249 death records it was possible to find a unique best link, and these matches are called *unequivocally matched records*; no further efforts will be implemented to ensure 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.

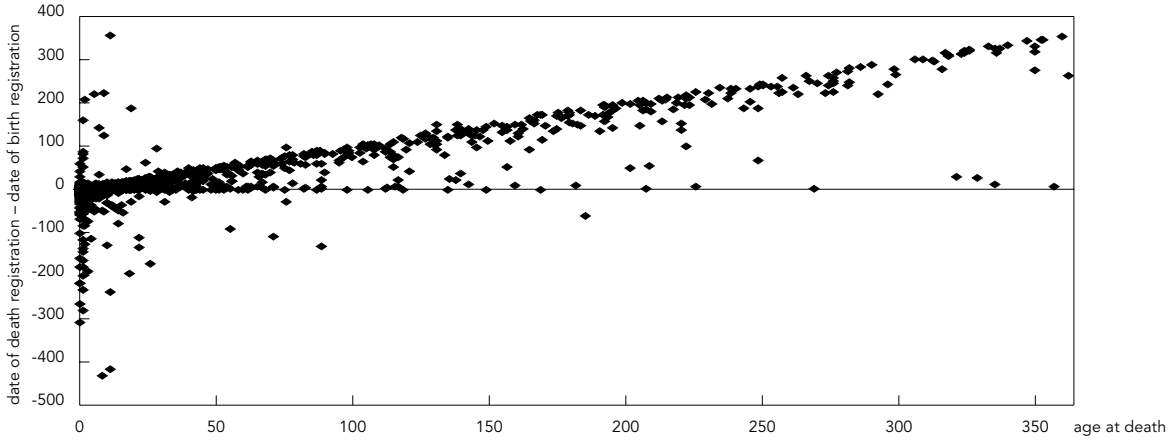
Resolution of ties

After a record linkage operation the researcher should seek other information in order to decide the matches and non-matches among ties [10]. 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: more than 80% of the death records (or 3,093 records) lack information on maternal education and more than 55% of the death records (or 2,128 records) lack information 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 & Ross [11] suggest that results from a record linkage operation obtained from an initial run through the data generally suggest opportunities for improving the linkage. Winkler & Scheuren [12] suggested a recursive method were firstly matched files provide information to a subsequent matching. Assuming that each of these 2,249 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, so the time elapsed between birth and death registration tended to be shorter than the actual time between the birth and death of the infant. Therefore, the idea was to use these 2,249 pairs as a "learning set of pairs" in order to calculate the ranges of birth and death registration for each of the 2,249 matched pairs. First we plotted the difference in dates of registration against the reported age at death of the child in number of days. Results are shown in Figure 1.

Most points follow or are slightly below the diagonal line, thus also indicating that the time elapsed between dates of registration may be

Figure 1

Scatterplot of the difference in dates of registration compared to reported age at death.
Deceased infants with unequivocally matched birth records.



slightly less than the time elapsed between the birth and the death of an infant. In fact, according to Brazilian 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 certificate for the burial. Therefore, it would be reasonable to expect that the time between registrations of the two events would lie 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 zero, which means that the birth and the death of the infant were registered on 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).

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 as the birth. 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 to 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 assigned the pair a score of one point (+1). Otherwise, a minus one point (-1) was assigned. If either the birth or the death did not possess information on date of registration, we assigned 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 for the birth and death registration was the same, we assigned the pair a score of one point (+1). A minus one point (-1) was assigned in case of discordant registrars. In

the absence of information on registrar for either the birth or the death, a null score of zero (0) was assigned. For post-neonatal deaths, we assigned 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 revisited pairs considered temporary in Table 2. Death record "200" belonged to an infant who died at 18 days (Table 6). Pair ("200"; "11,863") was selected as an unequivocally matched pair. However, sometimes we were unable to select only one pair, as for death record "277". The infant died at 13 days (Table 6).

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 13,443 to 3,917, 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 another birth record (or records). For example, death record "196" belonged to an infant who died in the first day of life, and death record "1,447" belongs to an infant who died at three months (119 days). These death records achieved a best link with other birth records at the combined weight of 0.29 (Table 7).

Since the only birth record considered to be matched to death record "196" was birth record "11,179", a one-to-one relationship was established. We then ruled out birth record "11,179" as an option for death record "1,447" because a

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: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].
Date of birth registration, in number of days, starting with January 1st, 1998 as day "one", is "X"; Date of death registration, in number of days, starting with January 1st, 1998 as day "one", is "Y"; Time elapsed between registrations in the second column relates to "Y-X".

Table 6

Example pair of resolved ties – death record "200" and death record "277".

Identification number		**Registrar's office**	**Time between birth and death registration**	**Score**		
Death	Birth			Registrar	Time between dates	Total
200	11,863	Same	4	+1	+1	+2
200	14,232	Not same	7	-1	+1	0
277	16,558	Same	2	+1	+1	+2
277	16,567	Same	0	+1	+1	+2
277	16,575	Same	1	+1	+1	+2
277	192,334	Not same	-262	-1	-1	0
277	32,173	Same	-10	+1	-1	0
277	17,649	Not same	17	-1	-1	0
277	16,491	Same	14	+1	-1	0

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

Table 7

Example pairs – death records "196" and "1,447" – evidence provided by one-to-one relationship in resolving temporary matched pairs.

Identification number Death	Birth	**Registrar's office**	**Time between registration**	**Score** Registrar	Time between dates	Total
196	11,179	Same	0	+1	+1	+2
196	16,181	Not same	-2	-1	-1	-2
196	19,630	Not same	-11	-1	-1	-2
196	20,610	Not same	7	-1	-1	-2
196	27,113	Not same	-14	-1	-1	-2
196	27,133	Not same	-15	-1	-1	-2
196	27,588	Not same	-37	-1	-1	-2
196	57,358	Not same	-59	-1	-1	-2
196	73,832	Not same	-88	-1	-1	-2
1447	11,179	–	119	+1	+1	+2
1447	16,181	–	117	+1	+1	+2
1447	19,630	–	108	+1	+1	+2
1447	20,610	–	126	+1	-1	0
1447	27,113	–	105	+1	+1	+2
1447	27,133	–	104	+1	+1	+2
1447	27,588	–	82	+1	-1	0
1447	57,358	–	60	+1	-1	0
1447	73,832	–	31	+1	-1	0

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

one-to-one relationship was established between death record "196" and "11,179", but not between "1,447" and "11,179". We notice, however, that if the only birth record left for "1,447" was "11,179", 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 six 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 "3,410", pertaining to an infant who died at the age of twenty-five days (Table 8).

All birth records that best-linked to death record "3,410" also best-linked to death record "3,121" with a higher composite weight. Death record "3,121" 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 "3,410". However, not all birth records linked to death record "3,121" were kept after we checked on information about registration dates and registrar's office. We ended up selecting only pairs ("3,121"; "42,431") and ("3,121"; "200,212") as definitive matches and the remaining birth records were allowed to be an option for death record "3,410". Finally, we evaluated the consistency between dates of registration for each pair and also the information on registrar office and chosen pair ("3,410"; "42,431") as the most likely to belong to the same infant.

A further 592 temporary matches were eliminated. The total reduction in the number of temporary matches was 76 % (from 14,029 to 3,319 pairs). Final results are in Table 9.

According to these results, 2,827 death records were unequivocally matched (74% of the death records), since 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.

Table 8

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?	
3,121	42,431	17.85	Yes	Yes	Yes
3,121	43,442	17.85	Yes	Yes	Yes
3,121	200,212	17.85	Yes	Yes	Yes
3,121	203,961	17.85	Yes	Yes	Yes
3,410	42,431	11.94	Yes	No	No
3,410	43,442	11.94	Yes	No	No
3,410	200,212	11.94	Yes	No	No
3,410	203,961	11.94	Yes	No	No

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

Table 9

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	2,847	2,847	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	6,166	3,842	100.0	100.0		

Source: Departamento de Informática do Sistema Único de Saúde [2]; Fundação Sistema Estadual de Análise de Dados [3].

Final considerations

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 2,249 to 2,827, which corresponds to and increases from 59 to 74% of the 3,842 matched death records. We also reduced the number of death records best-linked to at least four 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 are certain to have decreased the number of uncertain matches.

At least two limitations can be identified in this research. First, the assumption that the mother's district of residence at the time of her infant's birth was the same 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 them, 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% within two months of life; and only 10% after six months.

Second, even though we have reduced the number of non-uniquely matched records from 1,593 to 1,015 death records, we recognize that this number of records with uncertain matches is far from satisfactory, and that clerical review might have eliminated a number of temporary matches. Our aim, however, was to show that before undertaking 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 matches from 14,029 to 3,319 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 retrieve record matches more efficiently and with less cost.

Resumo

O relacionamento probabilístico permite que fontes de informações do mesmo registro e em bancos de dados distintos sejam unificadas. Apresenta-se um procedimento utilizado quando se espera que um registro de um banco de dados corresponda a apenas um outro num segundo banco. As fontes de dados foram os nascimentos e óbitos infantis da coorte de nascimentos de 1998, na cidade de São Paulo, Brasil. Os dados relacionados com o mais alto escore e relação unívoca foram utilizados como padrão-ouro e concorreram para a decisão sobre pares obtidos sem relação unívoca. Um comportamento esperado dos dados univocamente relacionados em termos da diferença nas datas de registro de óbito e de nascimento, e também dos locais de registro de nascimento e de óbito para óbitos neonatais foi observado e, aplicou-se esta relação aos demais dados. O número de pares com relação unívoca aumentou substancialmente, de 2.249 para 2.827, e diminuiu o número de nascimentos ligados a um óbito. Este procedimento deve ser associado à revisão manual (procedimento padrão na presença de incerteza) a fim de conseguir um pareamento eficiente.

Probabilidade; Registros; Estudo de Coortes

Contributors

C. J. Machado conceived the study, performed the probabilistic matching, and drafted the manuscript. K. Hill contributed by advising the student throughout the dissertation development process.

Acknowledgments

The first author's Doctoral studies at The Bloomberg School of Public Health, Johns Hopkins University, were fully funded by The Brazilian Agency for Post-Graduate Education (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 authors would also like to thank Dr Carlos E. C. Ferreira for thoughtful comments during the development of the first version of this manuscript as part of the dissertation.

References

1. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959; 30:954-9.
2. Departamento de Informática do Sistema Único de Saúde. Sistema de informações de nascidos vivos de 1998. http://tabnet.datasus.gov.br/cgi/sinasc/dados/ (accessed on 20/May/2000).
3. Fundação Sistema Estadual de Análise de Dados. Informações sobre mortalidade 1998/1999: São Paulo. São Paulo: Fundação Sistema Estadual de Análise de Dados; 2000.
4. Machado CJ. Early infant morbidity and infant mortality in Brazil: a probabilistic record linkage approach [PhD Thesis]. Baltimore: Bloomberg School of Public Health, The Johns Hopkins University; 2002.
5. Camargo Jr. KR, Coeli CM. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cad Saúde Pública 2000; 16:439-47.
6. Macleod MC, Bray CA, Kendrick SW, Cobbe SM. Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods. Comput Biomed Res 1998; 31:257-70.
7. Kendrick SW, Douglas MM, Gardner D, Hucker D. Best-link matching of Scottish health data sets. Methods Inf Med 1998; 37:64-8.
8. Ferreira CEC, Flores LPO. The dimensions of infant mortality in São Paulo. Brazilian Journal of Population Studies 1998; 1:145-64.
9. Tepping BJ. A model for optimum linkage of records. J Am Stat Assoc 1968; 63:1321-32.
10. Waien SA. Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data. Acad Emerg Med 1997; 4:1087-95.
11. Wadja A, Ross LL. Simplifying record linkage: software and strategy. Comp Biol Med 1987; 17:239-48.
12. Winkler W, Scheuren S. Recursive analysis of linked files that are computer matched. http://www.census.gov/srd/papers/pdf/rr96-8.pdf (accessed on 30/May/2002).

Submitted on 25/Mar/2003
Final version resubmitted on 01/Oct/2003
Approved on 28/Oct/2003