A literature review of record linkage procedures focusing on infant health outcomes

Procedimentos para relacionamento de registros: revisão bibliográfica com enfoque na saúde infantil

Carla Jorge Machado [1]

[1] *Departamento de Demografia, Centro de Desenvolvimento e Planejamento Regional, Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.*

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

Abstract

Record linkage is a powerful tool in assembling information from different data sources and has been used by a number of public health researchers. In this review, we provide an overview of the record linkage methodologies, focusing particularly on probabilistic record linkage. We then stress the purposes and research applications of linking records by focusing on studies of infant health outcomes based on large data sets, and provide a critical review of the studies in Brazil.

Child Welfare; Records; Record Linkage

Introduction

Public health researchers have used multiple opportunities for record linkage processes between datasets. The primary purpose of linking different data sets is to increase the amount of information available on a unit of analysis that can be an individual or other unit. As a general principle, any record file can undergo record linkage, as long as there is adequate identification in each record to be linked, but that is not always the case. In this review, we provide an overview of the record linkage methodology, focusing particularly on probabilistic record linkage. We then stress the purposes and research applications of linking records by focusing on studies on infant health outcomes. This review was presented as part of a doctoral dissertation submitted to the Department of Population Dynamics, Bloomberg School of Public Health, Johns Hopkins University, in November, 2002 [1].

Material and methods

We conducted this review from March 2000 to August 2002, as part of ongoing work on the doctoral dissertation. We searched for articles in PubMed and MEDLINE (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi), and a few documents (mostly technical reports or conference papers) on Google (http://www.google.com). For Brazil, we searched in SciELO (Scientific Electronic Library Online: http://www.scielo.br), with three exceptions: one article was found in the journal *Informe Epidemiológico do SUS* by manual search conducted in issues since 1992 at the Federal University in Minas Gerais – Library of the Center for Development and Regional Planning (CEDEPLAR/UFMG) in July 2001. Two other doctoral dissertations submitted at Universities in Brazil were obtained through personal communication. We selected the seminal studies on record linkage, and in order to obtain a "state-of-the-art" picture we chose recent studies (from the late 1990s onward) whenever possible. In Machado [1], 70 studies were reviewed, and the 42 most informative were chosen for this article. We selected studies that used large data sets and focused exclusively or primarily on infant health outcomes, for example, birth certificates linked to death certificates or medical records linked to death records. With the exception of Brazil, all studies located in the search were from developed countries.

Results

Introduction

The idea of record linkage is by no means new. Jenner's research on cowpox and smallpox vaccination in the late 18th century produced a detailed record system that linked all sorts of cows with human beings. Jenner's motivation to link data was the fact that there was something "special" about the material taken from the cows which, when injected into humans, produced protection against smallpox [2]. Jenner's linked records were used as evidence that the intervention was efficacious. Today, historians consider Jenner's breakthrough a turning point in Western medicine [3].

In the 20th century, the first time the term "record linkage" appeared in the literature was in Dunn [4,5]. There is a distinction between the two main types of record linkages. *Deterministic record linkage* links pairs of records on the basis of whether they agree on specific identifiers; *probabilistic record linkage* uses probabilities to determine whether a pair of records refers to the same individual. Deterministic record linkage can be undertaken whenever there is a unique identifier, such as a personal identification code. In addition, if the researcher possesses a fair amount of identifying information for individuals in two different files, these pieces can almost uniquely identify an individual and allow a deterministic linkage. A critique of the deterministic match rules is that they do not adequately reflect the uncertainty that may exist for some potential links [6].

Newcombe et al. [7] describe the problem of automatically linking records from separate hospital databases. This is the very first mention in the literature about the possibility of probabilistically linking records. The authors wanted to link records of individuals exposed to low levels of radiation to determine the causes of their eventual death, the impact on their fertility, and later genetic defects in the Province of British Columbia, Canada. Howard Newcombe is indeed considered the scientist who had crucial insights that led to computerized approaches to record linkage in the absence of a unique identifier. He observed that the frequency of occurrence of a characteristic, such as surname, among matches and nonmatches could be used to compute a *score* or *matching weight* associated with the linking of two records (links are pairs formed, not necessarily referring to the same individual or entity, while a match is a pair formed that refers to the same individual or entity). For example, an agreement involving surnames "Smith" and "Smith" is certainly more likely to happen among unmatched pairs than an agreement between "Rothschild" and "Rothschild" (phonetic indexes such as Soundex or NYSIIS have been used in the context of deterministic and probabilistic record linkage; their key feature is that they code names based on the way they sound rather than on how they are spelled). Newcombe also observed that matching weights over different variables, such as age, surname, and first name, should be computed and added to obtain an overall matching weight. Pairs with higher values would be considered matches. Fellegi & Sunter [8] built on Newcombe's intuition and introduced the mathematical and statistical foundation for probabilistic record linkage, which is extensively used to the present. Fellegi & Sunter's approach is an extension of the classical theory of hypothesis testing, and can be summarized as follows: (1) In order to establish a summation of the matching weights, they had to be assumed to be statistically independent. (2) It is not always possible to consider a pair as a *match* or as a *non-*

match, and therefore a third category, a *possible match*, should be introduced. Fellegi & Sunter [8] established decision rules based on fixed upper bounds of rates of false matches and false non-matches that minimize the number of cases that need to be manually reviewed. (3) Hypothesis-testing theory was used to assert that there are error probabilities involved in the decision to classify pairs as *matches* or *non-matches*. (4) A mechanism was established in comparing two sets of records in order to avoid having to deal with every single pair-wise comparison; the authors suggested the use of a *blocking* mechanism which essentially requires that to be suitable for comparison, records should agree exactly on a unique variable, such as place of residence or date of birth.

Much research has been done since this groundbreaking work. It has been recognized that when there is a high probability that one individual in one file is uniquely represented in the other file, the power of record linkage can increase immeasurably [9]. Once a best link is established between an individual in the first file and an individual in the second file, on the basis of having the highest probability weight, the process of searching the corresponding record in the second file is terminated for the record in the first file. Another area that has evolved enormously since the late 1960s is computer technology, which boosted the development of much software suitable for record linking. To date, the software developed by Matthews Jaro in 1985 [10], called Automatch©, has been used extensively in the United States by individual and group researchers and also by the Census Bureau. In Canada, considered well advanced since the early 1980s in linking national data systems, especially to monitor health [11], the software GRLS (Generalized Record Linkage System) was developed at Statistics Canada [12].

The use of the commercial software Automatch© was considered unfeasible in Brazil due to the high cost of its private license (US$ 12,500 in 1999 for the least expensive version)[13]. Meanwhile, Camargo Jr. & Coeli [13] developed the Recklink© software, which implements Fellegi & Sunter's theory and works much the same as Automatch©.

Research application on record linkage: infant health outcomes

The need to link birth records and death records to assess risk factors for perinatal and infant outcomes has been recognized by researchers in developed countries [12]. The rationale is that birth records provide information about the parents' characteristics, the mother's use of prenatal care, and neonatal health status, including birth weight and gestational age [14], whereas death records provide information on the age at death and causes of death. In Brazil, where data on birth weight and gestational age are collected on birth records and death records, the quality of information on death records is far inferior to that in birth records. Therefore there is a need to retrieve information from the birth records.

Combining information from birth records and death records allows one to assess several associations. For example, it allows measurements of birth weight-specific infant mortality rates (Wilcox & Russell [15] formulated a model of birth weight-specific infant mortality that is widely used in epidemiology). Based on a series of studies in 1983 and 1986, they showed that reduced birth weight is not sufficient in itself to increase mortality and hence should not be used as a surrogate of adverse outcome, since the optimum birth weight maintains a fairly constant distance from the mean weight. This suggests that a shift in the birth weight distribution yields an equivalent shift in the mortality curve. If all else is constant, such a shift produces no net effect on infant mortality [16]. In the field of epidemiology, it was only with the series of studies by Wilcox & Russell that the implications of such a phenomenon in populations started being recognized. Their work provided a powerful conceptual tool for understanding birth weight and mortality [17]. All of this was made possible by using a linked data set of birth and death records from 49,000 early neonatal deaths in the United States in 1960 [15]. In a later study they matched births and deaths in North Carolina from 1970 to 1973 and used this linked data to study birth weight differentials in black and white infants in relation to their respective birth weight distributions [18] and soundly established their theory. Previous studies had been severely weakened by their small data sets [17]. Today it is almost unacceptable not to review Wilcox & Russell's studies on birth weight and perinatal survival given the sound basis of their findings.

Studies in selected countries

- **United States**

The United States has compiled national linked birth and death files on a routine annual basis, starting with the 1983 birth cohort. To date, the combined analysis of birth and infant death certificates is considered a mainstay of maternal and child health epidemiology in the country [19]. Major issues underlie the infant mortality problem in the United States assessed by linking birth and death records. For example, the data set showed that infants born weighing less than 1,500 grams account for less than two percent of all births but half of all deaths countrywide [14]. The gap between black and white infant mortality rates is an additional feature in the U.S. infant mortality picture. Birth weight for blacks is more likely to be less than 2,500 grams (and more markedly, less than 1,500 grams) as compared to whites, but the neonatal mortality rate was actually lower among black infants weighing less than 1,500 grams as compared to whites, which is in accordance with Wilcox & Russel's framework. In contrast, the post-neonatal mortality rate is twice as high among infants 2,500 grams and above for black than for white infants. It was thus possible to identify one of the major elements associated with general infant mortality in the United States as well as the major racial disparities.

The State and national linked birth and infant death file is generated essentially by matching the death and the birth record of the deceased infant by using highly specific information, such as mother's and father's names. Other information is also used as needed. Combined, these several pieces of information can be seen as very specific and almost uniquely identify the infant. Most studies using birth and infant death records in the United States employ deterministic record linkage, which Scheuren [6] refers to as an operation that relies on multiple exact matches. However, we are aware of a few exceptions. Adams et al. [20] linked 1.4 million fetal deaths and birth certificates filed in Georgia from 1982 to 1990 in order to build birth histories for women. Mother's name was used, in addition to her social security number whenever available. The authors mentioned problems in the linkage related to the under-recording of fetal deaths and incorrect linkages of mothers' offspring in the case of twins (having similar names, same date of birth, etc.). Holian [21] also discussed the possibility of linking birth and infant death records probabilistically in each State and used the example of Cleveland, Ohio. The two studies were essentially methodological.

Attempts to use probabilistic record linkage in United States to study birth outcomes seem to be an option in case there is a need to extend the linkages to other databases. Holian [22] linked data from the comprehensive maternity and infant care project (MIC) in Cleveland and the adjoining suburbs of East Cleveland with a hospital database by deterministic record linkage, in order to add more personal identifiers to the MIC data. The second step was linking this newly constructed database with live birth and stillbirth certificates, by means of a combination of deterministic and probabilistic procedures. Eventually, 85 percent of all records were matched, but the author seemed to consider this percentage too low to derive conclusions from the data. The discussion focuses on the failure of linking infants to their respective birth (or stillbirth) registration rather than on a discussion of the substantive findings based on the matched data [22]. Bell et al. [23] linked data from 1.46 million Medicaid claims with 53,000 birth records from California vital statistics for very low birth weight infants, from 1980 to 1987. The task involved a large amount of missing data, errors, and variations in names, and the authors reported that without probabilistic record linkage, the analysis of this combined data would have been impossible or prohibitively expensive.

- **Canada**

By means of record linkage, Chen et al. [24] examined, in Quebec, differences in fetal and infant mortality by maternal and infant characteristics and concluded that marked differences existed between educational groups that were substantially reduced after controlling for gestational age, birth weight, and smoking. Higher mortality rates were found in women with less than 12 years of education, especially in the post-neonatal period and for non-congenital diseases. In another study, all births and infant deaths in 1985-87 and 1992-94 in Canadian provinces were analyzed to assess the impact of the increased use of labor induction for post-term pregnancies in recent years [25]. They found a reduction in post-neonatal mortality in 1992-94 in comparison to 1985-87 and suggested that the assessment of the efficacy of labor induction for post-term pregnancies should be expanded to include the post-neonatal period. These findings, obtained by record linkage, were in disagreement with most randomized clinical trials but are probably a better reflection of current clinical practice, and suggested new avenues for research. In addition, changes in stillbirth and infant mortality rates for triplets were analyzed for the period 1985-90 to 1991-96 and were shown not to experience the significant decline that the overall fetal and infant mortality experienced [26].

In all studies, probabilistic record linkage was used. According to Fair et al. [12] probabilistic linkage is considered the preferred linkage method in Canada. Indeed, the authors assessed the validity of probabilistically linking birth and infant deaths in Canada and showed that 99 percent of infant deaths in the Nova Scotia provincial data were successfully located in the Public Health Statistics Canada file. Furthermore, the fact that the theory of probabilistic record linkage was established by two Canadian statisticians, Ivan Fellegi and Allan Sunter, may play a role in the methodology's more widespread acceptance by researchers in that country.

- **Scotland**

In the field of health research, probabilistic matching has been used in Scotland to identify records belonging to an individual since a unique patient identification number has not been in general use. Over the last ten years, event histories for patients have been built through the use of this methodology [27]. Maternity records have been matched to birth records to verify whether maternal hypertensive disease alters the association between intrauterine growth retardation (IUGR) and neonatal mortality in preterm infants. In fact, there is considerable debate on this issue in the literature. Most authors believe that IUGR is a risk factor for neonatal mortality, while others believe that neonatal mortality is unaltered and still others believe that IUGR exerts a protective effect on neonatal mortality [28]. Maternity records were subsequently matched to mortality records [28] in an extended linkage operation. Maternity records provided information on maternal morbidities; birth records on birth weight and gestational age; and death records, on status of the infant (death in the neonatal period). IUGR was shown to be a risk factor for neonatal death and the presence of maternal hypertensive disease did not alter the incidence of neonatal death. The linkage allowed the authors to investigate a substantial number of infants and to achieve the necessary power to assess the associations. Small studies would not allow the assessment of associations in which the probability of the outcome measure is low, as well as the number of hypertensive morbidities, with such a degree of capacity for generalization.

Scotland has a number of registers under the supervision of its Information and Statistics Division and National Health Service [27]. Established in 1990, the Scottish Register of Children with a Motor Deficit of Central Origin (SRCMDCO) contains information on the corresponding cases born from 1984 onwards. In order to calculate cerebral palsy rates, routinely collected maternity data from the Scottish Morbidity Record series (SMR2) were probabilistically linked to cases of cerebral palsy in the SRCMDCO. In this record linkage operation, each cerebral palsy record was allowed only to make a best possible match to a corresponding SMR2 record [29], since it was known that each cerebral palsy record would correspond to only one maternity record from the SMR2. The aim of the study was also to investigate whether or not twinning is a risk factor in itself for cerebral palsy after controlling for other factors considered predictive of cerebral palsy. The authors confirmed that being a twin poses an independent risk for cerebral palsy.

- **Sweden**

Record linkage between registers in Sweden is immensely facilitated by the fact that identification numbers are given to every resident and are unique for each individual. These numbers are extensively used and accepted throughout society, which allows deterministic record linkages.

Four central health registers in Sweden were used to assess specific effects of maternal age, parity, educational level, and smoking for specified causes of neonatal deaths and stillbirths [30]. Those registers are the Medical Birth Registry, which contains medical data on 99 percent of all delivered infants; the Registry of Congenital Malformations; the Child Cardiology Register; and the Cause of Death Registry. A record linkage was made to the Registry of Education, where the mothers' level of education could be assessed. All linkages were made possible by using the infant's date of birth combined with the identification number of the mother or the infant. Among the authors' many conclusions, smoking aggravates the risk of death from *abruptio placentae*; in addition, the risk of death for IUGR infants appears to be greater among higher-parity as compared to lower-parity mothers.

- **Norway**

In Norway, as in Sweden, a unique identification number is assigned to all residents. Medical birth registries were established in Norway in 1967, with unique identification numbers, which facilitates deterministic record linkage between several data sets [31]. Stene et al. [32] designed a cohort study linking records in the medical birth registry with those in the national childhood diabetes registry, which contains all newly diagnosed cases of type-1 diabetes in children under 15. The authors were able to match 1,824 of the 1,863 cases of type-1 diabetes diagnosed from 1989 to 1998. Maternal and paternal age was obtained in the medical birth registry, and the birth order of the child was inferred from the number of previous live births. The authors concluded that the risk of diabetes in first-born children is not associated with maternal age, but increasing maternal age is a risk factor in children of second or higher order. This study was important because many other studies have failed to detect a significant association, and this is attributed partially to small sample sizes [32].

Another study aimed to assess the association between birth defects and paternal occupational exposure. All births in Norway from 1970 to 1993 were linked to the population censuses of 1970, 1980, and 1990. Presumably the paternal identification number in the child's birth register was linked to his identification number in the census. More than 1 million births were matched [33], and it was possible to make several specific associations. Some associations observed in the literature were confirmed (such as a tendency towards increased frequency of spina bifida among children of painters), but others were not (such as a tendency towards increased frequency of cleft lip/palate and neural tube defects in children of fathers in sales-related occupations). The authors discuss in detail the possibilities of misclassification of the exposure (occupation) and outcome (malformations) and the implications for such a study.

- **Denmark**

Since 1968, all Danes have been given a 10-digit personal registration number at birth that is used in all Danish data sources. It makes the linkage between registries simple and valid [34]. Linkages are thus performed with a deterministic methodology. Sorensen et al. [34] matched data on cognitive function obtained from the draft board with data from the Danish birth registry. The authors studied all men born in Denmark after January 1973 who were drafted at age 18 and resided in the fifth conscript Jurisdiction of Denmark (5,183 men). Low birth weight was found to be negatively associated with early adulthood cognitive performance. This study was not subject to self-selection bias, such as studies in which individuals are invited to participate. For example, in England individuals were identified in the birth registries and then invited to be interviewed; if they accepted, they underwent cognitive testing [35]. About 47 percent of the identified individuals did not want to be tested (1,576 out of 3,318 individuals). If individuals with higher cognitive performance were more likely to participate (a reasonable assumption) then a positive association would not be found. Indeed the study by Martyn et al. [35] failed to show an association between fetal growth and poorer cognitive performance. In the Danish study the population is larger and is not self-selected, which was an advantage obtained from the record linkage study.

More recently, using the same cohort of Danish men, it was possible to examine the associations between fetal growth indicators and sight and hearing. A birth weight of less than 3,000 grams and body mass index at birth of less than 3.4 were associated with reduced visual acuity and impaired hearing [36]. These studies are important for testing whether biologically plausible hypotheses hold in large populations and can thus suggest new pathways for research and explanations for findings.

- **England**

One study in England may well exemplify the importance of record linkage in studies associated with neonatal and maternal health in the era of HIV/AIDS. Interpreting seroprevalence trends is usually difficult because of the limited amount of demographic information retained in each sample. In trying to overcome that difficulty, Ades et al. [37] studied surveys in which computer records holding data on mother's age, ethnic status, and both parents' country of birth had been electronically linked to samples prior to anonymization tagging and testing. The neonatal seroprevalence surveys were matched electronically with data from child health computer systems (that had information on maternal age and ethnicity) and birth registration (parents' country of birth). The results showed that even though there is an increase in prevalence of HIV/AIDS in Southern Asia and Africa, the spread of HIV within the UK-born Southern Asian and Sub-Saharan African communities is very limited. A very low seroprevalence rate was observed in infants of mothers or fathers born in Southern Asia, in spite of cultural and travel ties to high-prevalence countries. The authors suggest that this kind of data linkage helps monitor the impact of the worldwide epidemic on prevalence and incidence locally [37].

- **Japan**

Due to excellent data quality in Japan, which allows very accurate matching and high matching rates [38], it has been possible to do some extended linkages using several databases. Miura et al. [39] investigated whether birth weight and also childhood growth (especially rate of height increase in childhood) were independently related to high serum cholesterol and/or high blood pressure. The authors matched clinical records from infants with records at one year, three years, and then at 20. A high number of individuals in a single city (5,127) had their records matched, using the names and date of birth, in a two-item deterministic record linkage. The authors found that birth weight and growth rate in early childhood were associated with subsequent cardiovascular disease risk factors. This population-based study adds considerably to the so-called "Barker hypothesis", which asserts that an infant's nourishment before birth and during infancy, as manifested in fetal and infant growth patterns, is a determinant of the development of risk factors for coronary heart disease [40]. There has been considerable controversy on this issue in the literature [41,42]. All relevant evidence so far comes from the retrospective study of cohorts of countries that possess good vital registration of events over a long period of time, such as Finland [42], Australia, and the United Kingdom [43].

- **Brazil**

In Brazil the idea of linking birth and infant death records in a more systematic way is relatively recent. It is closely related to the introduction of the Information System on Live Births (SINASC), which established routine physician reporting of variables like birth weight and gestational age in the birth records. Several methodological studies since the mid-1990s have explored the possibilities of linking birth and death records. Almeida & Mello-Jorge [44] linked birth and death records of infants who died in the first semester of 1992 for mothers who resided in the city of Santo André, São Paulo State, at the time of the birth. The linkage was deterministic. First, records were computer-linked using sex and date of birth. Then, within this set of possible birth records for each death record, there was a manual search of the corresponding birth record by mother's name. Since there were 3,225 birth records and 66 death records, there were initially an average of 49 birth records for each death record, later reduced after sorting by sex and date of birth; on this reduced scale, the matching was a feasible task. Almeida & Mello-Jorge [45] used the matched data in a later study and concluded that IUGR, even after controlling for categories of gestational age, significantly increased the risk of neonatal death. Older maternal age (35 or older) and lower maternal schooling (less than completed primary school) also posed extra risks to neonatal mortality in that population.

Fernandes [46] linked birth and death records of infants born in 1989, 1990, and 1991 in the Federal District of Brazil. The linkage was based primarily and almost solely on the name of the mother and was done manually and involved many clerks and research assistants. Eventually, of the 3,062 infant deaths, it was possible to match 2,943 to their respective birth records. An important result of this study was the recovery of much information from the birth records not otherwise known for infants who had died. The percentage of missing information on the death records for mother's age, delivery mode, and birth weight decreased after the record linkage, from 30 percent to 4 percent; from 29 percent to 7 percent; and from 43 percent to 18 percent, respectively. The recovery of information was substantial for all variables, and especially for maternal age.

Noronha et al. [47] automatically and deterministically linked birth and neonatal infant death records for mothers residing in the city of Rio de Janeiro. The infants belonged to the 1996 birth cohort (second semester). The authors were able to identify and match 556 deaths with their corresponding birth records by using almost exclusively the mother's name that was typed in both databases. There was no mention of using phonetic codes such as Soundex. Birth weight and sex were used in case of homonymous mothers. As in Fernandes [46], the authors were able to recover considerable information reported as missing from or not recorded on the death record. The matching process substantially reduced the percentage of missing information for all variables in death records. For example, for birth weight and gestational age the percentage of missing data decreased from 8% to 1%; for maternal education, from 21% to 1.3%; for multiple pregnancy, delivery mode, and maternal age it decreased respectively from 6.3%, 6.5%, and 19.1% to zero in all three variables. This study was essentially methodological, and the authors stressed the importance of using birth records to retrieve information to supplement death record information.

Morais Neto & Barros [48] used the same procedure described in Almeida & Mello-Jorge [44], a combination of automated and later manual search of records by mother's name, and matched birth and infant death records for mothers whose city of residence was Goiânia (Goiás State, in Central-West Brazil) at the time of childbirth. The authors matched 342 infant death records with their corresponding birth records for infants from the 1992 birth cohort. Birth weight under 2,500 grams, gestational age below 37 weeks, non-cesarean delivery, and birth in a public hospital were independently associated with neonatal mortality. The variables birth weight below 2,500 grams and birth in a public hospital were independently associated with post-neonatal mortality. Low maternal education was found to be significantly associated with post-neonatal mortality, and illiteracy posed extra risks for post-neonatal death, but not for neonatal death.

Cunha [49] used an automated and deterministic record linkage procedure to match birth and infant death records for infants from the 1997 birth cohort in the State of São Paulo. Maternal names were not available for this linkage. The author, in a first phase, used four variables to conduct the deterministic linkage: exact date of birth, sex, city of residence within the State of São Paulo, and maternal age. Records with missing values for any of the four variables were excluded from the record linkage operation. In addition, since the author wanted to study racial differentials in infant survival, records with missing values for the race variable were also excluded. The author found 7,260 possible birth records for 5,820 death records, or about 1.25 birth records for each death record. In order to decide which birth record was truly matched to each of the 5,820 death records, there was a comparison of all other common variables, such as birth weight or delivery mode, and a decision was made based on agreements and disagreements on these variables, using a deterministic approach. For the final analysis, the author decided that in case the information in birth and death records was in disagreement, for a single variable, the information on the birth records would be used, rather than that on the death record.

The author reported several important findings. As compared to white infants, black infants who died before the age of one year were more likely to have been born to multiparous mothers. Infants of mothers with less schooling showed a higher post-neonatal mortality rate, and the mothers were more likely to have declared at least one abortion. In a multivariate analysis in which the outcome variable was infant death (< 1 year of age), Apgar scores of less than seven at one and five minutes were the strongest predictors of infant mortality. Other factors independently predictive of infant mortality were: birth weight below 2,500 grams, gestational age below 37 weeks, race (black), maternal education below secondary level, maternal age less than 20 or more than 35, fewer than 7 prenatal care visits, non-singleton birth, and non-cesarean delivery [49].

Discussion and conclusion

Record linkage processes can be deterministic or probabilistic. In this review we aimed to provide evidence on what Scheuren [6] (p. 419) stated after 25 years of experience in record linkage: "*Record linkage can aid a society in achieving advances in the well being* [of] *its citizens (...).* [The] *epidemiological literature is full of health studies that use record linkage techniques to advance knowledge*". In countries where there is a longstanding tradition of a unique identifier for all citizens and residents, record linkage is a straightforward task; Sweden, Norway, and Denmark are among these countries. In Scotland, England, and Japan, a combination of good data quality and use of some personal identifiers has allowed researchers to conduct record linkages with a high degree of success and accuracy, using probabilistic and deterministic methods. In Canada much importance is placed on the record linkage potential, and researchers appear to see the probabilistic record linkage as the "road to follow" for the next century [12]. We have speculated that this is also related to a "milieu" in the research area of health and statistics in Canada. In the United States, some research uses probabilistic methodology – the most notable example is the Post-Enumeration Survey carried out to evaluate U.S. census coverage [6]. In the area of infant health and mortality, the vast majority of studies made use of multiple-item deterministic record linkage.

Brazil is in the very early stages of record linking. Most studies on infant health that used record linkage were based on a combination of deterministic automated and manual record linkage, while a few used probabilistic record linkage, a methodology that takes into account the uncertainty that may exist in personal identifiers. Based on our search, no other developing country has conducted such a body of studies using the method of record linking of large datasets. Researchers should consider utilizing record linkage procedures in order to link data to increase knowledge on infant health in Brazil.

Resumo

O relacionamento de dados é um instrumento metodológico importante que possibilita que diferentes fontes de informações sejam unificadas em um só registro. Este procedimento é utilizado por vários pesquisadores na área de saúde pública. Neste artigo, faz-se uma revisão das metodologias de relacionamento de dados, especialmente do relacionamento probabilístico de registros. Enfatiza-se os motivos e as aplicações de pesquisa do relacionamento de dados, focalizando estudos sobre saúde infantil que fizeram uso de bancos de dados com um grande número de registros. Finalmente, faz-se uma revisão crítica dos estudos de relacionamento de dados nesta área, no Brasil.

Saúde Infantil; Registros; Relacionamento de Dados

References

1. Machado CJ. Early infant morbidity and infant mortality in Brazil: a probabilistic record linkage approach [PhD Thesis]. Baltimore: Bloomberg School of Public Health, Johns Hopkins University; 2002.
2. Wendel HF. Medical record linkage – we need it now. J Clin Comput 1984; 13:72-9.
3. Watts S. Epidemics and history: disease, power and imperialism. New Haven/London: Yale University Press; 1997.
4. Dunn HL. Record linkage. Am J Public Health 1946; 36:1412-6.
5. Smith ME. Record linkage: present status and methodology. J Clin Comput 1984; 13:52-71.
6. Scheuren F. Linking health records: human rights concerns. Proceedings of an International Workshop and Exposition: Record Linkage Techniques; 1997 March 20-21; Arlington, United States. Washington DC: National Academy Press; 1999.
7. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959; 30:954-9.
8. Fellegi IP, Sunter A. A theory of record linkage. J Am Statist Assoc 1969; 64:1183-210.
9. 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.
10. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995; 14:491-8.
11. Beebe GW. Record linkage systems: Canada vs the United States. Am J Public Health 1980; 70:1246-8.
12. Fair M, Cyr M, Allen AC, Wen SW, Guyon G, MacDonald RC. An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. Chronic Dis Can 2000; 21:8-13.
13. Camargo Jr. KR, Coeli CM. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cad Saúde Pública 2000; 16:439-47.
14. Buehler JW, Prages K, Hogue CJR. The role of linked birth and infant death certificates in maternal and child health epidemiology in the United States. Am J Prev Med 2000; 19:3-11.
15. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: II. On weight-specific mortality. Int J Epidemiol 1983; 12:319-25.
16. Wilcox AJ. On the importance – and the unimportance – of birthweight. Int J Epidemiol 2001; 30:1233-41.
17. David R. Commentary: birthweights and bell curves. Int J Epidemiol 2001; 30:1241-3.
18. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: III. Towards a new method of analysis. Int J Epidemiol 1986; 15:188-96.
19. Buehler JW, Prages K, Hogue CJR. The role of linked birth and infant death certificates in maternal and child health epidemiology in the United States. Am J Prev Med 2000; 19:3-11.
20. Adams MM, Wilson HG, Castro DL, Berg CJ, McDermott JM, Gaudino JA, et al. Constructing reproductive histories by linking vital records. Am J Epidemiol 1997; 145:339-48.
21. Holian J. Live birth and infant death record linkage: methodological and policy issues. J Health Soc Policy 2000; 12:1-11.
22. Holian J. Client and birth record linkage: a method, biases, and lessons. Eval Pract 1996; 17:227-35.
23. Bell RM, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. Med Care 1994; 32:1004-18.
24. Chen J, Fair M, Wilkins R, Cyr M, Fetal-Infant Study Group of the Canadian Perinatal Surveil-

lance System. Maternal education and fetal and infant mortality in Quebec. Health Rep 1998; 10:53-64.

25. Wen SW, Joseph KS, Kramer MS, Demissie K, Oppenheimer L, Liston R, et al. Recent trends in fetal and infant outcomes following post-term pregnancies. Chronic Dis Can 2001; 22:1-5.
26. Joseph KS, Marcoux S, Ohlsson A, Kramer MS, Allen AC, Liu S, et al. Preterm birth, stillbirth and infant mortality among triplet births in Canada, 1985-96. Paediatr Perinat Epidemiol 2002; 16:141-8.
27. Walsh D, Smalls M, Boyd J. Electronic health summaries – building on the foundation of Scottish Record Linkage System. Medinfo 2001; 10: 1212-6.
28. Chard T, Penney G, Chalmers J. The risk of neonatal death in relation to birth weight and maternal hypertensive disease in infants born at 24-32 weeks. Eur J Obstet Gynecol Reprod Biol 2001; 95:114-8.
29. Bonellie SR, Currie D, Chalmers J. Comparison of risk factors for cerebral palsy in twins and singletons. Edinburgh: Napier University; 2000. (Applied Statistics Technical Report Number 9).
30. Winbo I, Serenius F, Dahlquist G, Källén B. Maternal risk factors for cause-specific stillbirth and neonatal death. Acta Obstet Gynecol Scand 2001; 80:235-44.
31. Bakketeig LS. Perinatal epidemiology: a Nordic challenge. Scand J Soc Med 1991; 19:145-7.
32. Stene LC, Magnus P, Lie RT, Sovik O, Joner G. Maternal and paternal age at delivery, birth order, and risk of childhood onset type 1 diabetes: population based cohort study. Br Med J 2001; 323:369.
33. Irgens A, Kruger K, Skorve AH, Irgens LM. Birth defects and paternal occupational exposure. Hypotheses tested in a record linkage based dataset. Acta Obstet Gynecol Scand 2000; 79:465-70.
34. Sorensen HT, Sabroe S, Olsen J, Rothman KJ, Gillman MW, Fischer P. Birth weight and cognitive function in young adult life: historical cohort study. Br Med J 1997; 315:401-3.
35. Martyn CN, Gale CR, Sayer AA, Fall C. Growth in utero and cognitive function in adult life: follow up study of people born between 1920 and 1943. Br Med J 1996; 312:1393-6.
36. Olsen J, Sorensen HT, Steffensen FH, Sabroe S, Gillman MW, Fischer P, et al. The association of indicators of fetal growth with visual acuity and hearing among conscripts. Epidemiology 2001; 12: 235-238.
37. Ades AE, Walker J, Botting B, Parker S, Cubitt D, Jones R. Effect of the worldwide epidemic on HIV prevalence in the United Kingdom: record linkage in anonymous neonatal seroprevalence surveys. AIDS 1999; 13:2437-43.
38. Iwasaki T, Miyake T, Ohshima S, Kudo S, Yoshimura T. A method for identifying underlying causes of death in epidemiological study. J Epidemiol 2000; 10:362-5.
39. Miura K, Nakagawa H, Tabata M, Morikawa Y, Nishijo M, Kagamimori S. Birth weight, childhood growth, and cardiovascular disease risk factors in Japanese aged 20 years. Am J Epidemiol 2001; 153:783-9.
40. Barker DJ. The fetal origins of adult hypertension. J Hypertens 1992; 10:S39-44.
41. Paneth N, Susser M. Early origin of coronary heart disease (the "Barker hypothesis"). Br Med J 310:411-2.
42. Barker DJ, Forsen T, Uutela A, Osmond C, Eriksson JG. Size at birth and resilience to effects of poor living conditions in adult life: longitudinal study. Br Med J 2001; 323:1273-6.
43. Phillips DI, Walker BR, Reynolds RM, Flanagan DE, Wood PJ, Osmond C, et al. Low birth weight predicts elevated plasma cortisol concentrations in adults from 3 populations. Hypertension 2000; 35:1301-6.
44. Almeida MF, Mello-Jorge MHP. The use of 'linkage' of information systems in cohort studies of neonatal mortality. Rev Saúde Pública 1996; 30: 141-7.
45. Almeida MF, Mello-Jorge MHP. Small for gestational age: risk factor for neonatal mortality. Rev Saúde Pública 1998; 32:217-24.
46. Fernandes DM. Concatenamento de informações sobre óbitos e nascimentos: uma experiência metodológica do Distrito Federal 1986-1991 [Tese de Doutorado]. Belo Horizonte: Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais; 1997.
47. Noronha CP, Silva RI, Theme-Filha MM. Concordância das declarações de óbitos e de nascidos vivos para a mortalidade neonatal no município do Rio de Janeiro. Informe Epidemiológico do SUS 1997; 4:57-65.
48. Morais Neto OL, Barros MB. Risk factors for neonatal and post-neonatal mortality in the Central-West region of Brazil: linkage between live birth and infant death data banks. Cad Saúde Pública 2000; 16:477-85.
49. Cunha EMP. Condicionantes da mortalidade infantil segundo raça/cor no Estado de São Paulo, 1997-1998 [Tese de Doutorado]. Campinas: Faculdade de Ciências Médicas, Universidade Estadual de Campinas; 2001.

Submitted on 10/Mar/2003
Final version resubmitted on 06/Aug/2003
Approved on 27/Nov/2003