Probabilistic linkage in household survey on hospital care usage

Relacionamento probabilístico em inquérito domiciliar sobre uso de serviços hospitalares

Cláudia Medina Coeli[a], Régis Blais[b], Maria do Carmo Esteves da Costa[a] and Liz Maria de Almeida[a]

[a]Núcleo de Estudos de Saúde Coletiva. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brazil. [b]Groupe de Recherche Interdisciplinaire en Santé. Université de Montreal. Montreal, Canada

Keywords
Record linkage. Household survey. Claims data.

Abstract

Objective
To evaluate the potential advantages and limitations of the use of the Brazilian hospital admission authorization forms database and the probabilistic record linkage methodology for the validation of reported utilization of hospital care services in household surveys.

Methods
A total of 2,288 households interviews were conducted in the county of Duque de Caxias, Brazil. Information on the occurrence of at least one hospital admission in the year preceding the interview was obtained from a total of 10,733 household members. The 130 records of household members who reported at least one hospital admission in a public hospital were linked to a hospital database with 801,587 records, using an automatic probabilistic approach combined with an extensive clerical review.

Results
Seventy-four (57%) of the 130 household members were identified in the hospital database. Yet only 60 subjects (46%) showed a record of hospitalization in the hospital database in the study period. Hospital admissions due to a surgery procedure were significantly more likely to have been identified in the hospital database. The low level of concordance seen in the study can be explained by the following factors: errors in the linkage process; a telescoping effect; and an incomplete record in the hospital database.

Conclusions
The use of hospital administrative databases and probabilistic linkage methodology may represent a methodological alternative for the validation of reported utilization of health care services, but some strategies should be employed in order to minimize the problems related to the use of this methodology in non-ideal conditions. Ideally, a single identifier, such as a personal health insurance number, and the universal coverage of the database would be desirable.

Descritores
Relacionamento probabilístico de registros. Inquérito domiciliar. Dados de faturamento hospitalar.

Resumo

Objetivo
Avaliar as potenciais vantagens e limitações do uso das bases de dados dos formulários de Autorização de Internação Hospitalar e da metodologia do relacionamento probabilístico de registros, para a validação de relatos de utilização de serviços hospitalares durante inquéritos domiciliares.

***Correspondence to*:**
Cláudia Medina Coeli
Núcleo de Estudos em Saúde Coletiva – UFRJ
Av. Brigadeiro Trompowsky, s/n 5° andar, Ala Sul
21931-590 Rio de Janeiro, RJ, Brazil
E-mail: coeli@nesc.ufrj.br

Supported by "Conselho Nacional de Desenvolvimento Científico e Tecnológico" (CNPq - Process n. 464042/2000-3).
Presented in the 2001 Conference on Epidemiology, Toronto, Canada, 2001.
Submitted on 4/3/2002. Reviewed on 6/9/2002. Approved on 21/10/2002.

Métodos

Um total de 2.288 entrevistas domiciliares foram realizadas em Duque de Caxias, RJ (Brasil). A informação sobre a ocorrência de ao menos uma hospitalização durante o ano que precedeu a entrevista foi obtida de um total de 10.733 moradores. Os 130 registros de moradores que relataram ao menos uma hospitalização em hospital público foram relacionadas com uma base de dados hospitalares contendo 801.587 registros, usando um processo automático combinado com uma extensiva revisão manual.

Resultados

Dos 130 moradores, foram encontrados registros de 74 (57%) na base hospitalar. Entretanto, somente 60 indivíduos (46%) estiveram hospitalizados na base hospitalar dentro do período de referência do estudo. Hospitalizações devido a procedimentos cirúrgicos foram significativamente mais prováveis de serem identificadas na base hospitalar. O baixo nível de concordância obtido no estudo pode ser explicado pelos seguintes fatores: erros no processo de relacionamento, efeito telescópio e um registro incompleto dos episódios na base hospitalar.

Conclusões

O uso de bases hospitalares administrativas e da metodologia de relacionamento probabilístico de registros pode representar uma alternativa metodológica para validação de relatos de utilização de serviços de cuidados de saúde. Entretanto, algumas estratégias devem ser empregadas com o objetivo de minimizar os problemas relacionados ao uso dessa metodologia em condições não ideais. Um identificador unívoco, como o número de seguro saúde e a cobertura universal do banco de dados, seria desejado.

INTRODUCTION

Record linkage is made relatively easy when a unique identifier, such as a health insurance number, is available in the databases to be linked. In the absence of a unique identifier, probabilistic linkage is the best alternative. Probabilistic linkage is based on similar variables available in the databases to be linked (e.g., name, sex, date of birth, area of residence). These personal identifiers are used together in order to determine how likely a pair of records refers to the same individual.[8]

Record linkage has been increasingly used in health research, such as etiologic studies[9,17] and health care research.[7] However, these studies have been conducted for the most part in developed countries. This can be explained, at least partially, by the lack of availability of computerized health databases in less developed countries.

In 1984, a prospective payment system was adopted in Brazil to reimburse hospitals, and since 1991, it has been utilized by all services that depend on public financing, which provide the majority of health care in the country.[16] The system is based on the use of the hospital admittance authorization forms (commonly known by their Portuguese acronym, AIH forms), which are computerized hospital claims. This data source has been increasingly applied in Brazil for the evaluation of hospital care utilization, but to one's best knowledge it has not yet been used linked to primary data.

Record linkage between primary data (e.g., population survey) and secondary data (e.g., administrative claims data) can be a powerful tool in epidemiological and health services research. It can be employed to test hypotheses using fine individual variables (e.g. education, self-rated health) that are not available in administrative databases. In addition, computerized administrative databases represent an alternative to medical records for evaluating the accuracy of the information about health care utilization obtained through surveys in the general population.

The accuracy of the probabilistic linkage process is strongly dependent on the number and quality of personal identifiers available to be compared. Low education and income, as well as high geographic mobility, are common characteristics observed among populations living in poor areas of less developed countries. Under such conditions, the information obtained and recorded in health databases can be expected to be inaccurate, incomplete and unreliable, making the process of probabilistic record linkage a real challenge.

The main aim of this study is to evaluate the potential advantages and limitations of the use of the AIH database and the probabilistic record linkage methodology for the validation of reported utilization of hospital care services. To achieve this goal it was carried out the linkage of the AIH form database with individual data from a population survey on hospital care use conducted in a poor area in Brazil. The concordance between a positive report of hospital admission in the year preceding the interview and the information recorded in the AIH database were assessed and the factors related to the data sources that might have influenced the concordance between them were analyzed.

METHODS

Data sources

The household survey

The survey on hospital care service use was carried out in the county of Duque de Caxias (Rio de Janeiro, Brazil) as part of a study called Guanabara Bay Decontamination Program (GBDP), which was designed to evaluate the prevalence of hepatitis A in the county. The details of the sample design of the GBDP have been reported elsewhere.[1] All members of the households where at least one person was randomly selected to undergo a serologic test for hepatitis A formed the study sample. From October to December 1997, a total of 2,287 household interviews (response rate of 96%) were conducted in permanent privately owned housing units, which had a median (25th and 75th centiles) of 4 (4, 5) members and 4 (4, 5) rooms. The median of household monthly income was equivalent to approximately US$ 329.00 at the time of study, or 3 times the Brazilian minimum wage.

One household respondent answered the survey questionnaires and reported the hospitalizations of all household members. A question about hospital admissions in the year preceding the interview was included in the main questionnaire (except for deliveries). Whenever a positive answer was reported, a second instrument was applied in order to collect the following data: number of hospital admissions in the previous year; reason for the last hospital admission; name of the facility where the last hospitalization took place and type of payment (out-of-pocket, private health insurance, government- financed) for the last hospital admission. The following variables were measured for the identification of household members: full name, sex, month of birth, year of birth and household address.

The information on the occurrence of at least one hospital admission in the year preceding the interview was obtained from a total of 10,733 individuals. Of these, 157 (1.5%) reported as having been hospitalized at least once in the study period. Four household dwellers paid out-of-pocket for their last hospital admission, 23 had the hospitalization paid by a health insurance company and 130 were hospitalized in hospitals financed by the government. The latter group was selected to be linked to the AIH forms database because hospitalization data were available only for hospitals financed by the Brazilian government. Despite this group includes all public hospitals and some contracted private hospitals, for the sake of brevity all of them will be referred as public hospitals.

Choosing a small subset of the household dwellers who additionally were a priori more likely to be found in the hospital database was an attempt to achieve the best possible conditions for linkage.

The AIH forms database

An AIH form is filled out for each hospital discharge by the hospital administrative staff based on the information abstracted from the medical record. AIH forms are not filled out for emergency room admissions, even when it requires an overnight stay. Normally patient names and full addresses are not available in the databases. But for the study analysis this information was obtained by special request to the Brazilian Ministry of Health - Department of Information Technology (DATASUS), after the research proposal had been approved by a local ethics committee of the Federal University of Rio de Janeiro.

In the study period, 392 hospitals in the state of Rio de Janeiro provided inpatient care services at the expenses of the Brazilian government. The AIH forms databases from hospitalizations that took place in these hospitals were processed and records were selected according to the following criteria: date of discharge from June 1996 to March 1998 and residence of the patient located in the county of Duque de Caxias or in one of the seven neighboring counties. Records of these latter counties were included because misclassification of the patient's county of residence could be expected, especially for those addresses near the county's limits. In addition, records with missing information about the patient's county of residence were also included. A file with 801,587 hospitalizations was generated for linkage with the survey database.

Record Linkage Process

The databases were preprocessed in order to achieve standardization and parsing of the name, date of birth and sex fields, which were selected to be used as matching and/or blocking variables. Table 1 shows the fields used in each step of the linkage process and in a clerical review carried out to establish the actual status of each pair.

The linkage process was conducted using the

RecLink[3] software, which implements the probabilistic record linkage methodology. In this study, five consecutive blocking steps using different sort fields (Table 1) were applied. The software performs both exact (character-by-character) and approximate string comparisons. The month and year of birth fields were compared through the exact procedure while the first name, last name and initials of middle names fields were compared through the approximate algorithm. For each pair of records within a block, a composite weight was calculated with the sum of the agreement or the disagreement weight for each field being compared. The agreement and disagreement weights were computed, respectively, as: $\log_2$ (m/u) and $\log_2$ [(1-m)/(1-u)], where m is the probability that a field agrees given that the pair is a true match and u is the probability that a field agrees given that the pair is a false match.[8] Table 2 shows estimates of the u and m probabilities.

Clerical Review

The clerical review was conducted by one of the authors (CMC). First, an automatic procedure was applied to classify the year of birth and month of birth fields in the following categories: full agreement, partial agreement (one year or one month difference, respectively) and disagreement. The entire file of pairs generated in the first and fourth blocking strategy was reviewed. For the remaining blocking steps the review was restricted to the pairs that presented a composite weight higher than -10.

Pairs that presented at least a partial agreement on the full name were subsequently analyzed in terms of address agreement. In the case of an agreement on the address, the pair was classified as a true match after confirming that the admission recorded in the AIH form database didn't refer to another member of the household (e.g. father and son with same names). In the case of a disagreement on the address, the decision to classify the pair as a true match was based on the presence of an agreement (full or partial) on at least two of the following fields: year of birth, month of birth, cause of admission and facility where the admission took place.

Table 1 - Fields utilized in each step of the linkage process and clerical review.

Blocking	
Step	Sort field
1	Soundex of the last name + soundex of the first name + sex
2	Soundex of the last name + sex
3	Soundex of the first name + sex
4	Soundex of the last name + soundex of the first name
5	Year of birth + sex
Matching	
First name (full or partial agreement)	
Last name (full or partial agreement)	
Middle names initials (full or partial agreement)	
Year of birth (full agreement)	
Month of birth (full agreement)	
Clerical Review	
Full name (full or partial)	
Address (full or partial)	
Year of birth (full or partial agreement)	
Month of birth (full or partial agreement)	
Reason for hospital admission (only if address disagrees)	
Facility where the hospital admission took place (only if address disagrees)	

Table 2 - Estimation of the matching probabilities.

Matching Fields	Probability of agreement given that a pair is a true match (m)	Probability of agreement given that a pair is a false match (u)
First name	0.99	0.01
Last name	0.99	0.04
Middle names initials	0.89	0.03
Year of birth	0.74	0.02
Month of birth	0.82	0.09

Data Analysis

Regarding the household survey, the following variables were described: 1) sociodemographic characteristics of the sample; 2) respondent's sex, age, years of schooling and relationship with the household member reported as having been hospitalized; 3) number of hospital admissions; and 4) place and reason of the last hospital admission. The completeness of the variables used in the linkage process was evaluated in each data source, as well as the agreement on these variables between the two data sources. The latter was evaluated among the true matches identified after the clerical review. In order to assess the concordance between the two data sources with regards to the report of hospital admission, the proportion of household members with a positive report of a hospital admission, identified in the AIH data source within the reference period of the study, was calculated. Finally, there were calculated odds ratios to assess the association between factors related to the household member, the respondent and the reason of hospital admission, with the identification of the hospital admission in the AIH database form.[10] Differences in proportions were tested for significance using χ^2 tests. All analyses described above were performed with Stata software (version 7.0).[15]

RESULTS

Household Survey

Table 3 presents the demographic characteristics of the sample. Overall there was a majority of women and individuals under the age of 19. The group of household members who were reported as having undergone

Table 3 - Demographic characteristics of the sample. Duque de Caxias, RJ, Brazil, 1997.

	Household members with a reported hospitalization in a public hospital				Total sample (N=10,733)	
	Yes (N=130)		No (N=10,603)			
Variables	N	%	N	%	N	%
Age group (years)*						
0-4	62	47.7	2,389	22.5	2,451	22.8
5-9	14	10.8	1,293	12.2	1,307	12.2
10-19	7	5.4	1,894	17.9	1,901	17.7
20-39	31	23.8	3,496	33.0	3,527	32.9
40-59	13	10.0	1,208	11.4	1,221	11.4
≥60	3	2.3	323	3.0	326	3.0
Sex**						
Male	74	56.9	5,019	47.3	5,093	47.5
Female	56	43.8	5,584	52.7	5,640	52.5

*Statistically significant (p=0.000) for difference between groups.
**Statistically significant (p=0.030) for difference between groups.

at least one hospitalization in a public hospital in the year preceding the interview showed a larger proportion of men and children under the age of 5.

At least one case of hospitalization in a public hospital was reported in 125 households. The distribution of the number of cases per household was as follows: one – 121 households; two – 3 households; and three – 1 household. Regarding the interview respondents in these households, 87% (113/125) were female, mean age of 32 (SD 11.8 years) and mean years of schooling of 6 (SD 3.1 years).

Information about the relationship with the respondent was not available for 4 cases. Of the remaining 126, the respondent was most often the mother (59/71; 83%) in the group of children under the age of 10, whereas in the group of adolescents and adults (age ≥10 years old; n=55) the most common respondent was the hospitalized individual (47%), followed by his/her spouse (22%) and mother (14%).

The great majority of cases (106; 81%) were reported as having had only one hospital admission in the study period, 16 (12%) were reported as having had two admissions and the remaining 8 (7%) were reported as having had three or more. In five cases the respondent could not recall the name of the facility where the last hospital admission took place. In addition, in 7 cases it was impossible to identify the hospital by the name provided by the respondent. Of the 118 remaining, the last hospital admission took place in 43 different hospitals. In all cases at least one symptom/ill-defined condition was reported as the reason for the last hospital admission. Sixteen (12%) hospitalizations were because of a surgical procedure and 114 (88%) because of a medical condition. The most frequent reasons reported were pneumonia (26; 20%), intestinal infectious diseases (16; 12%), bronchitis (13; 10%), and hernia (6; 5%).

Record Linkage

The first aspect to be considered is the completeness of the identifying items in the databases used in the linkage process. The only variable that was not available in all records of the AIH database was the patient's name, but even that the information was missing in only 10 records (0.001%). Regarding the household survey data, the month of birth was missing in two records (1.5%) and the name of the facility of the last admission was missing in 12 (9.2%) records.

Combining the automatic linkage process and the clerical review, it was possible be to identify 97 hospitalizations (true matches) of 67 household members. Among these, only 60 had at least one hospital admission in the year preceding the interview. In order to identify whether the 63 subjects who were not found in the AIH database had had their hospital admission outside the study period, it was performed the linkage of these subjects' records with a subset of records of the AIH form database, selected according to the same residence criteria and considering the discharge date from June 1995 to May 1996. Eight new pairs were classified as true matches and 7 new patients were identified. Considering thus the two-linkage process together, it was identified in the AIH form database 101 hospitalizations of 74 (57%) of the 130 subjects who were reported as having had a hospital admission in the year preceding the interview. Yet there were only 60 (46% of subjects) hospitalizations recorded in the AIH form database in the study period.

Table 4 shows the proportions of different levels of agreement between the information recorded in both data sources. A high proportion of disagreement was observed for the address field.

In the analysis of the name of the facility of the last hospital admission and reason for hospital admission fields, it was considered only the 74 pairs relative to the last hospitalization. An agreement on hospital of admission was observed in 50 pairs (67%), whereas perfect agreement on the reason for hospital admission was observed in 35 pairs (47%). Furthermore, in 25 other pairs (34%) the reason reported, though inconsistent with the diagnosis recorded in the AIH form database, could be considered as belonging to the same group of disorders (e.g. asthma and bronchitis; bronchitis and pneumonia).

Hospital admissions due to a surgery procedure were significantly more likely to have been identified in the AIH form database. No other variables were significantly associated with the identification of hospitalization in the AIH form database (Table 5).

DISCUSSION

This study shows that a low percentage of hospital admissions in a one-year- period were identified in the federal government hospital administrative database.

Norish et al[12] (1994), evaluating a four-year recall of hospital admissions through the use of a hospital computerized database, found minimal over-reporting but underreporting of 38%. However, opposed to that and to the present study, many others[6,11,13,14] showed a fairly accurate concordance between reported and recorded hospitalizations.

The low level of concordance seen in the study can be explained in several ways. First, as the linkage process was not based on a single identifier, some mistakes might have happened. Some pairs could have been erroneously classified as true matches (false matches), resulting in an artificial improvement of the agreement, whereas some true matches could have been missed (false non-matches), ensuing an apparent over-reporting. The fact that it was used different blocking steps and the automatic linkage process was combined with an extensive clerical review may have contributed to minimize the occurrence of such errors. Nevertheless, whenever a possible link showed a disagreement on address the decision to consider the pair as a true match had to be based on less reliable fields, and it was decided to adopt a more con-

Table 4 - Agreement between the information recorded in the survey data and the AIH database, among the 101 pairs identified as true matches. Duque de Caxias, RJ, Brazil, 1997.

	Full		Partial		Disagreement	
Field	N	%	N	%	N	%
First name	60	69.3	30	29.7	1	1.0
Last name	95	94.0	2	2.0	4	4.0
Middle names initials	85	84.2	0	0.0	16	15.8
Year of birth	77	76.2	14	13.9	10	9.9
Month of birth	80	79.2	7	6.9	14	13.9
Address	78	77.3	-	-	23	22.7

AIH = Hospital admittance authorization.

Table 5 - Factors associated with the identification of a reported hospital admission in the AIH database. Duque de Caxias, RJ, Brazil, 1997.

Characteristic	Crude OR	95%CI	p value
Household member with a hospital admission reported			
Age			
<10 years old	1		
≥10 years old	0.74	0.37-1.50	0.41
Sex			
Male	1		
Female	0.88	0.62-1.25	0.47
Respondent			
Age of the respondent			
<40 years old	1		
≥40 years old	1.21	0.51-2.87	0.66
Sex			
Female	1		
Male	0.74	0.44-1.25	0.26
Relationship with the household member			
Own individual/mother	1		
Other	1.37	0.92-2.05	0.12
Years of schooling of the respondent			
<8 years	1		
≥8 years	1.08	0.51-2.28	0.84
Hospitalization			
Reason for the hospitalization			
Medical condition	1		
Surgery	4.12	1.25-13.57	0.02

OR = *odds ratio*.
95% CI = 95% confidence interval.

servative approach. The option for an improvement of the specificity of the linkage process probably favored the occurrence of erroneous non-matches, resulting in an apparent over-reporting.

Second, the observed discrepancy can also be explained by a telescoping effect, i.e., the hospitalizations were recalled as having occurred more recently than they actually did. Of the 14 cases where the last hospital admission was identified outside the study period, 11 presented an agreement either on diagnosis or on the facility where the hospitalization took place. Moreover, 75% of the episodes occurred within eighteen months apart from the interview date. These findings strongly suggest that the hospital admissions identified outside the study period were in fact the same reported during the interview. The telescoping phenomenon has been reported in previous studies that assessed the recall of hospital admissions,[11] non-fatal injuries[5] and general practitioner consultations.[2]

Third, and most important, there may have been an incomplete recording of the episodes in the AIH form database, probably due to administrative problems or partial coverage of the services rendered. It was observed that eight hospitalizations were reported as having taken place in one of the public hospitals, but no hospitals admissions related to this particular hospital were found in the AIH form database for the period analyzed. According to official sources, this hospital did not provide claims forms in the study period. It turned out that three of these hospitalizations have taken place in other hospitals, but the remaining five were not found.

Hospital admissions that take place in emergency rooms are not recorded in the AIH form database, even when the admission requires an overnight stay. At least in two cases not identified in the AIH form database, the facility reported that the place where the admission took place was a hospital without an inpatient sector. Moreover, in several cases the reported cause of hospital admission (e.g. bronchitis, asthma, and diarrhea) was a condition that could have been treated at the emergency room. It is worth noting that a surgical procedure reported as the cause of the hospital admission was the only factor significantly associated with a better concordance between the two data sources. Ritter et al[12] found an clear tendency to over-report emergency room visits in comparison to the number recorded in a health maintenance organization (HMO) computerized database, which was probably explained by emergency rooms visits that took place outside the HMO. Coulter et al[4] (1985), checking the recall of surgical accounts against general practitioner records, also found suggestive evidence that in cases where a clear over-report occurred, inadequacies were found in the notes instead of the patient's self-report.

In the present study, it was decided to restrict the analysis to the subset of household members reported as having undergone a hospital admission in a public hospital. This file was about 80 times smaller than the entire household data set, allowing the generation of a relatively small number of pairs in each blocking step and facilitating the clerical review. The values of the matching parameters estimated here will be used to carry out the linkage of the remaining records of the survey data in subsequent studies. In addition, an improvement in the discriminatory power of the automatic linkage process could be achieved through the introduction of an algorithm to assess a partial agreement on the month and year of birth fields.

Indeed, there would be a greater gain with the use of the address field in the automatic process. Unfortunately the patient's residence address was stored in a free form field in the AIH database and the information was filled out without any standardization. In addition, during fieldwork, it was found a high rate of household members who moved to another address within the same area. The high proportion of disagreement on the field address observed confirmed the high geographic mobility of the population studied. These factors represent barriers to the future incorporation of the address in the automatic process.

This study had some limitations. First, the household survey was not primarily designed to test the concordance between the two data sources, and several variables that could have improved this assessment were not collected, such as date of discharge, length of hospital stay and the department where the hospitalization took place (inpatient or emergency). Second, it were only checked the cases with a positive history of hospital admission, which precluded the assessment of underreporting and the overall agreement between data sources. Finally, the small number of cases allowed only a simple analysis of the association between some putative factors and the identification of the hospitalization reported in the AIH database, with only one factor being analyzed at a time.

Nevertheless, to one's best knowledge, this is the first study conducted in a poor area of a less developed country to assess the concordance between hospitalization reports and administrative claims data. Despite the limitations mentioned above, the study results may contribute to a better understanding of the advantages and limitations of using hospital administrative databases and the record linkage methodology in non-ideal conditions. The population's high geographic mobility and the partial coverage of claims data of the services rendered were the most important problems found. The latter problem can be minimized through the inclusion in the survey questionnaire of additional questions (e.g., the department where the hospitalization took place) in order to obtain a better characterization of the hospital admissions. Concurring with Ritter et al[13] (2001) it is essential to determine the actual coverage of administrative database and to estimate the effect of incomplete coverage on the occurrence of discrepancies between reported and recorded hospitalizations.

Despite the problems found in the study, the use of hospital administrative databases for the accuracy evaluation of reports about hospitalizations is valuable. For instance, it would have been necessary to contact 392 different hospitals if it were decided to validate the entire survey data against medical records. Even if the analysis had been restricted to the positive reports of hospitalizations, as this study was, it would have been necessary to obtain data from 43 hospitals. Furthermore, it was found that the name of the facility where the hospital admission took place was not reported or was reported incorrectly in a high proportion of cases, which might have represented an additional barrier to the utilization of medical records. In order to minimize the problems presented by either hospital administrative databases or medical records for the validation of reported health care use data, it would be interesting to develop a two-stage validation study, using computerized database in the first step of the analysis, and limiting access to medical records for hospitalizations not found in the first step.

In conclusion, the use of hospital administrative databases and probabilistic linkage methodology may represent a methodological alternative for the validation of reported utilization of health care services. Yet some strategies should be employed to minimize the problems related to the use of this methodology in non-ideal conditions. In the long run, however, a single identifier, such as a personal health insurance number, and universal coverage of the database would be desirable to facilitate record linkage.

REFERENCES

1. Almeida LM, Costa MCE, Luiz RR , Collety PE, Azevedo Neto RS, Machado VA, et al. Soroprevalência da hepatite A no município de Duque de Caxias, Rio de Janeiro, Brasil. *Cad Saúde Coletiva* 1988;6(Supl 1):39-48.

2. Bruijnzeels MA, Van der Wouden JC, Foets M, Prins A, Van den Heuvel WJA. Validity and accuracy of interview and diary data on children's medical utilisation in the Netherlands. *J Epidemiol Community Health* 1998;52:65-9.

3. Camargo Jr KR, Coeli CM. RECLINK: Aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. *Cad Saúde Pública* 2000;16:439-47. Disponivel em URL: http://www.ensp.fiocruz.br/publi/cad_por.html [2001 Dez 6].

4. Coulter A, McPherson K, Elliot S, Whiting B. Accuracy of recall of surgical histories: a comparison of postal survey data and general practice records. *Community Med* 1985;7:186-9.

5. Harel Y, Overpeck MD, Jones DH, Scheidt PC, Bijur PE, Trumble AC, Anderson J. The effects of recall on estimating annual nonfatal injury rates for children and adolescents. *Am J Public Health* 1994;84:599-605.

6. Harlow SD, Linet MS. Agreement between questionnaire data and medical records. *Am J Epidemiol* 1989;129:233-48.

7. Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999;23:453-9.

8. Jaro MA. Probabilistic linkage of large public health. *Statist Med* 1995;14:491-8.

9. Jones ME, Swerdlow AJ, Gill LE, Goldacre MJ. Pre-natal and early risk factors for childhood onset diabetes mellitus: a record linkage study. *Int J Epidemiol* 1998;27:444-9.

10. National Center for Health Statistic. *Reporting of hospitalization in the health interview survey*. Washington (DC); 1965. [Vital and Health Statistics. Series 2, n 6].

11. National Center for Health Statistic. *Comparison of hospitalization reporting in three survey procedures*. Washington (DC); 1965. [Vital and Health Statistics. Series 2, n 8].

12. Norrish A, North D, Kirkman P, Jackson R. Validity of self-reported hospital admission in a prospective study. *Am J Epidemiol* 1994;140:938-42.

13. Ritter PL, Stweart AL, Kaymaz H, Sobel DS, Block DA, Lorig KR. Sel-reports of health care utilization compared to provider records. *J Clin Epidemiol* 2001;54:36-141.

14. Roberts RO, Bergstralh EJ, Schmidt L, Jacobsen SJ. Comparison of self-reported and medical record health care utilization measures. *J Clin Epidemiol* 1996;49:989-95.

15. StataCorp. Stata statistical software: release 7.0. College Station, TX: Stata Corporation; 2001.

16. Travassos Veras CM. Equity in the use of private hospitals contratcted by a compulsory insurance scheme in the city of Rio de Janeiro, Brazil, in 1986 [Tese de Doutorado]. London: Department of Public Administration, School of Economics and Political Science; 1992.

17. Whiteman D, Murphy M, Hey K, O'Donnell M, Goldacre MJ. Reproductive factors, subfertility, and risk of neural tube defects: a case-control study based on the Oxford Record Linkage Study Register. *Am J Epidemiol* 2000;152:823-8.