Record Linkage and Capture-Recapture Estimates for Underreporting of Human Leptospirosis in a Brazilian Health District

Liane Brum and Emil Kupek

Department of Public Health, Federal University of Santa Catarina; Florianópolis, SC, Brazil

Record linkage and capture-recapture models were used to estimate the number of cases of human leptospirosis in the health district of Santa Maria, RS in southern Brazil. Twelve months of laboratory, hospital and epidemiological surveillance data were matched by name, age, residence and the month of diagnosis. Only laboratory-confirmed cases were considered. The record linkage revealed more than 20 times more cases than the official estimate for the health district, indicating a leptospirosis epidemic, with an annual incidence of more than 3 per 1,000 inhabitants and a case fatality of 0.37%. Severe cases were predominantly found through hospital records, overlapping to some extent with the epidemiological surveillance data, whereas less severe cases were found almost exclusively through laboratory logs. Different combinations of data sources influenced the detection rate for low *versus* high severity cases. Based on log-linear capture-recapture models, stratified by case severity and taking into account possible dependencies between the data sources, an insignificant number of cases were missed by all sources.

Key Words: Leptospirosis, record linkage, capture-recapture, log-linear models, reporting bias.

Underreporting of infectious diseases, including those that are required to be reported by law, is a notorious problem in many developing countries. As a result, the basis for epidemiological planning and preventive actions is often imprecise or even misleading. Nevertheless, increasing use of electronic data bases in health administration allows for record linkage and for checking for completeness and accuracy within and between the data sources, as well as for the use of capture-recapture (CR) statistical methods [1-4] to obtain estimates of the number of cases missed by surveillance. Combined registers of chronic diseases and health problems have been established in many countries and have been submitted to extensive analysis of this type [5-8]. However, published examples of this practice are quite rare in Brazil and are restricted to chronic health conditions, such as lower limb amputation rate [9], brain tumor prevalence [10], diabetes [11], the number of street children [12] and the size of a mosquito population [13]. Among the infectious diseases, AIDS has received the most attention [14].

A major criticism of early CR methods applied to surveillance data was the lack of plausibility of the assumption that the data sources are statistically independent, based on the legal obligation and common practice of laboratories and hospitals to report certain diseases to the epidemiological surveillance agencies. However, methods that allow for intrinsic data dependency have been developed and tested in practice for a variety of health conditions during the last decade [6-8].

We used record linkage and modern capture-recapture methods to estimate the completeness of surveillance data on human leptospirosis in the health district of Santa Maria in Brazil. These estimates were used to correct important epidemiological statistics, including disease incidence, case fatality and mortality.

Received on 19 June 2005; revised 16 October 2005.
Address for correspondence: Dr. Emil Kupek. Departamento de Saúde Pública, CCS. Universidade Federal de Santa Catarina Campus Universitário, Trindade. Zip code: 88040-900 Florianópolis-SC, Brazil.

The Brazilian Journal of Infectious Diseases 2005;9(6):515-520

Material and Methods

The data collection was restricted to a one-year period, from May 1, 2001 to April 30, 2002, in the Santa Maria health district, Rio Grande de Sul state, Brazil, with a population of approximately 526,000 inhabitants living in 30 municipalities. During this period, both macroscopic and microscopic agglutination tests for leptospirosis were routinely performed, thus permitting a reliable diagnosis.

Health district surveillance data, regional (LACEN) and university (USFM) laboratories and hospitals in the region were the sources for the epidemiological data. It was believed that residents who were ill with leptospirosis were unlikely to seek medical assistance outside the health district, although no verification procedure was possible with this data.

The case definition followed that of the Brazilian Ministry of Health for the laboratory confirmed cases, i.e. a positive test result for either microscopic or macroscopic agglutination tests in individuals presenting symptoms compatible with leptospirosis [15]. The laboratory used "Simples Teste - Leptospirose", produced by Fundação Oswaldo Cruz, Bio-Manguinhos, Rio de Janeiro, for macroscopic agglutination. The UFSM laboratory cultivated 12 different sorovars for the microscopic agglutination test: *Leptospira australis, L. andamana, L. bratislava, L. butembo, L. canicola, L. castellonis, L. copenhageni, L. icterohemorrhagiae, L. patoc, L. pyrogenes, L. Pomona* and *L. sentot*.

We had two main objectives: a) to evaluate the completeness of important epidemiological data on case severity, municipality of residence, age, sex, month of diagnosis (seasonality) and serovar type, and b) to estimate the number of cases of human leptospirosis in the region based on record linkage and CR modeling. Case severity was considered high if the patient died or was hospitalized because of leptospirosis and low otherwise.

Individual records were matched by name, age, residence and the month of the year the diagnosis of leptospirosis was made. CR estimates of the total number of cases were obtained using log-linear models, which allows testing dependencies between data sources as interaction terms [3]. A variety of models with different breakdowns of cases by disease severity, seasonal influences and sex were tested and compared in terms of deviance (twice the log-likelihood) and corresponding degrees of freedom, as well as in terms of residuals. After this preliminary analysis, a stratified CR analysis was performed for the factors whose influence was found statistically significant. The F-distribution was used to test statistical significance for sequential sum of squares as in an analysis of variance. The number of cases missed by all three data sources was calculated by exponentiating the intercept parameter of the log-linear models. Stata software [16] was used.

Results

Matching the data resulted in 1,611 individual records. The principal difficulties in this process were spelling errors and the abbreviation of names. Completeness of patient information after record linkage showed large gaps for professional exposure, serovars, age and residence (Table 1). Among those whose age was reported, one quarter were less than 12 years old, another quarter was older than 40 years, and the remaining half were in between. Slightly more than half (53%) were men. Over 90% of the cases whose residence was reported were from the Santa Maria region. The months of August, September and October, during which there is seasonally-elevated rainfall, concentrated more than 60% of all cases.

Among the 608 patients with a known serovar, 20% had more than one serovar isolated. *Leptospira patoc* was the predominant serovar, found in 63% of patients with only one serovar and in half of all patients. Other frequent serovars were *L. sentot*, *L. bratislava*, *L. butembo* and *L. castellonis*, representing 14%, 9%, 7% and 6% of all isolated serovars, in the same order. *Leptospira icterohemorrhagiae* was found in only 1% of the patients.

The number of laboratory-confirmed cases increased from 75, in the health district epidemiological

Table 1. Percentage of important epidemiological information on leptospirosis unavailable after record linkage (N=1,611)

Information unavailable	N	%
Professional risk	1,571	95.33
Serovar	1,040	63.11
Age	783	47.51
Residence (municipality)	721	43.75
Sex	58	3.52
Hospital stay	56	3.40
Disease outcome	17	1.03
Hospitalization	1	0.06

Table 2. Cases of human leptospirosis for all combinations of data sources by case severity in the Santa Maria, Rio Grande do Sul health district

Data sources			Low case severity (n=1,510)		High case severity (n=101)		All cases (n=1,611)	
Laboratory	Hospital	Surveillance	N	%	N	%	N	%
No	No	Yes	58	3.84	0	0.00	58	3.60
No	Yes	No	0	0.00	17	16.83	17	1.05
No	Yes	Yes	0	0.00	8	7.92	8	0.48
Yes	No	No	1,418	93.91	0	0.00	1,418	88.02
Yes	No	Yes	32	2.12	0	0.00	32	1.99
Yes	Yes	No	2	0.13	50	49.50	52	3.23
Yes	Yes	Yes	0	0.00	26	25.74	26	1.61

Table 3. Final capture-recapture model parameters by case severity (stratified analysis)

Parameter	Low case severity			High case severity		
	Estimate (std. error)	F*	P>F**	Estimate (std. error)	F*	P>F**
Intercept	-12.090 (0.078)	_	_	-23.119 (0.025)	_	_
Laboratory	19.334 (0.077)	13.41	0.0003	1.112 (0.027)	1954.94	0.0001
Hospital	-6.662 (0.248)	16772.15	0.0001	26.000 (0.001)	8631.14	0.0001
Surveillance	16.146 (0.063)	0.03	0.8668	-0.678 (0.025)	795.89	0.0001
Laboratory & Surveillance	-19.939 (0.001)	31.74	0.0001	***	***	***

* F-distribution value for sequential sum of squares. ** Probability of obtaining a larger F-value. *** Not significant for this model.

surveillance register, to 1,611 after record linkage. Therefore the incidence of leptospirosis should be corrected accordingly, from 14 (official estimate) to 306 per 100,000 inhabitants. Although no deaths caused by leptospirosis were registered in the vital statistics mortality register for this period, six deaths from hospital records indicated a case fatality of 0.37% and a mortality of 1.14 per 100,000 inhabitants.

There was a striking difference in the distribution of cases according to the data sources for low *versus* high case severity (Table 2). While the former are overwhelmingly concentrated in laboratory data without overlapping with other sources, the latter had hospital patient records as the main data source, with almost half of them overlapping with laboratory data and a quarter of them captured by all three sources. This justifies a CR analysis stratified by case severity.

Log-linear CR models tested the influence of case severity, seasonal influences and sex on the leptospirosis case counts. Only case severity was statistically significant (data not shown). A stratified CR analysis for low and high case severity confirmed the predominant importance of laboratory data for less severe cases as compared to the hospital data for severe cases (Table 3). Surveillance data are an individually significant source for the low severity cases, as dividing the estimate by its standard error gives a large t-value, although it adds little to the information already obtained from the laboratory and hospital data, as indicated by the insignificant F-statistic for the sequential sum of squares. A significant negative estimate for the interaction of laboratory and surveillance sources means that being captured by one of them makes it very unlikely to be captured by the other. On the other hand, no significant interaction of data sources was found for severe cases. Different from the low severity group, being detected by epidemiological surveillance means being less likely to appear among severe cases, because exponentiating the estimate gives an odds ratio of 0.51 (95% confidence interval from 0.48 to 0.53). Each of the stratified models in Table 3 explained more than 90% of the variation in case counts (pseudo-R^2). The residuals were very small in magnitude, with the largest difference between observed and predicted being below 0.83, thus suggesting a very good fit. The main conclusions from the stratified analysis is that different combinations of data sources influence the detection rates for low *versus* high severity cases and the models accounting for this found that an insignificant number of the cases were missed by all three data sources after record linkage.

Discussion

Evaluation of completeness of reporting indicated serious gaps in the gathering of relevant epidemiological information (Table 1). Even basic information on age and residence, as well as more specific data on professional exposure and serovars, were often unavailable, making it difficult to develop timely preventive actions.

There was a striking lack of communication between the health district epidemiological surveillance and the laboratory and hospital administration, resulting in substantial underreporting of human leptospirosis (Table 2). Record linkage found 21 times more cases than the number registered by the health district. The incidence thus exceeds 3 new cases per 1,000 inhabitants, per year, reaching epidemic proportions. The situation was particularly aggravated in the period from August to October, when more the 60% of all cases were registered. The case fatality rate of 0.37% was considerably lower, and the incidence of 306 per 100,000 considerably higher than the national average of 2.45% and 2.43 per 100,000, respectively, reported for the year 2000 [15]. However, the latter is likely to be a gross underestimation of the real number of cases in the country and biased towards severe cases for which hospitalization increases the likelihood of notification. For this reason, the national case fatality average is likely to be biased upward.

Several limitations should be kept in mind when interpreting these findings. First, a case severity definition based only on hospitalization and case fatality may not be very precise, given the unknown variability of the hospitalization criteria between hospitals. More precise information on disease complications (e.g.

hemorrhage and renal failure) and treatments indicating such scenarios (e.g. intensive care) were generally missing from the patient records. However, the percentage of patients with severe disease estimated for the linked data (6.3%) is within the 5-10% range reported in the literature [15,17]. Second, we could not quantify the probability of laboratory results indicating an infection from previous years as opposed to a recent infection. In endemic areas such as this, the antibodies against *Leptospira* can persist in the human body for long periods of time [17]. On the other hand, the presence of disease symptoms that motivated laboratory investigation made this scenario considerably less likely. The use of an enzyme-linked immunosorbent assay for immunoglobulin types M, G and A, which can distinguish between old and recent infection, has already been recommended by the health authorities [15], but its use is still very restricted. In addition, net bias due to false positive and false negative diagnosis may be less severe for CR methods than with traditional counting methods [5]. Third, no adjustment could be made for the probability that the mildest cases with unspecific flu-like symptoms were not diagnosed at all as they did not seek medical assistance. Although they may have little clinical relevance, their epidemiological importance is obvious as they are the carriers of infectious *Leptospira*. Fourth, the CR method is sensitive to a small overlap between sources and the so-called 'variable catchability' [3], which in this context may reflect unknown differences in the accessibility of medical services and in assistance-seeking behavior. The latter depends on knowledge and beliefs about the disease, but it remains unaccounted for in the model. More sophisticated models capable of modeling these effects have been developed recently [18].

Despite the above limitations, the estimates were based on laboratory test results and can be interpreted as best 'informed guesses' in the case of the data at hand. Although the case definition based on a positive result of either of the two laboratory tests routinely used to diagnose leptospirosis may overestimate the number of recent infections in a screening, this is not likely given the disease symptoms which motivated the laboratory tests in the first place. Nevertheless, due to the current limitations of epidemiological surveillance in Brazil, estimates based on laboratory confirmed cases of leptospirosis are more precise than broader definitions based solely on clinical symptoms and possible epidemiological links with other cases [15].

A relationship between flush floods and the incidence of leptospirosis in southern Brazil has been described recently [19]. Apart from better drainage, which requires substantial investments, other modifiable risks include better dissemination of knowledge on disease prevention, diagnosis and treatment. A simple linkage of existing data sources can greatly enhance important epidemiological information to combat leptospirosis, with virtually no additional economic resources, and this linkage can provide a more realistic picture of the disease burden.

Acknowledgements

The authors thank all the administrative workers and directors of both regional and central laboratories at the epidemiological surveillance and collaborating hospitals in the region for providing the data used in the analyses.

References

1. Agresti A. Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics **1994**;50:494-500.
2. Hook E.B., Regal R.R. Recommendation for presentation and evaluation of capture-recapture estimates in epidemiology. J Clin Epidemiol **1999**;52:917-26.
3. Tilling K., Sterne J. Capture-recapture models including covariate effects. Am J Epidemiol **1999**;149(49):392-400.
4. Tilling K. Capture-recapture methods—useful or misleading? Int J Epidemiol **2001**;30:12-4.
5. Brenner H. The effects of misdiagnoses on disease monitoring with capture-recapture methods. J Clin Epidemiol **1996**;49:1303-7.
6. Hickman M., Cox S., Harvey J., et al. Estimating the prevalence of problem drug use in inner London: a discussion of three capture-recapture studies. Addiction **1999**:1653-62.

7. Semenciw R.M., Le N.D., Marrett L.D., et al. Methodological issues in the development of the Canadian Cancer Incidence Atlas. Stat Med **2000**;19:2437-49.
8. Gill G.V., Ismail A.A., Beeching N.J. The use of capture-recapture techniques in determining the prevalence of type 2 diabetes. Q J Med **2001**;94:341-6.
9. Spichler E.R.S., David Spichler D., Lessa I., et al. Capture-recapture method to estimate lower extremity amputation rates in Rio de Janeiro, Brazil. Pan Am J Public Health **2001**;10(5).
10. Argolo N., Lessa I. Estimativa da prevalência de neoplasia cerebral na faixa pediátrica pelo método de captura-recaptura. Arq Neuropsiquiatr **1999**;57(2-B):435-41.
11. Campos J.J.B., Almeida H.G.G., Iochida L.C., Franco L.J. Incidência de diabetes mellitus insulino dependente (Tipo 1) na cidade de Londrina, PR – Brasil. Arq Bras Endocrinol Metab **1998**;42(1):36-44.
12. Gurgel R.Q., da Fonseca J.D., Neyra-Castaneda D., et al. Capture-recapture to estimate the number of street children in a city in Brazil. Arch Dis Child **2004**;89(3):222-4.
13. Santos R.L., Forattini O.P. Marking-release-recapture methods for determining the size of the natural population of *Anopheles albitarsis* (Diptera:Culicidae). Rev Saude Publica **1999**;33(3):309-13.
14. Caiaffa W.T., Mingoti S.A., Proietti F.A., et al. Estimation of the Number of Injecting Drug Users Attending an Outreach Syringe-Exchange Program and Infection With Human Immunodeficiency Virus (HIV) and Hepatitis C Virus: the AjUDE-Brasil Project. Journal of Urban Health: Bulletin of the New York Academy of Medicine **2003**;80(1).
15. Ministério da Saúde, Fundação Nacional de Saúde. Leptospirose. In: Ministério da Saúde ed. Guia de Vigilância Epidemiológica. Brasília: Ministério da Saúde, Fundação Nacional de Saúde, **2002**.
16. Stata Corporation. Stata statistical software: release 5.0. College Station, TX: Stata Corporation, **1997**.
17. Da Silva M.V., Camargo E.D. Leptospirose. In: Ferreira A.W., Ávila S.L.M. eds. Diagnóstico laboratorial das principais doenças infecciosas e auto-imunes. Rio de Janeiro: Guanabara Koogan, **1996**.
18. Stanghellini E.., Heijden P.G.M. A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. Biometrics **2004**;60:510-6.
19. Kupek E., Faversani M.C.S.S., Philippi J.M.S. The relationship between rainfall and human leptospirosis in Florianópolis, Brazil, 1991-1996. Braz J Infect Dis **2000**;4:131-4.