Epidemiol. Infect. (2007), **135**, 1030–1036. © 2006 Cambridge University Press
doi:10.1017/S0950268806007667 Printed in the United Kingdom

Capture–recapture estimation of underreporting of legionellosis cases to the National Legionellosis Register: Italy 2002

M. C. ROTA*, A. CAWTHORNE, A. BELLA, M. G. CAPORALI, A. FILIA AND F. D'ANCONA, on behalf of the Legionellosis Working Group†

Centro Nazionale di Epidemiologia, Sorveglianza e Promozione della Salute, Reparto Epidemiologia delle Malattie Infettive, Istituto Superiore di Sanità, Roma, Italy

(*Accepted 24 October 2006; first published online 19 December 2006*)

SUMMARY

The objective of this study was to evaluate the degree of underreporting to the Italian National Legionellosis Register (NLR). For the year 2002, all cases of Legionellosis notified to the NLR were compared with cases recorded in the hospital discharge record (HDR) database. The number of unreported cases and the total number of cases in the population were estimated using the capture–recapture method with two independent data sources. Seventeen out of 21 Italian regions participated in the study. Overall, underreporting was estimated to be 21·4% and was found to be significantly greater in the Centre-South (28·2%) than in the North (20·0%). However, even after taking into account the higher degree of underreporting, a significantly lower incidence of the disease is registered in central-southern Italy. The hypothesis, which needs to be verified, is that, in addition to underreporting, under-diagnosis of legionellosis is more widespread in this geographical area.

INTRODUCTION

Legionella bacteria are widespread in the environment and can be found, usually in low numbers, in natural water sources such as rivers, lakes and reservoirs. Bacteria present in natural sources pass into sites that constitute artificial reservoirs [1].

In most hospital-based series, *Legionella* is implicated in 2–6% of community-acquired pneumonia cases [2–4]. Risk of legionellosis is related to exposure, increasing age, smoking, and impaired cell-mediated immunity such as in transplant recipients. Although rare in immunocompetent adults aged <30 years, *Legionella* is a major cause of lethal pneumonia, with mortality rates of 5–25% among immunocompetent hosts and substantially higher rates among immunosuppressed hosts [5].

A rate of 20 cases per million population per year is considered by the European Working Group on Legionella Infections (EWGLI) to be a good estimate of the expected rate of Legionnaires' disease (LD) in European countries. This estimate is based on rates consistently reported by Denmark, a small country that carries out high levels of testing for *Legionella*, and that also has a centralized Legionella reference laboratory [6].

In recent years, the number of legionellosis cases reported in Italy has considerably increased, going from 100 cases reported annually in the late 1990s to 325 cases in 2001 and over 600 cases annually in

* Author for correspondence: Dr M. C. Rota, Centro Nazionale di Epidemiologia, Sorveglianza e Promozione della Salute, Istituto Superiore di Sanità, 00161 Rome, Italy.
(Email: mariacristina.rota@iss.it)

† Members of the Legionellosis Working Group are listed in the Appendix.

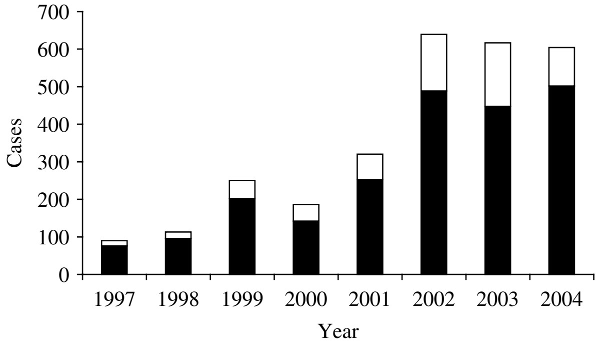

Fig. Number of cases of legionellosis reported annually from 1997 to 2004 in northern Italy (■) and central-southern Italy (□).

2002–2004. The incidence rate has increased, therefore, from 2 cases per million in the 1990s to ~11 cases per million from 2002 onwards [7, 8].

The number of reported legionellosis cases, however, has not been uniform across Italy; in fact, ~60–70% of cases that are notified annually are reported by northern regions, while there are several regions in southern Italy that consistently do not report cases. In addition, the marked rise in legionellosis incidence observed in Italy since 2002 is only partly accounted for by increased reporting by central and southern regions. In fact, the greatest increases in incidence rates have been observed in northern regions which already previously reported an elevated number of cases (Fig.) notwithstanding the fact that no large clusters of disease have been detected in the observed period, that may have modified the trend of the disease. For example one northern region reported 246 cases in 2002, approximately double the number of cases reported in the previous year, while several southern regions have continued in not reporting any cases at all.

On the other hand, according to a recent European study [9] regarding the distribution of travel-associated Legionnaires' disease (TALD) within selected European countries, most TALD cases in Italy occur in the North, but if rates of LD per 100 000 travellers are calculated, the burden of disease appears to be more standardized across the country.

This supports the hypothesis that differences in incidence rates observed between northern and southern regions are not a result of actual fluctuations in the distribution of the disease in Italy but rather, a result of differing attitudes towards disease reporting and/or diagnosis of legionellosis in the different regions. Assuming that a substantial degree of underreporting occurs, especially in central-southern regions, we attempted to quantify it through the capture–recapture method.

Capture–recapture studies are valuable tools in epidemiology for estimating the extent of incomplete ascertainment using population-based data from two independent, but overlapping sources. Originally developed in wildlife biology and demography, the method has been subsequently adapted for epidemiology, to provide population parameter estimates based on two or more incomplete sources [10–14].

The objective of the present study was to estimate the degree of underreporting of cases to the National Legionellosis Register (NLR) in 2002, using a capture–recapture approach and two different data sources.

METHODS

Sources of data

The following two sources of data were used for the study: the NLR database for 2002 and hospital discharge records (HDR) for the same year.

Source 1: NLR

Since 1983, when the NLR was established, it became mandatory for clinicians in Italy to report all confirmed or presumptive cases of legionellosis directly to the Istituto Superiore di Sanità (ISS) and to the Ministry of Health. To notify a case, a special surveillance form must be completed which includes information on symptoms and risk factors such as chronic illnesses, smoking status, alcohol consumption, previous hospitalization, and travel. All forms are entered into the computerized NLR database and analysed on a regular basis to evaluate disease trends and identify the presence of community, nosocomial or travel-associated clusters as well as risk factors for infection.

Case definition

The case definition of LD used by the NLR includes confirmed and presumptive cases according to the following criteria:

Confirmed case. A confirmed case of LD is defined as a case of radiologically confirmed pneumonia with laboratory evidence of acute *Legionella* infection including: (*a*) isolation of any species or serogroup of *Legionella* from respiratory secretions, lung tissue, or

https://doi.org/10.1017/S0950268806007667 Published online by Cambridge University Press

blood, (*b*) a four-fold or higher rise in specific serum antibody titre against *L. pneumophila* serogroup 1 by immunofluorescence or microagglutination in paired acute- and convalescent-phase serum specimens, or (*c*) detection of *L. pneumophila* antigen in urine.

Presumptive case. A presumptive case of LD is defined as a case of radiologically confirmed pneumonia with laboratory evidence of acute infection with *Legionella* including: (*a*) a four-fold or higher rise in specific serum antibody titre to *L. pneumophila* other serogroups or other *Legionella* species by immunofluorescence or microagglutination in paired acute- and convalescent-phase serum specimens, (*b*) a single high titre (>1:256) against *L. pneumophila* serogroup 1, (*c*) the detection of specific *Legionella* antigen in respiratory secretion or direct fluorescent antibody (DFA) staining of the organism in respiratory secretion or lung tissue using evaluated monoclonal reagents.

The NLR exclusively includes cases of legionellosis diagnosed in Italy even if the patient is a non-resident. Before being inserted in the register, data is validated, by checking that all diagnosis meets the case definition for a confirmed or presumptive case of legionellosis.

Source 2: Hospital discharge records

Hospitalized cases of legionellosis occurring in either Italian or non-Italian residents are recorded by each Italian hospital in a national database used primarily for administrative and planning purposes. Data registered by this system includes information about the patient, admission/discharge dates and diagnoses (up to six diagnosis codes are permitted), which are recorded using the 9th International Classification of Diseases – Clinical Modification (ICD-9-CM) [15]. As there is no legionellosis-specific ICD-9-CM code in the version currently used in Italy (version 17), for the present study, the following three codes were used to identify legionellosis cases: 482.83 – pneumonia caused by other Gram-negative bacteria (code recommended by Ministry of Health), 482.89 – pneumonia caused by other specific bacteria, 483.8 – pneumonia from other specific organisms. The latter two codes were added in order to increase the sensitivity of the HDR because they represent two broad categories which could, therefore, also include cases of legionellosis. Due the low specificity of the selected ICD-9-CM codes, however, cases extracted from the HDR database but not registered in the NLR were included in the study only if laboratory confirmation was documented.

For the purposes of this study, individual regions were requested to extract, from regional HDR databases, a list of all hospital admissions, with an admission or discharge date between 1 January and 31 December 2002, and with one of the above codes (482.83, 482.89 or 483.8) as a main or secondary discharge diagnosis. Data obtained for each possible legionellosis case included: patient name, surname, date of birth, gender, admission and discharge dates, main discharge diagnosis, secondary diagnoses if present, hospital of admission.

Capture–recapture method

The degree of underreporting was estimated for one year only and 2002 was chosen because it was the most recent year for which complete HDR data was available.

Legionellosis cases identified from the two aforementioned sources were matched through a stepwise procedure by first using the patient's name and surname, or patient's tax identification number, a unique code given to each person residing in Italy, when available. A list of all legionellosis cases appearing in only one of the two databases was then generated and the data fields where compared first automatically and then manually by two researchers, to ensure that data input errors did not account for failure to match. Second, date of birth, date of diagnosis and date of discharge were checked for matching in both data sources to verify that the records were related to the same person and to the same event. Data was linked in Access 2000.

This comparison allowed us to identify three groups of cases: cases registered in both databases, those registered in the NLR but not in the HDR and finally, cases registered in the HDR but not in the NLR.

No further investigations were performed for cases present in both databases, since these necessarily had to meet the case definition for legionellosis in order to be included in the NLR database.

When a case appeared in the HDR but not in the NLR, the diagnosis of legionellosis was confirmed by contacting the hospital laboratory and checking for laboratory confirmation, in order that the same case definition used by the NLR could be applied.

Finally, for cases registered in the NLR but which did not appear in the HDR database, a thorough search was performed in the HDR database to

https://doi.org/10.1017/S0950268806007667 Published online by Cambridge University Press

Table. *Number of cases of legionellosis in 2002, in Italy, estimated by capture–recapture analysis in the 17 participating regions*

	Cases identified (*n*)				Capture–recapture method (*n*)		% underreported cases		Incidence rate/ 1 000 000 as estimated by capture–recapture
	By NLR	By HDR	Matching cases	By both sources	Estimated total cases (95 % CI)	Estimated unreported cases	NLR	HDR	
Italy	566	556	437	685	720 (705–735)	35	21·4	22·8	14·35
North Italy	477	456	365	568	596 (583–609)	28	20·0	23·5	23·75
Centre-South Italy	89	100	72	117	124 (118–130)	7	28·2	19·4	4·94

NLR, National Legionellosis Register; HDR, hospital discharge record; CI, confidence interval.

identify other ICD-9-CM codes that may have been used to codify the discharge diagnosis in these specific cases.

Data analysis

Incidence was calculated by using as the denominator the total Italian population for each year (1997–2004), provided by National Institute of Statistics. For 2002, overall incidence and incidence by geographical area (North, Centre-South) were calculated, using as the numerator the number of cases estimated by the capture–recapture method in the two areas and using as the denominator, the population of the 17 regions that participated in the study.

The number of unreported cases and the total number of cases in the population and confidence intervals were estimated by using the Chapman & Seber formula [16].

For each data source, the degree of underreporting was calculated by dividing the number of cases of legionellosis not identified by that source by the number of cases estimated from the capture–recapture analysis. All the analyses were performed by using Epi-Info version 6.04d (CDC, Atlanta, GA, USA).

RESULTS

Seventeen out of 21 Italian regions participated in the study, these included 8 out of 9 northern regions and 9 out of 12 central-southern regions. Participating regions reported 566 cases of legionellosis to the NLR in 2002 (range 0–246 cases per region); of these, 1·9 % were foreign citizens and 1·8 % were patients admitted to a hospital located in a region different from the patient's region of residence. One hundred twenty-nine out of 566 cases were not found in the HDR by using the three ICD-9-CM codes chosen for the purposes of the study.

During the same time period, 556 cases with a clinically compatible illness that was laboratory confirmed as legionellosis, were recorded in the HDR database for the same regions (range 0–252 cases per region); 119 of these cases were not present in the NLR database.

Matching was possible in 437 cases; in total, therefore, 685 cases of legionellosis were identified and an agreement of 64 % was found between the two sources. We estimated, using the Chapman & Seber formula, that an additional 35 cases were not captured by either data source; the range for the number of cases not captured was found to be 0–26 in the different regions. The degree of underreporting to the NLR was estimated to be 21·4 %, while for northern regions it was estimated to be significantly ($P=0{\cdot}04$) lower (20·0 %), than for central-southern regions (28·2 %) (Table).

In 2002 the overall incidence of legionellosis for the 17 regions, as estimated by the number of cases notified to the NLR, was 11·3 cases per million population, while it was estimated to be 14·3 cases per million population by using data obtained with the capture–recapture method. Using the same method, the incidence rate in the northern regions was 23·7 cases per million, almost five times higher than that found in the central and southern regions (incidence rate = 4·9 cases per million) (Table).

DISCUSSION

In this study, we attempted to evaluate the degree of underreporting to the NLR surveillance system by using a two-source capture–recapture analysis.

https://doi.org/10.1017/S0950268806007667 Published online by Cambridge University Press

This analysis indicated that the NLR captures ~80% of cases from the estimated total from both sources. In addition, our results showed that, after including unreported case estimates, the pattern of legionellosis incidence in Italy did not significantly change, and substantial regional differences in ascertainment remained. These differences were found to be most pronounced between northern and central-southern regions, with the highest incidence rates observed in northern Italy.

From a methodological point of view, effective use of the capture–recapture technique is dependent on a number of assumptions which include: use of independent data sources, same case definition, correct identification of cases, and equal probability of inclusion [13].

In our study, case matching was accurately performed and no records were discarded due to data incompleteness. The same case definition was applied to each data source and laboratory confirmation of all cases was required. This means that all identified cases were 'true' cases. Equal probability of inclusion in both databases was assumed since almost all cases of legionellosis are admitted to hospital for treatment. In fact, 99% of cases of legionellosis reported to the NLR in 2002 required hospitalization. This may represent a slight over-estimation of the percentage of hospitalized cases as there are likely to have been cases of legionellosis diagnosed by general practitioners or by private sector physicians that were not reported. Nevertheless, the fact that such a high percentage of cases found in the NLR were hospitalized allowed us to use the HDR database as a second source of data, since both sources do not include the less severe cases that did not require hospitalization.

One possible limitation of using the HDR database as a data source is the absence of a specific ICD-9-CM code for legionellosis. Although laboratory confirmation was requested for all cases of pneumonia included in this study, we cannot rule out that some legionellosis cases may have been missed because codified, in the HDR database, with other ICD-9-CM codes not considered. Nevertheless, by using the three selected codes, we found a 64% concordance between the two databases, while 119 cases (17% of total cases included in the study) were present only in the HDR database.

It is important to consider that an initial search in the HDR database for cases codified with any of the three selected codes, identified 4186 possible cases of legionellosis that were not present in the NLR database. However, by carefully applying the same case definition used by the NLR to each possible case, the number of confirmed or presumptive cases of legionellosis was reduced to 119.

On the other hand, 129 cases (19% of total) reported to the NLR were not found in the HDR by using the three selected ICD-9-CM codes. A careful evaluation of these cases revealed that they were in fact present in the HDR database but codified with ICD-9-CM codes other than those considered for this study. This evaluation permitted us to identify other ICD-9-CM codes that are used to codify for legionellosis. These include: code 485 – bronchopneumonia not specified; 486 – pneumonia, agent not specified; 518.81 – acute respiratory insufficiency; 490.7 – bronchitis not specified. It is clear that, in the absence of a specific ICD-9-CM code for legionellosis, most clinicians codify the disease with one of the three codes used in this study, although other, more general and less appropriate codes are also used in ~20% of cases.

This finding underscores the importance of moving from the ICD-9-CM to the ICD-9-CM 2002 revision (version 19), as recommended by the Italian Ministry of Health with a decree passed in November 2005 [17]. Use of the new ICD-9-CM version, which introduces a specific code (482.84) for the diagnosis of LD would make the comparison between different sources easier and more straightforward, allowing a more accurate estimation of underreporting.

When using capture–recapture methods, the use of completely independent data sources is rarely possible [13]. For a disease like legionellosis, which is subject to statutory notification by clinicians and is primarily diagnosed in hospitals, truly independent sources of data are impossible to find. In this study, clinicians diagnosing legionellosis cases in hospital are legally required to report these cases to the NLR, while HDR codes are, in most cases not assigned by the same physician. It is only in a few cases that the diagnosing clinician may also be requested to attribute the HDR codes. When using only two sources of data, it is not possible to statistically test dependence of data. We assumed that positive dependence probably led to an overestimation of the sensitivity of the surveillance system and a conservative estimate of the incidence rate.

The results of this study showed that in 2002, the incidence of legionellosis in the participating regions, estimated by the capture–recapture method, is 23·7 cases per million population in the North, which is

https://doi.org/10.1017/S0950268806007667 Published online by Cambridge University Press

almost five times that estimated for the Centre-South (4·9 cases per million). Results also showed that underreporting of legionellosis does occur in Italy with a clear increasing trend from North to South but that the observed differences in incidence rates between the two geographical areas cannot be explained solely by differences in the degree of underreporting. Moreover, it is unlikely that the observed regional variations are a result of actual differences in the pattern of disease in the different regions. In fact, regions with high incidence rates border or are close to regions with low incidence rates and these differences do not reflect variations in lifestyle, population structure or exposure to potential risk factors.

We hypothesize that in addition to underreporting, under-diagnosis plays an important role in explaining the lower incidence rates observed in central-southern Italy, since many physicians, especially in these regions, diagnose pneumonia on the basis of clinical signs and symptoms alone without confirming its aetiology.

It is essential, therefore, that this problem be addressed, by improving awareness of LD among physicians and by encouraging them to seek aetiological confirmation in the diagnosis of pneumonia, for example by using urinary antigen testing, which over the past decade has proved to be an efficient means of diagnosis of *Legionella* infections. It is also fundamental to encourage physicians to notify all *Legionella* cases promptly. Early diagnosis of LD and timely notification by physicians would allow public health authorities to implement appropriate preventive measures and to accurately evaluate the real disease burden in Italy.

APPENDIX. Legionellosis Working Group

Piemonte: V. Demicheli, C. Di Pietrantoni, R. Raso; *Val D'Aosta*: L. Sudano; *Lombardia*: G. Bertani, L. Macchi; *P. A. Bolzano*: P. Kreidl; *Veneto*: A. Ferro, F. Michieletto, E. Verizzi; *Friuli Venezia Giulia*: G. Rocco; *Liguria*: P. Durando, S. Sensi; *Emilia Romagna*: A. Cappelletti, C. Ancarani; *Marche*: G. Grilli, E. Manzo; *Lazio*: F. Curtale, E. Ferroni, L. Alecci; *Abruzzo*: R. Cassini; *Molise*: L. D'Alò, G. Di Giorgio; *Campania*: R. Pizzuti, E. De Campora, A. Lombardo; *Puglia*: M. T. Montagna, C. Napoli, D. Tatò; *Basilicata*: G. Montagano, M. Gallo; *Sicilia*: S. Scondotto, A. Cernigliano, A. Nicolosi, A. Mira; *Sardegna*: G. Novelli, R. Masala.

DECLARATION OF INTEREST

None.

REFERENCES

1. **European Working Group for Legionella Infections.** European guidelines for control and prevention of travel associated Legionnaires Disease, 2002 (http://www.ewgli.org/data/european_guidelines/european_guidelines_jan05.pdf). Accessed 12 July 2006.
2. **Stout JE, Yu VL.** Legionellosis. *New England Journal of Medicine* 1997; **337**: 682–687.
3. **Marston BJ, *et al.*** Incidence of community-acquired pneumonia requiring hospitalizations: results of a population-based active surveillance study in Ohio. Community-Based Pneumonia Incidence Study Group. *Archives of Internal Medicine* 1997; **157**: 1709–1718.
4. **Fang GD, *et al.*** New and emerging etiologies for community-acquired pneumonia with implications for therapy: a prospective multicenter study of 359 cases. *Medicine* 1990; **69**: 307–316.
5. **Bartlett JG, *et al.*** Practice guidelines for the management of community-acquired pneumonia in adults. *Clinical Infectious Diseases* 2000; **31**: 347–382.
6. **Joseph CA.** Legionnaires' disease in Europe 2000–2002. *Epidemiology and Infection* 2004; **132**: 417–424.
7. **Rota MC, Castellani Pastoris M, Salmaso S.** Legionellosis in Italy in 1997. *Notiziario Istituto Superiore di Sanità* 1998; **11**: 1–5.
8. **Rota MC, Caporali MG, Ricci ML.** Legionellosis in Italy in 2004. Annual Report. *Notiziario Istituto Superiore di Sanità* 2005; **18**: 3–9.
9. **Ricketts KD, Joseph CA.** The distribution of travel-associated Legionnaires' disease within selected European countries, and a comparison with tourist patterns. *Epidemiology and Infection* 2006; **134**: 887–893.
10. **Seber GA.** *The Estimation of Animal Abundance and Related Parameters*, 2nd edn. London: Charles Griffin, 1982.
11. **El-Khorazaty M, *et al.*** Estimating the total number of events with data from multiple-record systems: a review of methodological strategies. *International Statistical Review* 1977; **45**: 129–157.
12. **Hook EB, Regal RR.** The value of capture-recapture methods even for apparent exhaustive surveys. *American Journal of Epidemiology* 1992; **135**: 1060–1067.
13. **Hook EB, Regal RR.** Effect of variation in probability of ascertainment by sources ('variable catchability') upon 'capture-recapture' estimates of prevalence. *American Journal of Epidemiology* 1993; **137**: 1148–1166.
14. **Hook EB, Regal RR.** Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 1995; **17**: 243–264.
15. **Ministry of Health, Health Care Planning Department.** *International Classification of Diseases. ICD9 CM,*

https://doi.org/10.1017/S0950268806007667 Published online by Cambridge University Press

Italian version, 1997. Rome: Istituto Poligrafico e Zecca della Stato, 2001.

16. **Chapman DG.** Some properties of the hypergeometric distribution with applications to zoological sample census. *University of California Public Statistics* 1951; **1**: 131–160.

17. **Ministry of Health.** DM 21 November 2005. Updated classification systems for codifying clinical information in hospital discharge records and for reimbursing hospital services. *Official Bulletin* (*Gazzetta Ufficiale*) no. 23, 5 December 2005.

https://doi.org/10.1017/S0950268806007667 Published online by Cambridge University Press