Epidemiol. Infect. (2008), **136**, 540–550. © 2007 Cambridge University Press
doi:10.1017/S0950268807008977 Printed in the United Kingdom

Incidence and completeness of notification of Legionnaires' disease in The Netherlands: covariate capture–recapture analysis acknowledging regional differences

N. A. H. VAN HEST[1,2]*, C. J. P. A. HOEBE[3], J. W. DEN BOER[4], J. K. VERMUNT[5], E. P. F. IJZERMAN[6], W. G. BOERSMA[7] AND J. H. RICHARDUS[1,2]

[1] *Department of Infectious Disease Control, Municipal Public Health Service Rotterdam-Rijnmond, Rotterdam, The Netherlands*
[2] *Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam, The Netherlands*
[3] *Department of Infectious Disease Control, South Limburg Public Health Service, Geleen, The Netherlands*
[4] *Department of Infectious Disease Control and the Environment, Kennemerland Public Health Service, Haarlem, The Netherlands*
[5] *Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands*
[6] *Regional Public Health Laboratory Kennermerland, Haarlem, The Netherlands*
[7] *Department of Pulmonary Diseases, Medical Centre Alkmaar, Alkmaar, The Netherlands*

(*Accepted 14 May 2007; first published online 22 June 2007*)

SUMMARY

To estimate incidence and completeness of notification of Legionnaires' disease (LD) in The Netherlands in 2000 and 2001, we performed a capture–recapture analysis using three registers: Notifications, Laboratory results and Hospital admissions. After record-linkage, 373 of the 780 LD patients identified were notified. Ascertained under-notification was 52·2%. Because of expected and observed regional differences in the incidence rate of LD, alternatively to conventional log-linear capture–recapture models, a covariate (region) capture–recapture model, not previously used for estimating infectious disease incidence, was specified and estimated 886 LD patients (95% confidence interval 827–1022). Estimated under-notification was 57·9%. Notified, ascertained and estimated average annual incidence rates of LD were 1·15, 2·42 and 2·77/100 000 inhabitants respectively, with the highest incidence in the southern region of The Netherlands. Covariate capture–recapture analysis acknowledging regional differences of LD incidence appears to reduce bias in the estimated national incidence rate.

INTRODUCTION

Any surveillance system is concerned with the quality of the data collected, including the degree of ascertainment of affected individuals [1]. A conventional surveillance system is notification, possibly containing false-positive cases and often incomplete for true-positive cases, as described for Legionnaires' disease (LD) [2, 3].

LD is a serious, possibly fatal, pneumonia caused by *Legionella* spp., occurring in sporadic cases and outbreaks [4, 5]. Under the present legislation regarding infectious diseases in The Netherlands, LD is placed in category B. This group of infectious diseases has to be notified within 24 h to the Municipal Public Health Service by the diagnosing physician. The Municipal Public Health Service forwards this information to the Register of Notifiable Infectious

* Author for correspondence: N. A. H. van Hest, M.D., M.Sc., Tuberculosis Control Physician/Epidemiologist, Division of Infectious Disease Control, Municipal Public Health Service Rotterdam-Rijnmond, PO Box 70032, 3000 LP Rotterdam, The Netherlands.
(Email: vanhestr@ggd.rotterdam.nl)

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

Diseases at the Office of the Health Care Inspectorate where national data are aggregated for analysis, monitoring, public health intervention or policy making. Since 1999 an average of 230 LD patients were notified in The Netherlands annually. The average national annual incidence rate was 1·4 LD patients per 100 000 inhabitants, almost three times higher than the average annual incidence rate in the United States and the United Kingdom [6, 7]. However, the incidence rate based on notifications varies considerably per province [8]. Under-diagnosis and under-notification are likely. This can obscure the true burden of LD, hamper the detection of clusters of LD patients and hinder good investigations into the possible source of legionella infections. The Dutch Health Council estimated an annual number of 800 LD patients. This number is based on the annual number of cases of pneumonia in The Netherlands (110 000) of whom 15% require hospital admission (16 000) of which 5% is caused by *Legionella* spp. (800) [9].

Record-linkage is important for assessing the quality and completeness of infectious disease registers, i.e. comparing patient data across multiple registers [10]. Completeness of notification can be assessed by comparison with case-ascertainment, i.e. the total number of patients observed in at least one register, or the estimated total number of patients obtained by capture–recapture analysis. The total number of individuals present in one or more registrations does not necessarily reflect a reliable approximation of the true number of cases. The purpose of capture–recapture analysis is to assess the number of cases that are not registered. In an article published in 1972, Stephen Fienberg demonstrated how this number of unobserved cases could be estimated, using log-linear analysis [11]. For capture–recapture analysis, according to Fienberg, the availability of data from at least three different, possibly incomplete, partially overlapping and preferably, but not necessarily, independent sources is needed [12–16]. The data can be put in a $2 \times 2 \times 2$ contingency table, indicating the absence or presence of a case in each of the registers. This table has one empty cell, corresponding to the number of cases never registered. Based on certain assumptions, which will be discussed later, capture–recapture analysis aims at obtaining an estimate of the unregistered number of patients in the empty cell from the available data in the other cells. This estimate can be found under the best fitting and most parsimonious log-linear model, as explained later. Finally, the total number of individuals is the number of registered cases plus the estimated number of non-registered patients. Capture–recapture methods have been used to estimate the total number of patients with LD and other infectious diseases [2, 3, 17].

The validity of capture–recapture analysis depends on possible violation of the underlying assumptions and one focus is to establish which method is most appropriate for specific datasets [15]. Usually, log-linear modelling of data from at least three linked registers is the preferred capture–recapture method because it can reduce bias due to inter-dependencies between two registers [13, 17] Stratified capture–recapture analysis according to categorical covariates associated with the probability of capture in a register can further reduce bias [11, 12, 14, 16]. An alternative is to include these covariates, e.g. demographic, diagnostic or prognostic variables, in a log-linear covariate capture–recapture model but these models have rarely been used to estimate human disease incidence [18, 19].

This study aims to estimate incidence and completeness of notification of LD in The Netherlands in 2000 and 2001 using record-linkage of three data sources and capture–recapture analysis.

MATERIALS AND METHODS

Data sources and patient identifiers

Three LD data sources were used:

(1) *Notification*. Patients notified by their physician to the Health Care Inspectorate. A uniform questionnaire collected additional information from local Public Health Services processing the notifications.

(2) *Laboratory*. Patients with a specified positive laboratory test result reported by the clinical microbiologists in a survey among all clinical microbiology laboratories after obtaining permission for this survey from the Dutch Society for Microbiology and supported by the Inspector-General for Infectious Diseases of the Health Care Inspectorate. Positive laboratory test results were classified as either confirmed (culture, urine antigen test or a fourfold rise in antibody titre [⩾128 IU] against *Legionella* spp. in paired acute and convalescent serum samples) or probable [PCR, a high titre (⩾256 IU) against *Legionella* spp. in one serum sample or direct fluorescent antibody staining], according to the European Working Group for Legionella Infections

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

(EWGLI) definitions. Patients with LD only known to the Hospital register were classified as cases with unknown laboratory verification.

(3) *Hospital*. Hospitalized patients recorded in the National Morbidity Registration by Prismant, covering all hospitals in The Netherlands with:

(*a*) an International Code for Diseases (ICD-9 code) for all forms of pneumonia (ICD-9 codes 480.0–487.0) for individuals known to Notification and/or Laboratory.

(*b*) An ICD-9 code 482.8 for individuals only known to Hospital.

ICD-9 has no specific code for LD and, as reported from other countries, in The Netherlands ICD-9 code 482.8 (pneumonia due to other specified bacteria) is used for LD patients [20]. Hospital records coded as 482.8 can therefore include false-positive cases, mainly patients with *Escherichia coli* pneumonia, a rare nosocomial disease, predominantly occurring among intensive-care patients. Data on the annual number of *E. coli* pneumonia patients in The Netherlands are not available. Based upon an estimated annual number of 60 000 intensive-care admissions and an estimated *E. coli* pneumonia incidence of 1/1000 intensive-care admissions (derived from a random survey among intensive-care consultants in The Netherlands), the estimated annual number of *E. coli* pneumonia patients is 60. This number is used to correct the number of patients only known to Hospital. Because proxy code 482.8 is used for cross-validation and collection of additional information, uniform questionnaires requested all chest physicians to report hospitalized LD patients in 2000 and 2001.

For all patients in each register it was attempted to collect date of birth, postal code or town of residence, sex and date of notification (and first day of illness), first laboratory sample or hospital admission as personal identifiers to be used in all record-linkage procedures. Duplicate entries in each register were deleted.

Case-definition and study period

LD patients are defined as all ascertained (notified, laboratory-reported or hospitalized) and unascertained LD patients. Notified LD patients with a first day of illness in 2000 and 2001 were included in the study. For inclusion of patients known to Laboratory and/or Hospital the laboratory sample date, hospital admission date or first known of both dates were used as proxy for first day of illness. Through examining the registers 1 month before and after the study period, all registers were corrected for late notification or laboratory results, as described previously [17].

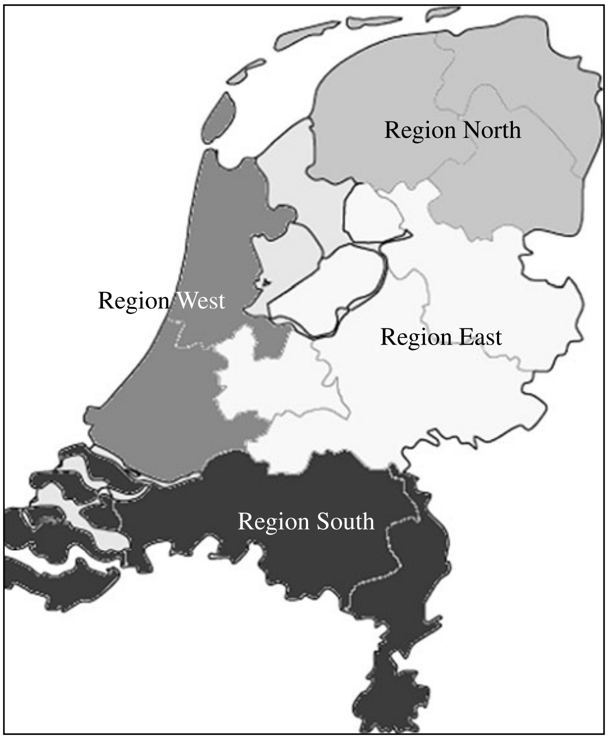

Fig. The four regions of The Netherlands.

Record-linkage and stratification

Record-linkage was performed manually using the patient identifiers, proximity of dates and geographical information found in the three registers. In case of doubt consensus was sought between two investigators. Because of expected geographical differences in incidence of LD, after record-linkage, on the basis of the provinces of The Netherlands, ascertained LD patients were stratified into four regions: North (1 671 534 inhabitants), East (4 467 527 inhabitants), West (5 955 299 inhabitants) and South (3 892 715 inhabitants) (Fig.). Correction for the estimated number of *E. coli* pneumonia patients in the different regions was proportional to the regional division of the total number of patients only ascertained in Hospital.

Coverage rates and capture–recapture analysis

The ascertained register-specific coverage rate is defined as the number of LD patients in each register

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

divided by the case-ascertainment, expressed as percentage. The total number of un-ascertained LD patients was estimated on the basis of the distribution of the ascertained cases over the three registers. For internal validity analysis we used two-source capture–recapture analysis, as explained elsewhere [21]. Briefly, by two-source capture–recapture analysis the estimated total number of cases, N_{est}, equals the number of cases on register A, N_{A}, times the number of cases on register B, N_{B}, divided by the overlap of the two registers, N_{both} ($N_{\text{est}} = N_{\text{A}} \times N_{\text{B}} / N_{\text{both}}$, also known as the Petersen estimator equation). Approximately unbiased estimates of N_{est} are expected when the registers are large. To correct for bias caused by small registers Chapman proposed the Nearly Unbiased Estimator, which can be expressed as $N_{\text{est}} = [(N_{\text{A}} + 1) \times (N_{\text{B}} + 1)/(N_{\text{both}} + 1)] - 1$ [13, 22, 23].

The independence of registers and other assumptions underlying capture–recapture analysis were described previously [17]. Specific interdependencies between the three registers, causing bias in two-source capture–recapture estimates, are probable. Using SPSS statistical software (version 13.0; SPSS Inc., Chicago, IL, USA), conventional total and stratified three-source log-linear capture–recapture analysis was employed taking possible interdependencies and heterogeneity into account, as previously described [17]. Alternatively to capture–recapture analysis stratified by region, a log-linear covariate capture–recapture model with one covariate, region, was specified [18, 19, 24]. Other covariates considered will be discussed later. The best-fitting models were identified using the likelihood ratio test (G^2). The null hypothesis in the likelihood-ratio goodness-of-fit test is that the specified model holds and the alternative is that it does not hold. If the null hypothesis does not need to be rejected (e.g. $P > 0{\cdot}05$) this means that there is no evidence that the specified model is in disagreement with the data. The lower the value of G^2 the better is the fit of the model. In the log-linear estimation procedure model selection follows model fitting, i.e. to identify the models that are clearly wrong and select from a number of acceptable models the most appropriate. For model selection we used Akaike's Information Criterion (AIC) which can be expressed as $\text{AIC} = G^2 - 2$ degrees of freedom (D.F.) [25]. The first term, G^2, is a measure of how well the model fits the data and the second term, 2 D.F., is a penalty for the addition of parameters (and hence model complexity). A second information criterion used was the Bayesian Information Criterion (BIC) which can be expressed as $\text{BIC} = G^2 - (\ln N_{\text{obs}})$ (D.F.), where N_{obs} is the total number of observed individuals [26]. Relative to the AIC, the BIC penalizes complex models more heavily. In general, in the log-linear capture–recapture estimation procedure the least complex, i.e. the least saturated (in other words the most parsimonious) model, whose fit appears adequate, is preferred [13]. Since the G^2 of the saturated model is zero and has no degrees of freedom left, the AIC and BIC are also zero and models with a negative AIC and BIC are preferred although this does not necessarily mean that the estimate is correct. The estimated register-specific coverage rate is defined as the number of LD patients in each register divided by the estimated total number of LD patients, expressed as percentage.

RESULTS

Notification system

In the notification register from the Health Care Inspectorate 358 LD patients were recorded. An additional 15 patients were reported through the questionnaires from local public health services processing the notifications, giving a total of 373 notified LD patients.

Laboratory survey

Questionnaires were received from 36 out of the 48 laboratories (response rate 75%). Based on population estimates the cooperating laboratories served 81·2% of the Dutch population. A total of 261 patients with a positive test for *Legionella* spp. were reported. Of these patients 186 (71·3%) were notified. Additional information on laboratory diagnosis was available for another 127 patients through Public Health Service or chest physician questionnaires, bringing the total number of patients with known laboratory results to 388.

Hospital records

From 385 chest physicians in The Netherlands 179 replies were received (response rate 46%), the majority indicating that the requested information could not be retrieved or no LD patients were admitted. Chest physicians reported 44 LD patients, all of them also known to Notification and/or Laboratory.

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

Out of 448 LD patients in Notification and/or Laboratory, 331 (73·9%) could be linked to the National Morbidity Registration pneumonia records. Of the linked LD patients 79 (23·9%) were classified as either 'pneumonia not specified' (ICD-9 code 486, 63 cases), 'pneumonia due to other specified organism' (ICD-9 code 483, nine cases) or 'pneumococcal pneumonia' (ICD-9 code 481, seven cases). The remaining 252 linked patients (76·1%) had ICD-9 code 482.8, the assigned code for LD. Another 452 patients, unknown to Notification and/or Laboratory, were identified in Hospital with ICD-9 code 482.8. This number was adjusted to 332 LD patients after deduction of an estimated number of 120 *E. coli* pneumonia patients in the two years studied, also recorded under ICD-9 code 482.8.

Epidemiological results

Table 1 shows the epidemiological characteristics of 447 LD patients in Notification and/or Laboratory (one patient had insufficient data). The mean age was 54 years (S.D. = 14 years). The recorded case-fatality rate was 5·6%. The mean duration between onset of disease and microbiological diagnosis was 12 days (median 6 days). The mean duration of hospital admission was 19 days (median 13 days).

Table 2 shows the number and proportion per region of the different laboratory tests for *Legionella* spp. There are differences between the four Dutch regions in laboratory diagnostic approach. In region North no culture results were reported. In region West a low proportion of fourfold rise in antibody titre and PCR results were reported and more patients had unknown test results, probably the result of non-participation of some larger laboratories. In region South a high proportion of a fourfold rise in antibody titre and PCR results were reported, probably the result of a major reference laboratory in that region.

Case-ascertainment

Table 3 shows the distribution of the 780 ascertained LD patients over the three registrations after record-linkage, in total and stratified by region. The ascertained register-specific coverage rate of Notification, Laboratory and Hospital was 47·8% (373/780), 33·5% (261/780) and 85·0% (663/780) respectively. The ascertained under-notification was 52·2%. Table 4 shows the number of notified and ascertained LD patients, the average annual incidence rate by notification and by case-ascertainment and the proportion of the ascertained patients notified, in total and stratified per region. The average national annual incidence rate by notification was 1·15/100 000 and by case-ascertainment 2·42/100 000. The regional annual incidence rates differ, with a 100% difference between the highest and lowest regional incidence rate based on notification, reducing to 50% difference after record-linkage. Based upon the notification data the low incidence rate in region North partly results from under-notification but the notified and ascertained incidence rates in region South were higher than in the rest of The Netherlands ($P<0·0001$).

Capture–recapture analysis

Internal validity analysis by two-source capture–recapture analysis on Notification and Hospital and on Laboratory and Hospital both estimate 865 LD patients through Chapman's Nearly Unbiased Estimator. The considerable lower capture–recapture estimate obtained with Notification and Laboratory (523 LD patients) indicates a larger positive association between this pair than between the other pairs, resulting in an estimate more biased downwards.

The best-fitting three-source log-linear capture–recapture model was the saturated model, i.e. the model including all two-variable associations and assuming absent three-way interaction, which yielded an estimate of 1253 LD patients [95% confidence interval (CI) 1019–1715]. Estimated under-notification was 70·2%. To acknowledge the geographical differences capture–recapture analysis stratified by region was performed. For all regions apart from region East a more parsimonious model, containing only one two-way interaction (between Notification and Laboratory), was selected as best-fitting model, with totals of 78, 327 and 277 LD patients and incidence rates of 2·33, 2·75 and 3·56/100 000 inhabitants for regions North, West and South respectively. For region East a saturated model was selected that estimated an unexpectedly high number of 650 LD patients with a wide 95% CI of 283–2382 patients.

As an alternative to the stratified capture–recapture analysis we specified a log-linear covariate (region) capture–recapture model. The covariate model that served as a starting point contained, apart from the main effects for region and the three registers, the Region-Notification, Region-Laboratory, Region-Hospital, Notification-Laboratory, Notification-Hospital, Laboratory-Hospital two-variable terms.

Table 1. *Epidemiological characteristics of 447 Legionnaires' disease patients in The Netherlands**

	Male ($n=319$)	Female ($n=128$)	Total ($n=447$)
Age category (yr)			
0–19	0·3% (1/318)	4·7% (6/128)	1·6% (7/446)
20–39	11·9% (38/318)	18·0% (23/128)	13·7% (61/446)
40–59	55·3% (176/318)	43·8% (56/128)	52·0% (232/446)
60–79	28·9% (92/318)	30·5% (39/128)	29·4% (131/446)
⩾80	3·5% (11/318)	3·1% (4/128)	3·4% (15/446)
Seasonal pattern: month of disease onset			
Jan.–Feb.	7·8% (25/319)	10·2% (13/128)	8·5% (38/447)
Mar.–Apr.	11·0% (35/319)	10·9% (14/128)	11·0% (49/447)
May–June	19·4% (62/319)	14·1% (18/128)	17·9% (80/447)
July–Aug.	26·6% (85/319)	25·0% (32/128)	26·2% (117/447)
Sep.–Oct.	21·9% (70/319)	31·3% (40/128)	24·6% (110/447)
Nov.–Dec.	13·2% (42/319)	8·6% (11/128)	11·9% (53/447)
Travel abroad during incubation period†			
Travel abroad: yes	53% (169/319)	50% (64/128)	52% (233/447)
Countries involved			
Turkey	20% (33)	30% (19)	22% (52)
France	23% (39)	8% (5)	19% (44)
Spain	12% (21)	13% (8)	12% (29)
Italy	8% (14)	11% (7)	9% (21)
Germany	7% (12)	9% (6)	8% (18)
Portugal	2% (4)	2% (1)	2% (5)
Greece	2% (4)	2% (1)	2% (5)
Belgium	3% (5)	0%	2% (5)
Rest of Europe	11% (18)	13% (8)	11% (26)
Americas	5% (9)	6% (4)	6% (13)
Asia	3% (5)	2% (1)	3% (6)
Africa	0%	3% (2)	1% (2)
Unknown	3% (5)	3% (2)	3% (7)
Legionella spp.			
L. pneumophila serogroup 1	61·2% (170/278)	54·5% (60/110)	59·3% (230/388)
L. pneumophila serogroups 2–12	2·5% (7/278)	1·8% (2/110)	2·7% (9/388)
L. non-pneumophila	3·2% (9/278)	0·9% (1/110)	2·6% (10/388)
Unknown	31·7% (88/278)	42·7% (47/110)	34·8% (135/388)
Laboratory confirmation‡			
At least two confirming tests	22·0% (61/277)	17·3% (19/110)	20·7% (80/387)
One confirming test	56·0% (155/277)	56·4% (62/110)	56·1% (217/387)
Only probable test	22·0% (61/277)	26·4% (29/110)	23·3% (90/387)

* From 447 patients sufficient data was available for analysis; sometimes one or two variables are missing.

† *Rest of Europe*: Austria, Croatia, Cyprus, England, Hungary, Ireland, Luxembourg, Moldavia, Poland, Slovakia, Switzerland, Czech Republic, Yugoslavia; *Americas*: Netherlands Antilles, Brazil, Canada, Dominican Republic, Mexico, Peru, USA, Venezuela; *Asia*: China, Indonesia, Japan, Kazakhstan, Malaysia; *Africa*: Morocco and Tunis.

‡ Confirmed laboratory diagnosis: positive culture, positive urine antigen test or a fourfold rise in antibody titre against *Legionella* spp. in paired acute and convalescent serum samples, ⩾128 IU. Probable laboratory diagnosis: positive PCR, a high titre in one serum sample against *Legionella* spp., ⩾256 IU, or direct fluorescent antibody staining of the organism.

In this model we allow for regional differences in the number of cases in the three registers, but not for interaction with other effects per stratum, as the association between the registers is assumed equal across regions. This model fits the data reasonably well ($G^2=22{\cdot}1$, D.F.$=9$, $P=0{\cdot}009$) and estimates 932 LD patients with a narrower 95% CI of 851–1106, reducing statistical uncertainty. Inspection of the misfit for individual cells showed a large adjusted residual for LD patients only known to Laboratory in region East. After including a separate parameter for this single cell we obtain a good fitting model

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

Table 2. *Number and proportion of the laboratory test results for* Legionella *spp. in The Netherlands in 2000 and 2001, in total and stratified per region*

	Confirmed laboratory test			Probable laboratory test			
	Culture (%)	Urine antigen test (%)	Fourfold rise in antibody titre (%)	Positive PCR (%)	High single titre (%)	DFA†	Unknown
All *Legionella* pneumonia (100% of population)*	71 (100%)	216 (100%)	92 (100%)	33 (100%)	119 (100%)	0	56/441 (13%)
Region							
North (11%)	0 (0%)	15 (7%)	14 (15%)	2 (6%)	8 (7%)	0	3/34 (9%)
East (18%)	16 (22%)	61 (28%)	23 (25%)	5 (15%)	33 (28%)	0	14/123 (11%)
West (41%)	31 (44%)	78 (36%)	17 (19%)	3 (9%)	30 (25%)	0	31/149 (21%)
South (30%)	24 (34%)	62 (29%)	38 (41%)	23 (70%)	48 (40%)	0	8/135 (6%)

* For 441 patients information of region was known.

† Direct fluorescent antibody staining.

Table 3. *Ascertained total number of Legionnaires' disease* (*LD*) *patients and number stratified by region of The Netherlands in three linked LD registrations in 2000 and 2001, after proportional adjustment for false-positive* Escherichia coli *pneumonia patients only known to the Hospital register*

	No. ascertained	Only NOT*	Only LAB†	Only HOSP‡	NOT and LAB	NOT and HOSP	LAB and HOSP	NOT and LAB and HOSP
All LD patients	780	56	30	332	31	131	45	155
Region								
North§	69	3	2	35	2	6	8	13
East§	185	13	13	62	3	42	7	45
West§	286	23	5	136	7	55	14	46
South§	234	13	9	99	19	28	15	51

* NOT, Notification register (373 patients).

† LAB, Laboratory register (261 patients).

‡ HOSP, Hospital admission register. The proportional correction for the *Escherichia coli* pneumonia patients in regions North, East, West and South is 13, 22, 49 and 36 patients respectively (663 patients).

§ For six LD patients the place of residence unknown.

($G^2=5{\cdot}7$, d.f. $=8$, $P=0{\cdot}686$). The estimated number of LD patients is 886 (95% CI 827–1022), similar to the two internal validity estimates with least assumed interdependence.

The estimated register-specific coverage rate of Notification, Laboratory and Hospital was 42·1% (373/886), 29·5% (261/886) and 74·9% (663/886) respectively. The estimated under-notification was 57·9%. The estimated average annual incidence rate of LD was 2·77/100 000.

A sensitivity analysis, assuming double or half the number of false-positive cases due to *E. coli* pneumonia only known to Hospital, estimated the number of LD patients to range between 727 (95% CI 689–813) and 966 (95% CI 896–1126).

DISCUSSION

After record-linkage and log-linear covariate capture–recapture analysis of three registers of LD in 2000 and 2001 in The Netherlands we found a notified, ascertained and estimated annual incidence rate of 1·15, 2·42 and 2·77 cases/100 000 inhabitants

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

Table 4. *Number of notified and ascertained Legionnaires' disease (LD) patients, the average annual LD incidence rate (n/100 000) and the proportion of the ascertained LD patients notified in The Netherlands and stratified per region*

	Notification (passive surveillance)		Record-linkage (case-ascertainment)		
	Number of notified LD patients*	Average annual incidence rate (*N*/100 000)	Number of ascertained LD patients	Average annual incidence rate (*N*/100 000)	Proportion notified
All LD patients (15 987 075 inhabitants)	373	1·15	780	2·42	47·8 %
Region North (1 671 534 inhabitants)	24	0·72	69	2·06	34·8 %
Region East (4 467 527 inhabitants)	103	1·15	185	2·07	55·7 %
Region West (5 955 299 inhabitants)	131	1·10	286	2·40	46·0 %
Region South (3 892 715 inhabitants)	111	1·43	234	3·01	47·4 %

* The information on region was missing for four LD patients.

respectively. Ascertained and estimated under-notification was 52·2 % and 57·9 % respectively. This indicates the need for more consistent notification, e.g. through treatment of LD by a limited group of clinicians, familiar with notification. The southern part of The Netherlands had a higher notified, ascertained and estimated incidence rate of LD.

Legionella pneumonia might be responsible for 0–14 % of all nosocomial pneumonias and for 2–16 % of all community-acquired pneumonias [27]. In The Netherlands legionella pneumonia is reportedly responsible for 7 % of all nosocomial pneumonias and 2–8 % of all community-acquired pneumonias in hospitalized patients [28–30]. Under-notification of LD is estimated at 67 % in France, 90 % in England and 95 % in the United States [3, 31–33]. At 57·9 % we estimated a lower under-notification in The Netherlands, possibly influenced by increased awareness after a major outbreak or increased use of the urine antigen test (although this use is proportionally still low compared to the average EWGLI data for Europe) [4, 31]. Among patients in the laboratory survey with positive legionella results under-notification was 28·7 %, much lower than reported in France [2]. Parallel to mandatory notification by clinicians, many Dutch laboratories report positive results voluntarily to the public health services, which reduces under-notification of LD and other infectious diseases. The ascertained and estimated register-specific coverage rates for the laboratories would be higher with a better response. Record-linkage improved completeness of information in the linked dataset but, unlike laboratories, clinicians are not a useful source of additional information.

Several assumptions must be met for valid results of three-source log-linear capture–recapture models and limitations of capture–recapture analysis are described by others [13, 16, 34–39]. Violation of the closed population assumption is assumed limited for LD as opportunities for notification, laboratory verification or hospitalization are largely determined within a short period of time, but could result in over-estimation of the number of patients. Due to lack of a unique patient identification number used in all registrations and incomplete information on personal identifiers in some records, imperfect record-linkage cannot be excluded but balanced misclassification can still result in unbiased numbers in each category. Limitations of capture–recapture studies due to lack of a uniform and unambiguous case-definition and variable specificity of registers are described elsewhere [36, 40]. The notification criteria in The Netherlands requires a clinical diagnosis of pneumonia and a confirmed or probable laboratory diagnosis. However, for 187 (50·1 %) notified patients and 463 (69·8 %) hospitalized patients no laboratory-verification was found, although part of these patients could be microbiologically diagnosed in a non-participating laboratory or abroad or, due to imperfect record-linkage, could not be linked to Laboratory. Likewise

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

Laboratory may contain cases without pneumonia and cases diagnosed on a single high antibody titre, a test with a low positive predictive value [3, 29]. The 79 linked patients in Hospital with another pneumonia ICD-9 code than 482.8 are probably miscoded but some could be false-positive cases. Violation of the perfect positive value of the hospital episode registers is always a reason for concern in capture–recapture studies on infectious diseases and should be addressed critically, even when specific disease codes are used, e.g. for tuberculosis in ICD-9 [41–44]. We have corrected for imperfect positive predictive value for Hospital. Possible bias as a result of correction for other hospitalized patients with ICD-9 code 482.8 is reflected in the confidence intervals of the sensitivity analysis. Conventional log-linear capture–recapture analysis for The Netherlands and region East selected the saturated model, with an unexpectedly high estimate in region East. When saturated capture–recapture models are selected by any criterion investigators should be particularly cautious about the associated outcomes [16, 44–46]. We selected the three-source covariate capture–recapture model with equal two-way interactions across the regions as the best-fitting model. Internal validity analysis and analyses stratified by region indicate dependence between Notification and Laboratory as the dominant interaction. Positive three-way interaction across sources, causing underestimation of the number of LD patients, cannot be incorporated in the selected model but is arguably limited. Regional heterogeneity in the probability of being captured in the different registers was expected and observed [3, 8]. Covariate capture–recapture models have been used only rarely to estimate disease incidence but appear to reduce bias due to heterogeneity and result in plausible estimates of the total number of cases, e.g. in simulations [18, 19]. Inclusion of other covariates than region in the model, such as age or method of laboratory diagnosis, could have further reduced bias. In France, apart from region, method of diagnosis was identified as a variable with heterogeneity of capture [3]. However, proportional correction for *E. coli* pneumonia patients in Hospital, as performed for the regional stratification, was not feasible. Bias due to exclusion of these and unobserved possibly relevant covariates from the model can not be excluded.

Different characteristics of diseases, the patients and their registers can introduce various degrees of register interdependence and population heterogeneity into capture–recapture analysis, influencing model preference. This study shows that in The Netherlands for LD there is considerable interdependence between Notification and Laboratory and confirms geographical heterogeneity. Log-linear covariate capture–recapture analysis with region as covariate appears to reduce bias in the estimated number of LD patients. To our knowledge this is the first covariate capture–recapture study performed for infectious disease surveillance. Further research is needed into the causes of the geographical differences of LD incidence rates.

ACKNOWLEDGEMENTS

We thank Dr Carol Joseph for reviewing an earlier version of the manuscript. Permission for this study was obtained from the Medical Ethics Committee of the Erasmus MC, University Medical Centre Rotterdam, Rotterdam, The Netherlands, and the data protection committees of the Legionnaires' disease registrations.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Nanan DJ, White F.** Capture–recapture: reconnaissance of a demographic technique in epidemiology. *Chronic Diseases in Canada* 1997; **18**: 144–148.
2. **Infuso A, Hubert B, Etienne J.** Underreporting of Legionnaires' disease in France: the case for more active surveillance. *Eurosurveillance* 1998; **3**: 48–50.
3. **Nardone A, *et al.*** Repeat capture–recapture studies as part of the evaluation of the surveillance of Legionnaires' disease in France. *Epidemiology and Infection* 2003; **131**: 647–654.
4. **Den Boer JW, *et al.*** A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. *Emerging Infectious Diseases* 2002; **8**: 37–43.
5. **Lettinga KD, *et al.*** Legionnaires' Disease at a Dutch flower show: prognostic factors and impact of therapy. *Emerging Infectious Diseases* 2002; **8**: 1448–1454.
6. **Ricketts KD, Joseph CA.** Legionnaires' disease in Europe 2003–2004. *Eurosurveillance* 2005; **10**: 256–259.
7. **Jajosky RA, *et al.*** Summary of Notifiable Diseases – United States, 2004. *Morbidity and Mortality Weekly Report* 2006; **53**: 1–79.
8. **Den Boer JW, Friesema IH, Hooi JD.** Reported cases of Legionnaires' disease in the Netherlands, 1987–2000 [in Dutch]. *Nederlands Tijdschrift voor Geneeskinde* 2002; **46**: 315–320.

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

9. **Health Council of The Netherlands**. *Controlling Legionnaire's Disease*. The Hague: Health Council of The Netherlands, 2003; publication no. 2003/12 (http://www.gr.nl/pdf.php?ID=727&p=1). Accessed 18 April 2007.
10. **Migliori GB, *et al.*** Validation of the surveillance system for new cases of tuberculosis in a province of northern Italy. Varese Tuberculosis Study Group. *European Respiratory Journal* 1995; **8**: 1252–1258.
11. **Fienberg SE.** The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika* 1972; **59**: 591–603.
12. **Bishop YMM, Fienberg SE, Holland PW.** *Discrete Multivariate Analysis*. Cambridge: MIT Press, 1975.
13. **International Working Group for Disease Monitoring and Forecasting.** Capture–recapture and multiple-record estimation I: History and theoretical development. *American Journal of Epidemiology* 1995; **142**: 1047–1058.
14. **International Working Group for Disease Monitoring and Forecasting.** Capture–recapture and multiple-record estimation II: Applications in human diseases. *American Journal of Epidemiology* 1995; **142**: 1059–1068.
15. **Chao A, *et al.*** The applications of capture–recapture models to epidemiological data. *Statistics in Medicine* 2001; **20**: 3123–3157.
16. **Hook EB, Regal RR.** Capture–recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 1995; **17**: 243–263.
17. **Van Hest NA, Smit F, Verhave JP.** Underreporting of malaria incidence in The Netherlands: results from a capture–recapture study. *Epidemiology and Infection* 2002; **129**: 371–377.
18. **Tilling K, Sterne JA.** Capture–recapture models including covariate effects. *American Journal of Epidemiology* 1999; **149**: 392–400.
19. **Tilling K, Sterne JA, Wolfe CD.** Estimation of the incidence of stroke using a capture–recapture model including covariates. *International Journal of Epidemiology* 2001; **30**: 1351–1359.
20. **Slobbe LC, *et al.*** Classification of diagnoses and procedures and application in new hospital episode statistics [in Dutch]. Bilthoven, The Netherlands. National Institute of Public Health and the Environment (RIVM), 2004. RIVM report 260201002/2004, p. 73 (http://www.cbs.nl/NR/rdonlyres/E9DC7CF9-0BDF-40EA-A52D-EE6FBEE1B904/0/rivmrapport260201002.pdf). Accessed 18 April 2007.
21. **Hook EB, Regal RR.** Internal validity analysis: a method for adjusting capture–recapture estimates of prevalence. *American Journal of Epidemiology* 1995; **142**: S48–52.
22. **Chapman CJ.** Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics* 1951; **1**: 131–160.
23. **Wittes JT.** On the bias and estimated variance of Chapman's two-sample capture–recapture estimate. *Biometrics* 1972; **28**: 592–597.
24. **Hope VD, Hickman M, Tilling K.** Capturing crack cocaine use: estimating the prevalence of crack cocaine use in London using capture–recapture with covariates. *Addiction* 2005; **100**: 1701–1708.
25. **Sakamoto Y, Ishiguru M, Kitigawa G.** *Akaike Information Criterion Statistics*. Tokyo: KTK Scientific, 1986, pp. 1–24.
26. **Agresti A.** *Categorical data analysis*. New York: John Wiley and Sons, 1990, p. 251.
27. **Kool JL.** Preventing Legionnaires' disease [Thesis]. Amsterdam: University of Amsterdam, 2000.
28. **Bohte R, Van Furth R, Van Den Broek PJ.** Aetiology of community-acquired pneumonia: a prospective study among adults requiring admission to hospital. *Thorax* 1995; **50**: 543–547.
29. **Braun JJ, *et al.*** Community-acquired pneumonia: pathogens and course in patients admitted to a general hospital [in Dutch]. *Nederlands Tijdschrift voor Geneeskunde* 2004; **148**: 836–840.
30. **Van der Eerden MM, *et al.*** Comparison between pathogen directed antibiotic treatment and empirical broad spectrum antibiotic treatment in patients with community acquired pneumonia: a prospective randomised study. *Thorax* 2005; **60**: 672–678.
31. **Joseph CA.** Legionnaires' disease in Europe 2000–2002. *Epidemiology and Infection* 2004; **132**: 417–424.
32. **Marston BJ, Lipman HB, Breiman RF.** Surveillance for Legionnaires' disease. Risk factors for morbidity and mortality. *Archives of Internal Medicine* 1994; **154**: 2417–2422.
33. **Marston BJ, *et al.*** Incidence of community-acquired pneumonia requiring hospitalization – results of a population-based active surveillance study in Ohio. *Archives of Internal Medicine* 1997; **157**: 1709–1718.
34. **Desenclos JC, Hubert B.** Limitations to the universal use of capture–recapture methods. *International Journal of Epidemiology* 1994; **23**: 1322–1323.
35. **Cormack RM.** Problems with using capture–recapture in epidemiology: an example of a measles epidemic. *Journal of Clinical Epidemiology* 1999; **52**: 909–914.
36. **Papoz L, Balkau B, Lellouch J.** Case counting in epidemiology: limitations of methods based on multiple data sources. *International Journal of Epidemiology* 1999; **25**: 474–478.
37. **Hook EB, Regal RR.** Accuracy of alternative approaches to capture–recapture estimates of disease frequency: internal validity analysis of data from five sources. *American Journal of Epidemiology* 2000; **152**: 771–779.
38. **Jarvis SN, *et al.*** Children are not goldfish-mark-recapture techniques and their application to injury data. *Injury Prevention* 2000; **6**: 46–50.
39. **Tilling K.** Capture–recapture methods-useful or misleading? *International Journal of Epidemiology* 2001; **30**: 12–14.
40. **Borgdorff MW, Glynn JR, Vynnycky E.** Using capture–recapture methods to study recent transmission of tuberculosis. *International Journal of Epidemiology* 2004; **33**: 905–906.
41. **Tocque K, *et al.*** Capture recapture as a method of determining the completeness of tuberculosis notifications.

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press

Communicable Diseases and Public Health 2001; **4**: 141–143.

42. **Baussano I, *et al*.** Undetected burden of tuberculosis in a low-prevalence area. *International Journal of Tuberculosis and Lung Disease* 2006; **10**: 415–421.

43. **Van Hest NA, *et al*.** Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture–recapture analysis? *Epidemiology and Infection* 2006. Published online: 7 December 2006. doi:10.1017/S0950268806007540.

44. **De Greeff SC, *et al*.** Underreporting of meningococcal disease incidence in the Netherlands: Results from a capture–recapture analysis based on three registration sources with correction for false-positive diagnoses. *European Journal of Epidemiology* 2006; **21**: 315–321.

45. **Regal RR, Hook EB.** Validity of methods for model selection, weighing for model uncertainty and small sample adjustments in capture–recapture estimation. *American Journal of Epidemiology* 1997; **145**: 1138–1144.

46. **Cormack RM, Chang YF, Smith GS.** Estimating deaths from industrial injury by capture–recapture: a cautionary tale. *International Journal of Epidemiology* 2000; **29**: 1053–1059.

https://doi.org/10.1017/S0950268807008977 Published online by Cambridge University Press