Epidemiol. Infect. (2006), **134**, 579–584. © 2005 Cambridge University Press
doi:10.1017/S0950268805005315 Printed in the United Kingdom

Estimation of minimum infection rates with *Legionella pneumophila* in an exposed population

H. C. BOSHUIZEN[1]*, N. J. D. NAGELKERKE[1], J. W. DEN BOER[2], H. DE MELKER[1], J. F. P. SCHELLEKENS[1], M. F. PEETERS[3], H. VAN VLIET[1] AND M. A. E. CONYN-VAN SPAENDONCK[1]

[1] *National Institute for Public Health and the Environment, Bilthoven, The Netherlands*
[2] *Municipal Health Service Kennemerland, Haarlem, The Netherlands*
[3] *Regional Laboratory of Public Health Tilburg, Tilburg, The Netherlands*

(*Accepted 5 August 2005, first published online 20 October 2005*)

SUMMARY

The distribution of antibody levels to *Legionella* (*L.*) *pneumophila* (serotypes 1–7) was compared between subjects who worked near the source of a large outbreak of Legionnaires' disease ($n=668$) and a population sample of comparable age ($n=480$). In a previous analysis of these data, it was estimated that 80% of those working near the source were infected with *L. pneumophila*. However, the estimation procedure implicitly assumes that the probability of infection does not depend on the antibody level of a person before exposure. This is questionable, as antibodies could protect against infection. We have now estimated the minimum value consistent with the data on the number of infected persons. We observed that a minimum of 40% [95% confidence interval (CI) 32–48] of those working near the source and 13% (95% CI 8–18) of those working further away were infected with *L. pneumophila*. Implications of these findings for design options in future research are discussed.

INTRODUCTION

In 1999 a whirlpool on display at a trade fair caused a large outbreak of Legionnaires' disease (LD) in The Netherlands [1]. In the wake of this outbreak, blood samples were collected from exhibitors at this fair who did not develop LD. This investigation showed that quantitative titres of IgM and IgG antibodies against *Legionella* (*L.*) *pneumophila* were higher in highly exposed exhibitors working near the whirlpool than in others working elsewhere. Moreover, the titres in both those near the source, and in those working further away, were statistically significantly higher than those in the general population [2]. These findings agree with other outbreak studies that have shown increased antibody levels in those exposed to *L. pneumophila*, but not developing LD [3–8].

However, in individual persons, although titre levels were often above the population average, the levels reached usually were not high enough to be deemed seropositive, indicating that the overlap in titre distribution between those infected and not infected is large. In a supplemental analysis, we used mixture methods to disentangle these distributions [9]. We define 'infected' here as an increase in antibody level due to exposure. In this approach one regards the population of exposed subjects as a mixture of infected and uninfected individuals. The probability $f(x)$ of having antibody level x in the exposed population is a mixture of the distribution of the antibody levels in uninfected $[g_U(x)]$ and infected $[g_I(x)]$ individuals:

$$f(x)=(1-\lambda)g_U(x)+\lambda g_I(x), \quad (1)$$

* Author for correspondence: Dr H. C. Boshuizen, National Institute for Public Health and the Environment, IMA, PO Box 1, NL-3720 BA Bilthoven, The Netherlands.
(Email: Hendriek.Boshuizen@rivm.nl)

where λ is the fraction infected. Assuming that the distribution of antibody levels in the uninfected is equal to that of the general population, λ can be estimated. According to this approach 80% of those working near the source, and 25% of those working in the next nearest place were infected with *L. pneumophila*, much higher than the attack rates of clinical LD in these groups (2·7 and 0·4%). However, the estimation procedure implicitly assumes that the probability of getting infected given exposure is the same for everyone, i.e. does not depend on the antibody level of a person before exposure. This assumption can be questioned, as the presence of antibodies could protect against infection, and thus the probability of infection might be higher in those who have lower antibody levels before exposure. Such a selective infection of those with low antibody rates would lead to selective depletion of the group of uninfected persons, causing the post-exposure g_U to be shifted to the right, i.e. to high titres, compared to the pre-exposure g_U. Ignoring this shift would lead to overestimation of infection rates. However, without further knowledge of the nature of the dependency of the probability of infection on antibody levels, no value for λ can be estimated. However, we will show in this paper that it is, nevertheless, possible to estimate a minimum value of λ that is consistent with the data, and we will calculate this minimum value on this same dataset as mentioned above.

METHODS

Material and methods

Calculation of minimum prevalence of infection

In Figure 1 the left curve represents the distribution of antibody titres before exposure [$g_U(x)$]. After exposure, the distribution is shifted to the right [$g_I(x)$, the right curve]. This can only happen when those in the area between both curves change their titre value. Those in area A have to move out of it, as this area is no longer there after exposure, while area C does not exist before exposure, therefore persons have to move into this area C. Thus, the number of persons that do not move (change their titre value) can never be larger than those in area B. (More can move, however, as those in area B could move within B, or some could move from A to B, taking the place of those who move from B to C.) Conversely, the minimum number of persons that have to have moved (i.e. those that have become infected) is equal to area A = area C. The surface area of A (or C) can be easily determined if $g_I(x)$ and $g_U(x)$ are known. In practice, $g_I(x)$ and $g_U(x)$ have to be estimated from data. We used two procedures to estimate A from the data. First, we used a parametric method by fitting lognormal curves to $g_I(x)$ and $g_U(x)$, and then calculated the non-overlapping area between those two curves (A = C). Second, we used a non-parametric approach. This means that one has to calculate the area between the curves up to the point where the curves cross (Fig. 1). If one calculates the cumulative difference between the curves for different x values, i.e. the area between the curves up to titre value x, then the cumulative difference will increase unto the x where both curves cross, and afterwards it will decrease (as the newly added difference between the curves is negative). Therefore, the maximum of the cumulative difference curve is equal to area A, the area we want to estimate. However, with real data, the curves might oscillate, and cross many times, and the maximum cumulative difference might be partly due to random fluctuations. We therefore smoothed the cumulative difference curve first (using LOESS [10]), before taking its maximum. We used bootstrapping to computed 95% confidence intervals.

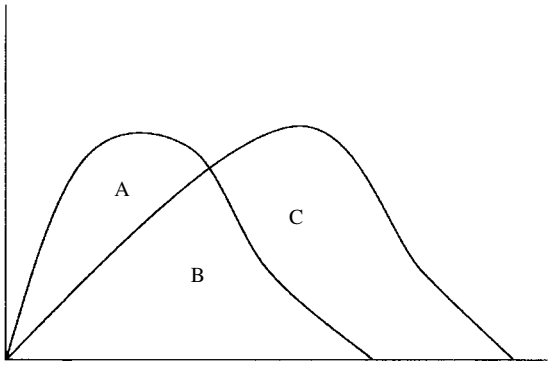

Fig. 1. Hypothetical antibody level distribution before exposure (left) and after exposure (right).

Subjects

A survey was conducted of exhibitors working on the 1999 West Friesian Flora stand, where a whirlpool spa on display caused a large outbreak of LD. The design of this study is described elsewhere [2]. In short, around 1 month after the Flora display, exhibition workers were approached with a questionnaire about their whereabouts during the exhibition and symptoms experienced afterwards and a request for blood samples for the detection of IgM and IgG antibodies against *L. pneumophila*. Participation rate

https://doi.org/10.1017/S0950268805005315 Published online by Cambridge University Press

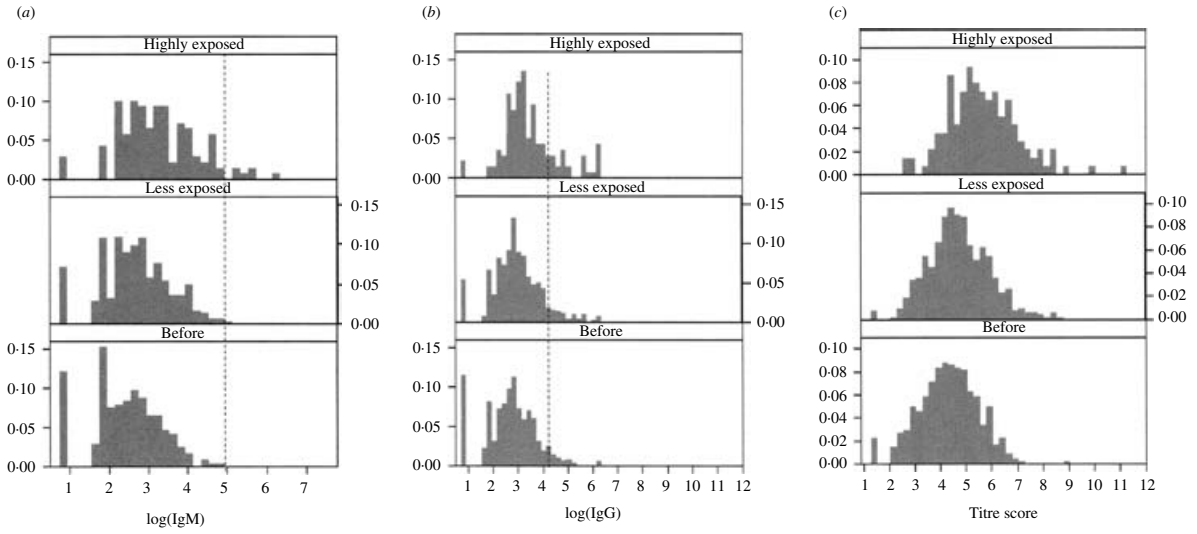

Fig. 2. Histograms of IgG (*a*), IgM (*b*) and titre score (combination of IgG and IgM) (*c*) for highly exposed, less exposed and general population. The vertical line gives the manufacturer's cut-off value for seropositivity.

was 56%. In this paper we use the data of 140 highly exposed exhibitors working in the exhibition hall where the whirlpool was located during the period when visitors of the exhibition contracted LD, and 528 less exposed workers working predominantly elsewhere at the exhibition during the same period.

As reference values for the non-exposed population, 480 blood samples were drawn from a bank of 8359 sera that was established in a cross-sectional population-based nationwide sero-surveillance study carried out in The Netherlands during 1995–1996. The design of this sero-surveillance study is described in detail elsewhere [11]. In short, in each of five geographical regions, with approximately equal numbers of inhabitants, eight municipalities were sampled proportionally to their size. Within each municipality, an age-stratified sample (0, 1–4, 5–9, …, 75–79 years) of 380 persons was drawn. Eligible individuals were asked to complete a questionnaire and to give a blood sample at a special clinic. The blood samples were stored in a refrigerator. The sera were harvested the next day and divided into portions that were stored at −86 °C. From this serum bank we randomly took for each municipality ($n=40$) one sample from those aged 15–24 years, 10 from those aged 25–64 years, and one from those aged 65–79 years, yielding a total sample ($n=480$) with an age structure similar to that of the exposed exhibitors.

Laboratory methods

IgM and IgG antibodies against *L. pneumophila* (serogroups 1–7) were determined by indirect ELISA with a commercially available assay (Serion classic ELISA, manufactured by Virion-Serion, Würzburg, Germany). According to the manufacturer, the inter-serial coefficient of variation is maximally 16%, while the intra-serial coefficient of variation is maximally 10% [12]. The following cut-off points are given by the manufacturer [13]:

IgM > 140 U/ml: positive; IgM 120–140 U/ml: borderline

IgG > 70 U/ml: positive; IgG 50–70 U/ml: borderline.

All titre values were log-transformed before analysis to achieve an approximately normal distribution. IgM and IgG values were combine into a titre score using the combination that best discriminated between the highly exposed and the serum bank sample:

$$\text{Titre score} = \log(\text{IgM}) + 0{\cdot}69\ \log(\text{IgG}).$$

RESULTS

Antibody levels in exposed persons

Initially, the distributions of antibody levels are shifted to the right with increasing exposure (Fig. 2). Twenty-one per cent (95% CI 8–19) of the highly exposed had either IgG or IgM titre values above the 99th percentile of the serum bank population. For the less exposed this was 4·9% (3·2–6·9%).

Figure 3 shows the cumulative difference between the titre score curves for the highly exposed and the general population sample. For each titre value are plotted the proportion of subjects in the highly

https://doi.org/10.1017/S0950268805005315 Published online by Cambridge University Press

Table. *Estimates of the minimum percentage of subjects infected in the group of highly exposed and less exposed exhibitors. The estimation method allows the infection rate to depend on the titre levels before exposure*

Estimation method	Highly exposed		Less exposed	
	Log normal	Non-parametric	Log normal	Non-parametric
IgM	35 (28–42)	31 (24–39)	11 (6–16)	10 (6–15)
IgG	29 (22–36)	28 (21–36)	10 (6–15)	9 (5–14)
Titre score	42 (34–48)	40 (32–48)	14 (9–19)	13 (8–18)

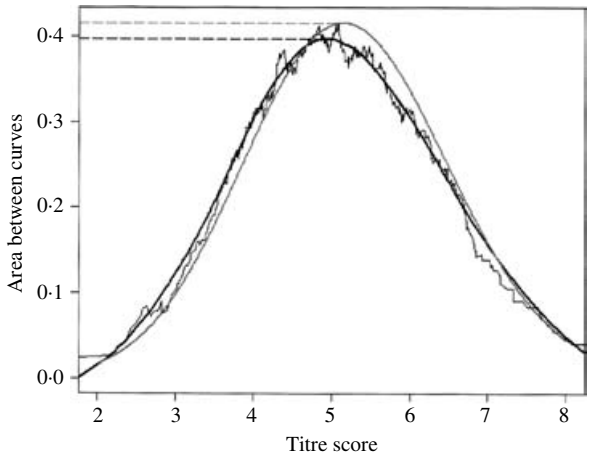

Fig. 3. Cumulative difference of the proportion with a certain titre score (combination of IgG and IgM) in the highly exposed and the general population. Wavy line: crude data; black line: the smooth version thereof; grey line: calculated by first fitting log-normal curves to Figure 2*c*. The dashed lines indicate the top of these curves.

exposed with a titre value at or below this value minus the proportion of subjects in the general population with a titre at or below this value. The tops of these curves (representing the highest cumulative difference) show the minimum prevalence of infection. The minimum infection rates resulting from these fittings (Table) show that at least a third of the highly exposed, and 10% of those less exposed are infected. The parametric method seems to overestimate the minimum prevalence slightly (Fig. 3). If information on IgG and IgM is used simultaneously (in the titre score), the percentages are 40 and 13%.

We used these minimum percentages to estimate the corresponding distribution of titre values in those infected (Fig. 4). These distributions represent the 'maximum' separation of the distributions, in the sense that the distribution in those infected is maximally different from the distribution in the uninfected. If the prevalence of infection would be higher than the minimum prevalence, more persons will have moved with (on average) smaller increases in titre, therefore, the distribution of titres in infected persons would have been closer to the pre-exposure distribution. The curves fitted for the infected from the highly exposed population, and those infected in the less exposed population were different ($P=0{\cdot}07$), suggesting that either the induced titre height might be dose-dependent, or that the prevalence of infection might exceed the minimum in the higher exposed population more than in the less exposed.

DISCUSSION

Assuming that the titre distribution in exhibitors at the outbreak site before being exposed was similar to that of the general population, our data show that 40% (but probably more) of the highest exposed and 13% of the other exhibitors had increased titre values due to the exposure. This is considerably higher than the percentage that would be considered serologically positive based on 99th percentile values of the general population (21 and 5% respectively). Previous calculations using the assumption that the rate of infection does not depend on pre-exposure antibody levels, yielded much higher rates of infection (80 and 25% respectively). Excluding the possibility that those with higher pre-exposure antibody levels have a higher probability of infection, the true infection rate will lie between the present and the previous calculations. Both absolute protection by higher pre-exposure antibody levels (necessary to yield the minimum prevalence of infection as calculated here), and no protection at all (assumed in the previous calculations) are not likely. However, even using our conservative approach, subclinical infection is more common in those exposed to *L. pneumophila* than would be thought by applying clinical criteria for seropositivity. However, our previous study showed that there was almost no correlation between symptoms and titre levels [2], therefore, most of these infections were probably asymptomatic.

https://doi.org/10.1017/S0950268805005315 Published online by Cambridge University Press

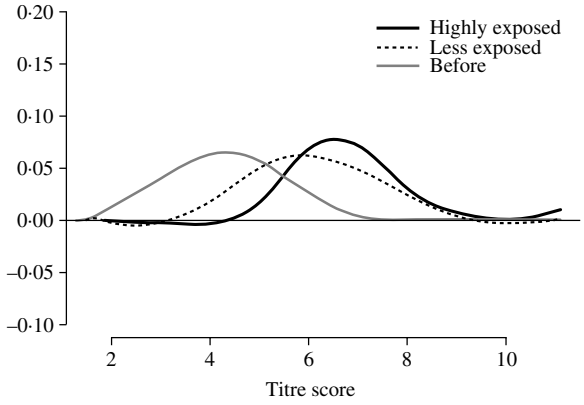

Fig. 4. Estimated distribution of titre score in highly exposed, less exposed and reference subjects when assuming the minimum fractions of infection as estimated (non-parametric) in the Table. The curves are estimated by taking the distribution of titres of those in area C of Figure 1 (non-parametric method) and smoothing this distribution.

We assumed that the titre distribution in exhibitors at the outbreak site before being exposed was similar to that of the general population. However, this might not be the case as exhibitors could have been exposed more frequently to *L. pneumophila* in their past, e.g. because they have spent more time on large-scale exhibitions or in hotels than the general population. However, the exhibitors were mostly working for firms or were volunteers for organizations residing in the region where the fair was held. Therefore, we believe that they are not especially more likely to travel and stay in hotels more often than the general population. Moreover, we surmise from the local character of the firms and organizations presenting themselves at the fair that most exhibitors were not working on exhibitions on a daily basis. Although our method in theory could be extended to allow adjustment for confounding factors, i.e. age, we did not explore this as the age distribution of the reference sample by design was similar to that of the exhibitors.

Moreover, if we used the titre distribution of the group of less exposed exhibitors as an indicator of distribution before exposure instead of that of the general population, the estimate of the minimum percentage of infected subjects remains fairly high. Based on the titre score the minimum percentage infected becomes 30% (95% CI 22–39) [this was 40% (95% CI 32–48) when using the general population], which still supports the conclusions drawn above.

Although log(IgG) and log(IgM) both increased with exposure, as did the correlation between them, the correlation between them nevertheless remained moderate (0·20 in the entire dataset), indicating that infection can manifest itself through different serological profiles. In order to make use of all the information in the data, we used a titre score, using weights that optimally discriminated between the highly exposed and the serum bank population. When estimating the probability of infection in the highly exposed population, using the serum bank distribution as reference, this might cause some overestimation as it is based on the same data that have been used to determine the optimal combination of IgG and IgM. However, this is not the case when we estimate infection probabilities for the less exposed, or for the estimates on the highly exposed using the less exposed as reference. The latter estimates yield a similar picture, as do the estimates using IgM and IgG separately and, therefore, we do not believe removal of this bias would change our conclusions.

Our findings suggest that it should be possible to investigate exposure to *L. pneumophila* in the population by studying titres in sero-surveys, thereby identifying high-risk groups that might be targeted for prevention. This option is available because long-term follow-up shows that not only IgG but also IgM antibodies against *L. pneumophila* persist in time, both in patients [14–16] and in those exposed to *L. pneumophila* but not having LD [17]; increased antibody levels are still present after several years. On the other hand, however, the presence of antibodies is determined not only by exposure to *L. pneumophila*, but also by host factors influencing the probability of developing antibodies when exposed, and also the amount of antibodies developing. Such a design, therefore, would be most suitable for studying environmental factors which can be believed not to be confounded by host factors (e.g. use of non-chlorinated drinking water).

Another lesson we learn is that the distributions of antibody levels in infected and (pre-exposure) uninfected populations largely overlap (Fig. 4). It is, therefore, possible to estimate the prevalence of exposure on a group level, but on an individual level it is not possible to indicate who is infected and who is not, especially as the maximally separated curves (Fig. 4) represent the best-case scenario. However, Figure 4 is based on persons who were exposed to *L. pneumophila* but who did not develop clinical disease. It is probable that the distribution of antibody titres in those with clinical disease is different, and better separated from the non-infected. However, any approach based on choosing antibody-level cut-off

points to separate infected and uninfected is destined to lead to a large amount of misclassification (Fig. 2). An approach directly using the original information by treating antibody levels as a continuous variable in the analysis is to be preferred. In such an approach one can conceptualize the (combination of) antibody levels as a proxy for the probability of being infected.

The fact that dichotomizing variables in cases like this implies wasting information is well known in epidemiology. Nevertheless, in serological studies this basic aspect of study design is often ignored. For example, in nine previous outbreak studies, seven reported only percentages of seropositives [3–5, 7, 18–20] and only two studies present the geometric mean titre values for the entire population [6, 8]. In clinical practice, there is a natural need for dichotomizing, as the decision to treat or not to treat is dichotomous. However, extending this to epidemiological research will unnecessarily decrease the power of studies. In outbreaks, the number of subjects that can be included is often limited, and maximizing power is, therefore, important. Directly using titre values in the analysis is thus to be preferred above using prevalence of seropositives.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Den Boer JW, Yzerman EP, Schellekens J, et al.** A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. Emerg Infect Dis 2002; **8**: 37–43.
2. **Boshuizen HC, Neppelenbroek SE, van Vliet H, et al.** Subclinical Legionella infection in workers near the source of a large outbreak of legionnaires disease. J Infect Dis 2001; **184**: 515–518.
3. **Bell JC, Jorm LR, Williamson M, et al.** Legionellosis linked with a hotel car park – how many were infected? Epidemiol Infect 1996; **116**: 185–192.
4. **Fisher-Hoch SP, Bartlett CL, Tobin JO, et al.** Investigation and control of an outbreaks of legionnaires' disease in a district general hospital. Lancet 1981; **1**: 932–936.
5. **Haley CE, Cohen ML, Halter J, Meyer RD.** Nosocomial Legionnaires' disease: a continuing common-source epidemic at Wadsworth Medical Center. Ann Intern Med 1979; **90**: 583–586.
6. **Marrie TJ, George J, Macdonald S, Haase D.** Are health care workers at risk for infection during an outbreak of nosocomial Legionnaires' disease.? Am J Infect Control 1986; **14**: 209–213.
7. **O'Mahony MC, Stanwell-Smith RE, Tillett HE, et al.** The Stafford outbreak of Legionnaires' disease. Epidemiol Infect 1990; **104**: 361–380.
8. **Saravolatz L, Arking L, Wentworth B, Quinn E.** Prevalence of antibody to the Legionnaires' disease bacterium in hospital employees. Ann Intern Med 1979; **90**: 601–603.
9. **Nagelkerke NJ, Boshuizen HC, de Melker HE, Schellekens JF, Peeters MF, Conyn-van Spaendonck M.** Estimating the incidence of subclinical infections with Legionella Pneumonia using data augmentation: analysis of an outbreak in The Netherlands. Stat Med 2003; **22**: 3713–3724.
10. **Cleveland WS.** Robust locally weighted regression and smoothing scatterplots. J Am Statist Assoc 1979; **74**: 829–886.
11. **De Melker HE, Conyn-van Spaendonck MAE.** Immunosurveillance and the evaluation of national immunisation programmes: a population-based approach. Epidemiol Infect 1998; **121**: 637–643.
12. **Scientific information on SERION ELISA classic (serogroups 1 to 7) IgG and IgM.** Leusden, The Netherlands: Clindia Benelux B.V., 1999. Version 1.0.
13. **Serion ELISA classic.** Legionella pneumophila 1–7. IgG, IgM/quant. Instructions. Würzburg, Germany: Institut Virion/Serion GmbH, 1998.
14. **Darelid J, Lofgren S, Malmvall BE, Olinder-Nielsen MA, Briheim G, Hallander H.** Legionella pneumophila serogroup 1 antibody kinetics in patients with Legionnaires' disease: implications for serological diagnosis. Scand J Infect Dis 2003; **35**: 15–20.
15. **Harrison TG, Taylor AG.** The diagnosis of Legionnaires' disease by estimation of antibody levels. In: Harrison TG, Taylor AG, eds. A laboratory manual for Legionella. Chichester: John Wiley, 1988: 130–136.
16. **Kallings I, Nordström K.** The pattern of immunoglobulins with special reference to IgM in Legionnaires' disease patients during a 2 year follow-up period. Zbl Bakt Hyg, I Abt Orig A 1983; **255**: 27–32.
17. **Lattimer GL, Rhodes LV, Salventi JS, et al.** The Philadelphia epidemic of Legionnaires' disease: clinical, pulmonary, and serologic findings two years later. Ann Intern Med 1979; **90**: 522–526.
18. **Benkel DH, McClure EM, Woolard D, et al.** Outbreak of legionnaires' disease associated with a display whirlpool spa. Int J Epidemiology 2000; **29**: 1092–1098.
19. **Darelid J, Hallander H, Lofgren S, Malmvall BE, Olinder-Nielsen AM.** Community spread of Legionella pneumophila serogroup 1 in temporal relation to a nosocomial outbreak. Scand J Infect Dis 2001; **33**: 194–199.
20. **Sanchez JL, Polyak CS, Kolavic SA, Brokaw JK, Birkmire SE, Valcik JA.** Investigation of a cluster of Legionella pneumophila infections among staff at a federal research facility. Mil Med 2001; **166**: 753–758.