Multiple regression as a preventive tool for determining the risk of *Legionella* spp.

Enrique Gea-Izquierdo


enriquegea@telefonica.net

Received: 23-09-2011; Accepted: 26-12-2011

Abstract

**Objective.** To determine the interrelationship between health & hygiene conditions for prevention of legionellosis, the composition of materials used in water distribution systems, the water origin and *Legionella pneumophila* risk. **Material and methods.** Include a descriptive study and multiple regression analysis on a sample of golf course sprinkler irrigation systems (n=31) pertaining to hotels located on the Costa del Sol (Malaga, Spain). The study was carried out in 2009. **Results.** Presented a significant lineal relation, with all the independent variables contributing significantly (p<0.05) to the model’s fit. The relationship between water type and the risk of *Legionella*, as well as the material composition and the latter, is lineal and positive. In contrast, the relationship between health-hygiene conditions and *Legionella* risk is lineal and negative. **Conclusion.** The characterization of *Legionella pneumophila* concentration, as defined by the risk in water and through use of the predictive method, can contribute to the consideration of new influence variables in the development of the agent, resulting in improved control and prevention of the disease.

**Key words:** *Legionella pneumophila*, legionnaires’ disease, multiple regression, hotel golf courses, sprinklers.

Resumen

La técnica de regresión múltiple como herramienta preventiva para determinar el riesgo de *Legionella* sp. **Objetivo.** Determinar el riesgo de *Legionella pneumophila* en relación a las condiciones higiénico- sanitarias para la prevención de la legionelosis, la composición de los materiales conductores de agua y el origen de la misma. **Material y métodos.** Incluyen un estudio descriptivo y análisis de regresión múltiple realizado sobre una muestra de sistemas de riego por aspersión de campos de golf (n=31) correspondientes a hoteles ubicados en la Costa del Sol (Málaga, España). El estudio se realizó en el año 2009. **Resultados.** Mostraron una relación lineal significativa, contribuyendo todas las variables independientes significativamente (p<0.05) al ajuste del modelo. La relación entre el tipo de agua y el riesgo de *Legionella* y de la composición del material con esta última, es lineal y positiva. En cambio, es lineal y negativa para la relación entre las condiciones higiénico- sanitarias y el riesgo de *Legionella*. **Conclusión.** La caracterización de la concentración de *Legionella pneumophila* definida a través del riesgo de la misma en el agua y mediante el empleo del método predictivo, contribuye a la consideración de nuevas variables de influencia en el desarrollo del agente y a un mejor control y prevención de la enfermedad.

**Palabras clave:** *Legionella pneumophila*, legionelosis, regresión múltiple, hotel con campos de golf, aspersores.

Resumo

A técnica de regressão múltipla como uma ferramenta preventiva para determinar o risco de *Legionella* sp. **Objetivo.** Determinar o risco de *Legionella pneumophila* em relação às condições higiênicas e sanitárias para a prevenção da legionelose, a composição dos materiais conducentes de água e o seu origem. **Material e métodos.** Incluem um estudo descritivo e análise de regressão múltiplo realizado sobre uma amostra de sistemas de rega por aspersão de campos de golfe (n=31) correspondentes a hotéis localizados na Costa do Sol (Malaga, Espanha). O estudo foi realizado no ano 2009. **Resultados.** Mostraram uma relação lineal significativa, contribuindo todas as variáveis independentes significativamente (p<0.05) ao ajuste do modelo. A relação entre o tipo de água e o risco de *Legionella* e da composição do material com esta última, é lineal e positiva. Em contraste, é lineal e negativa para a relação entre as condições higiênicas- sanitárias e o risco de *Legionella*. **Conclusão.** A caracterização da concentração de *Legionella pneumophila* definida através do risco da mesma em água e através do método predictivo, contribui para a consideração de novas variáveis de influência no desenvolvimento do agente e a um melhor controle e prevenção da doença.

**Palavras-chave:** *Legionella pneumophila*, legionelose, regressão múltipla, hotel com campos de golfe, aspersores.
Introduction

The Legionellaceae family consists of one single genus, Legionella, which contains more than 50 species and 70 different serogroups. Legionella pneumophila stands out among all the species with its 16 serogroups (1) and, based on the incidence in Europe of serogroup 1 (2), it is the most relevant. In Spain, the Legionnaires’ disease incidence has decreased since 2002, when it reached its maximum with 1,461 declared cases (an incidence rate of 3.54 per 100,000 persons), although in 2005 a small rise was recorded with 1,292 cases. In the year 2006 the rate per 100,000 population was 2.92 with 1,278 declared cases. Lethality during the 1989-2005 period was 4.6%, and it was higher for nosocomial outbreaks (24.6%) than for community-acquired infections (1.3%) or for those affecting tourists (Spaniards- 3.6% and EWGLINET- 9.8%) (3).

Legionellosis is a disease that has spread throughout the world and its incidence is higher in developed countries. It occurs in the form of epidemic outbreaks or sporadic cases. Although the origin of most cases of travel-associated legionellosis remains unknown, such cases have been related to hotels. In fact, approximately 60% of all the legionellosis cases associated with hotels are sporadic (1), which might point to the existence of a merely causal relationship (4). However, from an epidemiological perspective, the relationship between hotels and clusters has been microbiologically defined in different outbreaks (5,6) and it is precisely its repercussion on the hotel sector that illustrates the enormous economic impact this disease can have when it appears in tourist areas. Normally it occurs in the form of sporadic cases, but more and more frequently it is being detected in groups of cases and outbreaks.

Legionella species can be encountered distributed in low concentrations in natural aquatic habitats where they can survive to variations in pH, temperature and dissolved oxygen. The bacteria can colonize different water installations and spread through risk exposure in aerosols. Both its ecology and development have been subject to broad study (7-11). Globally, Legionella has been isolated in the water distribution systems of numerous hotels and buildings (4,12). The colonization rate is variable and can oscillate between 33% in the United Kingdom and 66% in Spain, while in some countries (Germany, Italy, Austria and United Kingdom) average rates of 55% can be found (13). At the national level, Spain has enacted legislation to prevent legionellosis and, depending on the count of Legionella (UFC/liter) (14), it imposes specific procedures applicable to evaporative cooling units. In some hotels, concentrations of Legionella ranging from 10^1 or 10^3 UFC/ml to a maximum as high as 10^6 UFC/ml (15) have been identified in water.

The presence of favorable conditions in water installations affects the bacterial proliferation, often leading to critical concentrations and thus posing risks to public health (16,17). Of particular interest is the type of water that the pathogen can colonize, as well as the water physicochemical conditions and the composition of materials used to pipe it. Legionella pneumophila can grow in water originating from diverse sources, for example, from waste water treatment plants, recycled water or well water. Depending on the composition and treatment of water from these different sources, the bacteria will either be eliminated or its development prevented, so theoretically any source of water is susceptible to contamination (18). The existence of other microorganisms, accompanied by certain levels of specific physicochemical parameters (pH, iron, turbidity and conductivity), are factors that can favor the bacteria encryption and presence (13). While the bacteria can grow in almost all materials some, such as iron, present a higher risk (19,20). In fact, the bacteria can be identified in biofilms found inside water piping systems, forming part of microbial “aggregates,” possibly influencing even water potability.

The aim of this study is to analyze the relationship between health-hygiene conditions for prevention of legionellosis in hotel irrigation systems, the composition of materials used in water distribution systems, and water type with the bacteria development.
Material and methods

Mandatory compliance with national health protocols on preventive maintenance, analysis and water disinfection is essential to control the development of *Legionella pneumophila* in risk installations. In order to analyze the relationship between one dependent variable (risk of *Legionella pneumophila*) and three predictor variables (composition of the material used to pipe the water, type of water, and the installations’ health-hygiene conditions), in order to develop a linear equation for predictive purposes. Both the material’s composition and water type are dichotomous variables. For the first variable, code 0 defines polyvinyl chloride, polyethylene, polybutylene and copper; code 1 defines lead, iron and stainless steel. For the second variable, 0 corresponds to recycled water and water originating from waste water treatment facilities; code 1 corresponds to well water. This dichotomization considers the prior risk of the pathogen development for the variables indicated, according to some authors (9,18,19). Health-hygiene conditions are defined in terms of the percentage of compliance with national health regulations for preventing legionellosis (14,18). Risk of *Legionella pneumophila* will be defined according to its value in a gradient degree measurement. The analysis is associated with a set of diagnostic strategies that provide information on the suitability and stability of the process, taking into consideration assumptions of linearity, independence, normality, homoscedasticity and non-collinearity. The terms of the regression model are population values, which must be estimated. The minimum quadratic estimates are obtained in an attempt to minimize the sum of the squared differences between the observed values and the predicted values. Goodness of fit and the regression equation are determined through an analysis of standardized regression coefficients and tests of significance. When fitting the model, a multiple correlation coefficient is obtained, as well as corrected and non-corrected squared, the typical error of residuals, and ANOVA. The *F* statistic tests the null hypothesis that the population value *R* is zero, making it possible to decide whether or not a significant linear relationship exists between the dependent variable and the set of independent variables taken together. The residuals are calculated to gather information on the degree of the prognoses accuracy, equally obtaining the prognoses maximum, minimum, and average values as well as the unbiased standard deviation, the residuals, and the typified prognoses and typified residuals. A detailed analysis of residuals makes it possible to obtain information on the first four previously mentioned assumptions. To obtain information on the residuals’ degree of independence, the Durbin-Watson statistic is calculated by estimating the equivalence of variances through typified prognoses and typified residuals, and especially by the Levene’s statistic. Statistics calculations to diagnose the presence of collinearity (tolerance levels and their inverse [VIF]) play a guiding role, indicating the degree of collinearity without being absolutely determinant. Condition indexes contribute to the reliability of the diagnosis, indicating equally the proportion of variance for each partial regression coefficient that is explained by each dimension or factor. All of the cases contribute to obtaining the regression equation, but not all of them do so with the same strength. The points of influence are cases that notably affect the value of the regression equation. Measures are determined to obtain an expression of the degree to which each case distances itself from the others: Cook’s distance and the centered leverage value. SPSS software (Copyright SPSS Inc., 1989-2006. Windows. Version 15.0.1. 22 Nov. 2006) was used for data analysis.

Results

The analysis led to the $R^2$ value (0.585), because of this the three independent variables included in the study account for 58% of the variance in the dependent variable (risk of *Legionella*). Since the number of variables is small in relation to the number of cases, the corrected value of $R^2$ is very similar to its non-corrected value (0.627). The critical level value from ANOVA associated with the regression yields significance = 0 which, being less than 0.05, indicates a significant linear relationship. It can be asserted that the hyperplane defined by the regression equation offers a good fit to the cloud of points. Therefore, based on the partial regression coefficients (Table 1) a minimum quadratic regression equation can be constructed as follows:

Risk of *Legionella* = 0.893 + 0.137 type of water - 0.252 health hygiene conditions + 0.094 composition of the material [1]

Estimates of the non-standardized partial regression coefficients (*B*), standardized (Beta) and individual *t*-tests for significance are obtained; confidence intervals are calculated for the partial regression coefficients (which are few) as well as for the *t*-tests and their critical significance levels (table 1), using the latter two to contrast the null hypothesis that one regression coefficient equals zero in the population and assuming that very small levels indicate that the hypothesis should be rejected. A matrix with the
Determining risk of Legionella spp.

covariances, correlations between the partial regression coefficients, and the matrix of bivariate correlations between the set of variables included in the analysis is obtained (Table 2); the mean, unbiased standard deviation of all the variables included in the analysis and the number of valid cases are also obtained. In addition to determining partial and semipartial correlation coefficients, correlations with zero-order appear (correlation coefficients that are calculated without taking into account the presence of third variables) (Table 3). The typical error of residuals presents a small value (0.313). It is especially noteworthy that the mean of the residuals equals zero and that the Durbin-Watson statistic is 2.383. Associated critical significance levels (Table 4) are calculated in relation to the Levene’s statistic. After eliminating the effect of the remaining independent variables, a linear and positive relationship exists between type of water and risk of Legionella, as well as between composition of the material and risk of Legionella. In contrast, the relationship between health-hygiene conditions and risk of Legionella is linear and negative. Table 3 shows the tolerance levels that correspond to collinearity diagnostic. The fact that the VIF values are low demonstrates the stability of the estimations of regression coefficients, consistent with the supposition of non-collinearity; this assumption was confirmed by the good tolerance level presented (high tolerance values) and condition indexes < 15. The variance proportions (Table 5) show that each dimension indicates a large amount of variance for one single coefficient, a characteristic of non-collinearity conditions. Since there is no dimension or factor with a high condition index that could contribute to explain the large amount of the variance for coefficients of two or more variables, no collinearity is presented. In relation to Cook’s distance, cases display values of less than 1 so they have no considerable weight in estimating the regression coefficients; when combined with the fact that the leverage values are less than 0.2, this fact demonstrates that they are not problematic.

### Table 1. Model coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized coefficients</th>
<th>Standardized coefficients</th>
<th>95% Confidence interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Standard error</td>
<td>Beta</td>
</tr>
<tr>
<td>Constant</td>
<td>0.893</td>
<td>0.182</td>
<td>4.908</td>
</tr>
<tr>
<td>Type of water</td>
<td>0.137</td>
<td>0.061</td>
<td>0.266</td>
</tr>
<tr>
<td>Health-hygiene conditions</td>
<td>-0.252</td>
<td>0.044</td>
<td>-0.693</td>
</tr>
<tr>
<td>Material composition</td>
<td>0.094</td>
<td>0.025</td>
<td>0.444</td>
</tr>
</tbody>
</table>

* Dependent variable: Risk of Legionella pneumophila.

### Table 2. Covariance matrix

<table>
<thead>
<tr>
<th></th>
<th>Risk of Legionella pneumophila</th>
<th>Type of water</th>
<th>Health-hygiene conditions</th>
<th>Material composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation</td>
<td>Risk of Legionella pneumophila</td>
<td>1.000</td>
<td>0.223</td>
<td>-0.591</td>
</tr>
<tr>
<td>Type of water</td>
<td>0.223</td>
<td>1.000</td>
<td>0.114</td>
<td>0.081</td>
</tr>
<tr>
<td>Health-hygiene conditions</td>
<td>-0.591</td>
<td>0.114</td>
<td>1.000</td>
<td>0.159</td>
</tr>
<tr>
<td>Material composition</td>
<td>0.356</td>
<td>0.081</td>
<td>0.159</td>
<td>1.000</td>
</tr>
<tr>
<td>Significance (unilateral)</td>
<td>Risk of Legionella pneumophila</td>
<td>.</td>
<td>0.114</td>
<td>0.000</td>
</tr>
<tr>
<td>Type of water</td>
<td>0.114</td>
<td>.</td>
<td>0.270</td>
<td>0.332</td>
</tr>
<tr>
<td>Health-hygiene conditions</td>
<td>0.000</td>
<td>0.270</td>
<td>.</td>
<td>0.196</td>
</tr>
<tr>
<td>Material composition</td>
<td>0.025</td>
<td>0.332</td>
<td>0.196</td>
<td>.</td>
</tr>
</tbody>
</table>
To ensure compliance, national regulations have been drafted that establish monitoring and control measures applicable to installations involved in the transmission of legionellosis. In the Autonomous Region of Andalusia additional preventive measures have been specified for installations of sports grounds (21). A special reference is made to spray irrigation systems, in which the aerosolized water can never have direct contact with people, requiring watering to be done when few people are on the grounds, preferably at night.
The characterization of *Legionella pneumophila* concentration is established as a function of the biological agent risk in water, its dispersion, the population exposure to it, the frequency of exposure, a person’s immunological state, and other variables, including meteorology. It is possible, however, to control the bacterial development including a lower determined risk threshold in the water. Preventive maintenance protocols and water quality controls are determinant factors in limiting risk. The water with distinct qualities is especially interesting due to the microorganism development and growth, possibly playing a determining role in the bacterial colonization of the medium. The mix of additional water sources may lead to secondary contamination that could affect pre-existing conditions of control. In other cases, the water origin and its characteristics (physicochemical and microbiological) are factors that can favor bacterial growth, which is why water treatment at origin is sometimes effective. The water distribution or its use for human purposes can alter the water conditions and favor new means of activation. In fact, the use of certain piping materials is being considered as one effective way to combat the bacteria, but other alternative water treatment methods are also necessary to ensure adequate, healthy conditions. Thus, it is easily understood that ensuring water quality depends as much on its own intrinsic conditions as on the elements involved in its distribution. Under Spanish regulation measures are in place to control and monitor health-hygiene conditions in irrigation installations as potential sources of risk distribution. The establishment of optimal health-hygiene conditions is one of the main elements recommended to combat the bacterial development and dispersion, and although it represents the main instrument for control, it is important to highlight the existence of many other influence variables. A priori, the prediction of *Legionella* risk can be defined exclusively under terms related to the control and monitoring of installations, although the study of other independent variables could be of considerable interest. A key role could also be played by the type of water being transported as well as the composition of materials used to pipe it. The potential for independent variables to influence one another has been considered to find out whether a possible relationship exists, using the few existing contributions available in this field. The construction of an equation that could serve as a preventive tool is presented as an additional means to limit bacterial concentration. A multiple regression method is used to predict the response from the explicative variables indicated; in fact, we have created a model that selects the variables that could influence the bacterial concentration, eliminating those that do not contribute with additional information. Detecting interactions among independent variables that affect the response variable is particularly relevant in order to observe the possibly greater than expected result of the independent variables sum; the presence of confounding variables was taken into account. In selecting the number of independent variables, more than 20 observations were taken into consideration for each independent variable that, a priori, was thought to be of interest for the model, contributing to the conclusions obtained and avoiding Type II errors. The identification and study of possible anomalous observations have no influence on the results obtained and therefore present no consequence in the analysis; significance is obtained from the regression model. Goodness of fit as interpreted by $R^2$ and the reduced number of independent variables introduced in the model help reduce the response variable uncertainty (variability). The proximity between $R$ and corrected $R^2$ demonstrates the model’s effectiveness with regard to the variables under consideration as well as the interest of those variables. Uncertainty (variance) regarding the response variable values is reduced by using information corresponding to the independent variables and randomly choosing a case for which no information is available; and, thanks to the linear regression model, a prediction can be made in which uncertainty decreases by 62.7% compared to the original. The matrix of correlations shows linear correlations between dependent and independent variables, the most extreme being the Pearson correlation coefficient corresponding to health-hygiene conditions for risk of *Legionella*, thus demonstrating the existence of an inverse linear association between this pair of variables.

For the other two dependent variables, specific correlations show that the relationship is direct with approximate growth. The coefficient value indicates how much the response variable can be expected to increase or decrease, depending on which variable is acted upon. The variable for health-hygiene conditions has greater weight in the regression equation than the other variables because its standardized regression coefficient presents the greatest absolute value, followed in importance by composition of the material. Considering the critical significance level associated with each $t$-test, all three variables contribute significantly to the model’s fit; in other words, to explain what occurs with the dependent variable. Since the confidence intervals of the partial regression coefficients are not very wide, the estimates obtained are precise and stable, thus presenting non-collinearity. The values obtained from the covariance matrix indicate that the partial regression coefficients are dependent. Comparisons among the zero-order, partial and semipartial coefficients (table 3) provide evidence that the relationship identified between the dependent variable and the three independent variables is not spurious, since risk of *Legionella* is explained through three variables. Given the low value obtained for typical error of the estimate, optimal prognoses are considered, meaning that the line of regression shows a better fit with points from the dispersion diagram. Considering the value obtained from the Durbin-
Watson statistic, it could be assumed that the residuals are independent, in other words, there is no reason to believe that the supposition of independence has not been met. The Levene test is used to contrast the hypothesis that the groups defined by the independent variables proceed from populations with the same variance and, given the critical level value, the hypothesis of homogeneity is accepted.

The presence of *Legionella pneumophila* and the critical impact its development can have on public health is characterized by the bacterial risk in water. The use of a predictive method can facilitate the study of new influence variables for the agent development and contribute to improved control and prevention of the disease. Further studies on the influence of predictor variables considered here and others are needed, as well as on possible interactions among them. Standardized processes of treatment and control to prevent legionellosis, as well as knowledge on installations and water characteristics, can serve as a theoretical predictive method to measure risk, reinforcing pertinent analytical controls. This method can also contribute to the reduction of health investments and facilitate quicker, more efficient actions related to the control of the state of installations. The inclusion of predictive tools in inspection routines will facilitate risk identification and the detection of different critical states.

**Financial support**

This study was partially financed by the Occupational Health and Safety General Directorate (EST 060/04), Regional Ministry of Employment, Government of Andalusia, Spain.

**Conflict of interest**

The author declares no conflict of interest regarding this study.

**References**


14. Real Decreto 865, de 4 de julio, por el que se establecen los criterios higiénico- sanitarios para la prevención y control de la legionelosis. B.O.E. núm. 171 de 18 de julio de 2003.

15. Habicht W, Muller HE. Occurrence and parameters of frequency of *Legionella* in warm water systems of hospitals


21. Decreto 287, de 26 de noviembre, por el que se establecen medidas para el control y la vigilancia higiénico-sanitarias de instalaciones de riesgo en la transmisión de la legionelosis y se crea el Registro Oficial de Establecimientos y Servicios Biocidas de Andalucía. B.O.J.A. núm. 144 de 7 de diciembre de 2002.