Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2016; 2(1): 47-51.

Orignal Article

Statistical analysis of various risk factors of tuberculosis in district Mardan, Pakistan

Salahuddin Falahuddin[1]*, Lakhkar Khan[2], Muhammad Iqbal[3], Najma Salahuddin[4]

[1] Department of Family and Community Medicine, College of Medicine, University of Dammam, Dammam, Kingdom of Saudi Arabia
[2] Department of Statistics, Government College, Mardan, Pakistan
[3] Department of Statistics, University of Peshawar, Peshawar, Pakistan
[4] Department of Statistics, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan

ARTICLE INFO

Received 14.02.2015
Revised 20.06.2015
Accepted 22.12.2015
Published 15.03.2016

Available online at:
http://jbe.tums.ac.ir

Key words:
logistic regression;
backward elimination procedure;
Brown method;
Wald statistic and risk factors

ABSTRACT

In this study, an effort has been made to determine the most important risk factors of tuberculosis (TB) in district Mardan. A total of 645 cases were examined, and their personal and medical data were collected. For each case, the phenomenon of TB was studied in relation to different risk factors. Statistical techniques of logistic regression and backward elimination procedure were used to analyze the data and to determine a parsimonious model. For both male and female cases, the final selected logistic regression model contain the risk factors: sex, residence, household population, diet, medical care, and close contact with infectious patients as well as a joint effect of two factors and three factors, namely, medical care and marital status; and economic status, and medical care and marital status. Separate logistic regression was then fitted for each sex using the same procedure. For male cases, the final selected logistic regression model contains risk factors: residence, diet, and close contact with infectious patients as well as a combined effect of two factors, namely, economic status and diet, medical care and diet. For female cases, the final selected logistic regression model contains the risk factors: household population, economic status, diet, and close contact with infectious patients as well as a combined effect of two factors, namely, medical care and close contact with an infectious patient.

Introduction

The word tuberculosis (TB) is derived from the Latin word tubercular, which means small lump, referring to small scars in tissues of infected individuals. TB is a bacterial disease that usually affects the lungs but it may also affect other parts of the body such as kidneys, lymph nodes, bones, and brains. TB is caused by a rod-shaped bacterium known as the tubercle bacillus or mycobacterium tubercle.

TB is an ancient disease, which is highly contagious. According to the WHO report (1), it takes the lives of more than two million people, more than any other single infectious disease worldwide annually, mostly in the third world countries. About 40-50 years back, the disease was almost uncontrollable even in the most advanced countries. The modern drugs, however, are quite effective against the disease, and if administered properly, the infection can be controlled in almost 90% cases.

TB is a barometer of poverty and race. Being a poor man's disease, TB is prevalent in countries of Africa, Latin America, and Asia, where the people are poor and hygienic conditions are substandard (2).

Acquired immune deficiency syndrome (AIDS) and human immunodeficiency virus (HIV) has accentuated the situation, as AIDS and HIV destroy the immune system of the

* Corresponding Author: Salahuddin Falahuddin, Postal Address: Institute of Management and Information Sciences, CECOS University of IT & Emerging Sciences, Sector H/3, Phase II, Hayatabad, Peshawar, Pakistan. Email: salahuddin_90@yahoo.com

body, thereby making it easier for the infection to attack, and difficult for the medicine to control the infection. Poverty, overcrowding, and HIV infection are the significant factors for the resurgence of childhood TB (3).

The annual risk of developing active TB in HIV-positive person is 100 times greater than HIV negative person. Poverty and disease are usually linked with each other. The association of TB with poverty is more established; therefore, TB is commonly known as the disease of poverty with 95% of cases and 98% of deaths occurring in developing countries. Today TB is accepted as a global health disaster (4).

TB is one of the leading causes of deaths in the developing countries. Pakistan ranks at number 8th among the countries with the highest burden of TB. It carries 44% of TB burden among the Eastern Mediterranean countries. According to WHO estimate, 181 per 100,000 populations have newly diagnosed cases of TB. This being the average, certain developed areas with relative prosperous population, the percentage will be less while in the far-flung poor areas, the percentage will be very high because the healthcare network is not well adequate there. There is no possibility to eliminate the disease completely until the socio-economic factors that influence the spread of TB infection are remedied. It is found that the risk of non-adherence to TB treatment is highly associated with unemployment, low status, low annual income, and cost of traveling to TB treatment facility (5).

China has the world's second largest TB epidemic, but progress in TB control was slow during the 1990s. The detection of TB had stagnated about 30% of the estimated total of new cases, and multi-drug-resistant TB (MDR-TB) was a major problem. In 2003, the spread of severe acute respiratory syndrome brought to light substantial weaknesses in the country's public-health system, and TB epidemic was brought under control. In 2005, the detection of TB had increased to 80% of the estimated total new cases and achieved 2005 global TB control targets (6).

A cross-sectional prospective survey of patients with pulmonary TB was carried out in Georgia; to assess the prevalence and risk factors for drug-resistant TB. It is found that MDR-TB has emerged as a serious public health problem in Georgia and will greatly impact TB control strategies (7).

TB is a reasonable issue because it has affected a large number of population and still a threat due to its contagious characteristics. TB had been ignored due to its long and expensive treatment resulting in the spread of the disease in society without any hurdle. Unawareness and lack of health facilities have not yet succeeded to get rid of TB

Methods

A total of 645 cases were examined from various hospitals of district Mardan, and their personal and medical data were collected. For each case, the phenomenon of TB was studied in relation to different risk factors such as age, sex, residence, household population, economic status of the household, marital status, drug, diet, medical care, living in a refugee camp, and close contact with the patient. Logistic regression analysis was applied to the collected data for all the cases as well as for sex-wise data. Backward elimination procedure was then used to determine a parsimonious model for all data as well as for sex-wise data.

Logistic regression model

Logistic regression requires binary dependent variable that is a categorical variable with two categories. Like dummy variables, these are coded (0, 1) and indicate if a condition is or is not present of if an event did or did not occur. For example, logistic regression might consider mortality as a response variable, which would be coded 0 (alive) and 1 (dead). There are only two values of dependent variable (which we will call occurrence or non-occurrence). Logistic regressions find the relationship between the response variable and a function of probability of occurrence. This function is the logit function (hence the name logistic function) also called the log-odds function. It is the natural logarithm of the odds of occurrence.

Using the log-odds creates an equation similar to the linear regression equation. Thus, by using

the log-odds instead of *Y* on the left-hand side of the equation, the right-hand side is identical.

Linear regression	Logistic regression
$Y = A + \beta X$	Log-odds $= A + \beta X$

By using SPSS (version 19, SPSS Inc., Chicago, IL, USA), we can calculate the coefficients, which are interpreted similar to linear regression coefficients.

Logistic regression is highly effective in estimating the probability that an event will occur. For this reason, it has been applied to medical research, where it is used to estimate the likelihood of individuals recovering from a disease.

The logistic regression model is defined as:

$$\text{Logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$$

Where, β_0, β_1, β_2, β_3 ……, β_k are the parameters that can no longer be estimated by least squares, but are obtained using the maximum likelihood estimation method (8, 9).

Results

Logistic regression model for all cases

Table 1 contains estimates of selected model parameters for all cases, their standard errors, calculated P values, estimated odds ratio, and confidence interval of the odds ratio. Since calculated value of $P < 0.050$ for all the regression coefficients of the model parameters, this shows that all the coefficients are highly significant. These estimates show the relationship between the explanatory variables and the response variable, where the response variable is on the logit scale. The final selected logistic regression model for all cases is:

$$\text{Logit}(p) = -6.168 + 1.701\,S - 1.02\,R + 0.836\,P + 1.003\,Dt - 2.297\,Mc + 3.51\,Cp - 0.866\,(Ms*Mc) + 0.444\,(Ms*Es*Mc)$$

Logistic regression model for male cases

Table 2 contains estimates of selected model parameters for only male cases, their standard errors, calculated P values, estimated odds ratio, and confidence interval of the odds ratio. Since calculated values of $P < 0.050$ for all the regression coefficients of the model parameters, this shows that all the coefficients are highly significant. The logistic regression model for male cases is as follows:

$$\text{Logit}(p) = -5.249 - 1.799\,R + 2.092\,Dt + 3.445\,Cp + 0.652\,(Dt*Es) - 1.681(Dt * Mc)$$

Logistic regression model for female cases

Table 3 contains estimates of selected model parameters for only female cases, their standard errors, calculated P values, estimated odds ratio, and confidence interval of the odds ratio. Since calculated values of $P < 0.050$ for the regression coefficients except variable "E" of the model parameters, this shows that all the coefficients are highly significant except variable "E" which is insignificant. The logistic regression model for female cases is as follows:

$$\text{Logit}(p) = -11.158 + 1.379\,P + 0.884\,E + 1.335\,Dt + 5.996\,Cp - 1.420\,(Cp*Mc).$$

Table 1. Estimates of selected model parameters for all cases

Variables	β	SE	Wald	df	Significant	Exp(β)	95% CI for Exp (β)	
							Lower	Upper
Step 1								
S	1.701	0.305	31.106	1	< 0.001	5.479	3.014	9.960
R	-1.020	0.297	11.799	1	0.001	0.361	0.201	0.645
P	0.836	0.289	8.351	1	0.004	2.308	1.309	4.070
Dt	1.003	0.342	8.606	1	0.003	2.727	1.395	5.331
Mc	-2.297	0.400	32.930	1	< 0.001	0.101	0.046	0.220
Cp	3.510	0.318	121.448	1	< 0.001	33.437	17.912	62.417
(Ms*Mc)	-0.866	0.356	5.909	1	0.015	0.421	0.209	0.846
Ms*Es*Mc	0.444	0.162	7.464	1	0.006	1.558	1.134	2.142
Constant	-6.168	1.116	30.562	1	< 0.001	0.002	-	-

SE: Standard error, CI: Confidence interval

Table 2. Estimates of selected model parameters for male cases

Variables	β	SE	Wald	df	Significant	Exp(β)	95% CI for Exp (β)	
							Lower	Upper
Step 1								
R	-1.799	0.438	16.850	1	< 0.001	0.165	0.070	0.391
Dt	2.092	0.873	5.738	1	0.017	8.102	1.463	44.872
Cp	3.445	0.456	57.080	1	< 0.001	31.357	12.828	76.649
(Dt * E)	0.652	0.299	4.749	1	0.029	1.920	1.068	3.453
(Dt* Me)	-1.681	0.268	39.379	1	< 0.001	0.186	0.110	0.315
Constant	-5.249	1.150	20.827	1	< 0.001	0.005	-	-

SE: Standard error, CI: Confidence interval

Discussion

Several studies have been conducted to investigate the incidence of TB and to determine the possible risk factors for this diseases. The main objective of the study was to determine the most likely risk factors of TB and to model the incidence of TB in cases arriving at different hospitals in district Mardan. The response variable in this study was TB, a binary variable taking the value "1" for TB positive case and "0" for TB negative case. The variables chose initially as predictors were A, S, R, P, E, Dg, Dt, Mc, Ms, Rc, and Cp. Since all variables were categorical, therefore, logistic regression analysis was applied to the data for both cases, male and female. Backward elimination procedure was then used to determine a parsimonious model for the collected data. The backward elimination procedure resulted in a logistic regression model containing various risk factors, two and three factors interactions terms: S, R, P, Dt, Mc, Cp, (Mc*Ms), and (Es*Mc*Ms). This shows that sex, residence, household population, diet, medical care, and close contact with infectious patients were the important risk factors for TB. The model also indicates that there is a joint effect of two factors on TB, namely, medical care and marital status. Similarly, the joint effect of three factors, namely, economic status, medical care and marital status jointly causes disease of TB.

A separate logistic regression model was then fitted for each sex using the same backward elimination procedure. For male cases, we got the final model with risk factors, namely, R, Dt, Cp, (Dt*Es) and (Dt*Mc). Thus, residence, diet, and close contacts with infectious patients were the important risk factors for TB. The model also shows that there is a combined effect of two factors on TB, namely, diet and economic status, diet and medical care.

For the female cases, the final model selected with risk factors namely P, E, Dt, Cp, and (Mc*Cp). Here, the model shows that household population, economic statuses, diet, and close contact with an infectious patient were the important risk factors for TB. The model also indicates that there is a joint effect of medical care and close contact with an infectious patient on TB for the female cases. After fitting the model, some diagnostics, namely, the index plots of Cook's influence D-statistic, leverage values and standardized residuals were used for assessing the validity of the models and to detect some outliers and influential observations. However, there was neither outlier nor influential observation found in all fitted models.

Table 3. Estimates of the model parameters for female cases

Variables	β	SE	Wald	df	Significant	Exp(β)	95.0% CI for Exp(β)	
							Lower	Upper
Step 1								
P	1.379	0.462	8.931	1	0.003	3.972	1.608	9.814
E	0.884	0.527	2.818	1	0.093	2.421	0.862	6.801
Dt	1.335	0.541	6.094	1	0.014	3.798	1.317	10.959
Cp	5.996	0.788	57.966	1	< 0.001	401.798	85.834	1.881E3
(Cp*Mc)	-1.420	0.307	21.403	1	< 0.001	0.242	0.132	0.441
Constant	-11.158	1.580	49.884	1	< 0.001	0.000	-	-

SE: Standard error, CI: Confidence interval

Our main conclusion from this analysis is that sex, residence, household population, diet, medical care, and close contacts with infectious patients are the significant risk factors of TB.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. World Health Organization. Global tuberculosis control: surveillance, planning, fi nancing [Online]. [cited 2008]; Available from: URL: www.unaids.org/sites/default/files/media_asset/who2008globaltbreport_en_0.pdf
2. Anderson NB. Levels of analysis in health science. A framework for integrating sociobehavioral and biomedical research. Ann N Y Acad Sci 1998; 840: 563-76.
3. Walls T, Shingadia D. Global epidemiology of pediatric tuberculosis. Journal of Infection 2004; 48(1): 13-22.
4. Gandy M, Zumla A. The return of the white plague: global poverty and the "new" tuberculosis. London, UK: Verso; 2003.
5. Mishra P, Hansen EH, Sabroe S, Kafle KK. Socio-economic status and adherence to tuberculosis treatment: a case-control study in a district of Nepal. Int J Tuberc Lung Dis 2005; 9(10): 1134-9.
6. Wang L, Liu J, Chin DP. Progress in tuberculosis control and the evolving public-health system in China. Lancet 2007; 369(9562): 691-6.
7. Mdivani N, Zangaladze E, Volkova N, Kourbatova E, Jibuti T, Shubladze N, et al. High prevalence of multidrug-resistant tuberculosis in Georgia. Int J Infect Dis 2008; 12(6): 635-44.
8. Cox DR, Snell EJ. Analysis of binary data. 2nd ed. Boca Raton, FL: CRC Press; 1989.
9. Tabachnick BG. Using multivariate statistics: Sas Workbook. 3rd ed. New York, NY: Pearson College Division; 1997.