Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 30

5-1-2007

Multinomial Logistic Regression Model for the Inferential Risk Age Groups for Infection Caused by *Vibrio cholerae* in Kolkata, India

Krishnan Rajendran
National Institute of Cholera and Enteric Diseases, West Bengal, India, rajenk20@yahoo.com

Thandavarayan Ramamurthy
National Institute of Cholera and Enteric Diseases, West Bengal, India, rama1murthy@yahoo.com

Dipika Sur
National Institute of Cholera and Enteric Diseases, West Bengal, India

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 324-330

1538 – 9472/07/$95.00

Multinomial Logistic Regression Model for the Inferential Risk Age Groups for Infection Caused by *Vibrio cholerae* in Kolkata, India

Krishnan Rajendran Thandavarayan Ramamurthy Dipika Sur
National Institute of Cholera & Enteric Diseases
West Bengal, India

Multinomial Logistic Regression (MLR) modeling is an effective approach for categorical outcomes, as compared with discriminant function analysis and log-linear models for profiling individual category of dependent variable. To explore the yearly change of inferential age groups of acute diarrhoeal patients infected with *Vibrio cholerae* during 1996-2000 by MLR, systematic sampling data were generated from an active surveillance study. Among 1330 *V.cholerae* infected cases, the predominant age category was up to 5 years accounting for 478 (30.5%) cases. The independent variables *V.cholerae* O1 ($p<0.001$) and non-O1 and non-O139 ($p < 0.001$) were significantly associated with children under 5 years age group. *V.cholerae* O139 inferential age group was > 40 years. The infection mediated by *V.cholerae* O1 had significantly decreasing trend Exp(B) year wise from 1996 to 2000 ($p < 0.001$, $p < 0.001$, $p < 0.001$, $p < 0.001$ and $p < 0.001$, respectively). MLR model showed that up to 5 year's age children are more vulnerable to infection caused by *V.cholerae* O1.

Key words: MLR, *Vibrio cholerae*, exp(B), explanatory, dependent, categorical

Introduction

Study design and data sources

Cholera is an epidemic disease in developing countries which has been the focus of intensive research for many years. This water borne disease is typified by severe watery diarrhea, vomiting and dehydration of the different serogroups of *V. cholerae,* serogroups O1, O139 and non-O1, non-O139 colonize in the small intestine and produce enterotoxin responsible for watery diarrhea. Until 19th century, cholera was confined in the Indian sub-continent (Epstein, 1993; Islam et al., 1994) and from this region, it has spread to many parts of the world causing seven pandemics. (Codeco, 2001; Faruque et al., 1998; Banerjee & Hazra, 1974). In 1992, a newly described non-O1 serogroup of *V. cholerae* designated O139 Bengal, caused unusual cholera outbreaks in India (Ramamurthy et al., 1993). Total eradication of this organism is very unlikely because of its propensity and acquaintance in the coastal ecosystem (Sack et al., 2004). In cholera endemic regions, severe cholera affects one in every 10-50 individuals, the highest attack rates of disease being in children of two to four years age (Cash et al., 1974). About 5.5 millions cases of cholera occur annually in Asia and Africa, 8% severe enough to be hospitalized, and 20% of the severe cases resulting in deaths, totaling approximately 120000/year (Mahalanabis et al., 1992; Noah & Mahony, 1998).

Classification and prediction are the more common practices in applied medical research. Mathematical model is widely used for prediction of disease outcomes. Discriminate analysis is mainly used for classification and

The authors thank all the scientists and Research Fellows of National Institute of Cholera and Enteric Diseases, Kolkata for helping to generate the laboratory data, and Infectious Diseases Hospitals, Kolkata, for taking part in the active surveillance program. Comments or question about this article should be directed to the first author at: rajenk20@yahoo.com or rajenk20@hotmail.com, or Division of Epidemiology, National Institute of Cholera & Enteric Diseases, P-33, Scheme-XM, CIT.Road, Beliaghata, Kolkata 700 010, West Bengal, India.

logistic regression and the dependent variable is binary or strict with two category. In a few studies, the relative predictivity of these methods were employed as an outcome variable that had more than two groups with unequal sizes. These models have been investigated when reducing bias by promoting the efficiency of the parameter estimation when the dependent variable has more than two groups. In this study Multivariate Logistic regression model was employed to identify inferential age group at greatest risk for diarrhea.

Materials And Methods

During 1996 to 2000, systematic sampling was done from every 5th hospitalized diarrhea patients attending the Infectious Diseases Hospital, Kolkata, India in two randomly selected days of the week. Samples were collected in the form of stool or rectal swab and sent to the laboratory for the isolation of common enteric pathogens within 3 hrs. The enteric pathogens were isolated and identified by standard laboratory methods (World Health Organization, 1987; Garg et al., 2000).

Data Management

The pre-designed proforma describing case demographics, symptoms etc. were checked manually and sent to data management center. The data were entered into pre-designed format of the proforma in EPI-info (6.0 version) with inbuilt entry validation checking facilitated program, by two trained data entry professionals in two separate computers. Data were randomly checked and matched to derive consistency and validity. The edited data was exported to SPSS version 4.0, and the final analysis was done using the SPSS.10. In this study, the inferential age groups was explored for three different serogroups of *V.cholerae,* O1, O139 and non-O1, non-O139 among culture positive cases by MLR and also to know the year wise changing pattern by parametric estimation through Odds Ratio(OR) ((Exp(B)). The proposed objective of the study was to determine the likelihood of age to have infection by *V.cholerae* O1, O139 and non-O1, non-O139, serogroups.

The age groups were classified into 6 categories viz. up to 5 years , above 5-10 years, above 10-20 years, above 20-30 years, above 30-40 years and more than 40 years and were coded as 1-6, respectively. The relationship between the risk dependent variable and each of the three categorical explanatory variables in the serogroup are shown in Table 1. Infection by any serogroup of *V.cholerae* was classified in numbers as 1 for organism present and 2 for its absence.

To describe categorical dependent variables and one or more categorical or dichotomous or continuous explanatory variables, Logistic regression was found suitable if dependent is strict with two categories. The conceptualized objective in this study was to employ MLR which may be more efficient and reliable to obtain the probability estimation of concerned patient population. In addition, MLR explores estimation of the net effects of a set of explanatory variables on the dependent variable (Cabrera, 1994; Demaris, 1992; Menard, 2000).

Data Analysis

The MLR model involves categorical dependent variable (more than two) Y. e.g. six categories of age group and 3 explanatory (*V.cholerae* serogroups) variables x_1 x_2 and x_3 (x_1=O1, x_2=O139 and x_3=non-O1, non-O139).

Let P_1 = the probability of up to 5 years age group at risk (Y=1), P_2 = the probability of above 5-10 years age group at risk (Y=2) P_3 = the probability of above 10-20 years age group at risk (Y=3), P_4 = the probability of above 20-30 years age group at risk (Y=4), P_5 = the probability of above 30-40 years age group at risk (Y=5) and P_6 = the probability of more than 40 years age group at risk (Y=6). The modality of MLR relates to the log of odds (or logit) of Y to the explanatory variable x_1 in linear form as

$$Pi = A+Px_i$$
$$Probit(P_i) = intercept + R.\ co\text{-}eff(x_i) \quad (1)$$

The model explores

$$Prob(y=j) = \frac{e^{\sum \beta_{jk} x_k}}{1+\sum e^{\sum \beta_{jk} x_k}} \quad (2)$$

Table 1: Distribution of *V.cholerae* O1, O139 and non-O1 , non-O139 among different age groups of patients during 1996-2000.

Age groups (in years)	*V.cholerae* serogroup	1996		1997		1998		1999		2000	
		No	%	No	%	No	%	No	%	No	%
Upto 5	(n=301) O1	70	23.3	43	14.3	93	30.9	54	17.9	41	13.6
	(n=55) O139	14	25.4	16	29.1	7	12.7	15	27.3	3	5.5
	(n=109) Non-O1 , Non-O139	8	7.3	13	11.9	29	26.6	38	34.9	21	19.3
>5-10	(n=90) O1	18	20.0	12	13.3	34	37.8	15	16.7	11	12.2
	(n=27) O139	6	22.2	12	44.4	2	7.5	3	11.1	4	14.8
	(n=14) Non-O1, Non-O139	2	14.3	3	21.4	5	5.7	2	14.3	2	14.3
>10-20	(n=87) O1	19	21.8	6	6.9	29	33.3	21	24.2	12	13.8
	(n=52) O139	8	15.4	24	46.1	6	11.5	11	21.2	3	5.8
	(n=39) Non-O1, Non-O139	6	15.4	15	38.4	12	30.8	3	7.7	3	7.7
>20-30	(n=100) O1	20	20.0	12	12.0	44	44.0	13	13.0	11	11.0
	(n=87) O139	20	23.0	29	33.3	18	20.7	18	20.7	2	2.3
	(n=59) Non-O1, Non-O139	15	25.4	18	30.5	14	23.7	7	11.9	5	8.5
>30-40	(n=40) O1	8	20.0	9	22.5	12	30.0	8	20.0	3	7.5
	(n=51) O139	11	21.6	19	37.2	8	15.7	7	13.7	6	11.8
	(n=29) Non-O1, Non-O139	7	24.1	6	20.7	7	24.1	5	17.2	4	13.8
>40	(n=63) O1	12	19.1	6	9.5	24	38.1	14	22.2	7	11.1
	(n=84) O139	20	23.8	25	29.8	18	21.4	16	19.0	5	6.0
	(n=43) Non-O1, Non-O139	7	16.3	9	20.9	13	30.2	11	25.6	3	7.0

$$P_{ij} = \log \frac{P(\Delta_i)}{P(\Delta_6)} = \text{intercept} + \text{parameter}(V_k) \quad (3)$$

i,j,k>0 i = age1 to age5, j = 1996 to 2000, k = O1, O139 and non O1, non O139, V = *V. cholerae* and Δ_6 = age6.

The intercept (initial level) terms are simple logit for positive *V.cholerae* O1. The first intercept is the log of ratio of the probability of a positive in up to 5 years to the probability of a positive in > 40 years. Hence, co-efficient for positive cases reveal the relationship between the logits and *V.cholerae* O1. Because the co-efficient is positive and significantly different from 0 that *V.cholerae* O1 positives are more likely associated with upto5 years age group as compared to >40 years age group.

Result

During 1996-2000, a total of 1330 *V.cholerae* stool culture positive cases formed the test set in this analysis, of which 681(51.20 %), 356 (26.8 %) and 293 (22%) were positive for *V.cholerae* O1, O139 and non-O1, non-O139 respectively. The age was coded as 6 categories in which 465 (35.0%), 131 (9.8%), 178 (13.4%), 246 (18.5%) 120 (9.0%) and 190 (14.3%) were in the age groups upto 5 years, >5-10 years, >10-20 years, >20-30 years, >30-40 years, >40 years respectively. The analysis was made to explore inferential age group. The predominant infected age category was upto 5 years age group. Overall, explanatory variables *V.cholerae* O1 ($p<0.001$, OR=3.48, 95% confidence interval (CI): 2.44, 4.90) was highly significant with children under five years of age. *V.cholerae* O139 was detected for more than 40 years age group ($p<0.001$, OR=2.99, 95% CI: 1.29, 4.98) and under five years old children were associated with *V.cholerae* non-O1, non-O139 ($p<0.001$, OR=2.50, 95% CI: 1.38, 4.55). As per the above equation 3 the result shows

Predicted logit (Y1≤5 years) = .302 + 1.247 (*V. cholerae* O1) (4a)

Predicted logit (Y1≤5 years)= 1.335 + (-)1.550 (*V. cholerae* O139) (4b)

Predicted logit (Y1≤5 years)= .898 +(-).055 (*V. cholerae* non-O1, non-O139) (4c)

Similarly, a prediction can be made for other age groups.

Overall, the chi-square test of proportional odds assumption was significant (degrees of freedom(df) = 5: $p<0.001$), indicating that the model is fit. Table 2 depicts only the predominantly affected age group with respective *V. cholerae* serogroup. Data on non-significant age group were not shown to avoid multiple tables. The explanatory variables are compared individually with dependent variable age. According to MLR models, the log of the Odds of an up to 5 years age group shows risk of infection positively related to the serogroup *V. cholerae* O1, in all years (Table 2). It was also shown a decreasing slope (rate of change) during the consecutive years. The respective years of OR for *V. cholerae* O139 has increased in 1998 and declined, though there was no significant association in all years with its inferential age group >40 years. In the case of *V. cholerae* non-O1, non-O139, year wise significance was not detected but the more vulnerable age group was <5 years.

Conclusion

Generally, Logistic Regression analysis (LR) is a common statistical technique which could be used to predict the likelihood of a categorical or binary or dichotomous outcome variables. In epidemiological studies, the dependent variable is presence or absence of a disease. The LR model has been applied in social science (Janik & Kravitz, 1994). Most of the microbiological laboratory generated data are not being utilized with proper statistical techniques owing to lack of appropriate guidelines for application. This study exploited the usefulness of MLR as a tool in statistical modeling and detecting the inferential risk age groups for *V. cholerae* mediated infection for 1330 culture positive cases from 1996-2000.

Table 2. Multinomial Logistic Regression Models exploring significant risk age group of cholera infection during 1996-2000.

V.cholerae (serogroup)	Years	Age group category(in year)						5 years age group (reference group (>40))		
		5	>5-10	>10-20	>20-30	>30-40	>40	P-values	OR	95%CI
O1	1996	70	18	19	20	8	12	<0.001	5.21	2.30, 11.78
	1997	43	12	6	12	9	6	0.002	4.43	1.70, 11.51
	1998	93	34	29	44	12	24	0.001	2.96	1.50, 05.64
	1999	54	15	21	13	8	14	0.007	2.89	1.33, 06.25
	2000	41	11	12	11	3	7	0.019	4.26	1.27, 14.33
								>40 years age group (reference group (>5-10))		
O139	1996	14	6	8	20	11	20	0.027	3.52	1.15, 10.75
	1997	16	12	15	29	19	25	0.147	2.08	.77, 5.62
	1998	7	2	6	18	8	18	0.005	9.00	1.95, 41.50
	1999	15	3	11	18	7	16	0.067	3.63	.91, 14.39
	2000	3	4	3	2	6	5	0.540	1.62	.31, 7.67

The MLR requires the dependent variables to be non-metric, dichotomous, nominal and ordinal, satisfy the level of measurement and independent variable to be metric or dichotomous. The minimum number of cases per independent variable is 10 using a guideline provided by Homen and Lameshow (2000), in which the MLR predicts and provides a set of co-efficient for each of the two comparisons. The co-efficient for the reference group are all zeros, similar to the co-efficient of the reference group for a dummy-coded variable. Dependent variable will be defined as groups, where the equations can be used to compute the probability and predict the groups associated with the highest probability. The predicted group membership can then be compared to the actual group membership to obtain a measure of classification accuracy.

The emphasis is given on MLR utility because (a) application for categorical outcomes in multivariate techniques are very few, including Logistic Regression, discriminant function analysis and log-linear models, (b) the MLR does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables (Hosmer & Lemeshow , 2000; Peng & Nichols, 2003; Clayton & Hills, 1993). (c) MLR does not impose these requirements, it is preferable to use discriminant analysis when the data does not satisfy these assumptions. (d) a more useful measure to assess the utility of MLR is classification accuracy and (e) because the laboratory data generally exist either in

dichotomous or an ordinal form of variable that can be explored in the form of Odds Ratio. The LR and MLR are best methods for the above format of data structure.

This study explained how effectively the MLR models are useful in the epidemiology of cholera and overall model evaluations. The likelihood Ratio was examined to improve the MLR model over null models. An intercept is the only model that serves as a good baseline with no predictors. According to MLR model, the test yielded significance and was more effective than the null model.

In the tests of individual predictors, the Wald chi-square statistic was tested using individual B coefficients to inclined relationship with dependent variables. The goodness of fit statistics assess fitness of logistic model against actual classification i.e. six levels of age group category in the MLR model. The two measures were almost similar in overall estimation, which is similar to Ordinary Least Square (OLS) regression. No equivalents of this concept for MLR explains variance, and for this reason the Pseudo R-Square reported to be complementary to others, which has more useful evaluative indices such as tests of individual regression coefficients (Peng et al., 2001).

The advantage of inferential test of the goodness of fit was suggested by Begg and Gray (1984) for multinomial logistic models. In the validation of predicted probabilities, the MLR model predicts the logit of levels of degrees of inferential risk age group from independent variables. The logit is probability/1-probability, which can be transformed later to the probability scale according to the equation 2 (Rabins & Dickinson, 1985; Peterson & Harrell, 1990; Greenland, 1987; Savitz, 1992). The predicted probability of inferential risk age group is evaluated and compared with actual risk age to determine various levels of age groups.

Reference category was fixed based on the occurrence of positive cases in the age groups in which the cases were low. The main aim of the selection of reference category was to interrogate the age group favored by the pathogen. In *V.cholerae* O1, greater than 40 years age was selected as reference category owing to less incidence rate. The interesting trend of *V.cholerae* O139 was higher incidence rate in older age group and lower in >5-10 years age group that served as reference category to explore existing relation. The parameter estimation of all age groups with different serogroups of *V.cholerae* was newly conceptualized by comparing reference category of age group with positive cases of respective serogroups. The above equation gives the ratio of comparing categories with reference category in the intercept.

The MLR supported the statistical significance of the three independent variables in different age groups for five consecutive years (1996-2000). Importantly, *V.cholerae* O1 infection mostly occurs upto5 years age group, which is highly significant. Infection caused by *V.cholerae* O139 showed the significant risk age group was >40 years, which is a more interesting trend. *V.cholerae* non-O1, non-O139 was not significantly associated with any age group, but the highest risk age group was less than 5 years age. The effectiveness of MLR model was supported by multiple indices, including models for overall test of all explanatory variables and significance test of each explanatory variables. In the categorical outcomes, logistic regression is more flexible and less restrictive than discriminant function analysis and log-linear models (Wacholder, 1986; Peng et al., 2002).

Few studies describes the application of Multinomial logistic regression methods. In this finding, we found that MLR is an effective model for profiling greatest risk age groups due to infection caused by different serogroups of *V.cholerae*. Microbiologists and epidemiologists can employ this model for laboratory data.

Reference

Banerjee, B. & Hazra, B. (1974). *Geoecology of cholera in West Bengal: A study in medical geography*. Jayati Hazra Publishers, Calcutta.

Begg, C. B. & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regression. *Biometrika, 71*, 11-18.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An

applied perspective. *Higher Education: Handbook of Theory and Research, 10*, 225-256.

Cash, R. A., Music, S. I., Libonati, J. P., Snyder, M. J., Wenzel, R. P. & Hornick, R. B. (1974). Response of man to infection with *Vibrio cholerae.* I. Clinical, serologic and bacteriologic responses to a known inoculum. *The Journal of infectious diseases*, *129*, 45-52.

Clayton, D. & Hills, M. (1993). *Statistical models in epidemiology*. Oxford, England: Oxford University Press.

Codeco, C. T. (2001). Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. *BMC infectious diseases, 1*, 1.

Demaris, A. (1992). *Logit modeling: Practical application*. Newbury Park, CA: Sage.

Epstein, P. R. (1993). Algal blooms in the spread and persistence of cholera. *Biosystems, 31*, 209-221.

Faruque, S. M., Albert, M. J. & Mekalanos, J. J. (1998). Epidemiology, genetics and ecology of toxigenic *Vibrio cholerae*. *Microbiology and Molecular Biology Reviews*, *62*, 1301-1314.

Garg, P., Chakraborty, S., Basu, I., Datta, S., Rajendran, K., Bhattacharya, T., Yamasaki, S., Bhattacharya, S. K., Takeda, Y., Nair, G. B., & Ramamurthy, T. (2000). Expanding Multiple antibiotic resistance among clinical strains of *Vibrio cholerae* isolated from 1992-7 in Calcutta, India. *Epidemiology and infection, 124*, 393-399.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analysis. *American Journal of Epidemiology*, *125*, 761-768.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons Inc.

Islam, M. S., Miah, M. A., Hasan, M. K., Sack, R. B. & Albert, M. J. (1994). Detection of non-culturable *Vibrio cholerae* O1 associated with a cyanobacterium from an aquatic environment in Bangladesh. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *88*, 298-299.

Janik, J. & Kravitz, H. M. (1994). Linking work and domestic problems with police suicide. *Suicide & life-threatening behavior*, *24*, 267-74.

Mahalanabis, D., Molla, A. M., & Sack, D. A. (1992). Cholera management. In; D Barua, WB Greenough III, eds. Cholera. New York: *Plenum Medical Book Company*, 129-154.

Menard, S. (2000). Coefficient of determination for multiple logistic regression analysis. *The American Statistician*, *54*, 17-24.

Noah, N. & Mahony, M. O. (1998). *Communicable disease epidemiology and control*. John Wiley & Sons Ltd.

Peng, C. Y., Manz, B. D. & Keck, J. (2001). Modeling categorical variables by logistic regression. *American Journal of Health Behavior, 25*, 278-284.

Peng, C. Y., So, T. S., Stage, H. F. K. & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education*, *43*, 259-293.

Peng, C. J. & Nichols, R. N. (2003). Using multinomial logistic models to predict adolescent behavioral risk. *Journal of Modern Applied Statistical Methods*, *2*, 538-1554.

Peterson, B. & Harrell, F. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics, 39*, 205-217.

Rabins, P. K. & Dickinson, K. P. (1985). Child support and welfare dependence: A multinomial logit analysis. *Demography, 22*(3), 367-380.

Ramamurthy, T., Garg, S., Sharma, R., Bhattacharya, S. K., Nair, G. B., Shimada, T., Takeda, T., Karasawa, T., Kurazano, H., Pal, A., & Takeda, Y. (1993). Emergence of novel strain of *Vibrio cholerae* with epidemic potential in southern and eastern India. *Lancet,* 703-704.

Sack, D. A., Sack, R. B., Nair, G. B. & Siddique, A. K. (2004). Cholera. *Lancet*, 363, 223-233.

Savitz, D. A. (1992). Measurements, estimates, and inferences in reporting epidemiologic study results [editorial]. *American Journal of Epidemiology* ,*135*, 223-224.

Wacholder, S. (1986). Binamial regression in GLIM: estimation risk ratios and risk differences. *American Journal of Epidemiology,123*, 174-84.

World Health Organization. (1987). Manual for laboratory investigation of acute enteric infection, CDD/33.3. WHO, Geneva, Switzerland.