IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 16, Issue 1 Ser. IV (Jan – Feb 2020), PP 26-31
www.iosrjournals.org

BinaryLogistic Regression Analysis of the Determinants of Survival of Cholera Patients

U.M. Hassan[1], Yusuf Abbakar Mohammed[2]
[1,2](Department of mathematical sciences, University of Maiduguri, Maiduguri, Nigeria)

Abstract: *Cholera remains a global threat to public health and key indicator of lack of social development. It has been classified as remerging global threat. The disease is primarily linked to insufficient access to safe clean water supplies, crowded living condition, poor hygiene and sanitation. The objective of the study is to model the significant factors that have effect on status (survival or death) of cholera infected person in order to predict the survival probability. Recorded data of 513 patients were obtained from UNICEF Cholera Hospital for Internally Displaced Persons within Maiduguri, Borno State of Nigeria. Descriptive and binary logistic regression were used to analyze the data. The result of binary logistic regression revealed that vaccinating a patient before the infection and mild degree of dehydration are statistically significant predictors that increases chance of survival of cholera infected person. To ensure an improved chance of cholera patient's survival effort should be made to hydrate the infected person and Vaccine (killed oral 01 with whole-cell with Bsubunit) should be administer whenever there is outbreak.*

Keywords-*binary logistic, cholera, odds ratio, logit model,survival probability*

Date of Submission: 12-02-2020 Date of Acceptance: 28-02-2020

I. Introduction

Worldwide, about 1.4-4.3 million cases and 28,000-142,000 death per year are due to cholera infection. In Nigeria, cholera infection is endemic and outbreaks are common [1].

A global strategy on cholera control with a target to reduce cholera deaths by 90% was launched in 2017. Before 1817, cholera was confined to India's Bay of Bengal. However, primarily following trade and migration between India and Europe, by the 1830s, cholera had spread internationally. The global spread of cholera was the driving force behind the first International Sanitary Conference in Paris, in 1851. The global health significance of cholera is underscored by its inclusion as one of four priority diseases in the 1969 and 2005 International Health Regulations.

In 2016 a total of 6600 cholera cases, including 229 deaths (CFR 3.47%) were reported from 94 Local Government Areas in 20 states, Borno state inclusive. Borno State has been in the forefront of most recent cholera outbreak in Nigeria [2].

There is difference in survival rate between the age groups, sex, vaccination status and degree of dehydration, patients of age (<1yr). Being a Female, being vaccinated before the infection and patients with mild degree of dehydration have better chance of surviving. Being a Female significantly increase the survival time by approximately 25% compared to male. Being vaccinated significantly increase the survival time by 75% compared to not vaccinate. The patients with severe (A) dehydration have shorter survival time compared to patients with moderate (B) dehydration, but the patients with mild (C) dehydration have longer survival time compared to patients with moderate dehydration [3].

This study was carried out to investigate the factors that determines the survival of cholera infected person and to model these factors in order to predict survival probability. Recorded data of 513 patients was obtained from UNICEF Cholera Hospital for Internally Displaced Persons within Maiduguri, Borno State of Nigeria to carry out the analysis.

The work will enable us to know the significant factors that increases the chance of survival of cholera infected person.

1.1 Binary logistic regression analysis

Binary logistic regression is a type of regression which is used when the response variable is dichotomous and the predictor variable are of any type. In binary logistic regression, a single outcome variable $Y_i (i = 1,2, \dots, n)$ follows a Bernoulli probability distribution that takes on the value of 1 with probability P_i and 0 with probability$1 - P_i$.

1.1.1 Odds ratio

Logistic regression analysis utilizes odds and odds ratio. The odds of success are simply the ratio of probability of success P (the probability that a patient will survive or $Y = 1$) to the probability of failure $1 - P$ (the probability that a patient will not survive or or $Y = 0$). That is

$$odds = \frac{P}{1-P}$$

The odds are non-negative with value greater than one when a success is more likely than a failure. The odds ratio denoted by OR, and is defined as the ratio of the odds for $Y = 1$ to the ratio of $Y = 0$ and is given by

$$odds\ ratio = \frac{P_1/_{1-P_1}}{P_0/_{1-P_0}}$$

Odds ratio is an estimate of the risk of an exposed group relative to a control group or unexposed (reference group). Odds ratio less than 1 indicates negative relationship and odds ratio greater than 1 indicates positive relationship and odds ratio equal to 1 indicates no difference between exposed and control group.

II. Methodology

2.1 Logistic Regression Analysis Model

Logistic Regression Analysis (LRA) model is used to study the relationship between a categorical or qualitative response (outcome) variable and one or more explanatory variable (independent or predictor) variables. In case of one predictor variable X and one dichotomous outcome variableY, the logistic model predicts the logit of Y from X. The logit is a natural logarithm of odds ofY. Simply the logit model can be written as [4]:

$$\boldsymbol{ln\left(\frac{P}{1-P}\right) = \alpha + \beta X} \qquad (\mathbf{1})$$

Hence,

$$\boldsymbol{P(X) = E(Y/X) = (p(Y = response(outcome)\ of\ interest)/X = x) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}} \qquad (\mathbf{2})$$

Where $P(X)$ the probability of the outcome of interest, α is the Y intercept (constant) and β is the slope parameter. X is predictor that can be categorical or continuous variable and Y is always categorical (dichotomous). The simple logistic regression model can be extended to multiple logistic regression as follow:

$$\boldsymbol{揭n\left(\frac{P}{1-P}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k} \qquad (\mathbf{3})$$

Therefore,

$$\boldsymbol{P(X) = E(Y/X) = [p(Y = response(outcome)\ of\ interest)/X_1 = x_1, X_2 = x_2, \ldots, X_k x_k)]}$$
$$\boldsymbol{= \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}} \qquad (\mathbf{4})$$

Where $P(X)$ the probability of the outcome of interest, α is the Y intercept (constant) and β_s are the slope parameters. $X's$ are the set of predictors. α and β_s are estimated by the MLE method.

2.2 Likelihood ratio test

The likelihood ratio test, a general test to compare two models, a full model and a simpler (reduced) model. It test that the parameters in the full model are equal to zero. The test uses the likelihood function. The maximum likelihood estimates maximize this function.

Let L_r is the likelihood of the reduced model and L_f is the likelihood of the full model (saturated) model. The formula for likelihood ratio test statistic:

$$\boldsymbol{G^2 = \left[(-2logL_r) - (-2logL_f)\right]} \qquad (\mathbf{5})$$

It compares the maximized values of the fitted/reduced model $(-2logL_r)$ and of the full/saturated model$(-2logL_f)$. Using minus twice itslog it is necessary to obtain a quantity whose distribution is known (approximately a chi-square distribution with degree of freedom equals the difference between number of parameters in the saturated model and the nested/smaller model for large sample), and can therefore be used for hypothesis testing purposes [5].

2.3 Hosmer-Lemeshow test

A commonly used test of the overall fit of a model to the observed data is the Hosmer-Lemeshow test. Hosmer-Lemeshow goodness of fit test divides subjects into deciles based on the predicted probabilities and construct a "goodness of fit" statistic by comparing the observed and predicted number events in each group. The differences between the observed number and expected number (calculated by summing predicted probabilities based on the model) in each group are then assessed using chi-square test.
The Hosmer-Lemeshow goodness of fit statistic is given by:

$$\widehat{\boldsymbol{C}} = \sum_{\boldsymbol{j}}^{\boldsymbol{g}} \frac{\left(\boldsymbol{O_j} - \boldsymbol{E_j}\right)^{\boldsymbol{2}}}{\boldsymbol{V_j}} \qquad (\mathbf{6})$$

Where $E_j = nP_j$, $V_j = nP_j\left(1 - P_j\right)$, g is the number of group, O_j is the observed number of events in the j^{th} group, E_j is the expected number of events in the j^{th} group and V_j is a variance correction factor for the j^{th} group.
If the difference between the observed number of events and what is expected by the model is large, then the statistic $\hat{C}$ becomes large and there will be evidence against the null hypothesis that the model is adequate to fit the data. Assuming that the null hypothesis that the model fits well $\hat{C}$ has an approximate chi-square distribution with $g - 2$ degrees of freedom [5].

2.4 Wald test

The Wald test is used for testing the significance of individual parameters in the logistic regression. That is, the Wald test is used to test:

$$H_0{=}\beta_i \quad \text{against} \quad H_1 \neq \beta_i i = 1,2,\dots,k$$

The Wald test statistics for testing the above hypothesis is:

$$\boldsymbol{w} = \frac{\widehat{\boldsymbol{\beta_i}}}{\boldsymbol{SE}(\widehat{\boldsymbol{\beta_i}})} \qquad (\mathbf{7})$$

$i = 1,2,\dots,k$

Under the null hypothesis $w{\sim}N(0,1)$. When the computed value of $|w| \leq Z_{\left(1-\frac{\alpha}{2}\right)}$ we do not reject the null hypothesis, while if $|w| > Z_{\left(1-\frac{\alpha}{2}\right)}$ then the null hypothesis can be rejected at the given alpha level [6].

III. Data Analysis And Results

The descriptive results (in absolute figures and percentage) and the chi-square results based on bivariate analysis are given in "Table 1" below. The bivariate analysis was used to determine the association between status of the patients (response variable) after treatmentand the factors (predictors) that determined the survival of the patients.

Table1: Descriptive and bivariate analysis results

Predictors	Categories	Status of the patients after treatment		Total 513(100%)	p-value
		Survived(1) 414(80.7%)	Not survived(0) 99(19.3%)		
Age(in years)*	<1	57(13.8%)	7(7.1%)	64(12.5%)	0.000
	1-20	134(32.4%)	19(19.2%)	157(29.8%)	
	21-40	104(25.1%)	19(19.2%)	123(24.0%)	
	41-60	71(17.1%)	29(29.3%)	100(19.5%)	
	>60	48(11.6%)	25(25.3%)	73(14.2%)	
Sex*	Female	214(51.7%)	41(41.4%)	255(49.7%)	0.042
	Male	200(48.3%)	58(58.6%)	258(50.3%)	
Vaccination status*	Not vaccinated	96(23.2%)	86(86.9%)	182(35.5%)	0.000
	Vaccinated	318(64.5%)	13(13.1%)	331(64.5%)	
Degree of dehydration*	Severe	10(2.4%)	72(72.7%)	82(16.0%)	0.000
	Moderate	157(37.9%)	25(25.3%)	182(35.5%)	
	Mild	247(59.7%)	2(2.0%)	249(48.5%)	
Marital status*	Single	224(54.1%)	30(11.8%)	254(49.5%)	0.000
	Married	161(38.9%)	41(41.4%)	202(39.4%)	
	Divorced	15(3.6%)	14(14.1%)	29(5.7%)	

	Widow	14(3.4%)	14(14.1%)	28(5.5%)	
IDP's camp	Muna	98(23.7%)	20(20.2%)	118(23.0%)	0.824
	Dalori	116(28.0%)	30(30.3%)	146(28.5%)	
	Bakasi	81(19.6%)	22(22.2%)	103(20.1%)	
	Mule	119(28.7%)	27(27.3%)	146(28.5%)	
Source of drinking water	Pipe	292(70.5%)	67(67.7%)	359(70.0%)	0.329
	Well	122(29.5%)	32(32.3%)	154(30.0%)	
LGA	Ghala	41(9.9%)	10(10.1%)	51(9.9%)	0.858
	Monguno	36(8.7%)	6(6.1%)	42(8.2%)	
	Marte	21(5.1%)	4(4.0%)	25(4.9%)	
	Bama	56(13.5%)	15(15.2%)	71(13.8%)	
	Konduga	39(9.4%)	9(9.1%)	48(9.4%)	
	Dikwa	21(5.1%)	6(6.1%)	27(5.3%)	
	Gwoza	41(9.9%)	7(7.1%)	48(9.4%)	
	Damboa	21(5.1%)	9(9.1%)	30(5.8%)	
	Chibok	19(4.6%)	6(6.1%)	25(4.9%)	
	Mafa	74(17.9%)	13(13.1%)	87(17.0%)	
	Kalabalge	20(3.9%)	6(6.1%)	26(5.1%)	
	Abadam	25(6.0%)	8(8.1%)	33(6.4%)	

The bivariate analysis result shows that the association between IDP's camp (p= 0.824), source of drinking water (p=0.329), local government area (p=0.858) and patients status are insignificant at 1% level of significance.

3.1 Goodness of fit test

Classification table is a method to evaluate the predictive accuracy of the logistic regression model. The classification table shows how well our full model correctly classifies cases. The overall percentage in the lower right corner of the table show that the classifications which predict patient's status is 93.4%.

Table2: Classification table for model predicting patient's status

Observed		Predicted		
		Status of the patients		Percentage correct
		Survived	Not survived	
Status of the patients	Survived	403	11	97.3
	Not survived	23	76	76.8
Overall percentage				**93.4**

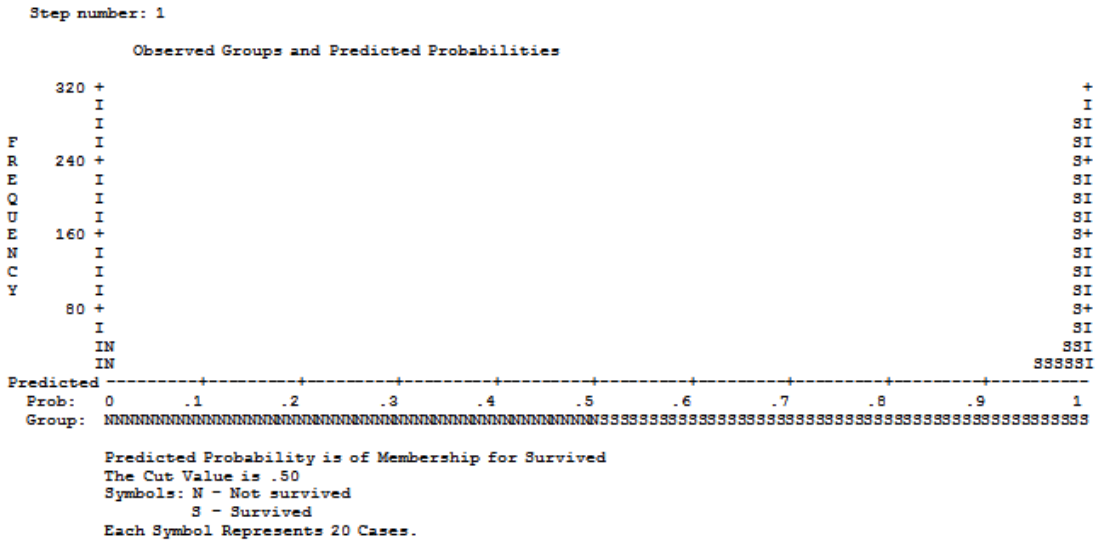

Figure 1: The Distribution Plot of the Patients Status verses the Probability

The distribution plot of the classification of the patients status verses the probability is shown in Figure 1 above. The graphical plot is also another technique of evaluating wrong and right predictions, this is carried-

out by plotting the survived and not survived case status. The cut-off value as used in this study is 0.5 (i.e. 50%); therefore, the decision of classifying the status of patient whether survived or not survived is heavily reliant on the cut-off value generated. Consequently, any patient having a score above 0.50 are termed as survived while, patient having less than 0.50 are classified as have not survived.
Thus, the classification table showed that 15.89% (i.e. 76 patients) under study fell below the cut-off point which implies that they have not survived the infection. While, 84.13% (i.e. 403 patients) went above the 0.50 cut-off point.

3.2 Likelihood ratio test (LRT), AIC and BIC

The likelihood ratio test statistic is approximately chi-square distribution with degrees of freedom equal to the difference between numbers of predictors between the nested model and the full model. It compares how the model fit the data between the empty and the full model by Alkaike's Information Criteria Value and Bayesian Information Criteria Value.

Table3: Results of Model Fit Statistic for the empty and full models

Goodness of fit measure	Empty model	Full model
Df	1	13
Log likelihood	-251.6361	-87.9359
AIC	505.2723	189.8717
BIC	509.5125	219.5536

The values of the likelihood ratio test statistic LR= 327.40 with P-value 0.0000. LR is approximately Chi-square distributed with degrees of freedom equal to the difference between numbers of predictors between the nested model (df=13-1=12). Since the P-value is very small, we can reject the null hypothesis of no significant difference between the two models. "The table 3" also shows that the AIC and BIC values for the full model are smaller than the empty model. Therefore, we conclude that the less restrictive model (full model) fit the data significantly better than the more restrictive model (empty model). So adding the predictor variables to the model has significantly increased our ability to predict the patient's status.

3.3 Hosmer - Lemeshow goodness of fit test

Table4: Test of significance of Hosmer-Lemeshow goodness of fit statistic

Chi-square	Df	Sig.
11.346	8	0.183

Hosmer-Lemeshow goodness of fit test is a check if the null hypothesis that the model adequately fits the data. From "table 4" the P-values for the model predicting patients status is 0.183, which are larger than 0.05. Therefore, we do not reject the null hypothesis, implying that the model fits the data at an acceptable level; this proves that the predicted data are not significantly different from the observed data.

The binary regression analysis results shows the parameter estimates using maximum likelihood estimate and reporting the odds ratio.

Table5: Binary logistic regression analysis results

Predictor	$\widehat{\beta}_j$	S.E	Wald	df	Sig.	EXP $\widehat{\beta}_j$	95% C.I for ($\widehat{\beta}_j$)	
							Lower	Upper
AGE			3.094	4	0.542			
<1yr	-0.191	0.839	0.052	1	0.820	0.826	0.159	4.279
1-20yrs	1.127	1.007	1.253	1	0.263	3.087	0.429	22.214
21-40yrs	0.710	1.069	0.441	1	0.507	2.034	0.250	16.533
41-60yrs	0.577	1.103	0.273	1	0.601	1.780	0.205	15.476
>60yrs (ref)								
SEX								
Female	0.530	0.462	1.318	1	0.251	1.699	0.687	4.200
Male (ref)								
VACCINATION								
Not vaccinated	-3.153	0.512	37.974	1	0.000	0.043	0.016	0.116
Vaccinated (ref)								
DEGREE OF DEHYDRATION			75.180	2	0.000			

Severe	-6.644	0.880	57.050	1	0.000	0.001	0.000	0.007
Moderate	-2.971	0.784	14.375	1	0.000	0.051	0.011	0.238
Mild (ref)								
MARITAL STATUS			4.564	3	0.207			
Single	1.425	0.984	2.098	1	0.148	4.156	0.605	28.566
Married	1.503	0.806	3.474	1	0.062	4.496	0.925	21.841
Divorced	0.285	1.259	0.051	1	0.821	1.330	0.113	15.671
Widow (ref)								
SOURCE OF DRINKING WATER								
well	0.084	0.468	0.033	1	0.857	1.088	0.857	1.088
pipe (ref)								
Constant	4.598	1.449	10.074	1	0.002	99.265		

"Table 5" presents the results of binary logistic regression analysis. The results shows that vaccination and degree of dehydration are predictors that affect status of the patients at 5% level of significance.

Vaccination has significant effect on status of the patient. Not vaccinated patients before the infection were 49% less likely (aOR= 0.512, 95% CI: 0.016-0.116) to survive compared to those patients whom were vaccinated before the infection.

Degree of dehydration has significant effect on status of the patient. Patients with severe and moderate dehydration are 99%, 95% respectively less likely (aOR= 0.001, 95% CI: 0.000-0.007, aOR= 0.051, 95% CI: 0.011-0.238) to survive compared to patients with mild dehydration.

The estimates of the fitted logistic regression models were obtained using maximum likelihood estimate from "Table 5" above.

The fitted logistic regression model:

$$\boldsymbol{P(patient's\ status = 1\ /\ X) = \frac{e^{4.598-3.153NV-6.644SD-2.971MD}}{1+e^{4.598-3.153NV-6.644SD-2.971MD}}} \qquad (\mathbf{8})$$

The logit model:

$$\boldsymbol{logit(patient's\ status = 1\ /\ X) = 4.598 - 3.153NV - 6.644SD - 2.971MD} \qquad (\mathbf{9})$$

IV. Conclusion

This study was carried out to investigate the factors that determines the survival of cholera infected person and to model these factors in order to predict survival probability. Recorded data of 513 patients was obtained from UNICEF Cholera Hospital for Internally Displaced Persons within Maiduguri, Borno State of Nigeria to carry out the analysis.

The work will enable us to know the significant factors that increases the chance of survival of cholera infected person and predict its survival probability based on significant factors (degree of dehydration and vaccination status).

The findings of this study identified vaccination and dehydration as the major factors that affect the status (survival or death) of cholera infected person. Vaccinated patients before the infection are 51% more likely to survive compared to not vaccinated patients, this shows that the vaccine administered before the infection have significant effect on the survival of the patients. Patients with moderate and severedegree of dehydration are 99%, 95% respectively less likely to survive compared to patients with mild degree of dehydration. The study recommend that much attention should be given to degree of dehydration and effort should be made to hydrate the infected person and vaccine (*killed oral 01 with whole-cell with Bsubunit)* should be administered whenever there is outbreak.

References

[1]. Pub med/Google scholar*World Health Organization*, 2016.
[2]. National Population Commission Nigeria. Accessed 25 December. 2016
[3]. Umar M Hassan and A.A. Abiodun,Survival analysis of cholera patients a parametric and non-parametric approach*SCIENCEDOMAIN international, Asian Journal of Probability and Statistics*,*5(4):* 2019, 1-18
[4]. Peng CYJ, So TSH. *Logistic regression analysis and reporting (*Erlbaum Associate Inc:A primer, Lawrence, 2000
[5]. Hosmer, W.D. and S. Lemeshow, *Applied Logistic Regression* (New York, United States of America: John Wiley and Sons, 2000).
[6]. Agresti .A,*An introduction to categorical data analysis* (New York, United States of America: A John Wiley and Sons, 1996)