RESPIRATORY SYNCYTIAL VIRUS INFECTION IN INFANTS: A COMPARATIVE STUDY USING DISCRIMINANT, PROBIT AND LOGISTIC REGRESSION ANALYSIS

Abstract

In babies, respiratory syncytial virus (RSV) is the most common cause of lung inflammation (pneumonia) or bronchiolitis (inflammation of the lungs' airways). This virus comes with several symptoms such as congested or runny nose, dry cough, low-grade fever, sore throat, sneezing, headache, difficulty in breathing, etc. The virus can cause death in babies if not properly managed and therefore calls for immediate investigations into finding the significant causes. Several research works have been done but the idea of investigating more potential predictor variables and the application of both regression and classification models have been grossly understudied. In this paper, the unpublished secondary data collected from three different hospitals in Port Harcourt, Rivers State, Nigeria on fifteen predictor variables that are potential causes of the respiratory syncytial virus is modeled using two categorical regression approaches – logistic and probit regression models and one classification model – discriminant function analysis. The models were compared using misclassification errors, ROC plot, concordance, sensitivity, specificity, and pseudo-r-square values. The linear discriminant function model outperformed both logit and probit models. The results showed that paternal history of asthma, maternal history of asthma, mother's occupation, mother's smoking habit and mother's education level were the most important variables to linearly classify seropositive and seronegative patients.

Keywords: Probit model, Binary Logistic Regression Model, Discriminant Analysis, ROC plot, Respiratory syncytial virus

1.0 INTRODUCTION

Respiratory syncytial virus (RSV) is a reorganized major contributor to infant hospitalizations in the whole world (Caroline et al. 2009). The symptoms of this respiratory virus are usually mild and cold-like. It can be severe in infants and aged ones and could develop bronchiolitis and pneumonia. According to the Centers for Disease Control and Prevention (CDC), RSV causes approximately 2.1 million outpatient visits (for babies < 5yrs), 58,000 hospitalizations (for babies < 5yrs), 177,000 hospitalizations among adults 65 and older, and 14,000 deaths among adults 65 and older in the United States and other areas with similar climates in a year.

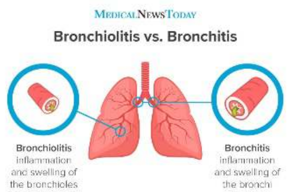
According to a report by National Centre for Immunization and Respiratory Diseases (NCIRD) in 2018, the symptoms take between 4 to 6 days to appear. These symptoms include Runny nose, loss of appetite, cough, sneezing, difficulty in breathing, fever, wheezing, etc.

This RSV infection leads to severe ones such as an inflammation of the small airways in the lung and acute respiratory infection that affects the lungs (Rha et al. (2015). It aggravates chronic health challenges. Asthmatic individuals may have asthma episodes as a result of RSV infection, and people with congestive heart failure may have more severe symptoms as a result of RSV infection (Falsey et al 2005). Acute lower respiratory infection (ALRI) is one of the primary causes of morbidity and mortality in babies, according to Shi et al. (2017), and human respiratory syncytial virus (RSV) is the most prevalent viral pathogen found in ALRI patients. Researchers are working on developing RSV vaccinations, but none are currently available, Ramilo and Mejias (2020).

Bronchiolitis is caused by viruses, with respiratory syncytial virus being the most frequent (RSV), RSV can infect babies under the age of three months, as well as those who have continuing illnesses such as heart or lung disease, were born prematurely (before 32 weeks of pregnancy), or were exposed to cigarette smoke, Megan and her colleagues (2009). Bronchiolitis and RSV are often confused with pneumonia, despite the fact that they share symptoms such as a stuffy or runny nose, dry cough, low-grade fever, sore throat, sneezing, headache, rapid breathing or difficulty breathing, short shallow and rapid breathing, struggling to breathe, cough, poor feeding, unusual tiredness, and so on.

Pneumonia is a lung infection brought on by a viral infection. When a healthy person breathes, small sacs in the lungs called alveoli fill with air. The alveoli get blocked with pus and fluid when a person has pneumonia, making breathing difficult and decreasing oxygen intake. Pneumonia is an infection-related illness in which the lungs become inflamed and congested, producing cough and shortness of breath and reducing oxygen exchange (Escobar et al, 2010). RSV is the most common cause of lung inflammation (pneumonia) or bronchiolitis (inflammation of the airways of the lungs) in newborns. Several issues might occur when RSV travels to the lower respiratory tract and produces pneumonia or bronchiolitis.

Although bronchiolitis and bronchitis are two different disorders, they share several characteristics. Bronchitis is defined as inflammation of the airways leading to the windpipe, whereas bronchiolitis is defined as inflammation of the bronchioles, which are small airways that branch from the bronchi. Bronchitis is a common infection that affects people of all ages, and RSV is the cause of bronchiolitis and BPD.



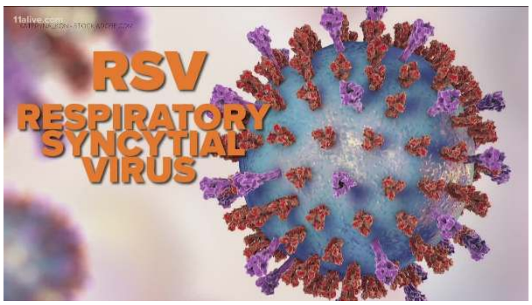


Image 1 : Respiratory syncytial virus

https://www.11alive.com/article/news/health/doctors-report-spike-in-rsv-cases-in-metro-atlanta-as-covid-restrictions-lift/85-37322ae6-1943-4efe-a492-b974ed1c136e

In bronchiolitis, the respiratory syncytial virus is the most prevalent infecting agent. Although it can happen at any time of year, it is more common during the winter. Di Carlos et al. (2007) investigated the development of a model for the prediction of Broncho-Pulmonary Dysplasia in seven-day-old newborns. The model properly classified the children 93.7 percent of the time, according to the findings. Light et al.(2008) looked at the relationship between respiratory syncytial virus (RSV) test results and RSV lower respiratory tract disease hospitalizations in Florida. According to the data, RSV circulated throughout the state year-round, with seasons ranging from 7-8 months in the southwest, northwest, and north to 11-12 months in the center and southeast. Caroline et al. (2009) conducted population-based monitoring of acute respiratory infections in children under the age of five using Chi-square tests and multivariable logistic regression models. According to the data, RSV infection in children in the United States is connected to considerable morbidity in both hospital and outpatient settings. Aliyu et al. (2015) conducted a serological survey among Zaria children. The data revealed that there was a statistical link between age and seropositivity. The seroprevalence was highest in the age group 49-60 months and lowest in the age group 0-12 months, with RSV infection being quite common among children aged 1-5 years in Zaria. Using demographic data from the year 2000, Philip et al. (2009) investigated variance in depth and identified factors for the duration of the RSV season. In a ZCTA, univariate and multivariate logistic regression analysis were utilized to find determinants of RSV season duration. They came to the conclusion that the length of the RSV season was impacted by demographic considerations. Using data from two tertiary health institutions in the North Central Zone, Sule (2013) used Discriminant Analysis and Logistic Regression models to predict the prevalence of Broncho-Pneumonia status (BPn) in babies. The Discriminant Model was found to be more effective in the zone than the Logistic Regression Model. They also discovered that, when compared to other predictor variables, the baby's weight at birth is the best in discriminating between the two groups, as it has the lowest Wilk's Lambda value. According to Obodai et al. (2014) in their study in Respiratory syncytial virus genotypes circulating in urban Ghana, nasopharyngeal aspirates (NPA) were obtained from children diagnosed with ALRI between February and November 2006 and screened for RSV using a nested multiplex reverse transcriptase polymerase chain reaction (RT-PCR) method for genotyping RSV. The results revealed that RSV was found in 60.4 percent (32/53) of the 53 NPA examined. Following genotypic analysis, RSV B infections accounted for 72 percent (23/32) of the 60.4 percent RSV infections, whereas RSV A and B co-infections accounted for 28 percent (9/32). The age group of children between the ages of 2 and 12 months was the most affected. Following a systematic literature review by RSV Global Epidemiology Network (RSV GEN) colleagues, Ting Shi et al. (2015) identified the risk variables for RSV-associated ALRI in young infants. There were substantial correlations between HIV and RSV-associated ALRI in all papers (regardless of quality grade). The connections between possible RSV risk variables were determined as relative risks with 95 percent confidence intervals by Caitriona et al. (2021). Shi et al. (2017) investigated the global, regional, and national illness burden estimates of acute lower respiratory infections caused by respiratory syncytial virus in children under the age of five. According to their findings, there were roughly 6 million occurrences of RSV severe ALRI in infants in 2015, with 51 percent dying in hospitals and 49 percent dying outside of hospitals. They also discovered that there are around 118,000 serious newborn deaths in poor countries. Yakubu et al. (2019) discovered that the baby's weight at birth, the baby's weight four weeks after birth, and the mother's occupation all have significant effects on the infant's bronchopneumonia status using Binary Logistic Regression Methods for Modeling Broncho-Pneumonia Status in Infants from Tertiary Health Institutions in North Central Nigeria. Using probabilistically linked perinatal, hospital, and laboratory records of 321,825 children born in Western Australia (WA) from 2000 to 2012, Gebremedhin et al. (2022) built a prediction model for RSV positivity in hospitalized children aged 5 years. The model's prediction capacity was evaluated using the cross-validated area under the receiver operating characteristic (AUROC) curve. RSV positivity was observed to be associated with a younger admission age, male gender, non-Aboriginal ethnicity, bronchiolitis diagnosis, and a longer hospital stay. In a descriptive cross-sectional study, Damilola et al. (2019) discovered that the proportion of newborns in the RSV-positive group was considerably greater than that in the RSV-negative group (82.9 percent vs 54.4 percent; p=0.002). RSV has remained a common cause of severe ALRI in babies, according to the researchers. Shi et al.(2020) conducted study on global respiratory syncytial virus disease burden estimations and discovered that RSV-ARI (respiratory syncytial virus-associated acute respiratory illness) is a significant disease burden in older persons aged 65 years. Oluwadamilare et al. (2021) conducted a study on Respiratory Syncytial Virus Infection in Children in Lagos, Nigeria. According to their findings, 200 people were recruited over the course of eight months. RSV (RNA) was found in 45 (22.5%) of the participants. RSV was most prevalent in children aged 2 to 6 months (32.4 percent). During three consecutive influenza seasons (2012–2015), Abdul et al. (2021) looked explored the characteristics and outcomes of respiratory syncytial virus (RSV) infection in people hospitalized with influenza-like illness (ILI) in six French hospitals. RSV was found in 4% (59/1452) of cases, while influenza virus was found in 39% (566/1452). Malignancy (adjusted OR 2.1, 95 percent CI 1.1–4.1, p 0.04) was a risk factor for RSV infection, and multivariate analysis of all patients with ILI revealed that cancer and immunosuppressive treatment were substantially linked with RSV detection. Ruimu & Jikui (2021) looked at clinical characteristics and differences between infants with single Bordetella pertussis (B. pertussis) infection and those with RSV coinfection in their study on the Clinical Impact of Respiratory Syncytial Virus Infection on Children Hospitalized for Pertussis. Sexe, body weight, preterm birth history, pertussis vaccination, symptoms, pneumonia presence, and lymphocyte count did not differ significantly between the two groups. According to univariate analysis, patients with RSV coinfection were older, more likely to be treated with -lactam antibiotics, had more wheezes and rales on chest auscultation, were more likely to be readmitted, and had a longer hospital stay. In a later binary logistic regression analysis, patients with RSV coinfection were more likely to have wheezes and be readmitted. Despite the constraints, their research demonstrated that while diagnosing and treating RSV infection, age, household smoking, gender, gestational age, and birth weight should all be taken into account.

The present work is different from other previous works such as Beki (2012) in which a combination of discriminant and binary logistic regression were used and considered only three variables (Sex, Weight at birth and Weight after four weeks) using two hospitals. Danbaba et al, (2013) investigated some variables considered from mother's aspect and use logistic regression. Yakubu et al. (2019) used a combination of Mother's and Baby's independent variables capable of having Broncho-pneumonia in infants. Baby's weight at birth, Baby's weight four weeks after, Baby's gender, Mother's age and Mother's occupation in two hospitals. Sule (2013) applied both Discriminant Analysis and Logistic Regression models data from two tertiary health institutions Five predictors which are (baby's weight at birth, baby's weight 4week after, sex, mother's age and mother's occupation) were considered. The above were in BNP and not RSV.

Again, several works had been done on the use of discriminant analysis and binary logistic regression for tracking the incidence of Broncho-Pulmonary Dysplasia among infants using one to three predictor variables (i.e. weight at birth, weight four weeks later and gender) such works are Beki (2012), Danbaba et al (2013) and Sindhu et al (2015). We want to compare Discriminant Function Analysis (Linear and Quadratic score functions), Probit Regression and Logistic Regression models using data from three hospitals (University of Portharcourt Teaching Hospital, Aluu Health Center and Braitwait Memorial Specialist Hospital) with fifteen predictor variables which are baby's gender (male or female), baby's weight at birth (low, normal or macrosoma), mother's smoking habit (yes or no), catarrh (yes or no), running nose (yes or no), maternal history of ashma (yes or no), paternal history of ashma(yes or no), breastfeeding method (exclusive or nonexclusive), IgG antibody level (low, normal or high), mother's age, baby's weight 4week after, baby's age (grouped in months), maternal education level (primary, secondary or tertiary), mother's occupation (business, civil servant or house wife) and the hospital (UPTH, BMS or AHC) in predicting the presence of RSV and not BND.

2.0 THEORETICAL FOUNDATIONS

2.1 Logistic Regression Model

The linear regression that uses the ordinary least squares method to minimize sum of the squared deviations is often used for predicting continuous Y variables while logistic regression is used for categorical Y variable classification.

It is inappropriate to use linear regression to model a categorical (dichotomous) Y variable since the resulting model will give predicted Ys outside 0 and 1. Also, other numerous assumptions such as normality of the errors may also get violated.

The response variable RSV is categorical with two levels (seropositive or seronegative) and therefore suggests that the appropriate logit model is the binary logistic regression model. Since the response variable is dichotomous, the mathematical foundation and representation for modeling the log odds of the event $\log_e\left(\frac{P}{1-P}\right)$ where P is the probability of an event are described by Orumie et al. (2021).

2.2 Probit Regression Model

Because the conditional probability function is assumed to be linear, the probit or logit models are favoured over the standard least squares regression model when the response variable is categorical (dichotomous). While logistic regression utilizes a cumulative logistic function for the estimate model, probit regression uses a normal cumulative density function.

In Probit regression, the cumulative standard normal distribution function $\phi(.)$ is used to model the regression function, that is, we assume

$$E(Y \mid X) = P(Y = 1 \mid X) = \varphi(\alpha + \beta X) \tag{1}$$

where
$\alpha + \beta X$ is regarded as a quantile Z.
α is the usual regression intercept term
β is the change in Z associated with a one unit change in X

Although the effect on Z of a change in X is linear, the link between Z and the response variable Z is nonlinear since ϕ is a nonlinear function of X.

For a binary Y variable, the model

$$Y = \alpha + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \tag{2}$$

with

$$P(Y = 1 \mid X_1, X_2, \ldots, X_k) = \phi\left(\alpha + \sum_{i=1}^{k} \beta_i X_i\right) \tag{3}$$

Is a population probit model with multiple regressors X_i, $i = 1, 2, 3, \ldots, k$ and $\phi(.)$ the cumulative standard normal distribution function.

2.3 The Discriminant Function Analysis (DFA)

This is another good model for application when the response variable is categorical and classification is the problem. The Logit and Probit models predict while the DFA classifies. It uses a linear or quadratic score function to classify observations into mutually exclusive groups within a categorical variable. It uses either continuous or categorical or both types of variables as the predictor variables. It can be a simple or multiple DFA. It is simple when only one predictor variable is used, otherwise it is multiple.

As described by Egbo and Bartholomew (2020), the discriminant function analysis used any of the three methods: equal, arbitrary or estimated approaches to find the class probabilities while building the model. The linear discriminant function uses a linear score function while a quadratic discriminant function uses a quadratic score function. The models and method of implementations is given by fisher (1936).

2.4 Methods of comparison

In this paper, the entire data is divided into two sets: training set and test set. The training set contains 70% of the entire sample size while the remaining is used as test set to validate the models. Simple methods like misclassification errors, ROC plot, concordance, sensitivity, specificity and pseudo r-square values are used for comparison purposes. At first, the probit and logit models is compared in the training set and the best model then compared with the best DFA (linear or quadratic).

3. EMPIRICAL RESULTS

The data used for this study is unpublished secondary data obtained from three hospitals mentioned earlier in the following manner (UPTH = 263 records, AHC = 233 records and BMH = 257 records). These gave a total sample size of 753 records. The missing data (120) rows with missing values were removed leaving 633 valid cases as the sample size. Out of the 633 valid cases, 285 are seronegative while 348 are seropositive to RSV. Information on the fifteen predictor variables was obtained for all.

3.1 Sampling method

The simple random sampling method was used to randomly sample 70% of 285(which is 199) and 70% of 348(which is 243) to make up the training data set of size 442 (199+243). The sampling was necessary in order to solve the problem of class bias. Ideally, the proportion of events (seropositive) and non-events (seronegative) in the

Y variable should approximately be the same but for the case of the data for this study, 285 is reasonably far from 348. The remaining 191 formed the sample size for the test set.

3.2 Data Structure

Out of the fifteen categorical variables, all except mother's age and baby's weight 4 weeks after are categorized. The following continuous variables were recoded: baby's weight at birth and baby's age (In months). Baby's weight at birth were categorized into low birth weight (less than 2.5kg), normal birth (between 2.5kg and 4.2kg) weight and macrosoma (over 4.2kg) as described by Bartholomew et al. (2021) while the baby's age was grouped into 5 age groups : (1 - 6months, 7 - 12months, 13 - 18months, 19 - 24months and 25 - 30months). For the categorical variables, one of the levels is chosen as the reference level as in baby's birth weight (ref = low), Gender (ref = Male), IgG level (ref = Normal), mother's occupation (ref = house wife), maternal history of ashma (ref = Yes), paternal history of ashma (ref = Yes) and for the response variable RSV (ref = seropositive).

3.3 Comparison of Logit and Probit Model Results

The logit and probit models were implemented using the training data set and the results are summarized in Table 1. and Table 2.

Table 1. Logistic and Probit Regression model outputs on training set

Training Data	Logit		Probit		
Coefficients:	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)	Sig
(Intercept)	0.759444	0.739286	0.513519	0.701804	NS
GenderFemale	-0.26296	0.288825	-0.14506	0.313336	NS
BWBNBW(2.5 - 4.2kg)	-0.11314	0.757531	-0.05052	0.812391	NS
BWBOW (> 4.2kg)	0.237745	0.699562	0.143445	0.687549	NS
MSHYES	0.122232	0.760897	0.08021	0.73347	NS
BBM-Non Exclusive	-0.05525	0.849773	-0.04184	0.804631	NS
IgG-Low	0.264255	0.319554	0.154463	0.315354	NS
IgG-High	-1.52729	0.000336	-0.88224	0.000322	***
Cataarh-YES	-0.69299	0.468306	-0.41223	0.463152	NS
RNYES	-0.30858	0.852254	-0.18974	0.846627	NS
`BA(Months)`7 - 12months	0.050546	0.875982	0.003332	0.985795	NS
`BA(Months)`13 - 18months	0.14371	0.670502	0.061412	0.754765	NS
`BA(Months)`19 - 24months	-0.00012	0.999772	-0.03571	0.88589	NS
`BA(Months)`25 - 30months	-1.04543	0.347422	-0.65232	0.317402	NS
MEL-Secondary	0.827623	0.02351	0.506965	0.017888	*
MEL-Tertairy	0.668238	0.045781	0.406907	0.037828	*
MO-Civil Servant	-0.25699	0.449345	-0.06892	0.723597	NS
MO-House Wife	1.982968	1.79E-11	1.204413	1.60E-12	***
MAHYES	1.862673	4.72E-10	1.088443	2.44E-10	***
PAHYES	3.654589	< 2e-16	2.149797	< 2e-16	***
Hospital-BMH	-0.1326	0.634778	-0.11135	0.493809	NS

Hospital-AHC	-0.14058	0.658759	-0.1142	0.536456	NS
MA	-0.0415	0.080406	-0.02675	0.050559	NS
BW4W	-0.10234	0.614151	0.117893	0.596511	NS

Key: NS – not significant, *** - significant at 0.05 and 0.01 alpha values, * significant at 0.05 alpha value

Table 2. Logit and Probit model validation results for test data

		Comparison Criteria (Test Set)			
Models	AIC	Concordance	Sensitivity	Specificity	Pseudo R-square
Logit	489.97	0.7323	0.7143	0.6744	0.2985
Probit	474.56	0.7297	0.7048	0.6628	0.2988

Confusion Matrix for Logit model

$$\begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 58 & 30 \\ 28 & 75 \end{bmatrix} \end{array}$$

Confusion Matrix for Probit model

$$\begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 57 & 30 \\ 29 & 74 \end{bmatrix} \end{array}$$

Table 3. Odd Ratio and its 95% Confidence Interval for Logit Model (sig. variables)

Coefficients	Odd Ratio	2.50%	97.50%
IgG-High	0.217124	0.09225	0.492684
MEL-Secondary	2.287873	1.124857	4.726974
MEL-Tertairy	1.950798	1.018132	3.791326
MO-House Wife	7.26427	4.128636	13.15493
MAH-YES	6.44093	3.636882	11.77849
PAH-YES	38.65162	17.76243	91.88824

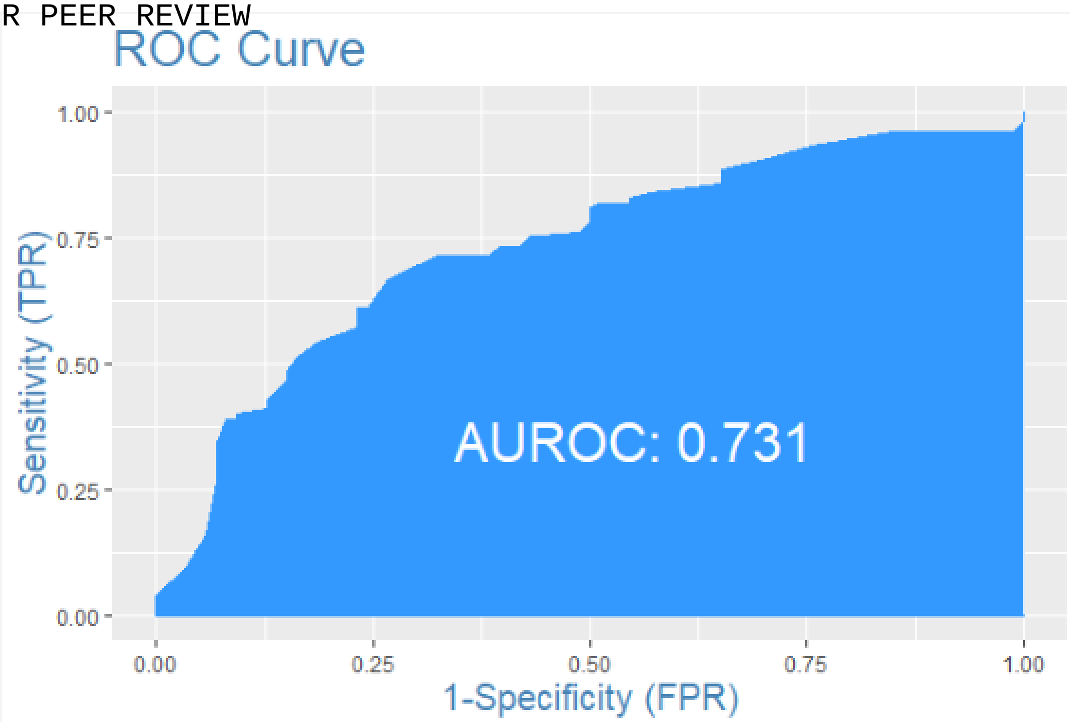


Figure1. ROC Curve for Logit Model (Test Data)

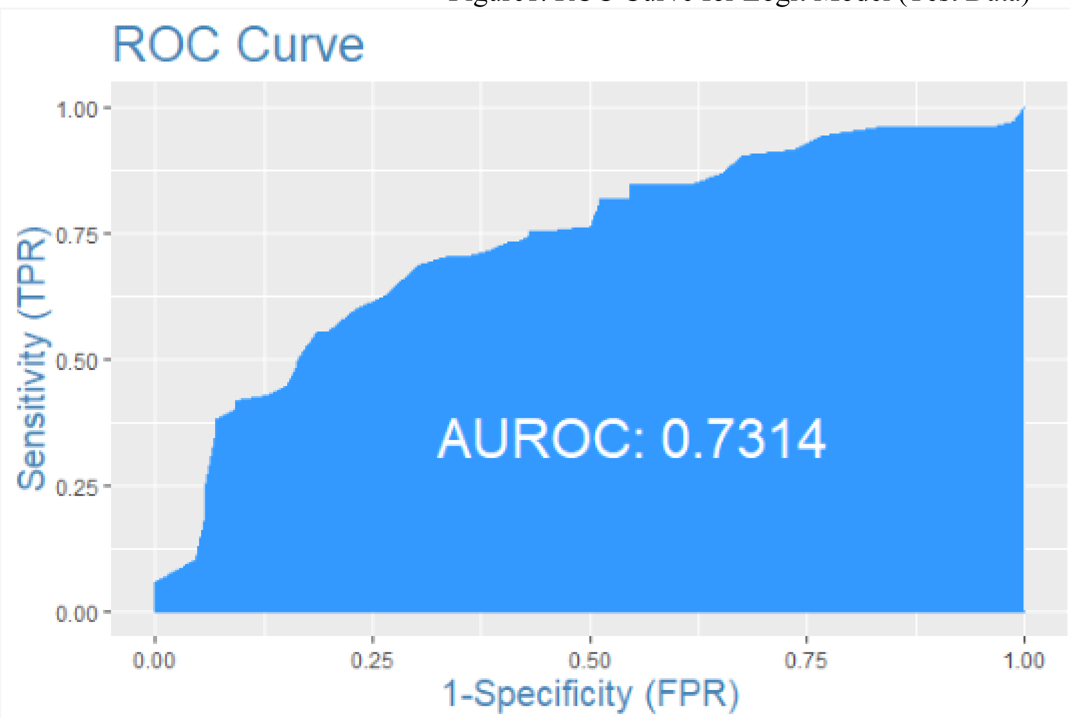


Figure 2. ROC Curve for Probit Model (Test Data)

3.4 The Discriminant Function Analysis (DFA) Results

3.4.1 Linear DFA Results

Because DFA presupposes multivariate normality, the data must be examined for significant deviations from normality before proceeding with the analysis. The first step, as always, is to plot the data to see if any outliers need to be removed or if any data transformations are required, particularly for the continuous predictor variables (baby's weight 4 weeks after birth and mother's age).

Table 4. Shapiro Wilk's Normality Test

	Baby's weight 4weeks after	Mother's age
W	0.99474	0.99434
p-value	0.1356	0.1013

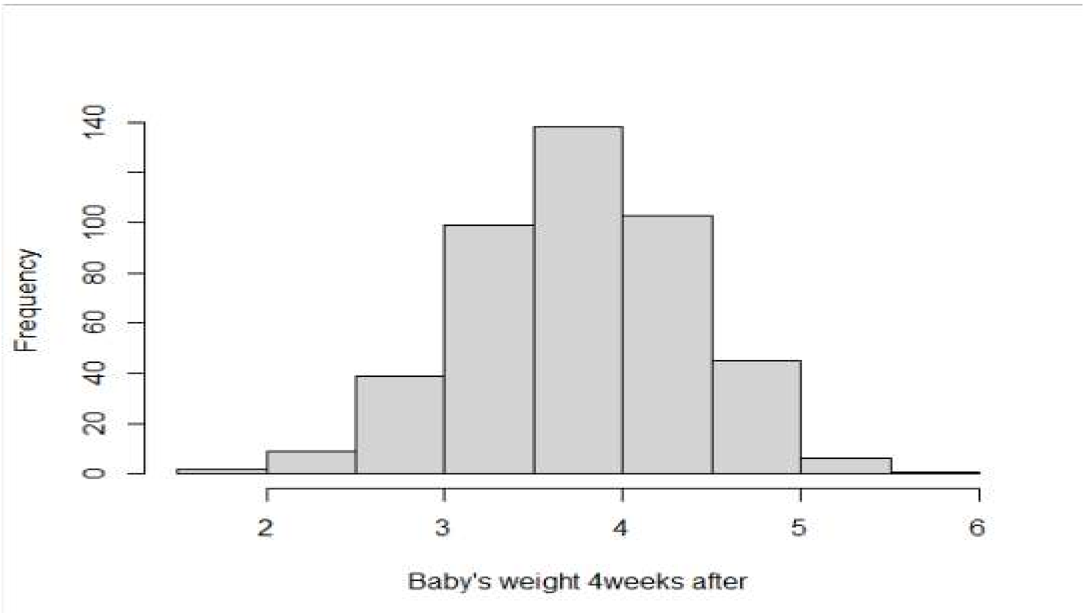


Figure 3 Histogram plot of baby's weight 4weeks after

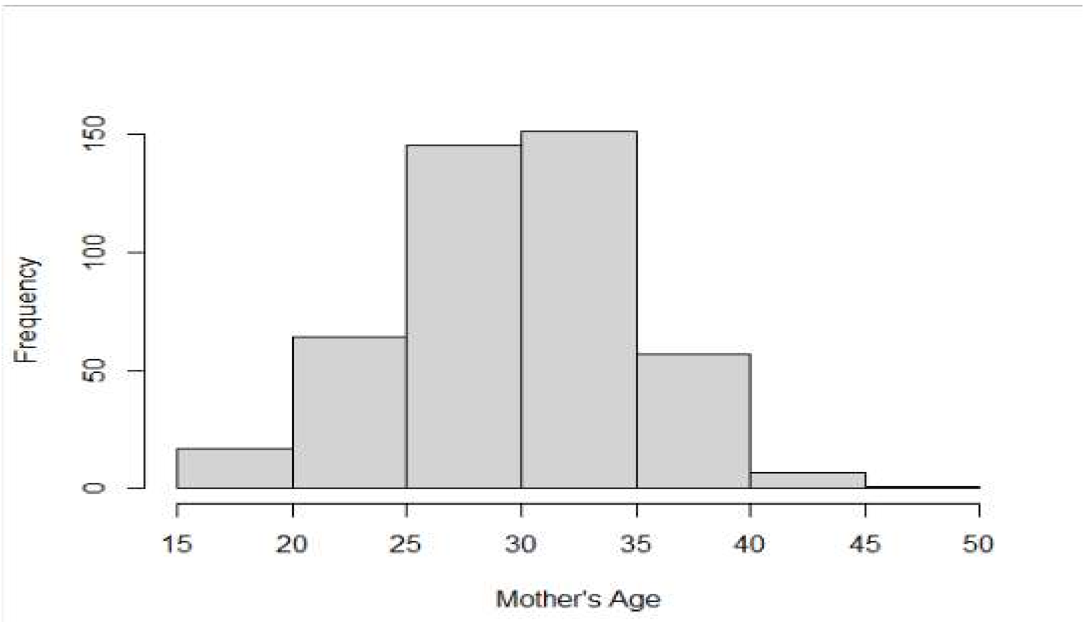


Figure 4 Histogram plot of mother's age

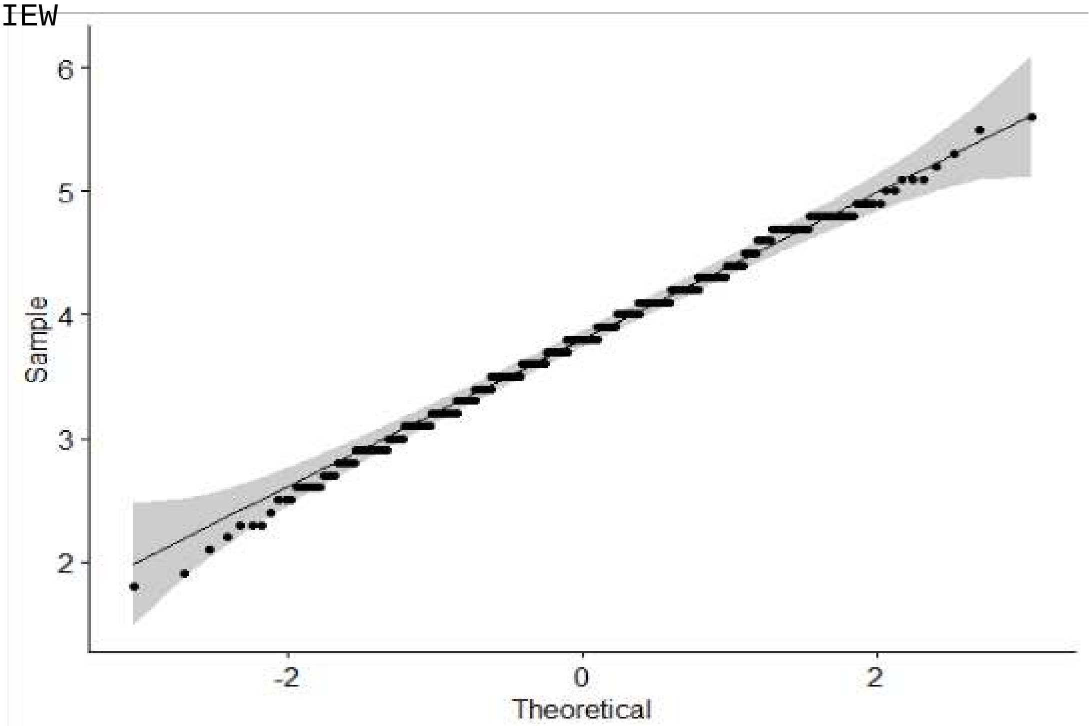


Figure 5 Quantile-Quantile plot of baby's weight 4weeks after

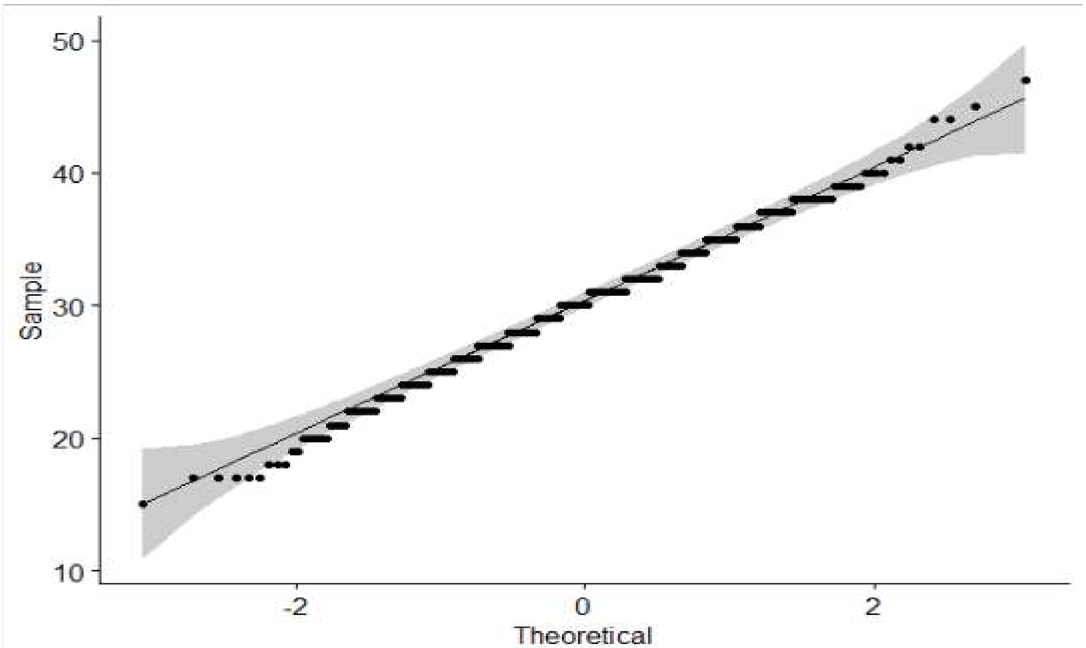


Figure 6 Quantile-Quantile plot of mother's age

Table 5 LDA Result for the training data

Call:
lda(RSV ~ Gender + BWB + MSH + BBM + IgG + Cataarh + `BA(Months)` +
MEL + RN + MO + MAH + PAH + Hospital + MA + BW4W, data = trainingData)

Prior probabilities of groups:

0	1
0.45	0.55

Group means:

	Gender	BWB	MSH	BBM	IgG	Cataarh	`BA(Months)`	MEL	MO	MAH	PAH	Hospital
0	1.56	1.91	0.10	1.21	1.65	0.99	2.34	2.35	1.68	0.18	0.07	1.91
1	1.53	1.95	0.12	1.24	1.58	0.97	2.22	2.37	0.12	0.37	0.35	1.86

	MA	BW4W	RN
0	30.43	3.82	0.99
1	29.84	3.76	1.00

Coefficients of linear discriminants:

	LD1
Gender	-0.20
BWB	0.10
MSH	0.18
BBM	-0.01
IgG	-0.32
RN	-0.42
Cataarh	-0.65
`BA(Months)`	-0.05
MEL	0.15
MO	0.66
MAH	1.17
PAH	2.28
Hospital	-0.05
MA	-0.03
BW4W	-0.08

Table 6 QDA Result for the training data

Call:
lda(RSV ~ Gender + BWB + MSH + BBM + IgG + Cataarh + `BA(Months)` +
MEL + RN + MO + MAH + PAH + Hospital + MA + BW4W, data = trainingData)

Prior probabilities of groups:

0	1
0.45	0.55

Group means:

	Gender	BWB	MSH	BBM	IgG	Cataarh	`BA(Months)`	MEL	MO	MAH	PAH	Hospital
0	1.56	1.91	0.10	1.21	1.65	0.99	2.34	2.35	1.68	0.18	0.07	1.91
1	1.53	1.95	0.12	1.24	1.58	0.97	2.22	2.37	0.12	0.37	0.35	1.86

	MA	BW4W	RN
0	30.43	3.82	0.99
1	29.84	3.76	1.00

Coefficients of Quadratic discriminants:

	QD
Gender	0.10
BWB	0.23
MSH	-0.03

BBM	0.18
IgG	0.09
RN	1.70
Cataarh	-0.52
\`BA(Months)\`	-0.11
MEL	0.09
MO	0.14
MAH	-0.05
PAH	0.02
Hospital	0.15
MA	0.00
BW4W	1.73

Table 7 Comparison of LDA and QDA on test set

	LDA	QDA
Model Accuracy	0.63	0.62
Actual number of observations in group 1	86	86
Number classified	81	93
Actual number of observations in group 2	105	105
Number classified	110	98

3.5 Comparison of LDA, Logit and Probit Models

Table 8 Comparison of LDA, Logit and Probit Models on test set

	LDA	Logit	Probit
Model Accuracy	0.63	0.71	0.70
Actual number of observations in group 1	86	86	86
Number classified	81	58	57
Actual number of observations in group 2	105	105	105
Number classified	110	75	74

4. DISCUSSION OF RESULTS

The Logit and Probit models were fitted using the training set and the results displayed in Table 1. Both models behaved likely the same. Out of the fifteen predictor variables considered in this study, both models identified IgG antibody level, Mother's education level, Mother's occupation, maternal history of ashma and Paternal history of ashma as the significant predictor variables for seropositive result when tested for RSV in babies. The two models were further compared based on the comparison criteria listed in this study in section 2.4. The values shown in Table 2 suggests that Logit and Probit models performed equally but the Logit model performed slightly better as seen in the confusion matrix. The higher the values on the diagonal of the confusion matrix, the better the model. As the prediction probability cutoff is lowered from 1 to 0, the Receiver Operating Characteristics Curve (ROC) tracks the percentage of true positives accurately predicted by a given logit model. As the cutoff is lowered, a successful model should mark more genuine 1s as positives and fewer actual 0s as 1s. When a result, for a decent model, the curve should rise steeply, suggesting that as the cutoff score decreases, the TPR (Y-Axis) climbs faster than the FPR (X-Axis). The greater the area under the ROC curve, the higher the model's prediction performance. The ROC curves for both models in Figure 1 and Figure 2 are nearly identical. Though, the probit model appears to be a better fit for the training set (Table 2. Probit AIC value of 474.56 is smaller than Logit AIC value of 489.97) but the logit model has more predict power (Table 2. Sensitivity value of logit is higher than that of probit), therefore the logit model is chosen as the best model and was used on the test set. The odd ratio and the corresponding 95% confidence intervals were displayed in Table 3. The following interpretations as described by Orumie et al. (2021) follow:

• Those with low IgG antibody level when compared with those with high IgG antibody levels are 0.217 times likely to test positive for RSV. That is about (1-0.217 = 0.783) 78.3% more likely to be seropositive for RSV.

• A baby whose mother’s highest education qualification is primary school when compared with another baby whose mother’s highest education level is secondary school and Tertiary education level is 2.28 and 1.95 times more likely to be seropositive to RSV respectively. That is (2.28 – 1.00 = 1.28) 128% and (1.95 – 1.00 = 0.95) 95% more likely to be attacked by RSV respectively. This is in line with the findings of Bartholomew et al. (2021), who found that the mother's educational level has a significant impact on the likelihood and/or productivity of health investment, as well as the financial resources available to the child, both directly and indirectly, through the choice of partner, fertility timing, and number of offspring.

• A baby whose mother is a House wife when compared with another baby whose mother is either a civil servant or business woman is 6.44 times more likely to be seropositive to RSV.

• With odds ratios of 6.44 and 38.7, maternal and paternal history of ashma has relatively the highest odds for RSV. This suggests that the variables maternal and paternal history of ashma were the most significant variables in predicting the incidence of RSV in the three hospitals. This finding is in agreement with the finding of Abdul momin Kazi et al. (2021).

The Linear Discriminant Function (LDF) and Quadratic Discriminant function (QDF) were fitted on the training set using the estimated priors to estimate class probabilities and the best discriminant model was obtained before comparing with logistic regression model and the results were displayed from section 3.4. The continuous predictor variables were first tested for normality assumptions since that is one of the basic assumptions of the DF model. Both graphical methods (Figure 3 through Figure 6) and Shapiro wilk’s normality test (Table 4.0, p-values 0.1356 and 0.1013 are less than 0.05) suggested that baby’s weight 4weeks after and mother’s age came from a normal distribution. LDA and QDA results were shown in Table 5 and Table 6 respectively. LDA performed slightly better than the QDA (Table 7, model accuracy) on the test set and was compared with Logit model in Table 8. The LDA is seen to perform better than both logit and probit models in predicting the incidence of RSV in the three hospitals (Table 8, more number of observations correctly classified). This finding is in agreement with Sule (2013). Each discriminant function should be scanned for the largest loadings, positive or negative, indicating the variables that contribute most to that discriminant function. Therefore, the variables that contributed as shown in Table 5 under coefficients of linear discriminants are Paternal history of ashma, Maternal history of ashma, mother’s occupation, mother’s smoking habit (as also revealed by Ting Shi et al. (2015)) and mother’s education level.

5 CONCLUSIONS

On the basis of the above findings, it is concluded that the logit model outperformed the probit model and the LDA outperformed QDA in predicting the incidence of RSV in the three hospitals considered. And LDA was chosen as the best model over the logit model. However, the three models has similar results as, maternal history of ashma, paternal history of ashma, mother’s education level and mother’s occupation were important contributors to a baby having RSV. Thus, it is recommended that

i. Babies whose either or both parents has(have) history of ashma should be quickly immunized to increase the IgG antibody level in the baby to fight RSV.

ii. House wives should improve their attention to health care practices for their babies and always seek medical attention from health practitioners. They should try to avoid self-medical practices. Mother's occupation have been identified as predictor of RSV by Yakubu et al. (2019).

iii. Mothers who smoke should stop immediately as this study has revealed that 38 out of the 348 (that is 11%) babies that were seropositive had smoking mothers.

iv. We also advise mothers to embrace education as educated mothers change their perceptions regarding how best to allocate resources for the betterment of children's health and fewer but healthier children.

References

Abdul Momin Kazi, Obianuju G Aguolu, Waliyah Mughis, Nazia Ahsan, Saima Jamal, Ayub Khan, Hanya M Qureshi, Inci Yildirim, Fauzia A Malik, Saad B Omer (2021) Respiratory Syncytial Virus-Associated Mortality Among Young Infants in Karachi, Pakistan: A Prospective Postmortem Surveillance Study. *Clinical Infectious Diseases*, Volume 73, Issue Supplement_3, 1 September 2021, Pages S203–S209,

Ahammed Y, S.S; Audu, I; (2019) Binary Logistic Regression Methods for Modeling Broncho-Pneumonia Status in Infants from Tertiary Health Institutions in North Central Nigeria. Appl. Sci. Environ. Manage. Vol. 23 (8) 1607-1614 August 2019

Aliyu, A, Detalle, L., Stohr, T., Palomo, C., Piedra, P.A., Gilbert, B.E., Mas, V., Millar, A., Power, U.F., Stortelers, C., Allosery, K. and Melero, J.A.,(2016). Generation and characterization of ALX-0171, a potent novel therapeutic nanobody for the treatment of respiratory syncytial virus infection. *Antimicrobial agents and chemotherapy*, *60*(1), pp.6-13.

Bartholomew, D.C., Oyinebifun, E.B & Arimie, C.O. (2021). Hierarchical Regression Modeling of some Factors Affecting Weight of Child at Birth, *Asian Research Journal of Mathematics*, 17(12): 11-27.

Beki, OD (2012).The Use of Discriminant Model and Logistic Regression for Tracking the Incidence of Broncho-Pulmonary Dysplasia among Infants. An unpublished Dissertation Submitted to Usman Danfodio University, Sokoto Nigeria.

Bonire, F.S., Umoh, V.J., Ado, S.A., Ellah, E.E. and ,Ojeleye, O.A. (2009) A Serological Survey of Human Parainfluenza Viruses (HPIVs) among Children in Kaduna Metropolis, Nigeria. *IOSR Journal of Pharmacy and Biological Sciences*,. Volume 10, Issue 1 Ver. IV (Jan -Feb. 2015), PP 60-65

Caroline Breese Hall, M.D., Geoffrey A., Weinberg, M.D., Marika K. I., Aaron, K.B., Kathryn, M.S., Edwards, M., Mary, M.D., Staat, A., Peggy Auinger, M.S., Marie R. G., Katherine A. P., (2009). The Burden of Respiratory Syncytial Virus Infection in Young Children, *The New England Journal of Medicine,* 360:588-598.

Centers for Disease Control and Prevention. Respiratory Syncytial Virus Seasonality — United States, 2014–2017. *MMWR*. 2018;67(2):71–76

Damilola M.O., Dimeji P.O., Rasheedat, M.I., Mohammed B.A., Rasaki A. R., Aishat A. G., Olarenwaju T. A., Abdul-Wahab B. R. J. (2019). Reappraisal of respiratory syncytial virus as an aetiology of severe acute lower respiratory tract infections in children younger than 5 years in Nigeria (Trans R Soc Trop Med Hyg. 2019 Aug 1;113(8):446-452.

Danbaba, A; Audu, M; Mahmud, AM (2013). Investigation of risk factors of low birth weight using multivariate logistic regression analysis. *J. of Hum, Sci. Technology*. A publication of Niger State Polytechnic. 3, 59-77

Bonnet D, Schmaltz AA, Feltes TF. Infection by the respiratory syncytial virus in infants and young children at high risk.(2005) *Cardiol Young*. 15(3):256-65.

Di Carlo P, Romano A, Plano MRA, Gueli A, Scarlata F, Mammina C.(2007) Children, parents and Respiratory Syncytial Virus in Palermo, Italy: prevention is primary. *Journal of Child Health Care*. 2010;14(4):396-407.

Egbo, M.N & Bartholomew, D.C. (2017). A Discriminant Function Analysis Approach to Country's Economy Status*, Journal of Advanced Statistics*, Vol. 2, No. 4, December 2017.

Escobar GJ, Ragins A, Li SX, Prager L, Masaquel AS, Kipnis P(2010). Recurrent Wheezing in the Third Year of Life Among Children Born at 32 Weeks' Gestation or Later: Relationship to Laboratory-Confirmed, Medically Attended Infection With Respiratory Syncytial Virus During the First Year of Life. *Arch Pediatr Adolesc Med.* 2010;164(10):915–922.

Falsey A.R, Hennessey P.A, Formica M.A, Cox C. & Walsh E.E. (2005). Respiratory syncytial virus infection in elderly and high-risk adultsexternal icon. *New England Journal of Medicine*; 352(17):1749-59

Fisher, R. A. "The Use of Multiple Measurements annals of Eugenics". 1936, 7:179–188.

Gebremedhin, A.T., Hogan, A.B., Blyth, C.C. et al. (2022). Developing a prediction model to estimate the true burden of respiratory syncytial virus (RSV) in hospitalised children in Western Australia. *Sci Rep* 12, 332 (2022).

Hall C.B., Weinberg G.A., Iwane M.K., Blumkin A.K., Edwards K.M. (2009). The burden of respiratory syncytial virus infection in young children, *New England Journal of Medicine*. 360(6):588-98.

Lia M. Hanes (2013). Progress and Challenges in RSV Prophylaxis and Vaccine Development. The Journal of Infectious Diseases, Volume 208, Issue suppl_3, December 2013, Pages S177–S183,

Light M, Bauman J., Mavunda K., Malinoski F., Eggleston M. (2008). Correlation between respiratory syncytial virus (RSV) test data and hospitalization of children for RSV lower respiratory tract illness in Florida. *Pediatr Infect Dis J.*, 27(6):512-8.

M Kaneko J Watanabe, M Kuwahara, E Ueno, M Hida, A Kinoshita, T Sone (2002) Impact of respiratory syncytial virus infection as a cause of lower respiratory tract infection in children younger than 3 years of age in Japan *J Infect . 2002* May;44(4):240-3

Megan A, Moreno, M.D, Fred Furtner, Frederick P. Rivara, M.D, MPH, Mohammed (2009) Bronchiolitis and Respiratory Syncytial Virus(Bronchiolitis and Respiratory Syncytial Virus. *Arch Pediatr Adolesc Med.* 2009;163(11):1072.

Obodai, Evangeline, Richard Asmah, Isaac Boamah, Bamenla Goka, John Kofi Odoom, and Theophilus Adiku.(2014) "Respiratory syncytial virus genotypes circulating in urban Ghana: February to November 2006." *The Pan African Medical Journal* 19 (2014).

Oluwadamilare A.O., Bamidele W. M., Abimbola A. (2021) Respiratory Syncytial Virus Infection among Children in Lagos, Nigeria. *Journal of Clinical Immunology & Microbiology*, Volume 2; Issue 1 (2021).

Orumie, U.C, Bartholomew, D.C, Obite, C.P & Kiwu, L.C. (2021) Likelihood Of Insurance Coverage on Damages Due To Level of Insecurity In Nigeria: Logistic Modeling Approach, *Financial Risk and Management Reviews*, *2021 Vol. 7, No. 1, pp. 50-59.*

Ramilo O. & Mejias A.P. (2020). Measuring the Burden of RSV in Children to Precisely Assess the Impact of Preventive Strategies. Jul;146(1):e20201727. doi: 10.1542/peds.2020-1727. Epub 2020 Jun 16.PMID: 32546584 No abstract available.

Rha B., Curns A.T., Lively J.Y., et al. (2015). Respiratory Syncytial Virus–Associated Hospitalizations among Young Children: 2015–2016. *Pediatrics*. 2020;146(1)

Ruimu Zhang & Jikui Deng (2021) Clinical impact of respiratory syncytial virus infection on children hospitalized for pertussis *BMC Infectious Diseases* volume 21, Article number: 161 (2021)

Shi T, Balsells E, Wastnedge E, Singleton R, Rasmussen ZA, Zar HJ, Rath BA, Madhi SA, Campbell S, Vaccari LC, Bulkow LR, Thomas ED, Barnett W, Hoppe C, Campbell H, Nair H.(2015) Risk factors for respiratory syncytial virus associated with acute lower respiratory infection in children under five years: Systematic review and meta-analysis. *J Glob Health. 2015* Dec;5(2):020416.

Shi T, *Angeline Denouel*, *Anna K Tietjen*, *Iain Campbell*, *Emily Moran*, *Xue Li*, *Harry Campbell*, *Clarisse Demont*, *Bryan O Nyawanda*, *Helen Y Chu*, *Sonia K Stoszek*, *Anand Krishnan*, *Peter Openshaw*, *Ann R Falsey*, *Harish Nair*. (2017). Global, regional and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling studyexternal icon. *Lancet. Journal 2017* Sep 2;390(10098):946-985. .

Shi T, McAllister DA, O'Brien KL, Simoes EAF, Madhi SA, Gessner BD, Polack FP, Balsells E, Acacio S, Aguayo C, Alassani I, Ali A, Antonio M, Awasthi S, Awori JO, Azziz-Baumgartner E, Baggett HC, Baillie VL, Balmaseda A, Barahona A, Basnet S, Bassat Q, Basualdo W, Bigogo G, Bont L, Breiman RF, Brooks WA, Broor S, Bruce N, Bruden D, Buchy P, Campbell S, Carosone-Link P, Chadha M, Chipeta J, Chou M, Clara W, Cohen C, de Cuellar E, Dang DA, Dash-Yandag B, Deloria-Knoll M, Dherani M, Eap T, Ebruke

BE, Echavarria M, de Freitas Lázaro Emediato CC, Fasce RA, Feikin DR, Feng L, Gentile A, Gordon A, Goswami D, Goyet S, Groome M, Halasa N, Hirve S, Homaira N, Howie SRC, Jara J, Jroundi I, Kartasasmita CB, Khuri-Bulos N, Kotloff KL, Krishnan A, Libster R, Lopez O, Lucero MG, Lucion F, Lupisan SP, Marcone DN, McCracken JP, Mejia M, Moisi JC, Montgomery JM, Moore DP, Moraleda C, Moyes J, Munywoki P, Mutyara K, Nicol MP, Nokes DJ, Nymadawa P, da Costa Oliveira MT, Oshitani H, Pandey N, Paranhos-Baccalà G, Phillips LN, Picot VS, Rahman M, Rakoto-Andrianarivelo M, Rasmussen ZA, Rath BA, Robinson A, Romero C, Russomando G, Salimi V, Sawatwong P, Scheltema N, Schweiger B, Scott JAG, Seidenberg P, Shen K, Singleton R, Sotomayor V, Strand TA, Sutanto A, Sylla M, Tapia MD, Thamthitiwat S, Thomas ED, Tokarz R, Turner C, Venter M, Waicharoen S, Wang J, Watthanaworawit W, Yoshida LM, Yu H, Zar HJ, Campbell H, Nair H; .(2017) RSV Global Epidemiology Network. Lancet. 2017 Sep 2;390(10098):946-958

Shi, Ting, et al. "Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study." *The Lancet* 390.10098 (2017): 946-958.

Sindhu Bharathi, S. (2016) "Prevalence of raised IgE levels and absolute eosinophil count in bronchiolitis in children aged 2 months to 2 years in tertiary care centre." PhD diss., Stanley Medical College, Chennai, 2016.

Sule A. S. (2013) Modeling Broncho-Pneumonia Status In Infants Using Discriminant And Logistic Regression Analysis https://eduproject.com.ng/statistics/modeling-bronchopneumonia-status-in-infants-using-discriminant-and-logistic-regression-analysis/index.html

Utsunomiya, Takeshi, Hibino, Akinobu MD, Taniguchi, Kiyosu, Nagai, Takao, Saito, Naruo , Tanabe, Ikumi; Odagiri, Takashi, Shobugawa, Yugo, Kaneko, Akira, Saito, Reiko, (2020) Factors Contributing to Symptom Duration and Viral Reduction in Outpatient Children With Respiratory Syncytial Virus Infection, The Pediatric Infectious Disease Journal: August 2020 - Volume 39 - Issue 8 - p 678-683

Zachariah P, Shah S, Gao D, Simões EA. (2009) Predictors of the duration of the respiratory syncytial virus season. *Pediatr Infect Dis J*. 2009 28(9):772-6.