This research is aimed at investigating the predictors and risk factors of diabetes using the log-normal approach for data of diabetic patients from two selected hospitals within Birnin Kebbi metropolis: Federal Medical Centre Birnin Kebbi and Sir Yahaya Memorial Hospital Birnin Kebbi. The data used in this study were obtained from two selected hospitals within the Birnin Kebbi metropolis consist of 750 people diagnosed with Diabetes out of which some happens to be diabetic patients while others are none. The data variables were coded as: Diabetes status (1 = Patient is having Diabetes, and 0 = Patient do not have Diabetes), Ages (less than 45years = 0, above 45years = 1), Sex/Gender (1 = male, 0 = female), History of Diabetes (1 = Yes, 0 = No), History of HBP/Hypertension (1 = Yes, 0 = No), Frequent urination/increased thirst (1 = Yes, 0 = No), Overweight/Obese (1 = Yes, 0 = No), and Fatigue or Muscle Pain (1 = Yes, 0 = No) and analysis was performed using Statistical Package NCSS. However, the results obtained revealed that Age has the highest impact on diabetes followed by gender, then frequent urination while Gender, history of HBP and obese/overweight have less impact with fitted model as Diabetes = 127.7988115605 + 2.91056748883139E-02*Age + 1.41071509860143*Gender -4.76836365190816* History of Diabetes -8.12517728481757* History of HBP -7.31335028559973* Frequent Urination -3.98943620318212* Obese/Overweight.

**Keywords:** Obesity, Diabetes, Log-normal, Predictors, Risk, Gender

1.0 INTRODUCTION
1.1 Background of the Study
Many experimental situations arise in which research observed the count of events within a set unit of time interval, measurements, experiments, systems failures, diseases, etc., and the data may be normally distributed or skewed. Many existing methods of statistical inference and analysis rely heavily on the assumption that the data are normally distributed. However, the normality assumption is not fulfilled when dealing with data that does not contain negative values or are otherwise skewed, a common occurrence in diverse disciplines such as finance, economics, political science, sociology, philology, biology, and physical and industrial processes Hussain and Naaz (2021). In this situation, a log-normal distribution may better represent the data than the normal distribution. Measurement data often have a lower limit, usually 0 (when there is no case detected) or the detection limit, but no distinct upper limit. Hence below the median no observations can be further away than the lower limit, but above the median there may be values that are many times further away, and this will give a positively skewed distribution.

Ellahi, et al. (2023) study the prevalence, demographic, anthropometric and diet-related predisposing factors for diabetes among urban and rural dwellers and assess participants’ self-awareness of their diabetes status. A cross-sectional survey involved 850 adult males and females age ≥18 years residing in urban and rural areas in the Hohoe Municipality of Ghana, randomly sampled using probability proportional to size. Data included demographic, anthropometric, physiologic (blood pressure, fasting blood glucose) and dietary information. Nutrient quantities were analysed using the Research to Improve Infant Nutrition and Growth (RIING Nutrient Database Software. All other data was analysed in SPSS (v25). Risk factors were estimated through ordinal logistic regression and the odds ratio (OR) with the corresponding 95% confidence level (CI) documented. Results shows More females participated than males (58.4% vs 41.6%), similarly rural compared to urban inhabitants (53.5% vs 47.5%, p=0.002) with a mean age of 47.3±16.1 years. Females generally
had higher adiposity, rural dwellers had higher BMI whereas urban dwellers had higher waist and hip circumferences. Overall, 4.4% and 18.5% were diabetics and pre-diabetics; while 20.4% and 12.1% were overweight and obese respectively. Of the 36.8% hypertensives, only 18.2% were aware of their status, with significant male (40.3%) female (59.7%); and urban (43.5%) rural (56.6%) differences. Males had higher intakes of energy and nutrients of public health importance to hypertension, similar as rural inhabitants except for cholesterol. Advancing age (95% CI: 0.02-0.05), being male (OR: 1.56 95% CI:0.12-0.81)and increased BMI (95% CI: 0.01-0.11) were independently associated with hypertension. The lognormal distribution is characterized by having only positive non-zero values, a positive skewness, a non-constant variance that is proportional to the square of the mean value, and a normally distributed natural logarithm. The probability density function for a lognormal distribution has an asymmetrical appearance, with a majority of the area below the expected value and a thinner right tail with higher values, while the probability density function for the normal distribution with the same expected value has a symmetrical bell-shaped curve.

This research aimed at investigating predictors and risk factors of diabetes with a view to determine set of suitable predictors for Diabetes, establish or model the relationship between this aforementioned predictors and response variables (status of being diabetes) and ascertain the impact and extent of each of these predictors on the response variables.

Lognormal distribution is a right skewed distribution with a long tail which is a close association with the normal distribution there by taking the natural logarithm of a random variable, the random variable then will have a normal distribution.

The PDF of the lognormal distribution is

\[ F(t) = \frac{1}{\sigma t \sqrt{2\pi}} \int_{0}^{\frac{1}{t}} e^{-\frac{1}{2} \frac{(ln(t) - \mu)^2}{\sigma^2}} d\theta \]

1.2

Ismail et al. (2021) identify the majority of the risk factors for the incidence / prevalence of type 2 DM on one hand and give a critical analysis of the cohort studies examine the impact of the association of risk factors on diabetes. Omar et al. (2019) discovered that the T2DM is high among the Sudanese population, especially in older people and those with a family history of DM. The high prevalence of uncontrolled DM in this setting is another hidden burden. Yan et al. (2023) study confirmed an interaction between age and other influencing factors, which may be important in explaining differences in risk factors for diabetes and predict diabetes in the middle-aged and elderly populations. Assimwe et al. (2020) study found that females and patients age 61-65 years were most affected by diabetes and the presence of family history for diabetes, overweight, and being obese increases the chances of acquiring type 2 diabetes. Belete et al. (2021) research advised that more attention should be paid to the prevention and control of the epidemic and for the reduction of morbidity and mortality associated with metabolic syndrome among type 1 DM patients.

A review of various statistical models and learning algorithms was presented by Hussain and Naaz (2021) in which random forest, Naïve Bayes, and neural network were compared for accuracy. For evaluating these machine learning algorithms, the Matthews correlation coefficient was used by the authors. Kumari et al. (2021) have worked on the Pima Indians Diabetes Dataset, applied Naïve Bayes, random forest, and logistic regression, and compared these three approaches with ensemble approach and model outperforms with ensemble approach with an accuracy of 79%. Gupta et al. (2021) worked with support vector machines and Naïve Bayes algorithms to classify the diabetic dataset. K-fold cross-validation model was used by the authors for training and testing purposes, and after applying both classification algorithms, the support vector machine classifier was performing better than the Naïve Bayes algorithm. Choubey et al. (2020) applied two feature selection methods named principal component analysis and linear
discriminant analysis to extract significant features from the Pima Indians Diabetes Dataset. A comparative analysis of the feature selection method was also presented in the article. Few machine learning algorithms, that is, radial basis kernel, k-nearest neighbour, and AdaBoost, were also applied to the dataset for classification purposes. Malik et al. (2020) worked on a local dataset taken from a hospital in Germany and applied decision trees, k-nearest neighbour, and random forest on this locally available dataset.

Tigga and Garg (2020) worked on the Pima Indians Diabetes Dataset. Important features extracted from the dataset are blood glucose levels, number of pregnancies, and body mass index. Logistic regression was used to predict the accuracy using RStudio, and the accuracy achieved was 75.32%.

Swapna et al. (2018) applied convolution neural network and convolution neural network long short-term memory networks on Electrocardiograms, a private dataset that consisted of 142000 samples and has achieved an accuracy of 90.9%. Neither pre-processing nor any feature selection technique was applied in this dataset.

According to Valasapalli et al. (2021) type 2 diabetes mellitus (DM) is a long-lasting disease whose incidence has been steadily increasing worldwide. In India, about 30 million individuals have diabetes, and many more are at risk. Thus, early detection is necessary to avoid diabetes and its associated complications. The rationale for utilizing various methods for hypothetical type 2 diabetes determination grounded on the indicative study is to extend the disease’s detection period by assessing suggestive features and regular habits, thereby enabling the estimation of type 2 diabetes without the use of clinical tests through predictive analysis. Today, an enormous amount of clinical knowledge is available about infections, their symptoms, the causes of illness, and their consequences for health. Due to the accuracy of these algorithms, the danger of type 2 diabetes may be predicted, which is critical for the medical sector.

Sun and Zhang (2019) have discussed a few deep learning methods and classification methods such as artificial neural network, decision trees, random forest, and support vector machine. Qawqzeh et al. (2020) have implemented a logistic regression classification technique for the classification of diabetes data. Training data includes 459 patients, and testing data includes 128 patients. Classification accuracy achieved by the authors was 92% using logistic regression. The major disadvantage of the model was that it was not compared with the other diabetic prediction models and hence could not be validated.

Sisodia and Sisodia (2018) have applied Naïve Bayes, decision trees, and support vector machine learning algorithms on the Pima Indians Diabetes Dataset, and the maximum accuracy to predict the diabetes was achieved by Naïve Bayes classifier. A tenfold cross-validation technique was used by Sisodia in which the dataset was divided into ten equal parts: 9 parts were used for training, and the remaining part was used for testing. Evaluation parameters on which the diabetes was predicted were accuracy, precision, recall, and area under the curve.

Lekha and Suchetha (2018) created their own dataset, which was based on breath signals and consisted of 25 patients, amongst which 11 were healthy, five were type 1 diabetic patients, and nine were type 2 diabetic patients. Leave one cross validation was used for validation purposes, and evaluation parameters were ROC curve, which was approximately 96%. According to Dremin et al. (2021) ageing and diabetes both result in protein glycation and malfunction of col-lagen-containing tissues. Collagen mechanical and practical alterations accompany the development of a variety of pathological abnormalities distressing the skin, plasma vessels, and nerves, resulting in a variety of problems, increased impairment hazards, and a danger to life. Indeed, there are presently no noninvasive techniques for assessing glycation and related metabolic processes in biotissues or for predicting potential skin consequences, such as ulcers, for endocrinologists and clinical diagnosis. To increase the accuracy of a diabetic prediction model, Dwivedi (2019) applied an adaptive neurofuzzy inference system, and a feature selection method named principal component analysis was used.
2.0 MATERIALS AND METHODS

2.1 Data for the Study

The data used in this study were obtained from two selected hospitals within Birnin Kebbi metropolis: Federal Medical Centre Birnin Kebbi and Sir Yahaya Memorial Hospital Birnin Kebbi consist 750 people diagnosed for Diabetes out of which some happens to be diabetic patients while others are none. From the data, the following variables were derived that is coded as:

- Diabetes status (1= Patient is having Diabetes, and 0 = Patient do not have Diabetes).
- Ages (less than 45years=0, above 45years=1)
- Sex/Gender (1 = male, 0 = female).
- History of Diabetes (1 = Yes, 0= No).
- History of HBP/Hypertension (1 = Yes, 0= No).
- Frequent urination/increase thirst (1 = Yes, 0 = No)
- Overweight/Obese (1 = Yes, 0 = No)
- Fatigue or Muscle Pain (1 = Yes, 0 = No)

2.2 The Log-Normal Distribution

Lognormal distribution is not new, it is a distribution which is skewed to the right, whose probability density function starts at zero, increases to its mode and decreases thereafter.

Formally, a random variable $X$ is said to follow a lognormal distribution if $\log (X)$ follows a normal distribution. When a variable $X$ can only take positive values, the arithmetic mean, median and mode may not be the same, and in particular, the arithmetic mean is affected heavily by the presence of large values in the data. In this case, $X$ is said to follow the lognormal distribution, and the geometric mean typically represents the median value, while the arithmetic mean exceeds the median leading to a right skew in the distribution. When we use normal distribution, using arithmetic mean as a measure of central tendency is acceptable because in a symmetric distribution arithmetic mean is same as its median. However, for lognormal distributed data, geometric mean is more suitable because it represents the centre of the distribution of the logarithms (which is symmetric) and corresponds to the median. (Google Schoolar, 2023)

The probability density function of the lognormal distribution is

$$F(t) = \frac{1}{\sigma t \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{\log X - \mu}{\sigma} \right)^2}$$

The Cumulative density function of the lognormal distribution is

$$F(t) = \frac{1}{\sigma t \sqrt{2\pi}} \int_0^t \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{\log t - \mu}{\sigma} \right)^2} d\theta$$

2.3 Bivariate Log-Normal Distribution

The standard bivariate log-normal distribution is given by the joint probability density function

$$f_{X,Y}(x,y) =$$

2.4 Model Selection

It is important that we have one or more a criterion to consider the best results and choose the appropriate model for data representation. There are several methods that provide a measure for selecting the appropriate model. We fitted all of the distributions by the method of maximum likelihood. The best distribution will be chosen according to the following information criteria:

So, we will use the following methods to selecting best model.

- $L$: The maximum likelihood function of the model.
- $K$: Number of model parameters.
- The model with minimum AIC value is chosen as the best model to fit the data.
2.4.2 Bayesian Information Criterion
The Bayesian information criterion (BIC) is another estimator for evaluating model fit for a given data among different types of models, and its formula is given as
\[ \text{BIC} = -2\log L + k \log n \]
where
- \( L \): The maximum likelihood function of the model.
- \( k \): Number of model parameters.
- \( n \): The no. of observations, or the sample size.

The model with minimum BIC value is chosen as the best model to fit the data.

2.4.3 The Hannan Quinn Criterion (HQC) Test
The Hannan Quinn Criterion (HQC) defined by
\[ \text{HQC} = -2\log L + 2k \log n \]
where
- \( L \): The maximum likelihood function of the model.
- \( k \): Number of model parameters.
- \( n \): The no. of observations, or the sample size.

The smaller the values of these criteria the better the fit.

3.0 RESULTS/RESULT AND DISCUSSION
3.1 Tabular Presentation Of Data
For convenience, the following variables were derived that is coded as:
- Diabetes status (1= Patient is having Diabetes, and 0 = Patient do not have Diabetes).
- Ages (less than 45years=0, above 45years=1)
- Sex/Gender (1 = male, 0 = female).
- History of Diabetes (1 = Yes, 0= No).
- History of HBP/Hypertension (1 = Yes, 0= No).
- Frequent urination/increase thirst (1 = Yes, 0 = No)
- Overweight/Obese (1 = Yes, 0 = No)
- Fatigue or Muscle Pain (1 = Yes, 0 = No)

The data is presented in appendix 1

### Table 3.1: Response Analysis

<table>
<thead>
<tr>
<th>Categories</th>
<th>Count</th>
<th>Rows</th>
<th>Prior</th>
<th>R-Squared</th>
<th>% Correctly</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0.00533</td>
<td>0.00466</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>746</td>
<td>80</td>
<td>0.99467</td>
<td>0.00466</td>
<td>100.000</td>
</tr>
<tr>
<td>Total</td>
<td>750</td>
<td>84</td>
<td>99.467</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The result in table 3.1 reveals that the response analysis i.e the analysis of status of being diabetic only four (4) out of seven hundred and fifty (750) patients that are non-diabetic from the study. Therefore about 99.5% were known to be diabetic from the results which mean approximately 100% correctly classification on the categories of being diabetic.

### Table 3.2 Parameter Significance Tests Section (Reference Group: Diabetes Status)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Regression Coefficient (B or Beta)</th>
<th>Standard Error</th>
<th>Wald Z-Value (Beta=0)</th>
<th>Wald Prob. Level</th>
<th>Odds Ratio Exp (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0: Intercept</td>
<td>27.79881</td>
<td>1.99022</td>
<td>13.968</td>
<td>0.00000</td>
<td>31.69957</td>
</tr>
<tr>
<td>B1: Age</td>
<td>0.02911</td>
<td>56.42211</td>
<td>0.001</td>
<td>0.99959</td>
<td>1.02953</td>
</tr>
<tr>
<td>B2: Gender</td>
<td>1.41072</td>
<td>0.99857</td>
<td>1.413</td>
<td>0.15773</td>
<td>4.09889</td>
</tr>
<tr>
<td>B3: Hist. of Diab.</td>
<td>-4.76836</td>
<td>1.99018</td>
<td>-2.396</td>
<td>0.01658</td>
<td>0.00849</td>
</tr>
<tr>
<td>B4: Hist. of HBP</td>
<td>-8.12518</td>
<td>1.99018</td>
<td>-4.083</td>
<td>0.00004</td>
<td>0.1854</td>
</tr>
</tbody>
</table>

3.2 Data Analysis
The collected data on the Diabetes status, age, gender, history of diabetes, history of HBP, frequent urination and obese/overweight from Sir Yahaya Memorial hospital and Federal Medical Centre, Birnin Kebbi was analyzed using log normal model to determine the suitable predictors of Diabetes in patients. Diabetes status was considered as the dependent variable and the remaining variables namely age, gender, history of diabetes, history of HBP, frequent urination and obese/overweight were taken to be the independent variables for the study. All the independent variables were coded except for age which was considered as numeric variable. The data was analyzed using statistical software package NCSS and the following results were obtained:
The results in Table 3.2 reveal the estimation of parameters in the lognormal regression analysis, from the results it was discovered that history of diabetes, high blood pressure are significant at 5% level i.e symptoms mention are likelihood 0.1854, 0.0177, and 0.00849 respectively this indicate that they less likely expose to diabetes at 99% level of significance.

Estimated Log normal Regression Model(s)


Table 3.3: Result from the Deviance Analysis

<table>
<thead>
<tr>
<th>Term Omitted</th>
<th>DF</th>
<th>Deviance</th>
<th>Deviance (Chi-Square)</th>
<th>Prob. Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>6</td>
<td>49.84886</td>
<td>3.70438</td>
<td>0.71661</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>46.25851</td>
<td>0.11402</td>
<td>0.73561</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>47.89807</td>
<td>1.75359</td>
<td>0.18543</td>
</tr>
<tr>
<td>Hist. of Diab.</td>
<td>1</td>
<td>46.16954</td>
<td>0.02506</td>
<td>0.04422</td>
</tr>
<tr>
<td>Hist. of HBP</td>
<td>1</td>
<td>47.09207</td>
<td>0.94758</td>
<td>0.33034</td>
</tr>
<tr>
<td>Freq. Urin.</td>
<td>1</td>
<td>46.17225</td>
<td>0.02777</td>
<td>0.02765</td>
</tr>
<tr>
<td>Obese/Over W.</td>
<td>1</td>
<td>46.15515</td>
<td>0.01066</td>
<td>0.01776</td>
</tr>
<tr>
<td>None (Model)</td>
<td>6</td>
<td>46.14448</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results in the table 3.3 also reveal the deviation analysis results. From the results it was observed that History of Diabetic in the family, frequent urine, and obesity are major contributors to influence the status of being diabetic. Were as Age, gender, High Blood Pressure were not significant at 5% level of significance.

4.0 DISCUSSION OF THE RESULTS

Results obtained from the binary log normal likelihood regression analysis revealed that a total of 750 observations were collected for all the variables and no missing value was realized for all the variables included in the analysis as shown in Table 4.1. The results obtained also indicated that the dependent variable coded as 0 for not having diabetes and 1 for having diabetes yielded all the
750 patients to suffer diabetes as indicated in Table 4.2. The log normal regression model fitted to the data is obtained as:

\[
\text{Diabetes} = 127.7988115605 + 2.91056748883139 \times 10^{-2} \times \text{Age} + 1.41071509635190816 \times \text{Gender} - 4.76836365190816 \times \text{History of Diabetes} - 8.12517728481757 \times \text{History of HBP} - 7.31335028559973 \times \text{Frequent Urination} - 3.98943620318212 \times \text{Obese/Overweight}.
\]

From the fitted model, it could be observed that the constant term and the coefficients for all the independent variables in the model are obtained as shown in Tables 4.3-4.5. It could be observed from the fitted model that the constant term in the model signifies that in the absence of all the six independent variables included in the model, at least 128 cases of diabetes will be recorded in the study area. Likewise, it could be deduced from the model that only age and gender have positive coefficients indicating they are the only variables that influence diabetes positively. All the other variables in the model namely, history of diabetes, history of high blood pressure, frequent urination and obese/overweight have negative coefficients indicating they don’t significantly influence diabetes. Out of the two variables that positively influence diabetes, it could be observed that age has the highest impact on diabetes followed by gender.

It could also be observed from the results in Table 4.6 that the variables, age, history of diabetes and frequent urination have values 0.04422, 0.0276 and 0.0177 respectively are each less than 0.05 meaning that at 5% level of significance the null hypothesis for each of them will be rejected and on the basis of that it could be concluded that there is enough evidence to show that these variables are each not equal to zero at 95% confidence interval. This means that these variables are each important to be included in the model. Therefore, there is enough evidence to conclude that these variables are relevant predictors in identifying diabetes in the study area. From the same table, it could be revealed that the set of variables gender, history of HBP and obese/overweight have probability values 0.7356, 0.1854, and 0.3383 respectively meaning that at 5% level of significance, the probability values are greater than the level of significance meaning that the null hypotheses will be accepted and therefore we can conclude that there is sufficient evidence to indicate that each of these variables equals to zero. This shows that these variables were not important to be included in the model and hence were not relevant in predicting diabetes in the study area.

Furthermore, table 4.5 gave the odd ratios for the variables included in the model. From the table it could be observed that the strongest variable of the outcome of diabetes patients was age with an odd ratio of 4.09889 (at 95% confidence interval). This indicated that patients who had been checked and referred to as having old age are more likely to estimate the success of diabetes as against those that are young included in the model. The odd ratio 1.01301 for gender indicates that for every treatment per patient there was more diabetes due to the gender of patients for some time caused by changes in gender.

All the 750 observations were analyzed on the basis of their arrangement comparing the actual values obtained with regards to the diabetes status of patients with the residuals obtained from the fitted model to determine the deviance of the residuals from the actual values obtained from each patient. These were used to indicate the actual values of patients without diabetes that were actually considered in the model.
5.0 CONCLUSION

The study showed that there is a positive relationship between diabetic family history and prevalence of diabetes mellitus. Some studies stated that there is a positive family history of diabetes. In this study, it was revealed that there is a negative relationship between diabetes mellitus status and other predictors: Age, gender, HBP, Obesity compared to other existing research. Base on the analysis as well as the results obtained, the following are conclusion reach:

a) Age has the highest impact on diabetes followed by gender
b) Age, history of diabetes and frequent urination are relevant predictors in identifying diabetes in the study area but contributed at least percentages.

c) Gender, history of HBP and obese/overweight were not important to be included in the model and hence were not relevant in predicting diabetes in the study area.

d) The log normal regression model fitted to the data is obtained as

\[
\text{Diabetes} = 127.7988115605 + 2.91056748883139E-02 \times \text{Age} + 1.41071509860143 \times \text{Gender} - 4.76836365190816 \times \text{History of Diabetes} - 8.12517728481757 \times \text{History of HBP} - 7.31335028559973 \times \text{Frequent Urination} - 3.98943620318212 \times \text{Obese/Over weight}
\]

6.0 ACKNOWLEDGEMENTS

The authors are deeply indebted to the editor and learned referees for their constructive suggestions leading to improving the quality of the contents and presentation of the original manuscript.

REFERENCES


